Eukaryotic Signature Proteins:
Guides to pathogenic eukaryotic parasites

A thesis presented in partial fulfilment of the requirements of the degree of

PhD

in

Genetics

At Massey University, Palmerston North,

New Zealand

Jian Han

2012

Abstract

Eukaryotic Signature Proteins (ESPs) are proteins that delineate the eukaryotes from the archaea and bacteria. They have no homologues in any prokaryotic genome, but their homologues are present in all main branches of eukaryotes. ESPs are thus likely to have descended from ancient proteins that have existed since the first eukaryotic cell. This project looks at ESPs of some eukaryotic parasites and human (*Homo sapiens*) as their host organism and focuses on *Giardia lamblia*, a fresh water pathogenic basal eukaryote. The ESP datasets from *Giardia* and two other parasites, *Trichomonas vaginalis* and *Plasmodium falciparum*, as well as the host human were calculated in light of available genomic data and the datasets contained a range of proteins associated with membrane, cytoskeleton, nucleus and protein synthesis.

ESPs have great potential in phylogenetic studies since these proteins are present in all eukaryotes and are expected to have a slow and constant rate of evolution. Phylogenetic analyses were performed on the 18 eukaryotic organisms including some basal eukaryotes, and also for mammals, using orthologues of the all ESPs from these organisms. Strategies such as concatenating sequences and constructing consensus networks were tested to evaluate their potential with large numbers of ESP alignments. The results were promising, and ESPs hold great potential for their use in future phylogenetic analyses of eukaryotes.

RNA interference is hypothesised to be an ancient mechanism for gene regulation and like the ESPs, it is typically found in all main branches of eukaryotes. High throughput sequencing data from *Giardia* and *Trichomonas* small RNAs (15-29mers) were re-analysed showing two length peaks for *Giardia* RNAs: a "larger peak" and an "ultra small peak", the former of which is likely to be the product of the enzyme Dicer, which processes miRNA. The "ultra small peak" but not the "larger peak" was also found in *Trichomonas*. The two peaks possibly represent two different mechanisms of RNA interference (RNAi) in these parasites, but analysis of potential target sites from the Dicer-processed RNAs has not yet shown any indication that ESPs are regulated any differently from other parasite proteins.

Sugar metabolic pathways including glycolysis and citric acid cycle were searched for ESPs, this was done to determine the relationship between the conservation of

eukaryotic metabolic pathways and conservation of individual proteins. However no ESPs were identified from these pathways because *Giardia* has enzymes that show more similarity to those from prokaryotes than eukaryotes. These enzymes are significantly different from that of the host's, and these alternative enzymes offer potential as novel drug targets. In addition, ESPs that are present from host but lost in some parasites were analysed, and these ESPs are involved in many understudied pathways. It is these differences which can provide a guide in determining which pathways we should examine when designing drug targets.

Overall, numerous proteomic similarities and differences in ESPs were identified between host and parasite. These proteins show potential for future evolutionary studies, and will guide future directions in ancestral eukaryotic regulation and metabolism.

Acknowledgements

It has been a very challenging yet rewarding journey towards the completion of this thesis. I am thankful to everyone who helped me throughout my work, and kept my life interesting during my study.

Foremost I would express my sincere gratitude towards my supervisor, Dr. Lesley Collins. Bioinformatics was never my forte during my undergraduate studies. But thanks to Dr. Collins, I have learnt many bioinformatics skills during the four years of my doctorial study. These skills will be very useful in my future career.

I would thank my co-supervisor Dr. Patrick Biggs, who has been tirelessly commenting on my "broken" English, it was very fortunate to have someone who can speak English like the Queen. Dr. Biggs has also been outstanding in helping me setting up databases.

I also thank my other co-supervisor David Penny, who has been helping me with the writing and cracking jokes from time to time.

I express my warm thanks to my colleagues who helped me during different parts of the project. Dr. Tim White, who is an absolute computing genius, has provided generous computing assistance; Dr. Simon Hills and Bojian Zhong have given a helping hand on my phylogenetics studies.

I thank my friends/flatmates (Nick, Sophie, Ping, Bryn, Sam, Ryan, Suz, Nat and Justin) and others, my weekends would be very dull without you guys. Also thanks to my snowboarding buddy Max for keeping me alive from giant snowballs.

Special thanks to my parents, both doctors, for their financial and emotional support. They provided me with plenty of encouragement.

Finally, thanks to everyone in the boffin lounge, the environment and work ethic has been wonderful here. Thanks to Massey University, Palmerston North for providing the working space.

This work was funded by Health Research Council (HRC) - Emerging Researcher Grant (Dr. L. Collins) 07/168. Eukaryotic Signature Proteins - Guides to Modern Eukaryotic Parasites.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Terminology

3'UTR (three prime untranslated region): Region of mRNA downstream of the termination codon. In metazoans this region is where miRNA binds to regulate gene expression.

5'UTR (five prime untranslated region): Region of mRNA upstream of the starting codon that often contain regulatory elements such as ribosome binding sites.

Akaike Information Criterion (AIC): Measure used in model testing based on the goodness of fit to a statistical model

Basal Eukaryote: A unicellular eukaryotic which is believed to have diverged early during the evolution of eukaryotes, e.g. *Giardia lamblia*.

Bayesian inference: A tree searching method which is statistically similar to maximum likelihood, the aim is to find the tree with maximum posterior probability; can allow complex models of evolution to be implemented.

BLAST (Basic Local Alignment Search Tool): Software which enables comparison of amino acid or nucleotide sequences.

Blastp: A BLAST program which compares protein queries with protein databases.

Cellular signature structure (CSS): Cell organelles or complex found in eukaryotes but not prokaryotes, e.g. mitochondria, Golgi apparatus, spliceosome.

Excavata: A eukaryotic supergroup that contains the morphological feature of a ventral feeding groove. This supergroup includes Diplomonads (*Giardia lamblia*) and Parabasalia (*Trichomonas vaginalis*).

Eukaryotic Signature Protein (ESP): A protein with no homologues in prokaryotic (archaea and bacteria) genomes, but it has homologues which are present in all the major branches of eukaryotes.

Gene Ontology (GO): A project aimed to unify the representation of gene attributes across all species by using a controlled vocabulary to assign their functions. Website: http://www.geneontology.org.

Long branch attraction (LBA): A phenomenon observed when highly divergent lineages are grouped together, regardless of their true evolutionary relationships. The long branches of a tree will group together regardless of the true tree topology.

Maximum likelihood (ML) inference: A tree searching method which aims to find the tree with highest probability to produce the observed data.

Messenger RNA (mRNA): RNA transcribed from DNA, after mRNA processing (e.g. Capping, intron splicing) the mature mRNA is translated into protein by the ribosome.

Micro RNA (miRNA): ~21-22 base pair (bp) single stranded RNA processed by the Dicer or Drosha proteins, which regulates gene expression by means of complimentary binding to the target mRNA.

Non-coding RNA (ncRNA): RNA that does not code for proteins, but may have a function such as regulating, modifying or processing other RNAs.

Perl: A dynamic programming language. Able to perform various bioinformatic tasks especially data mining and can connect with MySQL databases to enable fast and automated database management and queries.

Small interfering RNA (siRNA): ~21-26 bp double stranded RNA processed by Dicer which regulates gene expression by means of complimentary binding to the target mRNAs.

Chapter 1: Introduction

1.1 Eukaryotic signature proteins

Eukaryotic signature proteins (ESPs) are signature proteins that delineate Eukarya from Archaea and Bacteria. They have no homologues in prokaryotic genomes, but their homologues are present in all the main branches of eukaryotes. They are also involved in most core functions of a eukaryote and provide landmarks to track the origin and evolution of eukaryote genomes (Kurland *et al.* 2006).

The approach of searching for signature proteins for a cellular domain was first used by Graham *et al.* when they searched for archaeal signature proteins (Graham *et al.* 2000). Their study in 1999 found 351 clusters of proteins found only in Euryarchaeota species. Their definition of signature proteins, however, differs from that of ours, because their set of proteins was not conserved in all Euryarchaeota species.

Turning to eukaryotes, Hartman *et al.* then collected *Giardia* ESPs in 2001 by searching yeast protein homologues against all three domains of life (archaea, bacteria and eukaryotes). Homologues were considered to be proteins with primary amino acid sequence similarities to those yeast proteins. Their analysis procedure was as follows (Hartman *et al.* 2002):

- Initially the *Saccharomyces cerevisiae* genome was used to identify a potential ESP dataset, which contained 6271 proteins.
- Then they removed proteins without homologues in *Caenorhabditis elegans, Drosophila melanogaster* and *Arabidopsis thaliana*. The homologue searches were performed using BLAST with a bit-score cut-off of 55, which is approximately equivalent to an e-value of 10^{-6} for the largest *Giardia* and bacterial database used. They stated that the cut-off was very conservative, but it would ensure that there were no false positive results for homologous proteins.
- After that, they removed proteins that have homologues in any of the 44 bacterial and archaeal species (there were only 44 available complete bacterial and archaeal genomes at the time).
- Lastly they removed proteins without homologues in *Giardia lamblia*.

By using this procedure Hartman *et al.* were left with 347 proteins, and they named this dataset the Eukaryotic Signature Proteins aka ESPs of *Giardia*. The main point of

Hartman *et al.*'s paper was to form a novel hypothesis on the formation of eukaryotic cells. Previously, a large number of researchers had hypothesised that eukaryotes originated from an engulfment or symbiotic event between a member of the Archaea and a member of the Bacteria kingdom (e.g. (Lake *et al.* 1994)). From the finding of these 347 ESPs, Hartman *et al.* argued that the presence of proteins without any bacterial and archaeal homologues meant that they must have come from a cell of a distinct lineage. They hypothesised that there was a third cell type, which they called a "chronocyte", which was a progenitor of the eukaryotic cell, and the nucleus of a eukaryotic cell was formed from the endosymbiosis of an archaeon and a bacterium in the chronocyte (Hartman *et al.* 2002).

Hartman *et al.* then predicted a partial picture of the chronocyte from the functions of the ESPs, in that it had a plasma membrane and a cytoskeleton, which provided competence for it to phagocytise archaea and bacteria. The chronocyte also had a complex inner membrane system for protein synthesis and breakdown, indicated by the presence of ER proteins, GTP-binding proteins, ubiquitins and ribosomal proteins in the ESP dataset. Interestingly, they also found a RNA-directed RNA polymerase to be present in all eukaryotes analysed except for *Drosophila melanogaster*, but absent in all archaea and bacteria. This enzyme is involved in replication of RNA interference (RNAi), an RNA based system that controls gene activation (Vasudevan *et al.* 2007) and silencing (Sen *et al.* 2007). Hence this finding suggested to them that the chronocyte was an RNA based cell.

Subsequently the same research group also collected ESPs for the microsporidium *Encephalitozoon cuniculi* (Fedorov *et al.* 2004), the organism with the smallest sequenced eukaryotic genome. They found 401 ESPs for *E. cuniculi*, which consisted of 238 ESPs in common with *Giardia* ESPs. This high level of similarity has indicated that even a minimal eukaryotic cell still preserved most of the ESPs, which agrees with their earlier hypothesis that these ESPs must come from a cell of distinct lineage.

Hartman *et al.*'s paper has served as a reference for some other studies. Staley *et al.* compared the list of 347 eukaryotic signature proteins (ESPs) with genomes of two bacterial species, *Prosthecobacter dejongeii* of the Verrucomicrobia phylum and *Gemmata* sp. Wa-1 of the Planctomycetes phylum (Gillin *et al.* 1996). The Verrucomicrobia and Planctomycetes phyla possess a number of phenotypic and molecular features typical of eukaryotes. For example, *Prosthecobacter* have genes for

tubulin, which is a cytoskeletal element normally found only in eukaryotes (Jenkins *et al.* 2002). Jenkins *et al.* hypothesised that Verrucomicrobia and Planctomycetes were direct ancestors of eukaryotes (Jenkins *et al.* 2002). However, later on when Staley *et al.* BLASTed the proteome of these two species against the 347 ESPs obtained by Hartman *et al.*, they could only manage to detect 17 and 10 significant ESP homologues from the two bacterial species respectively[1], and this low number raised doubts over Jenkins *et al.*'s earlier hypothesis (Staley *et al.* 2005).

Kurland *et al.* linked ESPs with cellular signature structures (CSSs), which are cellular compartments that distinguish eukaryotes from prokaryotes (Kurland *et al.* 2006). Examples of CSSs are mitochondria, nucleoli and spliceosomes. There are substantial numbers of ESPs present in the CSSs except for mitochondria (mitochondria are descended from α-proteobacteria (Andersson *et al.* 1998), thus most mitochondrial proteins would have bacterial homologues). The presence of ESPs and CSSs indicated that eukaryotes form a unique primordial lineage. They also designed a new model of how eukaryotes originated, quite similar to that of Hartman *et al.*'s but without the requirement of prokaryotic progenitors. From a community of saprotrophic, autotrophic and heterotrophic cells, a phagotrophic unicellular "raptor" emerged and then acquired a bacterial endosymbiont/mitochondria lineage, to become the common ancestor of all eukaryotes (Kurland *et al.* 2006).

ESPs are thus a group of essential proteins because they are conserved by all eukaryotes. Parasitic eukaryotes, on the other hand, typically undergo reductive evolution, which gives the parasites an advantage in replication by permitting them to reproduce much faster than if they have a bigger genome. So, would these essential proteins be found even if the organisms have undergone severe reductive evolution? Which ESPs can parasites live without? They might hold the key to understanding the crucial differences between parasites and their host, and this is what is required for the discovery of new drug targets.

[1] The cut off used by Staley *et al.* was a 10^{-6} e-value, which is lower than the 55 bit-score used by Hartman *et al*., given *Giardia* and the two bacteria had small genomes.

1.2 Parasites involved in the project

The protozoans *Giardia lamblia*, *Trichomonas vaginalis* and *Plasmodium falciparum* are all obligate intracellular parasites of humans. These parasites cause the diseases giardiasis, trichomoniasis and malaria, respectively. These infections are amongst the leading causes of morbidity and mortality worldwide, and the nature of the complex life cycles of these organisms, as well as their highly adaptable gene expression mechanisms make it difficult to effectively treat infections caused by these protists (Adam 2001; Vedadi *et al.* 2007). Funding from the New Zealand Health Research Council (HRC) enabled this project to investigate using ESPs as conserved proteins to connect metabolism between humans and their protist parasites as a first step in uncovering new potential drug targets.

1.2.1 *Giardia lamblia*, a unique organism

The main parasite organism of this study is *Giardia lamblia* (also known as *Giardia intestinalis* or *Giardia duodenalis*, Figure 1). *Giardia* is a flagellated unicellular eukaryotic microorganism that commonly causes waterborne diarrheal disease in a variety of vertebrates, including humans (Adam 2001). It has two stages in its life cycle: cyst and trophozoite. The cyst is inert, and turns into a trophozoite, which is the vegetative form (Figure 2) after exposure to the acidic environment of the stomach. The complete cycle of *Giardia* cannot at present be replicated in laboratories. It was believed that *Giardia* reproduce asexually, however, recent studies indicate that homologues of genes specifically required for meiotic recombination are clearly present (Birky 2005; Logsdon 2008).

Figure 1. *Giardia* trophozoites as viewed by an electron microscope

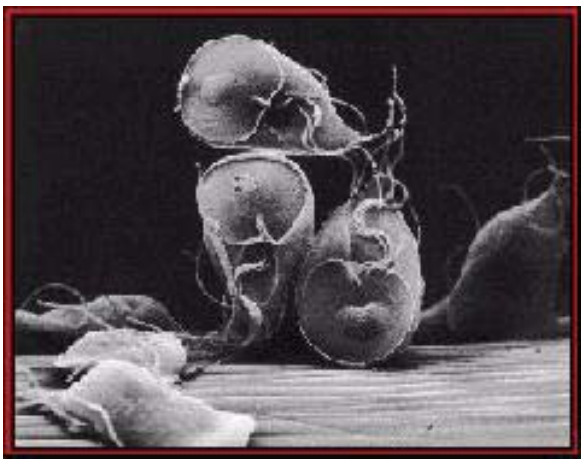


Figure reproduced from webpage: www.sierranaturenotes.com/naturenotes/Giardia.htm

Figure 2. *Giardia* life cycle

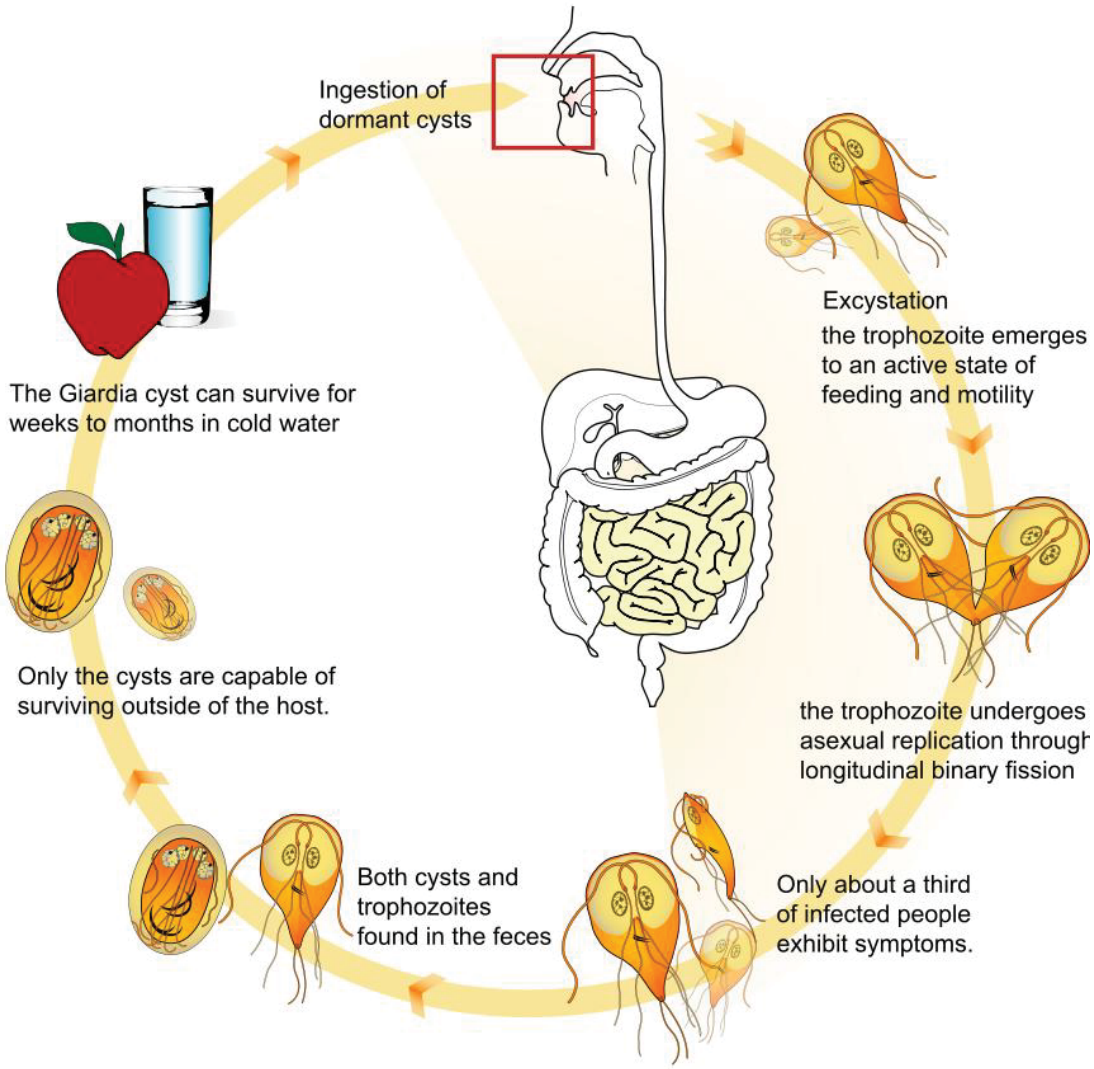


Figure is from web page: http://sharinginhealth.ca/pathogens/parasites/giardia_lamblia.html

New Zealand has a higher incidence rate of giardiasis than other developed countries (Snel *et al.* 2009) and the numbers of cases are on the rise. The reported cases of *Giardia* was 470 between January and March in 2009, this figure rose to 555 in the same quarter in 2010[2].One of the main drugs for treating *Giardia* infection is metronidazole (Mz), a synthetic 5-nitroimidazole (NI) derivative (Harris *et al.* 2001; Valdez *et al.* 2009). Metronidazole is activated when its 5-nitro group is reduced by ferredoxin that has in turn been reduced by pyruvate:ferredoxin oxidoreductase (PFOR), generating toxic free radicals, and these free radicals that cause lethal damage to the parasite. Humans have an alternative pathway to PFOR, the pyruvate dehydrogenase

[2]Source: http://www.stuff.co.nz/national/health/3811421/Concern-over-giardia-outbreak-as-cases-rise

complex, and will be less harmed by Mz. However, Mz treatment fails in 20% of patients (Upcroft *et al.* 2001) and there are other issues including developing resistance to 5-NI compounds from *Giardia* (Dunn *et al.* 2010), and that Mz is inactive against *Giardia* cysts (Adam 2001).

There are several other groups of compounds active against *Giardia*. Sodium nitrite is another respiratory inhibitor and it acts by destroying the iron-sulphur centre of PFOR (Adam 2001). Benzimidazoles act by interacting with the colchicine site in tubulin, thereby disrupting microtubules assembly and disassembly (Lacey 1988). This drug however is not very effective (Harris *et al.* 2001). Paromomycin acts by binding to a unique Giardial rRNA sequence, and inhibits protein synthesis (Harris *et al.* 2001). Quinacrine's mechanism of action is unclear, but reports suggest that quinacrine acts on either flavin components of some enzymes (Paget *et al.* 1989), or binds to DNA and inhibits nucleic acid synthesis (Thompson *et al.* 1993). Furazolidone's mechanism of action is also unclear with possibilities that it acts as electron acceptors of PFOR in a way similar to metronidazole, generating DNA damaging free radicals (Crouch *et al.* 1986), or by inhibition of DNA synthesis and completion of the cell cycle (Hoyne *et al.* 1989). Selective toxicity of all the above drugs is achieved through preferential absorption by the parasite, or minimally absorbed by the host intestine, and a higher dose also results in various unpleasant side-effects for the host (Harris *et al.* 2001). Due to the many weaknesses of current drugs treating infections, seeking treatments from a molecular biology angle is a potentially useful approach. The discovery and development of new therapeutics is important to expand the arsenal for controlling parasitic infection.

Table 1. Antigiardial drugs and their targets

Drugs	**Targets**
Metronidazole	pyruvate:ferredoxin oxidoreductase
Sodium nitrite	pyruvate:ferredoxin oxidoreductase
Benzimidazoles	colchicine site in tubulin
Paromomycin	rRNA
Quinacrine	DNA
Furazolidone	pyruvate:ferredoxin oxidoreductase or DNA

Giardia has two seemingly identical nuclei (hence it belongs to the group diplomonads), and the heterozygosity of genetic content of the two nuclei has been estimated to be less

than 0.01% (Morrison *et al.* 2007). Each nucleus contains five chromosomes (Adam 2001) which are slowly being assembled in genomic studies. The *Giardia* genome size is ~12 megabases (Mb) containing ~5000 protein coding genes (Aurrecoechea *et al.* 2009). The *Giardia* database GiardiaDB (http://www.giardiadb.org) (Aurrecoechea *et al.* 2009) provides the latest genomic resource for the organism. *Giardia*'s DNA organisation has most of the features expected for eukaryotic cells, as they are contained within linear chromosomes flanked by telomeres. The chromosomal DNAs are packed by Histone proteins (H2a, H2b, H3 and H4). The linker histone (H1) however is not present (Yee *et al.* 2007). It was suggested H1 is not needed since the genome of *Giardia* is small and gene-rich (77% of the genome are genes) (Morrison *et al.* 2007). Another significant protein missing in *Giardia* is myosin, which is a motor protein typically used during muscle contraction.

Giardia's reduced DNA synthesis, transcription, RNA processing and cell cycle machineries are often considered 'simple'. *Giardia* is largely anaerobic (Brown *et al.* 1998), with a 'limited' metabolic repertoire. For example, questions have been raised on the presence of the citric acid cycle and *de novo* purine and pyrimidine biosynthesis pathways (Morrison *et al.* 2007). In addition, *Giardia*'s amino acid and lipid metabolisms are also considered limited. With *Giardia* potentially having less redundancy in its metabolic pathways, it is important to understand which of these pathways it has in common with its host, and which are missing.

It is believed that *Giardia* is a basal eukaryote which diverged during the early days of eukaryotic evolution (Vanacova *et al.* 2003). This also makes *Giardia* an interesting organism for evolutionary studies since it has many distinguishing characteristics. Notably *Giardia* has no mitochondria, which lead some to believe that *Giardia* is an "archezoa". Cavalier-Smith first suggested the Archezoa theory (Cavalier-Smith 1987), which states free-living protists, called archezoa, were ancestors of eukaryotes. These ancient protists then acquired mitochondria and gave rise to modern eukaryotic cells. However, a true "archezoa" (i.e. a eukaryotic organism diverged prior to the endosymbiotic origin of mitochondria) however, has yet to be found (Brinkmann *et al.* 2007). The theory of *Giardia* being an archezoa was quickly dismissed by the finding of mitosomes (an organelle appeared to be descended from mitochondria) and some mitochondrial related proteins in its nuclear genome, which led to the suggestion that the *Giardia* (and other amitochondrial protists) have secondarily lost their mitochondria

(Roger *et al.* 1998). The Golgi apparatus is another organelle thought to be missing from *Giardia*, but again a reduced apparatus has now been found (Dacks *et al.* 2003). This organelle reduction is one of the reasons that *Giardia* as a representative of an early (perhaps the earliest) eukaryotic lineage (Morrison *et al.* 2007), is considered to have an evolutionary history distinct from other eukaryotes (Baldauf 2003).

Overall, the uniqueness and its impact on human disease have made *Giardia* a very interesting organism to study and it was chosen to be the main human parasite of this project.

1.2.2 *Trichomonas* and *Plasmodium*

Trichomonas vaginalis (Figure 3) and *Plasmodium falciparum* (Figure 4) are the two other parasites studied in the project. Comparison of the ESPs from these organisms was made with *Giardia* and human ESPs.

Like *Giardia*, *Trichomonas vaginalis* is also an amitochondriate belong to supergroup Excavata. The *Trichomonas* trophozoite is oval shaped and flagellated (Figure 3). *Trichomonas* also lacks mitochondria and necessary enzymes to conduct oxidative phosphorylation, and it also primarily has an anaerobic lifestyle (Seema *et al.* 2008).

Figure 3. Electron microscopy of *Trichomonas*

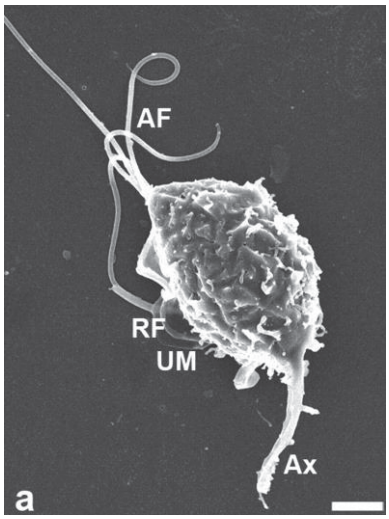


Figure reproduced from a research article (Lee *et al.* 2009).

Invasion of *Trichomonas* causes trichomoniasis, which is an extremely common sexually transmitted disease (Harp *et al.* 2011), with more than 160 million people worldwide infected by this protozoan annually. Trichomoniasis is treated with metronidazole (Mz), and tinidazole (Nanda *et al.* 2006). There are several cases of resistance to metronidazole reported in New Zealand (Lo *et al.* 2002).

Trichomonas has a large genome in comparison with other eukaryotic parasites (e.g. *Giardia* with 12Mb). The genome size of *Trichomonas* is ~160 Mb, organised into six chromosomes, with ~60,000 protein genes and ~1100 RNA coding genes (Aurrecoechea *et al.* 2009). The *Trichomonas* database TrichDB

(http://www.trichdb.org) (Aurrecoechea *et al.* 2009) is the database dedicated to latest genomic information for the organism.

Plasmodium falciparum invades human red blood cells causing the tropical disease known as malaria (Cowman *et al.* 2002). Transmission of these parasites to humans occurs via *Anopheles* (mosquito) vectors. Malaria has a wide geographic distribution, which puts almost half of the world's population at risk of contracting this tropical disease (Aurrecoechea *et al.* 2009). The year 2010 saw an estimated 216 million cases of malaria including 655,000 deaths worldwide (http://www.cdc.gov/MALARIA/). Although not endemic in NZ, travellers from overseas can come back to the country with malaria.

Figure 4. *Plasmodium* (trophozoite ring form) inside erythrocytes

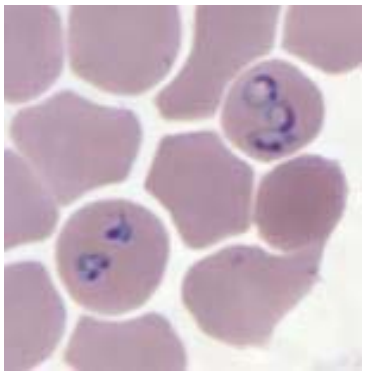


Figure from web page: http://dpd.cdc.gov/dpdx/html/Frames/M-R/Malaria/falciparum/body_malariadffalcring.htm

Plasmodium has a very complex life cycle, which takes it through multiple stages and multiple cell types (in the vertebrate host's liver, erythrocytes and in the arthropod vector) during which the parasite undergoes multiple developmental changes (Cowman *et al.* 2006). The trophozoite stage of its life cycle is illustrated in Figure 4. Some species of *Plasmodium* are also capable of invading other mammals, as well as birds and lizards (Cowman *et al.* 2006). Many antimalarial drugs are available, including chloroquine, amodiaquine and artemisinins (White 2004). However, the extensive deployment of these antimalarial drugs, in the past fifty years, has provided a tremendous selection pressure on human malaria parasites to evolve mechanisms of resistance.

Distant from *Giardia* and *Trichomonas*, *Plasmodium* belongs to the supergroup Chromalveolata, and the phylum Apicomplexa. The genome of *Plasmodium* is organised into 14 chromosomes (Gardner *et al.* 2002). This ~24 Mb genome is extremely AT rich (~80%) and contains ~5000 genes (Gardner *et al.* 2002). The *Plasmodium* database PlasmoDB (http://www.plasmodb.org) (Aurrecoechea *et al.* 2009) offers the latest genomic database for the organism.

1.2.3 Current RNA work on *Giardia* and *Trichomonas*

Some non-coding RNAs (ncRNAs) have also been shown to be eukaryotic signatures in the same manner as the ESPs. RNA processing has become an increasingly important area of research, and numerous ncRNAs have been uncovered in all the model eukaryotic organisms. The term ncRNAs will refer here to those transcribed RNAs which are not translated into proteins. These ncRNAs include transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), as well as small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), small interfering RNAs (siRNAs) and piwi-interacting RNA (piRNAs). The functions of these RNAs are summarised in Table 2.

Table 2. Some types of ncRNAs

Name	Abbreviation	Function
Micro RNA	miRNA	Processed by Dicer, associates with RISC complex and regulates gene expression, single stranded precursors
Small interfering RNA	siRNA	Processed by Dicer, target mRNAs, aids gene silencing, double stranded precursors.
Small nuclear RNA	snRNA	Involved in RNA splicing, regulating RNA polymerase
Small nucleolar RNA	snoRNA	Guides chemical modifications of snRNAs, rRNAs and tRNAs
Piwi interacting RNA	piRNA	Chromatin regulation and transposon silencing
Ribosomal RNA	rRNA	Ribosomal components, involved in translation
Transfer RNA	tRNA	Transfers of amino acids, involved in translation, decodes mRNA into proteins

Giardia possesses some different RNA processing components from those found in other eukaryotes (Chen *et al.* 2007). One major role of some ncRNAs is its involvement in RNA interference (RNAi), which is a system by which RNA is used to control the expression of genes. The high divergence of *Giardia* ncRNAs has created an interesting field to study. I am fortunate that I am associated with a research group at Massey University that is investigating ancestral RNAs and have done computational analysis and sequencing of ncRNAs from *Giardia* and *Trichomonas* (Chen *et al.* 2007; Chen *et al.* 2008; Chen *et al.* 2009). In 2007, Chen *et al.* found 31 ncRNAs from *Giardia* (Chen *et al.* 2007). Although only 5 of these have been characterised, spliceosomal RNA analysis found the uridine-rich snRNA U5 snRNA in small quantities. U6 snRNA, the

most conserved spliceosomal snRNA across all the eukaryotes was not found at the time, despite an intense effort to search for it. However, later on Chen *et al.* found a U2 candidate, which subsequently helped to identify candidates for U6, as well as U1 and U4.

Using Illumina sequencing (aka Solexa sequencing) and genome wide analysis of small RNAs from *Giardia*, Chen *et al.* identified another 10 miRNA candidates from *Giardia* and 11 from *Trichomonas* (Chen *et al.* paper in preparation). In addition, Chen *et al.* also characterised five unusual long tandem repeated double stranded RNAs that were named Girep-1 to Girep-5. Sequence alignments confirmed these five RNAs belong to the same group, and they share high degrees of sequence similarity with a number of variant-specific surface proteins (VSPs). VSP gene expression is crucial for the surface antigenic variation of *Giardia* trophozoites. By displaying different VSPs on the surface, *Giardia* is able to evade the host's immune system (Nash *et al.* 2001). Chen *et al.* suspected Gireps are precursor siRNAs and have a strong potential to be involved in the regulation of VSP expression.

These results overall suggest that ncRNAs and especially RNAi-associated ncRNAs existed in the last common ancestor of eukaryotes, and like ESPs, they are a characteristic unique to eukaryotes. The ancestral proteins are likely to have been under ancestral regulation. In addition, ncRNAs are becoming increasingly useful as markers for disease diagnostics in humans (e.g. (Fanini *et al.* 2011; Ferracin *et al.* 2011)). Understanding how ESPs as essential proteins are regulated in *Giardia* (and *Trichomonas*) could possibly uncover other differences in host and parasite metabolism.

Comparing the ancestral research on RNAs with ancestral proteins (ESPs) is a part of a bigger project that examines ncRNA in basal eukaryotes. By using some of the same techniques from the ESP work (e.g. databases, Perl scripts, comparing genomic data), and combining the ESP results with the *Giardia* and *Trichomonas* ncRNA results, I was also able to participate in this wider project.

1.3 Thesis structure

This PhD project used an integration of genomic, phylogenetic and biological approaches, aimed to gain an understanding of the molecular and cellular differences between hosts and parasites, with the main focus on protist parasite *Giardia lamblia* and humans. Two other parasites *Plasmodium falciparum* and *Trichomonas vaginalis* have

also been analysed as comparisons. This project calculated ESPs of human and parasites, and by grouping these essential proteins and comparing the differences between the host and parasites, interesting proteins can be identified for further research purposes, especially those that can be used as potential drug targets.

ESPs can potentially be very useful in further phylogenetic studies due to their conservation in virtually all eukaryotes and their consistent slow evolution rate. The ESP datasets for these eukaryotic parasites were then examined to assess their evolutionary significance, metabolic function and possible interactions with ncRNAs.

1.3.1 Generating a new ESP dataset – Chapter 2

The previously calculated ESP datasets (mainly Hartman *et al.*'s) can be considered a little outdated. Now that many more genomes with much better annotation are available, ESP datasets for the host organism *Homo sapiens*, parasitic organisms *Giardia lamblia*, *Plasmodium falciparum* and *Trichomonas vaginalis* were re-calculated. This was a crucial first step to obtain more accurate sets of ESPs.

The procedure used here was similar to that of the previous work (see section 1.1), but more species were included and parameters were analysed in-depth to ensure the correct selection of proteins. Selecting suitable organisms was an important part of the procedure, because ideally all branches of the three domains (Archaea, Bacteria and Eukarya) should be covered, and yet the total number of species should not form a computational barrier (i.e. too many species and the analysis takes too long). The methodology for obtaining the ESP data is described as a protocol for future studies, so that when genomes are updated or if new genomes become available, new ESP datasets can be readily obtained.

1.3.2 Phylogenetic analysis using ESPs – Chapter 3

The current taxonomic system classifies eukaryotes into five supergroups based on molecular and morphological/cell-biological evidence (Keeling *et al.* 2005; Keeling 2007). The five supergroups are Unikonta (note: this supergroup is often divided into Opisthokonta and Amoebozoa (Simpson 2003)), Plantae, Rhizaria, Chromalveolata, and Excavata. This system is significantly advanced from the classic "six kingdoms" eukaryotic tree of the 1980s, but is still not without controversy. For example the

monophyly of supergroups Chromalveolata, and Excavata is constantly under debate (Parfrey *et al.* 2006).

Previously, molecular based phylogenetic studies between distantly related species (such as between different supergroups of eukaryotes) was done by using 18S rRNA, or based on a single gene when these happened to be sequenced. This approach can tend to give misleading results if the gene in question has undergone rates of change different from what is 'typical' for that species, or if there if there has been more than one change per site. A wider variety and larger quantities of molecular data is needed to accurately build the eukaryotic trees that can correctly place protist species such as *Giardia*.

ESPs are a set of proteins conserved in all eukaryotes, so this potentially makes them great candidates for phylogenetic studies. ESPs provide large number of proteins to build phylogenetic relationships, and to ensure there is no missing data from any eukaryotic taxa. We would also expect the ancient proteins to have a slow and constant evolutionary rate (to keep them conserved), which also make them ideal for studying phylogenetic relationships between distant organisms. Therefore the potential for ESPs as candidates for phylogenetic analyses was examined in this project. The phylogenetic relationships between the 18 eukaryotic organisms used during ESP calculation process were analysed. Two methods were employed to deal with the large number of discrete protein sequences. The first approach, by using consensus networks, is where individual trees built from each protein are combined into a single network to show the consensus of all signals. The second approach was by concatenating all ESP sequences of each taxon, and then constructing a tree from the concatenated sequences. The first approach had not yet been tested for such a large amount of data. The second approach has been employed in other studies (e.g. (Hampl *et al.* 2009)), but their studies have selected genes with homologues not present in some of the organisms, thus their alignments had missing data. By using ESPs, there should be very little or no missing data since all orthologous proteins should be present in each taxon. In addition, the rate of evolution for every ESP should be more consistent (by being slow) and hence could be more reliable than using other sets of proteins.

However, there are some difficulties with the ESP approach. There are very few excavates and chromalveolates genomes completely sequenced, and thus taxa from these two supergroups used in this analysis (*Giardia* and *Phytophthora*) are likely to form long branches. In addition, many species of these two supergroups have undergone

reductive evolution, and even with their genomes completed, some ESPs could be missing from these genomes. In light of this, a second study using an established metazoan group of species was also conducted, so that the results could be compared to other published studies to indicate the accuracy of using ESPs for phylogenetic analysis. Establishing ESPs as good candidates in a closely related group of species is important as they could hold a greater potential than just being used to study excavates and chromalveolates.

1.3.3 Metabolic analysis of *Giardia* – Chapter 4

ESPs are conserved and hence essential proteins are expected to predict which metabolic pathways are conserved (or otherwise) between host and parasite. This analysis was important (and hence part of the funding requirements) because differences in metabolism is what makes drugs against protists effective. The standard drug treatment metronidazole (Mz) knocks out a key enzyme in *Giardia* metabolism, but harms humans less because there are alternative pathways. However, Mz and other protist-targeting drugs are not always effective and can have severe side effects, so there is a real need to find more targets for drug development. *Giardia* is evolutionarily distant from other eukaryotes and thus relatively little is known about its core metabolic pathways. KEGG (Kyoto Encyclopaedia of Genes and Genomes, http://www.genome.jp/kegg), the widely referenced site for providing information of metabolism does not yet include many enzymes from *Giardia* species, therefore developing a new method to look at metabolic pathways using the information from other organisms is needed. *Giardia*'s core sugar metabolism was analysed to develop this new approach utilising data from the ESP calculations and metabolic information from KEGG.

Here by comparing *Giardia* proteomes with known enzymes from other species, candidates for enzymes in the glycolysis pathway, as well as some enzymes involved in the TCA cycle and oxidative phosphorylation and amino acid metabolic pathways were identified, and differences between the parasitic and host enzymes observed. The enzymes from the *Giardia* glycolysis pathway have been previously reported to be more similar to those from bacteria (Morrison *et al.* 2007), and this was also investigated in this study. By identifying in more detail enzymes that are different in parasites in comparison with those found in mammals, the host organisms for *Giardia*, there is a

real possibility that these bacteria-like enzymes could indeed be novel drug targets for treating *Giardia* infections.

1.3.4 Small RNAs in *Giardia* and *Trichomonas* – Chapter 5

Both ESPs and RNAi (RNA interference) are found in all branches of eukaryotes, thus they are expected to be present in the last eukaryotic ancestor. So will there be any correlation between the ancient proteins and an ancient mechanism? Presence does not necessarily mean correlation and the connection between proteins and their possible regulation is another avenue in which humans may differ from their protist parasites.

RNAi involves small RNA molecules including micro RNAs (miRNA) and small interfering RNAs (siRNA). Typical miRNAs and siRNAs which are processed typically with the proteins Dicer and Argonaute are ~21-22 nucleotides (nt) in length. However, given that the Dicer protein from *Giardia* has been reported to cut differently (25-27 nt) (MacRae *et al.* 2006), we expect that the small ncRNAs including miRNAs from deep branching eukaryotes might be different from those found in metazoans and plants. Even essential proteins need to be regulated (i.e. to be in step with the cell cycle) and thus there is the question as to whether ESPs show any trends in their regulation. Comparing miRNAs between all eukaryotes was beyond the scope of this project so here I used data that was being produced by another study from members of our group.

Illumina sequencing of *Giardia* and *Trichomonas* small RNAs was performed as part of a study of ncRNA evolution in eukaryotes at Massey University, Palmerston North (Chen *et al.* 2009). Being ncRNA data it was assembled differently than what is typical for genomic sequencing. Computational analysis and sequencing was also done by the group previously.

Here a different way of analysing from the above is presented. By working with unassembled data, large quantities of smaller RNAs are harvested, and interesting information about the ncRNA of the two parasites could be discovered. The data was used in the overall analysis with the examination of the putative Dicer-processed ncRNAs (i.e. 26-27nt). Where these ncRNAs were located in relation to the coding region of the gene was analysed in relation to whether there was a trend for 3'UTR or antisense coding region based regulation. In the end, this work was taken beyond just looking at ESPs and included all *Giardia* proteins. During the course of the ncRNA analysis another group of ultra-small RNAs was discovered in *Giardia*. As co-

discoverer I will touch on this group only briefly since this is currently being researched in more detail.

1.3.5 Summary

The focus of this project is the study of ESPs, an interesting group of proteins which delineates eukaryotes from archaea and bacteria. The properties make these proteins interesting to study is its consistency and they have crucial functions in eukaryotes. ESPs can be used to directly or indirectly guide our way to discovering key differences between humans and their protist parasites. These differences in the past have led to the development of some drugs to combat infection but it is clear that new drugs are needed. The calculation and analysis of this unique set of proteins is but the first step in this pathway of discovery, but it is an important one to aid in the uncovering of whatever potential lies in the genomics currently being undertaken in this area. Studying ESPs will help our understanding of eukaryotic evolution, as they give insights on how eukaryotes first became distinguishable from other prokaryotes at the early days of their evolution. Genome reduction has also played an important role in parasitic evolution (Morrison *et al.* 2007). Even core proteins such as ESPs are often missing from some parasites, and the loss of ESPs can give insights to parasitic reductive evolution. In addition, ESPs can serve as guides to analyse eukaryotic phylogenetics, due to their conservation in all eukaryotic organisms. ESPs can be combined with other studies of other eukaryotic features, such as RNAi and intron splicing. Relationships between these ancient mechanisms may hold many interesting facts about eukaryotes. The thesis concludes with an overall conclusion and a look at future perspective in Chapter 6.

Chapter 2: Collecting Eukaryotic Signature Proteins

2.1 Introduction

In this chapter, updated Eukaryotic Signature Protein (ESP) sets for each of the three parasites, *Giardia lamblia*, *Plasmodium falciparum* and *Trichomonas vaginalis*, and the host for these three parasites – human (*Homo sapiens*) have been characterised. An ESP dataset for *Giardia* has been calculated before, and a summary of previous eukaryotic signature protein (ESP) work is given in Chapter 1, section 1.2.1 (pages 2-5).

The re-calculation was needed as few genomes were available at the time of the analysis by Hartman *et al.*. Then, only 44 prokaryotic and five eukaryotic genomes were used to represent all major groups of prokaryotes and eukaryotes. At present there are many more options. In addition, rather than start the search with the yeast proteome, as Hartman *et al.* did, I used a more straightforward approach to start directly with the proteome of the organism of interest (either human or the parasites). This is because the parasites' genomes are better annotated now than at the time Hartman *et al.* performed their research. In addition, *Giardia* is the focal organism in this project, and a set of *Giardia* ESPs serve better for the purpose of host and parasite comparisons than a set of yeast proteins.

It is very difficult to obtain an exact list of ESPs due to factors such as distant homologues or sequencing error in the proteome of certain organisms (discussed later in Section 2.3.1). The new ESP dataset is believed to be much more precise than that of Hartman *et al.*'s. This will be a set of functionality important proteins, since all eukaryotes maintain them; this will also be a set of evolutionarily conserved proteins, because they are not present in any prokaryotes, and they may hold the key for this debatable transition from relatively simple prokaryote cell to more complex eukaryotic cell. The new ESPs list will serve as primary work for the remainder of the thesis. In future, this set of ESPs is planned to be uploaded to GiardiaDB (http://www.giardiadb.org) for public use.

2.1.1 BLAST statistics

BLAST is used extensively in this chapter for homology searches, thus it is worth mentioning the algorithm of the software. BLAST is a suite of software developed by Altschul *et al* at the National Center of Biotechnology Information (NCBI) (Altschul *et al.* 1990). It is a heuristic program, thus not guaranteed to find every local alignment that passes its reporting criteria, and there is an array of parameters that control the shortcut it takes. These parameters influence BLAST's trade-off between speed and sensitivity, and they are of the least importance for a user to understand because, except for the occasional appearance or disappearance of a weak similarity, they do not greatly affect the program's output. What is important is the correct choice of scoring systems and interpretation of statistical significance (Korf *et al.* 2003).

BLAST is a searching tool for significant alignments of a query sequence (or sequences) from a database. The BLAST program "blastall" has five running modes designed for the comparison of proteins with proteins, nucleotides with nucleotides and combinations of the two. The program used for ESP calculation is "blastp", which compares protein query sequences with protein sequences in a database. The sequence alignments are performed using the default scoring matrix BLOSUM62 (blocks substitution matrix 62). The BLOSUM matrix was constructed by extracting ungapped sequence segments (blocks) from a set of multiply-aligned protein families. Then, counting the relative frequencies of each amino acid and its substitution probabilities, scores for each substitution were calculated. The number 62 indicates that all the sequences clustered have at least 62 percent similarity (Henikoff *et al.* 1992). BLOSUM is empirical and derived from a larger dataset and it is preferred over the PAM (percent accepted mutation) matrix (Korf *et al.* 2003).

For each significant alignment produced by a BLAST search, a score and an expect value (e-value) are given. The score is computed from the scoring matrix and gap penalties (a cost for inserting a gap in a sequence alignment). A higher score indicates greater similarity. Raw score has no unit and it can be normalised, and normalisation is indicated by the unit "bits" added to the raw score.

The e-value indicates the number of alignments expected at random given the size of the search space and the score of alignments. The lower the e-value the less likely it is that this similarity is random.

The Karlin-Altschul equation: $E = kmne^{-\lambda S}$

The above equation states the relationship between e-value, query size (m), database size (n) and normalized score (λS) during a sequence database search. The constant (k) undergoes a minor adjustment depending on the position of optimal alignment and usually has little effect. The e-value is inversely and exponentially related to the normalized score (λS). This means a small increase in score will lead to a large decrease in e-value, and therefore a more significant hit. Query size (m) and database size (n) do not directly refer to the actual length of the query or database, rather they refer to the "effective length". The sequence must reach a particular length before it can produce an alignment with a significant e-value; this minimum length is referred to as the "expected HSP (high scoring pair) length". The "effective length" is the actual length minus the "expected HSP length". In a large search space, the "effective length" of query (m) may be negative; in this case, m will be set to 1/k to remove the effect of short sequence to the e-value (Korf *et al.* 2003).

2.2 Material and methods

2.2.1 Selection of species for analysis

Ideally, the more species involved in the eukaryote-wide search, the more robust the ESP dataset would be; on the other hand, the time the analysis takes also increases as the number of species increase. Therefore, the number of species used has to be compromised to an extent. The ESP results will be biased due to his species selection so in order to minimise this bias, selected species should cover as wide a range of organisms as possible.

The "interactive tree of life (iTOL) (Letunic *et al.* 2007) was downloaded, because this is a tree of all species with a complete genome. Although this tree is slightly out of date, it provided a very good starting point to choose species for the ESP calculation described here. Using this tree, species which would best represent major branches of bacteria and archaea were chosen for analysis. A detailed description of species selection from each of the three kingdoms follows.

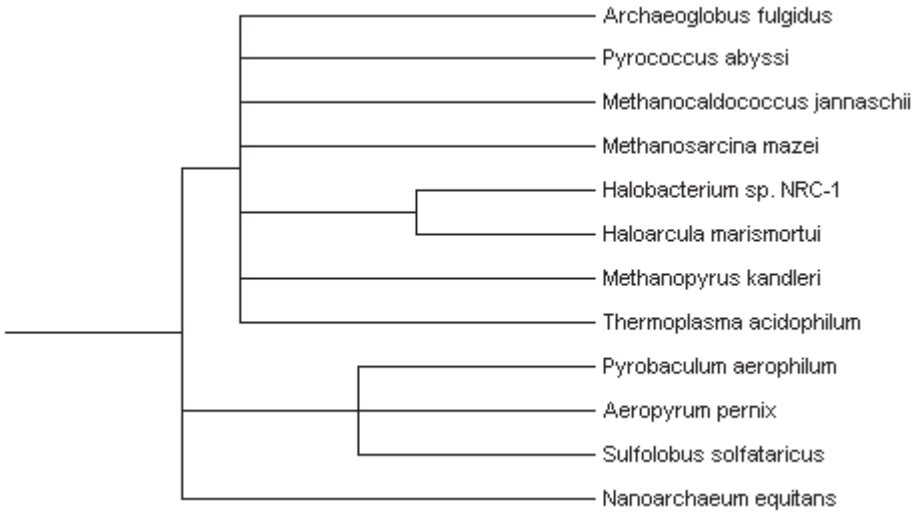
From Archaea, most of the culturable and well-investigated species are members of two main phyla, the Euryarchaeota and Crenarchaeota (Robertson *et al.* 2005). Recently a

new phylum also has been described, which is the Nanoarchaeota. This phylum currently has only one representative, *Nanoarchaeum equitans* (Huber *et al.* 2002). Eight Euryarchaeota and three Crenarchaeota species, as well as *Nanoarchaeum equitans* were selected for the analysis (see Table 1 for a complete list of archaeal species used). A guide tree (Figure 1) was also created by using the iTOL utility, and TreeViewX (Page 2002) was used to visualise the this tree. Please note this guide tree and other guide tree in Figure 2 are only acceptable as indicators of phylum coverage rather than true phylogenetic trees, which would be time consuming to construct and are always controversial.

Table 1. List of archaeal species used in study

Species	**Phylum**	**NCBI Taxonomy ID**
Sulfolobus solfataricus P2	Crenarchaeota	2287
Pyrobaculum aerophilum str. IM2	Crenarchaeota	13773
Aeropyrum pernix K1	Crenarchaeota	56636
Thermoplasma acidophilum DSM 1728	Euryarchaeota	2303
Pyrococcus abyssi GE5	Euryarchaeota	29292
Methanosarcina mazei Go1	Euryarchaeota	2209
Methanopyrus kandleri AV19	Euryarchaeota	2320
Methanocaldococcus jannaschii DSM 2661	Euryarchaeota	2190
Halobacterium sp. NRC-1	Euryarchaeota	64091
Haloarcula marismortui ATCC 43049	Euryarchaeota	2238
Archaeoglobus fulgidus DSM 4304	Euryarchaeota	2234
Nanoarchaeum equitans Kin4-M	Nanoarchaeota	160232

Figure 1. Phylogenetic relationship of selected archaeal species



This tree was created from iTOL, TreeViewX was used for creating an image view. This is a guide tree, not an accurate depiction of the true phylogenetic relationship, branch lengths are suggestive only.

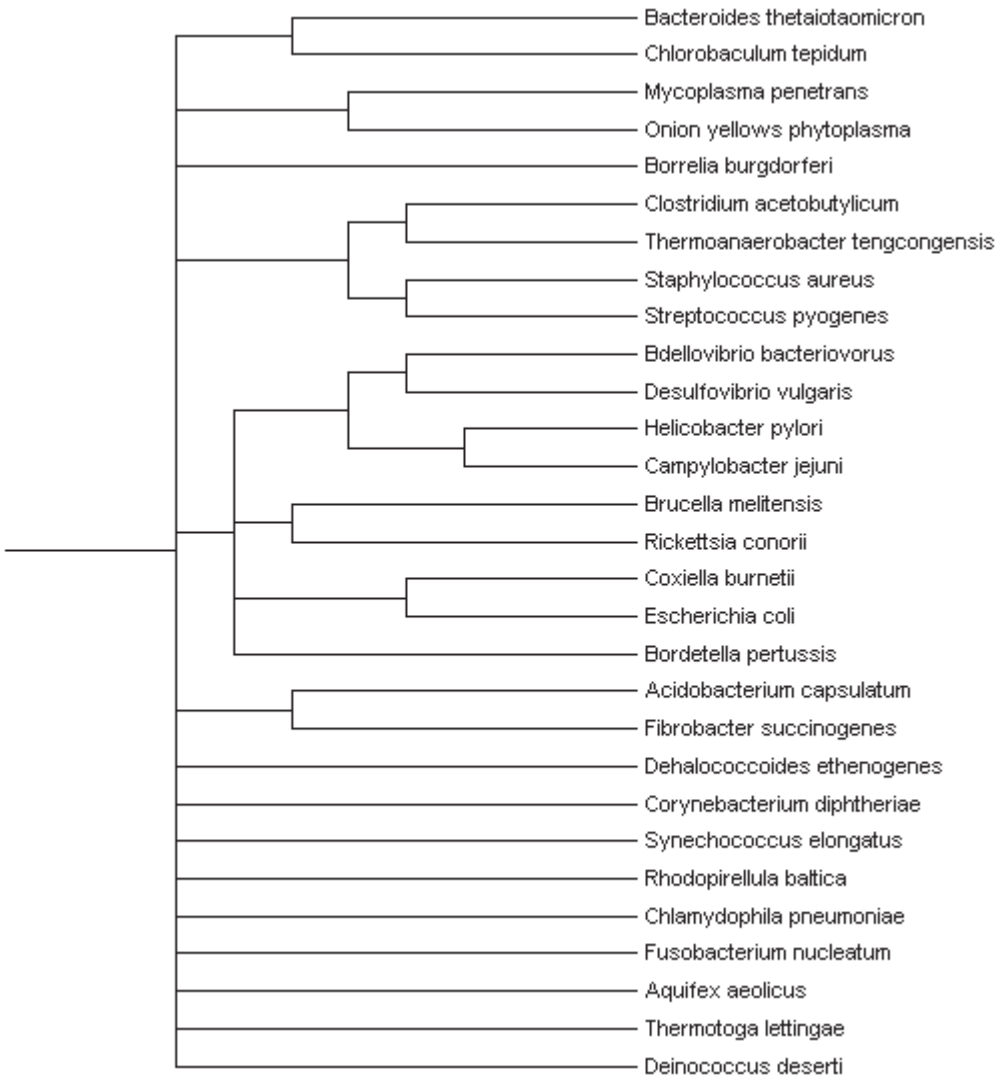
From the Bacterial kingdom, culturable bacteria are divided into 29 phyla (Euzeby 1997). Most phyla had at least one representative species in this analysis. Some phyla did not have any species with genomic sequencing information, for example, the phylum Caldiserica (which has only a few genes sequenced for few species), thus they were not used in the study. Some phyla had more than one species available. For example, the phylum Firmicutes contains a diverse groups of bacteria, as indicated by iTOL, thus four species (*Clostridium acetobutylicum, Staphylococcus aureus, Streptococcus pyogenes, Thermoanaerobacter tengcongensis*) from this phylum were used; two species from phylum Epsilonproteobacteria were included, *Helicobacter pylori* and *Campylobacter jejuni*. As mentioned earlier, the addition of these extra species is not expected to harm the results since including more species is beneficial. Planctomycetes have some eukaryotic characteristics (Fuchsman *et al.* 2006), which could confuse results in the initial screening, and therefore *Rhodopirellula baltica*, a species of planctomycete was excluded from the ESP protein analysis for now. Given its characteristics it is a species of interest for later work. See Table 2 for a complete list of bacterial species used, see Figure 2 for the guide tree.

In the end, 28 bacterial and 12 archaeal species were selected, which is a diverse selection of prokaryotic organisms, all of these proteomes were downloaded from NCBI (http://www.ncbi.nlm.nih.gov).

Table 2. List of eubacterial species used in study

Species	**Phylum**	**NCBI Taxonomy ID**
Acidobacterium capsulatum ATCC 51196	Acidobacteria	33075
Corynebacterium diphtheriae NCTC 13129	Actinobacteria	1717
Brucella melitensis 16M	Alphaproteobacteria	29459
Rickettsia conorii str. Malish 7	Alphaproteobacteria	781
Aquifex aeolicus VF5	Aquificae	63363
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes/Chlorobi	818
Chlorobium tepidum TLS	Bacteroidetes/Chlorobi	1097
Bordetella pertussis Tohama I	Betaproteobacteria	520
Chlamydophila pneumoniae J138	Chlamydiae/Verrucomicrobia	83558
Dehalococcoides ethenogenes 195	Chloroflexi	61435
Synechococcus elongatus PCC 6301	Cyanobacteria	32046
Deinococcus deserti VCD115	Deinococcus-Thermus	310783
Bdellovibrio bacteriovorus HD100	Deltaproteobacteria	959
Desulfovibrio vulgaris DP4	Deltaproteobacteria	881
Campylobacter jejuni RM1221	Epsilonproteobacteria	197
Helicobacter pylori 26695	Epsilonproteobacteria	210
Fibrobacter succinogenes subsp. succinogenes S85 (project was incomplete)	Fibrobacteres	833
Clostridium acetobutylicum ATCC 824	Firmicutes	1488
Staphylococcus aureus subsp. aureus MW2	Firmicutes	1280
Streptococcus pyogenes M1 GAS	Firmicutes	1314
Thermoanaerobacter tengcongensis MB4	Firmicutes	119072
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	Fusobacteria	851
Coxiella burnetii CbuK_Q154	Gammaproteobacteria	777
Escherichia coli str. K-12 substr. MG1655 K12	Gammaproteobacteria	562
Rhodopirellula baltica SH 1	Planctomycetes	265606
Borrelia burgdorferi B31	Spirochaetes	139
Thermotoga letting*ae TMO*	Thermotogae	177758
Mycoplasma penetrans HF-2	Other Bacteria	28227
Onion yellows phytoplasma OY-M	Other Bacteria	100379

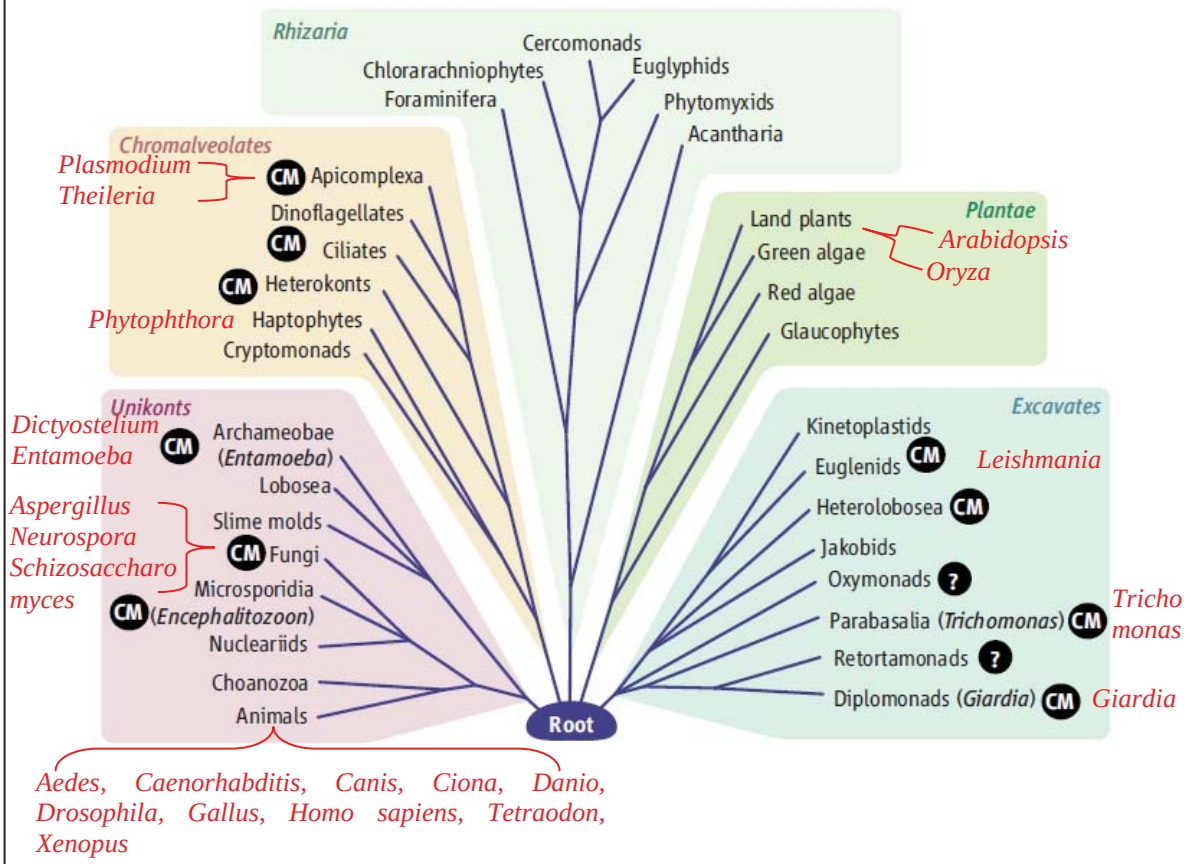
Figure 2. Phylogenetic relationship of selected bacterial species



This tree was created from iTOL, TreeViewX was used for creating an image view. This is a guide tree, not an accurate depiction of the true phylogenetic relationship, branch lengths are suggestive only.

The selection of eukaryotic organisms was based on the same principle as for the prokaryotic organisms, i.e. use species from as diverse organisms as possible. Based on this principle, 20 eukaryotic organisms in addition to the host (*Homo sapiens*) and parasites (*Giardia*, *Plasmodium*, *Trichomonas*) were downloaded. This representation of eukaryotic organisms is a significant improvement on that of Hartman *et al.* who only worked with five eukaryotic species.

Figure 3. Phylogenetic position of eukaryotic organisms chosen for this project



The positions of organisms chosen are indicated on the eukaryotic tree. CM indicates the presence of cryptic mitochondria (hydrogenosomes or mitosomes). A question mark indicates that no organelle has yet been found. This eukaryotic tree is from Keeling (Keeling 2007) with extra annotations added here.

The phylogenetic representation between the chosen eukaryotic organisms is shown in Figure 3. The monophyly of five supergroups of Eukaryotes, which includes Plantae, Rhizaria, Excavata, Opisthokonta and Amoebozoa (the last two branches are shown as "Unikonta" in Figure 3) are supported by recent phylogenetic studies (Parfrey *et al.* 2006; Keeling 2007; Hampl *et al.* 2009). However, clear evidence of Chromalveolata being monophyletic is still lacking. Some lineages are not currently represented by a complete proteome (e.g. Rhizaria and red algae) so they were not chosen for the calculation. Incompletely sequenced genomes could be detrimental to ESP calculations, because this could cause false negative results due to proteins not being sequenced rather than them not being present, so only those genomes at an advanced stage were included. Proportionally more animal (Metazoa) species have been chosen, because

there are more completely sequenced genomes in this branch and the use of more genomes provides more robust ESP datasets. Three organisms (*Entamoeba histolytica*, *Leishmania braziliensis* and *Theileria annulata*) possess what is called a "reduced genome". Although these three organisms are one of the few sequenced genomes in a large clade, due to the nature of their "reduced genomes", false negative results could be caused by their presence, and they were later removed from the ESP dataset calculation. This is also the reason behind my exclusion of the model yeast *Saccharomyces cerevisiae*, because it is also a reduced genome and not the most "typical" eukaryote (Drinnenberg *et al.* 2011), *Schizosaccharomyces pombe* was selected as a fungi representative.

Eukaryotic proteome databases were downloaded from best source available for each genome. For example, the *Giardia* database was downloaded from GiardiaDB (http://www.giardiadb.org), because this site provided the most updated version of the parasite's protein and annotations. Ensembl (http://www.ensembl.org) is a very trusted source for animal proteomes, and its online tool Biomart (Kinsella *et al.* 2011) allows users to perform useful tasks such as tracking down the nucleotide sequences with ease. Other sources used in the study included NCBI (http://www.ncbi.nlm.nih.gov), PlasmoDB (http://www.plasmodb.org), TrichDB (http://www.trichdb.org), Dictybase (http://dictybase.org/), the Broad Institute (http://www.broadinstitute.org), AspGD (http://www.aspgd.org), Swiss-prot (http://au.expasy.org/sprot) and the Sanger Institute (http://www.sanger.ac.uk) (see Table 3 for more details).

Table 3. List of Eukaryotic species used in study

Species	Supergroup	NCBI Taxonomy ID	Notes
From Ensembl			
Aedes aegypti	Opisthokonta	7159	
Caenorhabditis elegans	Opisthokonta	6239	
Canis familiaris	Opisthokonta	9615	
Ciona intestinalis	Opisthokonta	7719	
Danio rerio	Opisthokonta	7955	
Drosophila melanogaster	Opisthokonta	7227	
Gallus gallus	Opisthokonta	9031	
Homo sapiens	Opisthokonta	9606	Release 59
Mus musculus	Opisthokonta	10090	
Tetraodon nigroviridis	Opisthokonta	99883	
Xenopus tropicalis	Opisthokonta	8364	
From Dictybase			
Dictyostelium discoideum	Amoebozoa	44689	
From Broad Institute			
Neurospora crassa	Opisthokonta	5141	
Phytophthora infestans	Chromalveolata	4787	
From NCBI			
Arabidopsis thaliana	Plantae	3702	
Oryza sativa	Plantae	4530	
Schizosaccharomyces pombe	Opisthokonta	4896	
From Aspgd			
Aspergillus nidulans	Opisthokonta	162425	
From the Sanger Instititue			
Entamoeba histolytica	Amoebozoa	5759	Reduced genome
Leishmania braziliensis	Excavata	5660	Reduced genome
Theileria annulata	Chromalveolata	5874	Reduce genome
From Giardiadb			Version 1.3
Giardia lamblia	Excavata	5741	Reduced genome
From Plasmodb			Release 6.5
Plasmodium falciparum	Chromalveolata	5833	Reduced genome
From Trichdb			Version 1.1
Trichomonas vaginalis	Excavata	5722	Reduced genome
From Swiss-prot			
Homo sapiens		9606	

2.2.2 ESP calculations

Basic local alignment search tool (BLAST) (Altschul *et al.* 1990) is set of programs for searching homologous proteins or DNA sequences (Altschul *et al.* 1990). BLAST employs a heuristic approach and therefore has adequate speed. ESP calculations only take a day for parasites and a week for human using this well established method. BLAST also has an option which allows the user to select a different output format. The tabular format (with BLASTall parameter -m 8) is a useful format, because Perl scripts can easily work through these files to find significant matches and list accession numbers of appropriate matches. The BLAST program BLASTall using "blastp" as comparison option was used for all protein comparisons. Although there is now a new version of blast available from NCBI (BLAST+ (Camacho *et al.* 2009)), this work was done with the previous standard BLAST. I see no issues with using the new BLAST+ for ESP calculations in the future.

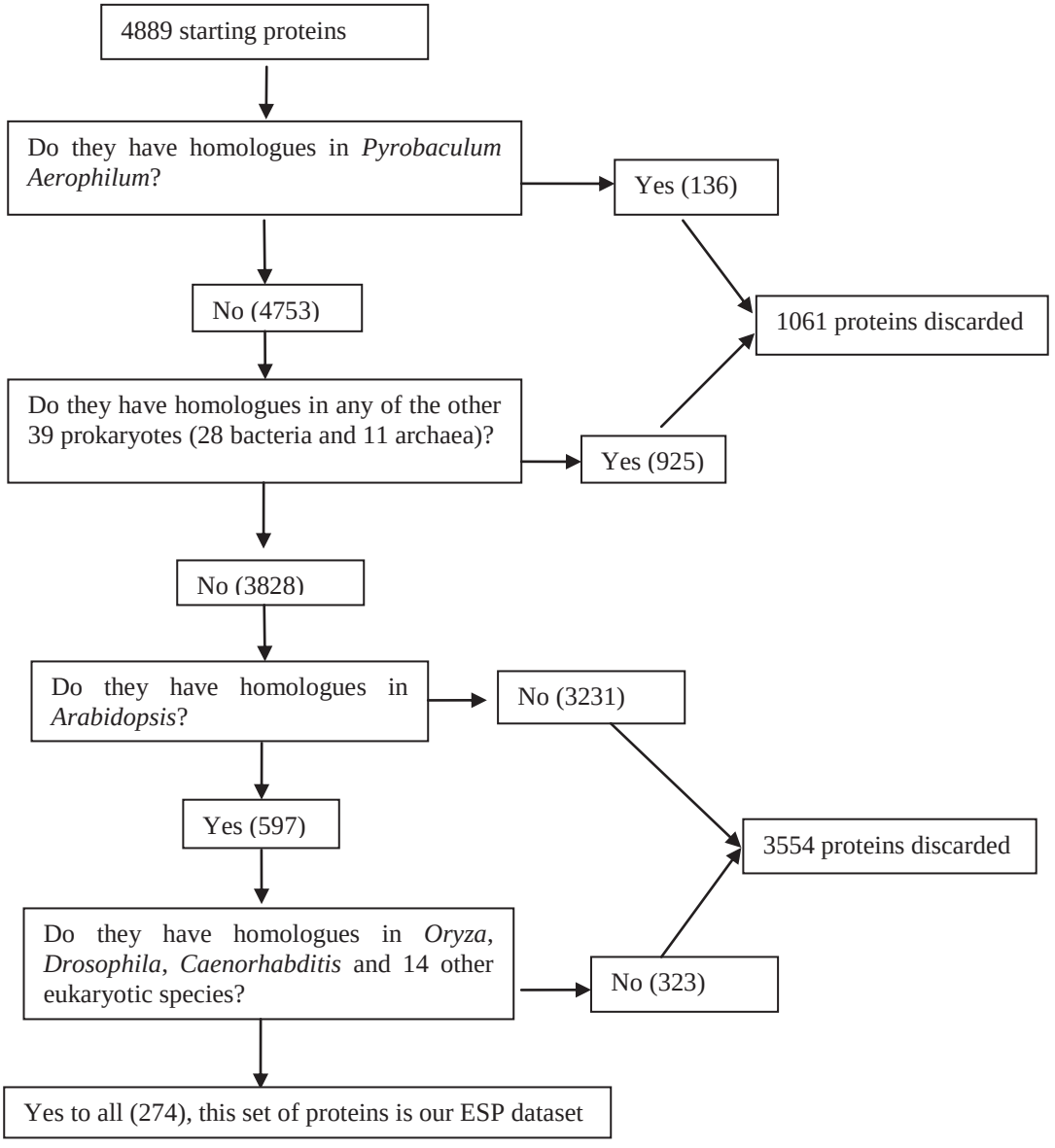
ESP datasets were calculated for the three parasites (*Giardia lamblia*, *Plasmodium falciparum* and *Trichomonas vaginalis*) and their host (*Homo sapiens*), under the following procedure (Figure 4):

The analysis began with all annotated proteins of the organism to be analysed (either *Giardia*, *Trichomonas* or human proteome), first proteins that had homologues in any of the 28 bacterial and 12 archaeal species were discarded; then proteins that did not have homologues in any of the 17 eukaryotic species were removed. The remaining proteins are termed ESPs. Initially, the ESP calculation procedure also included screening against *Rhodopirellula baltica* (a prokaryote with some eukaryotic characteristics, a species of Planctomycetes) and three reduced eukaryotes, *Entamoeba histolytica*, *Leishmania braziliensis* and *Theileria annulata*. These steps were decided to be excluded because they may cause false results. In all ESP calculations, BLAST hits with a bit-score ≥ 55 were considered as "homologues". The cut-off has been modified to test the robustness of ESPs (see section 2.3.3).

With this procedure, incomplete archaeal and bacterial proteomes (i.e. since some sequencing projects are still in progress) were still useful, because proteins with homologues in this prokaryote's genome could still be excluded. Incomplete eukaryotic genomes, however, can give false negatives, as some true ESPs may be excluded simply because their homologues were not listed in the incomplete genome it BLASTed against

or because sequence was not yet available. Note: "incomplete genome" denotes to an organism whose genome sequencing is not yet finished, different from "reduced genome", which means the genome sequencing might or might not be finished, but due to the parasitic life style of the organism, its genome have shrank and some genes have been lost through its evolution.

Figure 4. Procedure used for calculating ESPs



Note: numbers in brackets indicate number of proteins left from the calculation for *Giardia* ESPs.

Figure 4 is a flow diagram of the procedure and results used to calculate *Giardia* ESPs. Procedures for collecting *Plasmodium and Trichomonas* ESPs are the same as that of *Giardia*, except the search was started with 5446 *Plasmodium* proteins from PlasmoDB,

and 59672 *Trichomonas* proteins from TrichDB respectively. There were two sets of human ESPs calculated, the first set used Swiss-Prot human proteome database (20322 proteins) for the calculation, and second set used the Ensembl human proteome (79063 proteins). These two sets each have different advantages, Swiss-Prot has a minimal level of redundancy and hence a smaller number of proteins, as well as allowing a high level of integration with other databases; Ensembl data has cross-referencing to protein function and assignments to Gene Ontology (GO) term, as well as multiple transcripts from the same gene, e.g. proteins arising from such as alternative splicing of the same pre-mRNA transcripts.

Perl scripts played an integral part in the refinement of the ESP dataset. They were used to select BLAST hits that are above the threshold, as well as preparing tables for loading into MySQL databases. The ESP calculation procedure is summarised by the flow diagram in Figure 4.

From this work, I now have a protocol and procedure so that ESPs can be calculated for any collection of taxa with relative ease. The Perl scripts written for the procedure allow speedy calculations for future research (see supplementary data S2.1 for the protocol and examples of Perl scripts). This package has been uploaded to the DVD supplied with this thesis and given to my supervisor (Dr. Lesley Collins) to be used in future work.

2.2.3 Assigning Gene Ontology terms

Gene Ontology (GO, http://www.geneontology.org) is a collaborative effort to address the need for consistent descriptions of gene products in different databases and useful for clustering of results based on function. The GO assignments used in this work however are tentative and used primarily for clarifying ESP states.

Assigning GO terms for human ESPs was performed by using the Ensembl online tool Biomart (Kinsella *et al.* 2011). By entering the accession numbers of our ESPs, GO terms were easily derived for each human ESP from the collated knowledge in this large database.

Assigning GO terms for the parasites' proteins was more complicated, because these proteins have not typically been annotated with GO terms. The parasites' protein sequences were first BLASTed against the *Saccharomyces cerevisiae* genome, this is because GO term have been well annotated for *Saccharomyces* proteins. Then by using

Biomart, GO terms were assigned to the yeast proteins to which the parasite proteins were putatively homologous to. An issue is that one *Giardia* protein might have many homologues in the yeast genome; on the other hand, *Saccharomyces cerevisiae* also has a reduced genome, some of *Giardia* ESPs might have no yeast homologues at all (this is the case in 261 out of 274 *Giardia* ESPs); finding GO term for the homologue of proteins instead the actual proteins themselves can also create issues at times. This method had many weaknesses, so *Giardia* ESPs GO designations were discarded in the project, and the dataset was categorised manually according to function (see result section Table 4 and supplementary data S2.2 for more detail).

2.2.4 Database construction and management

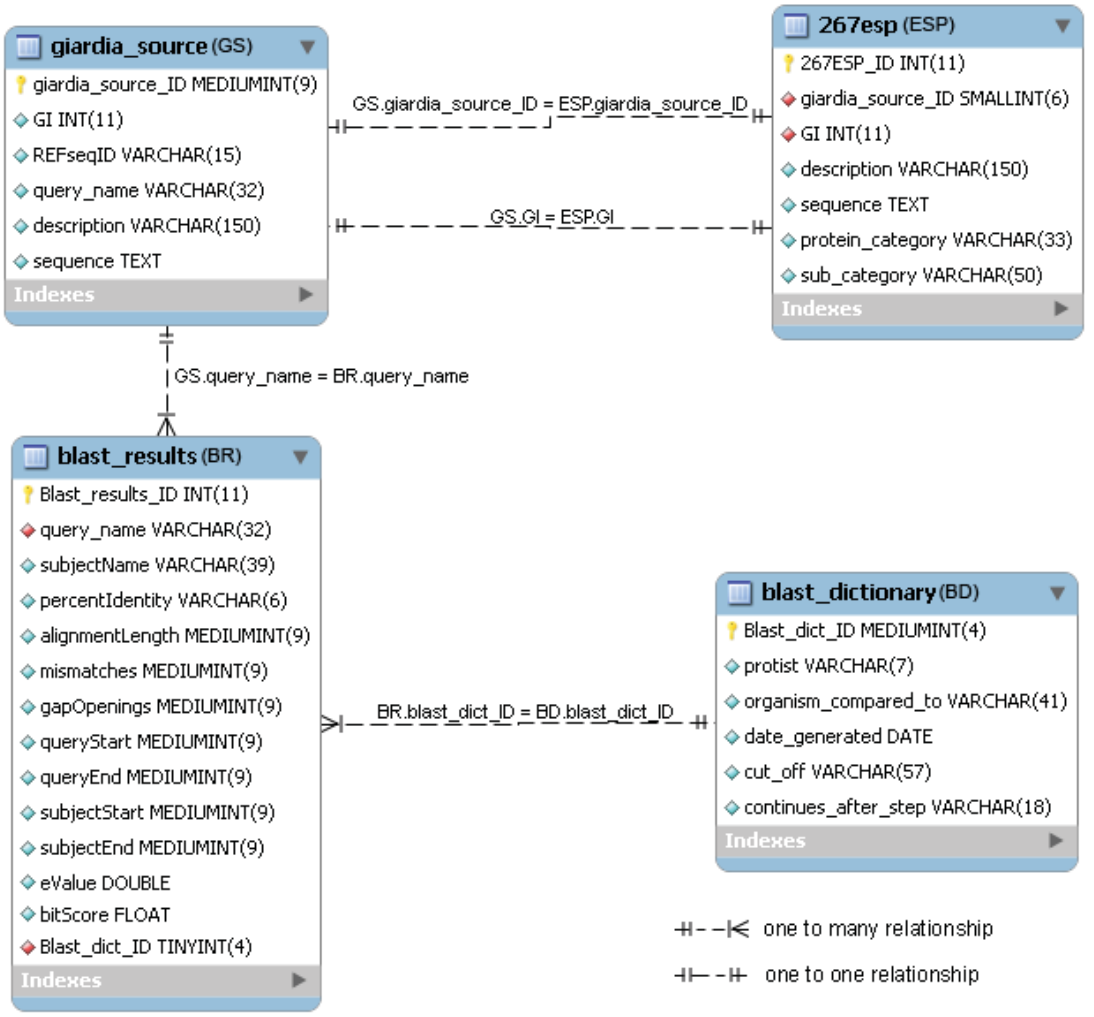
Database construction and the management of genome sized datasets is another crucial part of this project. The advantage of using databases is that they allow fast and organised information retrieval, and easier updating when newer parasitic genomes/proteomes become available. In addition, the relational database management system allows large volumes of information to be efficiently stored and retrieved. Thus, MySQL databases were used for the storage of my ESP data.

The *Giardia* database is illustrated in Figure 5 (created using MySQL Workbench version 5.0.22; OSS community edition). Each box corresponds to a table, the lines connecting tables show the relationship between the columns of the two tables, whether it is a one to one (one entity is related to only one occurrence in another table, indicated by a "⫣⫣") or one to many (when one entity is related to many occurrences in another table, indicated by a "⫣⋞") relationship, and the captions indicate which two columns correspond to each other. MySQL databases enabled the managed storage of a variety of information: for example the "Giardia_source" table contained detailed information about the *Giardia* proteins from GiardiaDB; the "BLAST_results" table stored all the information about the BLAST steps, which enabled to locate precisely when individual proteins were excluded from the datasets. MySQL databases for *Plasmodium*, *Trichomonas* and human were constructed in the same manner as for the *Giardia* database. These databases have been passed onto my supervisor for future work and can be easily updatable using my accompanying script "package".

By using MySQL commands, the comparison of multiple datasets can be readily performed; the homology between ESPs to proteins from other organisms can also be

viewed from the "BLAST_results" table. Additionally the use of a database greatly assisted statistical analysis such as assessing the number of GO terms. MySQL can be combined with Perl to perform various bioinformatic tasks effectively e.g. the fetching of homologous proteins from different organisms was performed in this manner in the phylogenetic section of the project (see Chapter 3).

Figure 5. Illustration of *Giardia* database layout



2.3 Results and Discussion

2.3.1 The *Giardia* ESP dataset

There were 274 *Giardia* ESPs obtained from the ESP analysis, these comprised of 267 distinctive proteins. Although the *Giardia* genome has been annotated, a large percentage of proteins are still designated as hypothetical proteins. This meant that a different strategy for assigning function had to be developed. It should be noted that this function assignment is putative only and indicative only of sequence similarity. The 274 ESPs were divided into seven protein groups according to their predicted conserved functions based on description and homologies to *Saccharomyces cerevisiae* proteins. The seven protein groups are: proteins related to the plasma membrane and endocytosis (34 proteins); those associated with the cytoskeleton (39 proteins); those are involved in the signalling system (97 proteins); those in the nucleus (45 proteins); those involved with protein synthesis and breakdown (15 proteins); those with unknown function (34 proteins) and hypothetical proteins (10 proteins). Table 4 has listed all ESPs by these categories (also see supplementary data S2.2 for the list of all ESPs). Some ESPs have multiple gene copies, and thus numbers of distinctive ESPs (i.e. not including the repeated ones) are also included in brackets.

There are some protein families in the *Giardia* ESP dataset. Protein families are proteins which are descended from a common ancestor. Proteins belong to the same family typically have sequence and structural similarity and they perform similar functions. Examples of such in the dataset are the histone family, which consist of H2A, H2B, H3 and H4 (H1 not an ESP since H1 is not found in *Giardia*); members of the tubulin family are alpha, beta, gamma, delta and epsilon tubulin, all of which are ESPs.

Table 4. Categories of *Giardia* ESPs

Protein category	**# proteins (distinct)**	**Sub category**	**# proteins (distinct)**
Cytoskeleton	37(34)	actins	4
		microtubule related	1
		tubulins	8(5)
		kinesins	24
Membrane	34	cell adhesion	2
		clathrin related	11
		endocytosis	1
		ER and Golgi	9
		lipid attachments	4
		vacuole	7
Nucleus	45(41)	DNA polymerase	1
		histones	11(7)
		histone-associated	4
		LIM related	4
		ribonucleoproteins	2
		RNA enzymes	9
		topoisomerase	1
		transcriptional factors	5
		transcriptional transactivators	2
		zinc fingers	6
Protein synthesis and breakdown	17	ribosome biogenesis proteins	4
		large ribosomal proteins	4
		small ribosomal proteins	3
		proteasome associated	2
		translation factors	4
Signalling system	97	14-3-3 protein	1
		calmodulins	5
		cell cycle related	9
		GTP-binding proteins	20
		kinases and phosphatases	35
		Phosphatidylinositol proteins	7
		ubiquitins	2
		ubiquitin conjugation enzymes	15
		ubiquitin proteases	5
Others	33	others	33
Hypothetical proteins	10	hypothetical proteins	10

Interestingly, there are five ESPs that possessed multiple gene copies in the *Giardia* genome, all of which are members of histone and tubulin family. Alpha tubulin, histone H2A and histone H2B each possess at least two copies in *Giardia*; beta tubulin and histone H4 each possess at least three copies (see Table 5).

Table 5. Proteins with multiple copies in ESP dataset

Protein description	**Number of copies in *Giardia***	**Protein 1**	**Protein 2**	**Protein 3**
Alpha tubulin	2	GL50803_103676	GL50803_112079	
Beta tubulin	3	GL50803_101291	GL50803_136020	GL50803_136021
Histone H2A	2	GL50803_14256	GL50803_27521	
Histone H2B	2	GL50803_121045	GL50803_121046	
Histone H4	3	GL50803_135001	GL50803_135002	GL50803_135003

There were 39 ESPs designated to the cytoskeleton, including a number of actins (proteins that make microfilaments and thin filaments), tubulins (proteins that make microtubules), kinesins (protein motors) and a microtubule-binding protein. The cytoskeleton is thought to be a eukaryotic cellular signature structure (CSS) that defines eukatyotes, but recently a prokaryotic cytoskeleton has been identified (Shih *et al.* 2006; Watters 2006). It has been reported that the eukaryotic actin and tubulin genes possess weak similarity to FtsA and FtsZ, both of which are part of the bacterial cell division machinery (Jimenez *et al.* 2011). The 3-dimensional structure of FtsA and FtsZ are remarkably similar to that of actin and tubulin, respectively, but their primary structures (i.e. sequence) have little similarity (Desai *et al.* 1998; van den Ent *et al.* 2000). It is unknown whether these are cases of convergent evolution of different proteins, or the proteins are indeed diverged from a common ancestor.

The majority of membrane associated ESPs are involved in the transportation of macromolecules. They contained a large number of clathrin (involved in forming coated vesicles), endoplasmic reticulum (ER) and Golgi apparatus related proteins, vacuolar proteins, proteins involved in attachment, and one protein involved in endocytosis. If indeed the eukaryotic cell arose by engulfing other cells, progenitors of these ancient proteins might once have functioned to enable the proposed “raptor” cell (Kurland *et al.* 2006) or chronocyte (Hartman *et al.* 2002) to engulf ancestral bacterial and archaeal cells.

ESPs associated with the nucleus included histones, RNA associated enzymes and proteins in the DNA replicating machinery. Histones are responsible for packing DNA into chromatin structures. Prokaryotes and archaea do not possess complicated DNA packaging systems, thus histones H2A, H2B, H3 and H4 are expected to be ESPs. H1, the linker of chromatin, is an exception because it is less conserved than its histone orthologue and it is absent in some eukaryotes such as *Giardia*. Although archaeal (euryarchaeotes) genomes also contain ancient histone homologues (Spitalny *et al.* 2008), the similarity is more at a structural level rather than the sequence level, in a similar manner to that of actin and tubulin. RNA enzymes include proteins involved in RNA editing, which has been proposed as an ancient mechanism (Collins *et al.* 2009). The presence of ESPs could suggest that RNA editing in eukaryotes existed since the last eukaryotic common ancestor. Finally, the DNA replication process of eukaryotes is much different from that of the prokaryotes', as different polymerases and different transcription factors (including some proteins annotated as "zinc finger proteins") were utilised, as expected DNA replication proteins were well represented in the ESP dataset.

Several ESPs appear involved in protein synthesis. The eukaryotic 80S ribosome is different to the prokaryotic 70S ribosome. Several ribosomal proteins, translational factors fulfilled the criteria of ESP indicating that although *Giardia* has a smaller ribosome than eukaryotes, it is clearly eukaryotic rather than prokaryotic. There are also two proteasome related ESPs indicating the protein degradation mechanism is universal to all eukaryotes.

Signalling system ESPs contain many kinases and phosphatases. These are enzymes performing a variety of functions by adding and removing phosphate groups to a molecule (such as proteins or ATP). Phosphatidylinositol kinases and phosphatases are involved in cellular functions such as cell growth, proliferation, differentiation, motility, survival and intracellular trafficking. GTP-binding proteins are prominent; they function as "molecular switches", and give more sophisticated regulation of enzymes, ion channels, transporters, controlling numerous cell activity such as transcription, motility, contractility, and secretion (Neves *et al.* 2002). Ubiquitin related proteins are very abundant; they are involved in directing protein degradation. Five calmodulins were found as ESPs, indicating regulation by means of calcium-binding is a distinct mechanism in eukaryotes.

There are 9 proteins associated directly with the cell cycle, such as cyclins and cyclin-dependent kinases (CDKs). Cyclins and cyclin-dependent kinases form complexes, which upon activation will dictate which phase the cell will go through. This is a unique scenario in eukaryotes because prokaryotes do not possess nuclei, and their cell division is relatively simple. Interestingly a protein annotated "notchless" was found as an ESP. The similarity between *Giardia* notchless and other eukaryotic notchless is very high (e.g. a bit-score of 296 to the *Drosophila* notchless protein), which suggested high confidence in the annotation. Notchless is a regulator of the notch pathway, which plays a central role in the control of cell fate decisions in a wide variety of cell lineages during invertebrate and vertebrate development (Royet *et al.* 1998). It is unknown why homologues of the gene for this protein would be present in *Giardia lamblia*, *Phytophthora infestans* and *Dictyostelium discoideum*, the three single celled eukaryotes used in this study.

ESPs which fell into the category of "others" are the ones which have not been well annotated. Their annotations typically only have suggestion to their sequence or predicted 3D structure, for example the "Glycine-rich protein" or "WD-40 repeat protein". Some proteins also have suggested functions such as "ATPase". Lastly, ESPs in the "hypothetical protein" category all have "Hypothetical proteins" as their annotation and were not able to be resolved further.

Kurland *et al.* suggested that all ESPs could be divided into three categories: proteins arising *de novo* in eukaryotes; proteins so divergent to homologues of other domains that their relationship is largely lost; or finally, descendants of proteins that are lost from other domains, surviving only as ESPs in eukaryotes (Kurland *et al.* 2006). In an evolutionary sense, group A ESPs (proteins arising *de novo* in eukaryotes) are of the most interest because these proteins hold the key for understanding of the difference between eukaryotes and prokaryotes. The *Giardia* ESPs were examined, taking Kurland's hypothesis into consideration, and it was difficult to divide ESPs into these categories. For example, weak homologues of actin, tubulin and histones are all present in some group of prokaryotes, at least at the protein structural level, but it is hard to say this similarity is the because of homology or convergent evolution (i.e. two separate protein lineage has evolved to be similar appearance due to similarity in their functions). Therefore dividing the ESPs could not be achieved for the *Giardia* ESP dataset. The

"proteins so divergent to homologues of other domains that their relationship is largely lost" scenario possibly has occurred in many groups of ESPs.

2.3.2 Comparison with Hartman's dataset

Hartman *et al.* initially researched *Giardia* ESPs in 2001 and obtained 347 ESPs for *Giardia*. Their approach used *Saccharomyces cerevisiae* as the starting point (see section 2.2.1 for detail of Hartman *et al.*'s method). Comparison between the ESP dataset obtained in this study and Hartman's dataset was made in order to find out the similarities and differences between the two datasets.

The results showed out of my 274 *Giardia* ESPs, 203 proteins had homologues in Hartman's dataset, and 71 did not (more detail from each protein category is shown in Table 6). The reverse BLAST search has also been performed (i.e. the use of Hartman's ESPs dataset as the input and search against my set of ESPs), and out of the 347 Hartman's ESPs, 237 had homologues in my ESP dataset, and 110 did not.

Table 6. *Giardia* ESPs with homologues from Hartman dataset

Protein category	**Number of ESP with homologues in Hartman's data**	**Number of ESP without homologues in Hartman's data**
Membrane	22	12
Cytoskeleton	36	1
Signalling system	73	25
Nucleus	37	8
Protein synthesis and breakdown	10	7
Others	18	15
Hypothetical protein	7	3
Total	203	71

When the 347 Hartman's ESPs were BLASTed against the whole *Giardia* genome, 326 had homologues; This is quite different from the prediction (all 347 Hartman ESPs should have homologues), as Hartman *et al.* calculated their dataset with a step of BLASTing against *Giardia*, this is probably because the *Giardia* database has been vastly updated and some redundant proteins have been removed. Conversely if the whole *Giardia* genome was used to BLAST against Hartman's data (using a cut off bit-score of 55), 351 proteins had homologues.

Overall the datasets are very similar, because the principle was the same in both sets – to find proteins conserved across eukaryotes and not found in prokaryotes. The main cause of the slight variation is the difference in the way the ESPs are calculated, and of course the new databases played a role. Hartman *et al.* started with the *Saccharomyces* proteome because the *Giardia* genome was very poorly annotated at that time (it is better annotated now), and performed BLAST searches with these yeast proteins against the only 44 prokaryotes' genomes available at the time, then only four eukaryotes, and lastly *Giardia*. I used a more straightforward approach and started my BLAST searches directly with *Giardia* proteins, giving the benefit of obtaining a set of the parasites protein in the end rather than a set of yeast proteins. In fact, *Saccharomyces* was not used at all in this project, because it also has a reduced genome and is no longer considered the most "typical" eukaryote. More species were also used in my study, which include 28 bacterial, 11 archaeal and 17 eukaryotic species. More genomes would definitely offer a more robust dataset.

2.3.3 Decision of using 55 bit-score as cut-off

The BLAST bit-score was used in my analysis as a cut-off value to delineate matches as being acceptable or not. The "e-value" is another parameter from the BLAST output that can be used to indicate the significance of the match. So would the e-value be a better indication of how good the match is? Also, is the cut off be too strict or too loose for deciphering homologues? To test this, an e-value of 10^{-7}, which is roughly the same as a 55 bit-score cut off if the proteome database file is ~15 megabytes[3] was used as a cut-off to calculate ESPs. The resulting dataset contained 248 ESPs, a similar number to my previous ESP datasets that used bit-score of 55 as a cut-off. There were 241 ESPs that appeared in both datasets, the other 33 proteins in this dataset did not appear in the original dataset calculated using a bit-score cut off of 55. The resulting datasets were not hugely different. Table 7 summarises the differences between the two datasets.

To understand the differences between the datasets requires an understanding of how the scores are calculated. The bit-score is directly computed from the scoring matrix and gap penalties, and then normalised to give the unit "bits" to the raw score; the e-value indicates the number of alignments expected at random given the size of the search

[3] Note the 15 megabytes is the size of the database file, not the genome. Most eukaryotic protein database files are about this size, plant database files are larger, and some fungi proteomes are smaller.

space and the score of alignments, the lower the e-value the less likely it is that this similarity is random. The e-value can be calculated by the Karlin-Altschul equation "$E=kmne^{-\lambda S}$", where "m" is query size and n is the database size. Based on this knowledge the following test was performed to confirm relationships between database size, bit-score and e-value:

Table 7. Comparison between using E-value and bit-score as cut-offs

Protein category	**Number of ESP in both datasets**	**Number of ESP in the original but not in this dataset**
Membrane	30	4
Cytoskeleton	37	2
Signalling system	82	15
Nucleus	42	3
Protein synthesis and breakdown	12	3
Others	28	6
Hypothetical protein	10	0
Total	241	33

A protein from *Giardia* was used as a query sequence, and the database contained only one protein from *Arabdopsis* (we call this protein A). When the query was BLASTed against the database, the BLAST output suggested the closest homologous *Pyrobaculum* protein was had e-value of $2x10^{-11}$ and bit-score of 52.

For the second BLAST search, the same query sequence was used, but the database is now the entire proteome of *Arabdopsis* (a total of 95500 proteins). The resulting bit-score remained at 52 bits, and so did other values such as percentage identity, alignment length, mismatches and gap openings. The only value which has changed was that of the e-value, which changed from $2x10^{-11}$ to $1x10^{-6}$, which is 50,000 fold larger. The effect of increased e-value in bigger databases can also be seen during the ESP calculating procedure, as e-values in BLAST results of eukaryotic steps are generally many magnitudes larger than results of archaeal/bacterial steps, when the bit-scores are roughly the same. This result suggested that when querying smaller databases (e.g. bacterial genomes), e-values will decrease, causing the result to be above the cut-off, suggesting the two proteins are homologues when the bit-score suggests otherwise, thus causing both false positives and false negatives in the ESP dataset. For the above

reason, bit-scores, a consistent scoring system was used in preference to e-value in this research.

Having decided to use the bit-score, the next question is whether the 55 bit-score (used by Hartman *et al.*), was a good value for the cut-off. In this study, a comparison of proteins was made between distantly related species, and the cut off set reasonably low; on the other hand, it should not be too low that false homologues would be found. Histone H4 offered some guidance in deciding the cut-off value: when this protein was compared with *Archaeoglobus fulgidus* (an archaea), the best hit for H4 had a bit-score of 38.5 and e-value of $4x10^{-4}$. This corresponds to the aforementioned archaeal homologue for histones in section 2.3.1. This homologue was considered insignificant in the above section. Except for H1, which are not found in organisms such as *Giardia*, all other histone proteins are present in all eukaryotes, they are robust ESPs. In addition, Hartman's dataset also included H4. Therefore H4 should be included in the dataset as a negative control. The ideal cut-off should be set above this value, but there is no absolutely right or wrong answer to this issue. The 55 bit-score cut off used this study was deemed to suffice and obtained good datasets.

The length the query sequence may also cause some issue in deciphering homology. BLAST will list alignment of each of the segment of a protein to the database sequences separately, and each of these alignments will be given bit-score by BLAST. *Giardia* proteins are ranged from 33 to 8166 amino acid residues in length (though 97% proteins have <2000 residues), longer sequences have more chance to contain a segment of matching a database sequences by random, i.e. if a segment of 50 residues from the query protein aligned a database protein with 55 bit-score, the whole protein is considered "to have a homologue". To fix this problem, there are 4 solutions:

1 The best idea would be set minimal "alignment length" to a fixed ratio of the query protein length, however BLAST does not have this parameter.

2 Set bit-score threshold higher for the large proteins, this can cause further controversy due to the nature of different criteria will be applied for each alignment.

3 Split these large proteins in to smaller segments, this might cause some matches being missed out if the sequence similarity occurs at the position where sequence was split.

4 Leave it the way it is and use 55 bit-score for all proteins.

We have chosen option 4, because after screening against a large number of species, the chance of a long sequence remaining in the dataset by random is minimised.

2.3.4 The *Plasmodium* and *Trichomonas* ESP datasets

The ESP datasets for these *Plasmodium falciparum* and *Trichomonas vaginalis* were calculated in a similar manner as that for the *Giardia* dataset (refer to Figure 4), starting the search with 5446 *Plasmodium* proteins from PlasmoDB, and 59672 *Trichomonas* proteins from TrichDB respectively (Table 8). For *Plasmodium*, 436 ESPs were obtained; and for *Trichomonas*, 2134 ESPs were obtained. These two parasitic datasets were calculated only for comparisons with *Giardia* and human ESPs, and thus were not analysed in great detail.

Table 8. Summary of the number of ESPs obtained for each organism

Organism name	Number of ESPs	Total number of proteins in database
Giardia lamblia	274	4889
Plasmodium falciparum	436	5446
Trichomonas vaginalis	2134	59672
Homo sapiens Swiss-Prot	2585	20322
Homo sapiens Ensembl	8000	79063

2.3.5 Human (*Homo sapiens*) ESP dataset

There are a number of different proteomic databases available so two human ESP datasets were calculated during this study. The first, using the human proteome from Swiss-prot (http://au.expasy.org/sprot), comprised 2585 ESPs. The second set was calculated using the Ensembl human proteome database (http://www.ensembl.org), and exactly 8000 ESPs were obtained. The number of human ESPs is significantly larger than that of parasites, partly because that human genome might possess more copies of the same gene whereas parasites in their reduced state, has lower copy numbers of each gene.

The GO term and nucleotide data for the Ensembl dataset was obtained by using Biomart, but 2325 ESPs did not have any GO term associated with them. Since the 8000 human ESPs would be difficult to manually group according to function like the way the *Giardia* dataset was categorised, GO terms can give reasonable indications on protein function.

However, the GO term annotation at Ensembl was very quickly updated. Ten months after GO terms were first assigned in September, 2010 (i.e. July 2011), the same 8000 ESPs were assigned GO terms again using Biomart, but this time the results were vastly

different from the GO terms assigned initially. In the updated version, there were only 477 ESPs without any GO term; 794 proteins did not exist anymore or had their accession number changed and Biomart simply ignored them, because the current release (release 62, released in April, 2011) is different from the version of the database used for ESP calculation (release 59, released in July, 2010).

In order to understand the functional difference between ESPs and other proteins, the ESP GO terms needed to be compared with those of other proteins from the human proteome. The Ensembl human protein database release 59 contained 79063 proteins. For convenience matter, 8000 proteins were selected at random by a Perl script for comparison. The results from each set are summarised in Table 9 and Table 10.

Table 9. The 15 most abundant GO terms for human ESPs (updated version)

	GO term accession	GO term name	c.f. random set[4]	# of proteins
1	GO:0005515	protein binding	►	3079
2	GO:0005622	intracellular	▲5	2041
3	GO:0005634	nucleus	▼1	1956
4	GO:0008270	zinc ion binding	▲3	1739
5	GO:0003677	DNA binding	▲5	1243
6	GO:0005737	cytoplasm	▼2	1182
7	GO:0046872	metal ion binding	▲4	1124
8	GO:0003676	nucleic acid binding	▲4	990
9	GO:0006355	regulation of transcription, DNA-dependent	▲4	859
10	GO:0000166	nucleotide binding	▲4	633
11	GO:0016020	membrane	▼8	632
12	GO:0005829	cytosol	▲5	564
13	GO:0005524	ATP binding	▼4	543
14	GO:0005525	GTP binding	▲40	511
15	GO:0007165	signal transduction	▲1	491
	Proteins did not have any GO terms			477
	Proteins do not exist (or ID have been changed) in the new release			794

In looking at the results, ESPs have been assigned to more GO terms than the random set, as the number of proteins for each GO term is higher, and the number of proteins that did not have any GO terms was lower. This is no surprise because ESPs are meant

[4] This column refers to the ranking of the GO term when compared with a random set of proteins. ▲1 indicates the ranking for this GO term in ESPs is higher than its ranking in the random set by 1 place; and ▼1 indicates the ranking for this GO term in ESPs is lower than its ranking in the random set by 1 place. ► indicates the ranking of this GO term is the same in both sets of proteins.

to be essential proteins for eukaryotic life and are expected to be better annotated than random sets of proteins.

From the comparison with the random selection, we can see which GO terms were more abundant in the ESP set. The GO terms more abundant in ESP sets were: Intracellular, nucleic acid binding (and terms alike), metal ion binding and GTP-binding.

GO terms appeared less in ESPs are membrane (and terms alike), which is expected since membrane is a feature in both eukaryotes and prokaryotes. Also GO terms of extracellular region are less abundant in ESPs.

Table 10. The 15 most abundant GO terms for 8000 random human proteins

	GO term accession	GO term name	c.f. ESP set	# of proteins
1	GO:0005515	protein binding	►	2299
2	GO:0005634	nucleus	▲1	981
3	GO:0016020	membrane	▲8	886
4	GO:0005737	cytoplasm	▲2	814
5	GO:0016021	integral to membrane	▲16	795
6	GO:0008270	zinc ion binding	▼3	594
7	GO:0005622	intracellular	▼5	590
8	GO:0005886	plasma membrane	▼14	517
9	GO:0005524	ATP binding	▲4	453
10	GO:0003677	DNA binding	▼5	448
11	GO:0046872	metal ion binding	▼4	438
12	GO:0003676	nucleic acid binding	▼4	343
13	GO:0006355	regulation of transcription, DNA-dependent	▼4	341
14	GO:0000166	nucleotide binding	▼4	340
15	GO:0005576	extracellular region	▲188	314
	Proteins did not have any GO terms			1863
	Proteins do not exist (or ID have been changed) in the new release			898

2.3.6 Human ESPs in parasites

ESPs can give a clue to differences between parasites and host. Parasites have a reduced genome and thus will maintain a small genome which would give them advantages in replication (Fedorov *et al.* 2004). The parasites can lose essential proteins so long as they are substituting that function in some way, for example, an amino acid synthesis pathway can be lost if the parasite can obtain it from the host. My results from the previous section raise some interesting questions:

1. For all these very important proteins - which ESPs can a parasite survive without?

2. Which ESPs are the absolutely essential ESPs, for which even organisms with a minimal genome cannot survive without?

To begin to look at these questions, the 8000 total human ESPs yielded from the Ensembl database were compared with the *Giardia*, *Plasmodium*, and *Trichomonas* genomes, from which the ESPs were divided into 3 groups: lost in all parasites, maintained in one or two parasites, or maintained in all parasites (the grouping "parasites" here meaning the three parasites mentioned above).

There were 2929 (36.6%) ESPs maintained by all parasites, and 1043 (13.0%) ESPs were absent from all of the three parasites. The other 4028 (50.3%) proteins are either maintained by some parasites but absent in other parasites. This agreed with the conclusion from section 2.3.4, and showed that these parasites can survive without some key proteins which are considered essential to free-living eukaryotes.

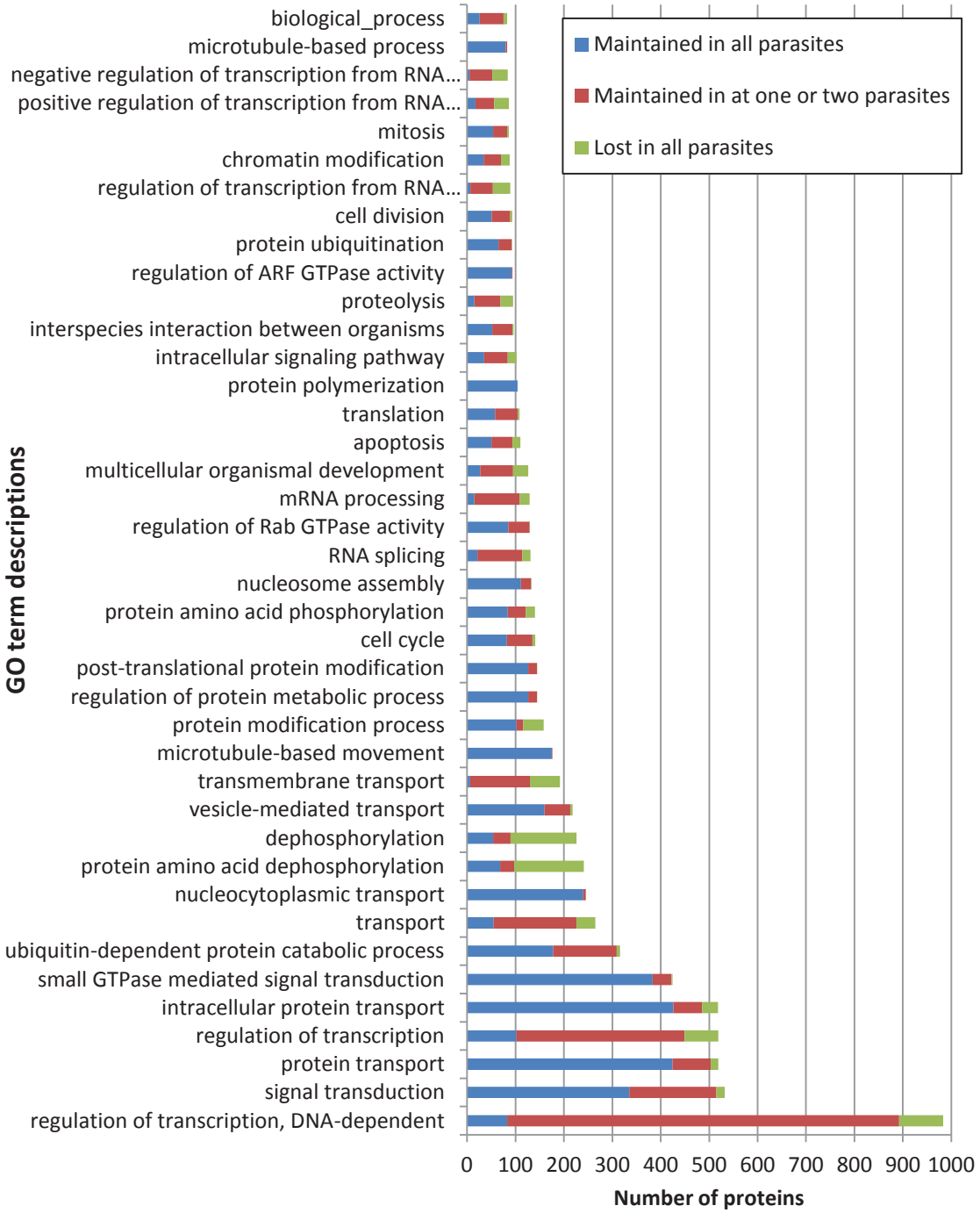
The proteins were all assigned with GO terms which illustrated which functional groups of ESPs are maintained by the parasites, and which ones are lost. Table 11 and Figure 6 shows the most abundant 40 GO terms and also indicates whether proteins belonging to these GO terms are maintained or lost in parasites. Note: as mentioned in section 2.2.3 the GO terms for parasites are putative only in that they represent sequence homology and assume that functional domains have remained intact. However, the GO terms for humans are directly from their database of origin and should represent a much clearer idea of function.

The results here show that many functional groups (i.e. proteins with the same GO terms) are mostly maintained by the parasites. For example, microtubules serve as structural components within cells and are involved in many cellular processes including mitosis (Sgro *et al.* 2011) and vesicular transport (Vale 2003). More than 95% of the microtubule related GO terms are maintained by all parasites (80 out of 83 proteins with GO term "microtubule-based movement" were maintained; similarly 175 out of 176 proteins with GO term "microtubule-based process" were maintained). Nucleosome assembly contains many histone proteins and the majority of them are conserved in the parasites. Some transport systems (nucleocytoplasmic transport, intracellular protein transport) are also mostly conserved. Other functional groups that are conserved in the three parasites are post-translational protein modification, protein polymerization, regulation of ARF GTPase activity and small GTPase mediated signal transduction.

Table 11. The 40 most abundant GO terms for human ESPs

GO term accession	GO term name	Total	Maintained in all parasites	Lost in all parasites
GO:0006355	regulation of transcription, DNA-dependent	983	83	91
GO:0007165	signal transduction	532	335	17
GO:0015031	protein transport	519	424	16
GO:0045449	regulation of transcription	519	102	70
GO:0006886	intracellular protein transport	518	426	33
GO:0007264	small GTPase mediated signal transduction	424	383	2
GO:0006511	ubiquitin-dependent protein catabolic process	316	178	7
GO:0006810	transport	265	55	39
GO:0006913	nucleocytoplasmic transport	245	240	0
GO:0006470	protein amino acid dephosphorylation	241	68	144
GO:0016311	dephosphorylation	226	54	136
GO:0016192	vesicle-mediated transport	218	160	5
GO:0055085	transmembrane transport	192	6	62
GO:0007018	microtubule-based movement	176	175	0
GO:0006464	protein modification process	158	102	42
GO:0051246	regulation of protein metabolic process	145	126	1
GO:0043687	post-translational protein modification	145	126	1
GO:0007049	cell cycle	141	82	6
GO:0006468	protein amino acid phosphorylation	140	84	19
GO:0006334	nucleosome assembly	133	111	1
GO:0008380	RNA splicing	131	21	17
GO:0032313	regulation of Rab GTPase activity	130	85	1
GO:0006397	mRNA processing	129	15	21
GO:0007275	multicellular organismal development	126	27	31
GO:0006915	Apoptosis	110	50	16
GO:0006412	Translation	108	58	3
GO:0051258	protein polymerization	104	104	0
GO:0023034	intracellular signalling pathway	102	35	18
GO:0044419	interspecies interaction between organisms	96	52	2
GO:0006508	Proteolysis	95	15	26
GO:0032312	regulation of ARF GTPase activity	94	92	0
GO:0016567	protein ubiquitination	93	65	1
GO:0051301	cell division	93	51	4
GO:0006357	regulation of transcription from RNA polymerase II promoter	89	7	36
GO:0016568	chromatin modification	88	34	18
GO:0007067	mitosis	86	54	3
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	86	17	30
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	84	5	32
GO:0007017	microtubule-based process	83	80	0
GO:0008150	biological process	83	26	8

Figure 6. Human ESP and GO term



Some functional groups are not well maintained in parasites, such as mRNA processing, RNA splicing, transmembrane transport, regulation of transcription (both DNA-dependent and RNA-dependent). Parasite genomes in general contain fewer introns, and the number of mRNA processing and RNA splicing components may be significantly reduced (L. Collins, personal communication). Parasites may also regulate their transcription differently from the host, or their transcription machinery may also have

been simplified. Transmembrane transport is the only transport system that appears not to be maintained in the three parasites, whereas other transport systems such as intracellular protein transport, nucleocytoplasmic transport and vesicle mediated transport all appear to be well maintained by these parasites.

An initial aim was to identify metabolic pathways with missing ESPs, and to determine how parasites accommodated such loss (either obtaining the lost proteins from the host or modifying appropriate biochemical pathways). The idea is that a new drug could be developed from the alternative pathways of the parasite. In depth comparison between the host and parasites' ESPs as individual proteins was not performed during this project, because this requires more complete annotation for the parasites than is presently available. It would also require more complete GO terms for the parasite proteins and more protein interaction data.

2.3.7 Differences and similarities between parasite ESP datasets

The ESP datasets of the three parasites were then compared to each other (summarised in Table 12). Quite remarkably, ESPs from one parasite are not necessarily present in the datasets of the other two parasites. The results can be explained by hypothesising a scenario that a free-living eukaryote might have more ESPs than any of these parasites, the parasites can survive without some key proteins which were considered essential to free-living eukaryotes (i.e. ESPs), they do not maintain them because this will give them a smaller genome and advantage in replication; their pattern in the conservation of essential proteins are different in each parasite due to their different niche.

Table 12. ESP between parasites

	Giardia	***Plasmodium***	***Trichomonas***	**Both parasites**
Giardia	274	207	225	195 (71.2%)
Plasmodium	211	436	298	194 (44.5%)
Trichomonas	1502	1642	2134	1334 (62.5%)

Another test for ESP conservation was performed by comparing ESP datasets with the genome of the parasites instead of the ESP datasets, the results were summarised by Table 13. The results from this test were different from the previous one, indicating in some cases, homologues of the ESPs were present in the other parasites proteome, but the homologues were not in the ESP datasets of the other parasite. The reason may be

these proteins are "borderline" ESPs, and have been lost during the ESP calculation steps of some divergent species. This indicates that the loss of ESPs may not be the same in every parasite. There are likely to be evolutionary differences in which proteins are lost and more importantly which essential proteins (i.e. ESPs) can be lost. This pattern of loss would depend on the parasitic life-style and which nutrient resources are available from the host.

Table 13. ESP vs other parasites' proteome

	Giardia	***Plasmodium***	***Trichomonas***	**Both parasites**
Giardia	274	242	259	236 (86.1%)
Plasmodium	246	436	343	235(53.8%)
Trichomonas	1619	1758	2134	1465(68.6)

2.3.8 Other groups of proteins

Besides ESPs, there are several other important and interesting groups of proteins calculated according to their conservation in the three domains of life. "Proteins conserved in all eukaryotic species" were calculated using *Giardia* proteins as a starting point, and during each step, proteins without homologue (55 bit-score cut off) excluded, the same 17 eukaryotic organisms from ESP calculation procedure were used. There were 849 proteins which fulfilled these criteria.

The group "*Giardia* proteins conserved in all organisms" contains the most ancient proteins with homologues in all extant phyla from prokaryotes to eukaryotes. The same 28 bacterial species, 12 archaeal species 17 eukaryotic species from the ESP calculation procedure were used. There were only 37 proteins conserved in all these organisms (listed in supplementary data S2.3). These proteins included numerous protease proteins (6 proteins), ATP-binding cassette transporters (13 proteins).

"Archaeal signature proteins" were calculated similar to the manner of ESP calculation (i.e. using the 536 *Nanoarchaeum equitans* proteins as the starting point, identified proteins that are conserved in all archaeal species, and then excluded proteins with homologues in bacteria and eukaryotes). However no archaeal signature proteins were found using our standard cut-offs. There were 28 proteins conserved in all the archaea species, but this number may be low due to the fact that *Nanoarchaeum* has a small genome. All of these 28 proteins have bacterial homologues and also eukaryotic

homologues. This result differs from that of Graham *et al.* (Graham *et al.* 2000), who found 351 proteins unique to archaea. An important difference in studies is that Graham *et al.*'s definition of "archaeal signature proteins" is different from this study, since they have included proteins which are found in 2 or more taxa of Euryarchaeota, whereas this analysis only included proteins which are found in all taxa of archaea.

Calculating "bacterial signature proteins" calculations (i.e. using 8308 *Escherichia coli* proteins as starting point, collecting proteins conserved in all bacterial species, and then excluding proteins with homologues in archaea and eukaryotes), resulted in 278 proteins conserved in all bacteria. Homologues of these 278 proteins were found in eukaryotes. There were however, 44 proteins out of the 278 conserved bacterial proteins that had no homologues found in any archaea (listed in supplementary data S2.4). These 44 proteins included almost all transcriptional and translational machinery, indicating that the bacterial transcription and translation systems are unique from those of archaea. It is also noticed that the aforementioned 28 proteins conserved in all archaea also contained a number of transcriptional and translational proteins. Therefore it seems that bacteria share the ancestry of transcriptional and translational machinery with Archaea, but have also since then developed different systems of their own.

2.4 Conclusions

2.4.1 ESP calculation conclusions

The ESP datasets for *Giardia lamblia* and *Trichomonas vaginalis* and human have now been re-calculated and databases containing and connecting these ESP datasets have been constructed. I have also examined the difference in ESPs from host and parasites, and results showed interesting patterns of the trend of loss of ESPs (section 2.3.6).

The current dataset is significantly more robust than Hartman *et al.*'s (Hartman *et al.* 2002) due to many more organisms being used, but it may still have small amount of false positives due to the lack of any completed genomes from some branches of prokaryotes and eukaryotes. ESPs are not a complete set of ancestral proteins. Some eukaryotic organisms may have lost ancestral protein. For example, Dicer is an ancestral enzyme, but some lineages such as yeast have secondary loss of this protein. In future studies a protocol that captures proteins such as Dicer would be useful.

The protocol for ESP calculation is flexible and permits future re-calculations to include newly sequenced genomes as well as updated genomic information. Increased computer

power means that more species can now be more readily included. The Perl scripts for forming ESP databases are planned for inclusion in a future manuscript (in preparation).

2.4.2 Database updates

Proteome databases are constantly being updated. This is problematic because outdated ESP results do not work on new databases very well due to change in the accession numbers, inclusion of more annotated proteins and exclusion of deprecated proteins from the newer database versions. For example, from the time our original human ESPs were calculated, three new versions of the human proteome were have been by Ensembl, the current version being version 62 (May 2011), whereas version 59 was used when human ESPs were calculated ten months earlier. The new databases have made changes to a number of proteins. The GO database has also been updated constantly, as the results from previous sections differed significantly when one analysis was performed ten months after the other (see section 2.3.4). Database use is therefore essential for managing data in this dynamic situation.

For future ESP analysis, the dataset should be periodically updated in order to keep up with these changes in input data. Perl scripts and command line scripts have already been prepared to make the process of ESP re-calculation straightforward, and one only has to run (and perhaps slightly modify) these scripts to have an updated dataset. If ESPs are re-calculated in the future, more organisms can be included in the calculations especially when more complete genomes become available. For example, species from the bacterial phylum Caldiserica, or species from eukaryotic supergroup Rhizaria could be included, since these species were not used in this study because there are no complete genomes yet. The continuing increase in computing power will thus permit even more robust ESP calculations.

2.4.3 Implications for current models of evolution

The presence of ESPs has identified weaknesses in the popular genome fusion model (Rivera *et al.* 2004) (in general, prokaryotes combine genomes becoming eukaryotes). This is because the model is uninformative about the existence of ESPs and cellular signature structures (CSSs) which are not found in prokaryotes (Kurland *et al.* 2006). However, the fact that all bacterial and archaeal essential proteins (proteins conserved in all archaea or all bacteria) are found in eukaryotes, does suggest that eukaryotes may

have acquired archaeal and bacterial endosymbionts in the early days of eukaryotic evolution.

The other two hypotheses on the origin of eukaryotes by Hartman *et al.* and Kurland *et al.* are very similar. Hartman *et al.* hypothesised that there was a third cell type, which they called a "chronocyte", which was a progenitor of the eukaryotic cell, and the nucleus of a eukaryotic cell was formed from the endosymbiosis of an archaeon and a bacterium in the chronocyte (Hartman *et al.* 2002). Kurland *et al.*'s hypothesis is a slight variation from that of Hartman *et al.*'s. Kurland *et al.*'s hypothesis is that eukaryotes originated from a community of saprotrophic, autotrophic and heterotrophic cells, where a phagotrophic unicellular raptor emerged and then acquired a bacterial endosymbiont/mitochondria lineage, and became the common ancestor of all eukaryotes (Kurland *et al.* 2006). The two hypotheses are similar in the way that they both suggest a phagotrophic cell has acquired archaeal and bacterial endosymbionts and formed the last eukaryotic common ancestor. The major difference between the two hypotheses is in the nature of that phagotrophic cell (i.e. the "chronocyte" and the "raptor"). The chronocyte is considered a complete separate lineage from archaea and bacteria, and is an RNA based cell with a complex membrane system which was needed for phagocytosis; whereas the "raptor" was just another archaeal or a bacterial cell which lived in the same community as other early archaea or bacteria.

The chronocyte hypothesis does have complications: archaea and bacteria diverged 3500 million years ago (Glansdorff *et al.* 2008), the eukaryotic cell arose 1850 million years ago (Knoll *et al.* 2006), and this makes it not very likely a third lineage of organisms existed between the 1650 million years in between the two events. Thus the "raptor" scenario is favourable. A "complex membrane system" was not necessarily needed for uptake of other cells as it was suggested by Hartman *et al.*. Prokaryotic cells can at least uptake genetic material through mechanisms such as transformation. This study found ABC transporters (able to efflux various macro-molecules) are conserved in all organisms, suggesting that it is an ancient protein, and implies some the earliest organisms (more bacteria like) are capable of engulfing others organisms.

In addition, it is proposed that the "raptor is more an archaea-like cell rather than bacteria-like, because the eukaryotic 18S ribosomal RNA is more similar to archaea (Woese *et al.* 1990), and literature suggests that bacteria diverged from the archaeal/eukaryotic lineage (Brown *et al.* 1997; Glansdorff *et al.* 2008). This group of

cells evolved clathrin like proteins which enabled phagocytosis to take place, and more complex membrane system was evolved to facilitate phagocytosis. This phagocyte could have engulfed many bacterial and archaeal cells and during its evolution, and maintained genetic material from the engulfed organisms. Early archaeal histone homologues enabled the "raptor" to sustain the large amount genetic material obtained from the engulfed cells and this group of proteins developed rapidly to the histones we know. This "raptor" was the common ancestor of all eukaryotes. This hypothesis can also successfully explain the existence all three groups of ESPs suggested by Kurland *et al.*. The "proteins appeared arising *de novo* in eukaryotes" were proteins unique to the "raptor" lineage at the time. The "proteins so divergent to homologues of other domains that their relationship is largely lost" were the proteins which played minor role in bacteria and archaea, but then proliferated to have major roles in the newly formed eukaryotic cell after they have been acquired by the "raptor". Lastly the "descendants of proteins that are lost from other domains, surviving only as ESPs in eukaryotes" was due to the role change of prokaryotes after eukaryotes have evolved, the reductive evolution might have driven the loss of these proteins in prokaryotes.

In conclusion, ESPs can be readily re-calculated and may hold other clues to early eukaryotic evolution when their functions are analysed further. Currently they are being used in a Marsden funded project looking at this very issue.

Supplementary material for Chapter 2

S2.1 ESP calculation protocol and Perl scripts

In a Windows environment, the commands for calculating are run in the following procedure:

From D drive, create a folder for calculating ESPs (e.g. GiardiaESP).

In folder GiardiaESP, create a folder for each organism used during screening process (e.g. organism1, organism2 etc). Each of these folders should contain the FASTA file for the annotated protein database for that organism, formatted by using BLAST command:

```
Formatdb –i organism1DB.fasta –o
```

In addition, organism1 folder also contain the protein database file for the organism whose ESP dataset is being calculated (e.g. Giardia_annotated_protein_db.fasta).

After the above steps, the following commands can be run in the command line interface. Provided that have the right setup, all command lines can be pasted into the command line interface and will be executed one after another and new ESP dataset will be calculated. The –m 8 command BLAST output format to be in spread sheet format, which would be convenient to work with. Command “perl script4.pl” executes the Perl script “script4.pl”. The Perl scripts are listed on the next four pages.

```
cd D:\GiardiaESP\organism1
blastall -p blastp –I giardia_annotated_protein_db.fasta -m8 -d
    organism1db.fasta –o homolog.xls
perl script4.pl
copy remainingesp.txt D:\GiardiaESP\ oranism2
cd D:\GiardiaESP\oranism2
rename remainingesp.txt previousesp.txt
blastall -p blastp -i previousesp.txt -m8 -d oranism2.fasta
    >homolog.xls
perl script4.pl
copy remainingesp.txt D:\ GiardiaESP\oranism3
cd D:\GiardiaESP\oranism3
rename remainingesp.txt previousesp.txt
blastall -p blastp -i previousesp.txt -m8 -d oranism3.fasta
    >homolog.xls
```

```
perl script4.pl
copy remainingesp.txt D:\ GiardiaESP\oranism4
etc...
```

Script 1. This Perl script allow selection of BLAST hits above a certain threshold of bit-score from a spread sheet formatted BLAST output (BLAST parameter "–m 8"), and print their accession number into output file, the output file was called 'GIhomolog.txt' in this study. The output listed accession numbers of all proteins with significant homologues from the screened organism.

```
use strict;
use warnings;
my $cut = 55; #adjust $cut to desired cut off.
my $input = "homolog.xls";
open (INPUT, "$input") or die "Input file not opened";
my $output = "GIhomolog.txt";
open (OUTPUT, ">$output") or die "output file not opened";
#load every line from input into @array:
my @array;
while (<INPUT>) {
my $hit = $_;
push @array, $hit;
}
my $i;
foreach $i(@array) {
    $i=~/(gb\|.*)\tgi.*\t.*\t.*\t.*\t.*\t.*\t.*\t.*\t.*\t.*\t(.*)/;
    if ($2 >= $cut){
    print OUTPUT "$1\n";
    }
}
```

Script 2. This Perl script allows the removal of accession numbers that occurred more than once in "GIhomolog.txt" generated from script 1, the output file is called "GInorepeat.txt". This script is needed because then script 3 can perform its designated task.

```
use strict;
use warnings;
my $input = "GIhomolog.txt";
open (INPUT, "$input") or die "Input file not opened";
my $output = "GInorepeat.txt";
my $output2 = $output . "_excl.txt";
open (OUTPUT, ">$output") or die "output file not opened";
open (OUTPUT2, ">$output2") or die "output2 file not opened";
#load accession numbers into array to use for searching
my @array;
my $idCount = 0;
while (<INPUT>) {
my $identifier = $_;
chomp $identifier;
push @array, $identifier;
++ $idCount;
}
my $i;
my @repeated;
my @norepeat;
for ($i = 0; $i < @array; ++$i) {
    if ($array[$i] eq $array[$i+1]) {
    push @repeated, $array[$i];
    }
    else {
    push @norepeat, $array[$i];
    }
}
foreach my $GI (@norepeat){
    print OUTPUT "$GI\n";
}
foreach my $GI2 (@repeated){
    print OUTPUT2 "$GI2\n";
}
```

Script 3. This Perl script can find protein sequences which have homologues from the screened organism, and put into one output file #1, and put proteins do not have homologues from the organism into output file #2. When screening against a

prokaryotic organism, output file #2 is called “remaininingESP.txt”, since this is the file carried onto the next step; and when screening against a eukaryotic organism, output file #1 is called “remainingESP.txt”, and carried onto the next step. Note Perl package “Bioperl” is needed for this script.

```
use strict;
use warnings;
use Bio::SeqIO;
my $idFile = "GInorepeat.txt";
open (IDFILE, "$idFile") or die "Identifier file not opened";
my $databaseFile = "previousesp.txt";
chomp $databaseFile;
print "Output File for included sequences: ";
my   $outFile   =   "remainingESP.txt";#when   screening   against   a
    eukaryotic organism, this file is called "nohomolog.txt";
my $outfile2 = "hasHomo.txt"; #when screening against a eukaryotic
    organism, this file is called "remainingESP.txt";
print "Output file for excluded sequences is $outfile2\n";
my @identArray;
my $idCount = 0;
while (<IDFILE>) {
my $identifier = $_;
chomp $identifier;
push @identArray, $identifier;
++ $idCount;
}
print "Number of sequences to remove: $idCount\n";
my @identFound;
my $found = 0;
my  $in  =  Bio::SeqIO->new('-file'  =>  "$databaseFile",  '-format'  =>
    'fasta');
while ( my $seq = $in->next_seq() ) {
    my $key = $seq->id;
    chomp $key;
    # change $key to the match the format of accession number of the
    organism
    $key =~ /(gb\|GL50803_.\d*)/;
    my $key2 = $1;
```

```
    foreach my $ident (@identArray) {
          if ($key2 eq $ident) {
                $found = 1;
                my $out = Bio::SeqIO->new('-file'  => ">>$outfile2",
    '-format' => 'fasta');
                $out->write_seq($seq);
                push @identFound, $ident;
          }
    }
    if ($found == 0) {
          my  $out  =  Bio::SeqIO->new('-file'  =>  ">>$outFile",  '-
    format' => 'fasta');
          $out->write_seq($seq);
    }
    else {$found = 0;}
}
my $thisCount = @identFound;
print "Number of sequences removed to $outfile2: $thisCount";
if ($thisCount == $idCount) {
    print "\nAll nominated sequences removed\n"
}
else { print "\nNot all sequences removed - please check output\n";}
print "Program Complete\n";
```

Script 4. This Perl script allows scripts 1-3 to be run, it is used purely for convenience reasons.

```
system "perl script1.pl";
system "perl script2.pl";
system "perl script3.pl";
```

S2.2 List of 274 *Giardia* ESPs

Cytoskeleton

Actins

GL50803_8589	Suppressor of actin 1
GL50803_16299	Sda1, severe depolymerization of actin
GL50803_15113	Actin
GL50803_40817	Actin related protein

Microtubule related

GL50803_14048	EB1

Tubulins

GL50803_5462	Delta tubulin
GL50803_136021	Beta tubulin
GL50803_136020	Beta tubulin
GL50803_101291	Beta tubulin
GL50803_103676	Alpha-tubulin
GL50803_6336	Epsilon tubulin
GL50803_112079	Alpha-tubulin
GL50803_114218	Gamma tubulin

Kinesins

GL50803_16945	Kinesin-13
GL50803_4371	Kinesin-8
GL50803_15134	Kinesin-6 like
GL50803_16224	Kinesin-related protein
GL50803_102455	Kinesin-6
GL50803_6262	Kinesin-3
GL50803_102101	Kinesin-3
GL50803_112846	Kinesin-3
GL50803_16456	Kinesin-2
GL50803_112729	Kinesin like protein
GL50803_11442	Kinesin-related protein
GL50803_10137	Kinesin-9
GL50803_16650	Kinesin-4
GL50803_16425	Kinesin-5

GL50803_8886	Kinesin-14
GL50803_17333	Kinesin-2
GL50803_14070	Kinesin-like protein
GL50803_17264	Kinesin like protein
GL50803_7874	Kinesin-16
GL50803_16161	Kinesin-16
GL50803_15962	Kinesin-7
GL50803_13797	Kinesin-14
GL50803_6404	Kinesin-9
GL50803_13825	Kinesin-1

Membrane proteins

Cell adhesion

GL50803_16882	Bystin
GL50803_92673	CHL1-like protein

Clathrin related

GL50803_102108	Clathrin heavy chain
GL50803_15339	Adaptor protein complex large chain subunit BetaA
GL50803_89622	Mu adaptin
GL50803_21423	Beta adaptin
GL50803_17304	Alpha adaptin
GL50803_8917	Mu adaptin
GL50803_3256	EH domain binding protein epsin 2
GL50803_5328	Sigma adaptin
GL50803_16364	Gamma adaptin
GL50803_91198	Sigma adaptin
GL50803_14373	Dynamin

Endocytosis

GL50803_42048	ABC transporter

ER and Golgi

GL50803_88082	Coatomer beta subunit
GL50803_11953	Coatomer alpha subunit
GL50803_4502	ER lumen protein retaining receptor
GL50803_17065	Sec24

GL50803_17164	Sec24-like
GL50803_9593	Coatomer beta' subunit
GL50803_15413	RER1-like protein-retention of ER proteins
GL50803_17192	Protein transport protein Sec7
GL50803_29487	Protein disulfide isomerase PDI1
Lipid attachments	
GL50803_10019	Phospholipid-transporting ATPase IA, putative
GL50803_137725	Phospholipid-transporting ATPase IIB, putative
GL50803_101810	Phospholipid-transporting ATPase IIB, putative
GL50803_17082	Rab geranylgeranyltransferase
Vacuole	
GL50803_100864	Vacuolar protein sorting 26, putative
GL50803_23833	Vacuolar protein sorting 35
GL50803_8559	Vacuolar ATP synthase 16 kDa proteolipid subunit
GL50803_10530	Vacuolar ATP synthase 16 kDa proteolipid subunit
GL50803_13000	Vacuolar ATP synthase subunit d
GL50803_14961	Vacuolar ATP synthase subunit H
GL50803_15598	Vacuolar ATP synthase 16 kDa proteolipid subunit
Nucleus	
DNA polymerase	
GL50803_6980	DNA pol/primase, large sub
Histones	
GL50803_3367	Histone H3
GL50803_135002	Histone H4
GL50803_135003	Histone H4
GL50803_14212	Histone H3
GL50803_20037	Histone H3
GL50803_135231	Histone H3
GL50803_27521	Histone H2A
GL50803_121045	Histone H2B
GL50803_121046	Histone H2B
GL50803_135001	Histone H4
GL50803_14256	Histone H2A

Histone-associated	
GL50803_14753	Histone acetyltransferase type B subunit 2
GL50803_10666	Histone acetyltransferase GCN5
GL50803_17263	Histone methyltransferase MYST1
GL50803_2851	Histone acetyltransferase MYST2
LIM related	
GL50803_9162	Nuclear LIM interactor-interacting factor 1
GL50803_14905	Nuclear LIM interactor-interacting factor 1
GL50803_4063	Nuclear LIM interactor-interacting factor 1
GL50803_4235	Nuclear LIM interactor-interacting factor 1
Ribonucleoproteins	
GL50803_16173	U3 small nucleolar ribonucleoprotein protein IMP4, putative
GL50803_17112	U3 small nucleolar ribonucleoprotein protein IMP4, putative
RNA enzymes	
GL50803_5661	RNA binding protein
GL50803_6054	RNA binding putative
GL50803_24860	Nonsense-mediated mRNA decay protein 3
GL50803_10840	DNA-directed RNA polymerases I and III 16 kDa polypeptide
GL50803_14763	Exonuclease
GL50803_24133	5'-3' exoribonuclease 2
GL50803_17325	Pumilio-family RNA-binding protein, putative
GL50803_113365	5'-3' exoribonuclease 2
GL50803_14702	RRNA biogenesis protein RRP5
Topoisomerase	
GL50803_16285	Topoisomerase I-related protein
Transcriptional factors	
GL50803_8209	CCR4-NOT transcription complex, subunit 7
GL50803_8427	Transcriptional repressor NOT4Hp, putative
GL50803_7231	CCAAT-binding transcription factor subunit A
GL50803_4125	Transcription factor IIIB 70 kDa subunit BRF
GL50803_10606	CCR4-NOT transcription complex, subunit 7
Transcriptional transactivators	
GL50803_5347	Myb 1-like protein

GL50803_8722 Myb 1-like protein

Zinc fingers

GL50803_8619 Zinc finger domain

GL50803_9529 Zinc finger domain

GL50803_6733 Zinc finger domain

GL50803_1908 DHHC-type zinc finger domain-containing protein

GL50803_16928 Zinc finger domain

GL50803_96562 Zinc finger domain

Protein synthesis and breakdown

Large ribosomal proteins

GL50803_8462 Ribosomal protein L27

GL50803_14622 Ribosomal protein L13

GL50803_19436 Ribosomal protein L7

GL50803_16387 Ribosomal protein L18a

Small ribosomal protein

GL50803_10367 Ribosomal protein S24

GL50803_6135 Ribosomal protein S17

GL50803_14329 Ribosomal protein S7

Ribosome biogenesis protein

GL50803_102722 Ribosome biogenesis protein BMS1

GL50803_16718 Partner of Nob1

GL50803_3589 Ribosome biogenesis protein Brix

GL50803_8361 SOF1 protein

Translation factors

GL50803_13561 Translation elongation factor

GL50803_8708 Eukaryotic translation initiation factor 1A

GL50803_93275 Translational activator GCN1

GL50803_13661 Eukaryotic translation initiation factor 3 subunit 2

Proteasome associated

GL50803_16823 Non ATPase subunit MPR1 of 26S proteasome

GL50803_7896 26S proteasome non-ATPase regulatory subunit 7

Signalling system

14-3-3 proteins

GL50803_6430	14-3-3 protein
Calmodulins	
GL50803_13652	Calmodulin
GL50803_6744	Centrin
GL50803_13231	Calmodulin
GL50803_5333	Calmodulin
GL50803_104685	Caltractin
Cell cycle related	
GL50803_102890	BUB3
GL50803_11044	Mob1-like protein
GL50803_5772	CDC72
GL50803_4008	Mob1-like protein
GL50803_3977	G2/mitotic-specific cyclin B
GL50803_9778	Tem-1-like protein
GL50803_17103	Orc1/CDC6
GL50803_15248	Spindle protein, putative
GL50803_13667	Notchless
GTP-binding proteins	
GL50803_13109	RabA
GL50803_12157	RabB
GL50803_16636	Rab2b
GL50803_1695	Rab11
GL50803_16979	Rab32, putative
GL50803_7569	GTP-binding protein Sar1
GL50803_13478	ARL1
GL50803_13930	ARF3
GL50803_7562	ARF2
GL50803_15567	Rab2a
GL50803_22454	ARF GAP
GL50803_11495	Rab GDI
GL50803_940	RabD
GL50803_2834	ARF GAP
GL50803_8497	RabF

GL50803_8496	Rac/Rho-like protein
GL50803_4192	ARL2
GL50803_15869	GTP-binding nuclear protein RAN/TC4
GL50803_17561	ARF GAP
GL50803_9558	Rab1a
Kinases and phosphatases	
GL50803_11554	Kinase, NEK
GL50803_5554	Kinase, NEK-frag
GL50803_5553	Kinase, NEK
GL50803_10313	Kinase
GL50803_3414	5'-AMP-activated protein kinase, gamma-1 subunit
GL50803_87928	Kinase, NEK
GL50803_14044	Kinase, NEK
GL50803_9365	Kinase, NEK
GL50803_14648	Kinase, NEK-frag
GL50803_14650	Ser/Thr phosphatase 2C, putative
GL50803_32398	Protein phosphatase PP2A regulatory subunit B
GL50803_9117	CAMP-dependent protein kinase regulatory chain
GL50803_115572	Kinase, Wee
GL50803_17406	Phosphoinositide-3-kinase, class 3
GL50803_137730	Kinase
GL50803_11740	Ser/Thr phosphatase 2C, putative
GL50803_103838	Kinase, ULK
GL50803_2538	Kinase, NAK
GL50803_16443	Protein phosphatase 2A B' regulatory subunit Wdb1
GL50803_113456	Kinase, VPS15
GL50803_9293	Protein phosphatase 2C-like protein
GL50803_4079	Protein phosphatase PP2A regulatory subunit B
GL50803_12095	Kinase, NEK-frag
GL50803_14404	Phosphatase
GL50803_17335	Kinase, CMGC SRPK
GL50803_36783	Kinase, NEK
GL50803_10612	Phosphotyrosyl phosphatase activator protein, putative

GL50803_15112	Dual specificity phosphatase, catalytic
GL50803_96616	Kinase, CMGC CDKL
GL50803_7588	Serine/threonine protein phosphatase 7
GL50803_21502	Kinase, putative
GL50803_21116	Kinase, CMGC CMGC-GL1
GL50803_137695	Kinase, CMGC DYRK
GL50803_4288	Ser/Thr phosphatase 2C, putative
GL50803_5999	Kinase, NEK
Phosphatidylinositols kinases and phosphatases	
GL50803_11897	Phosphatidylinositol-4-phosphate 5-kinase, putative
GL50803_14975	Phosphatidylinositol-glycan biosynthesis, class O protein
GL50803_14855	Phosphoinositide-3-kinase, catalytic, alpha polypeptide
GL50803_9077	Inositol 5-phosphatase 4
GL50803_35180	GTOR
GL50803_16558	Phosphatidylinositol 4-kinase
GL50803_14787	Type II inositol-1,4,5-trisphosphate 5-phosphatase precursor
Ubiquitins	
GL50803_7110	Ubiquitin
GL50803_8843	Ubiquitin
Ubiquitin conjugation enzymes	
GL50803_15162	Ubiquitin-conjugating enzyme E2-17 kDa
GL50803_5921	Ubiquitin-conjugating enzyme E2-28.4 kDa
GL50803_3171	UBCE14
GL50803_3978	Ubiquitin-conjugating enzyme E2-17 kDa
GL50803_31576	Ubiquitin-conjugating enzyme E2-21.2 kDa
GL50803_4083	Ubiquitin-conjugating enzyme E1
GL50803_6524	Ubiquitin-conjugating enzyme E2-28.4 kDa
GL50803_3994	Ubiquitin fusion degradation protein 1
GL50803_15252	Ubiquitin-conjugating enzyme E2-17 kDa 3
GL50803_8638	Ubiquitin-conjugating enzyme E2-28.4 kDa
GL50803_12950	Ubiquitin-conjugating enzyme E2-17 kDa
GL50803_27055	Ubiquitin-conjugating enzyme E2-17 kDa 3
GL50803_24068	UBC3

GL50803_2876	UBC2, putative
Ubiquitin proteases	
GL50803_16090	Ubiquitin carboxyl-terminal hydrolase 4
GL50803_5533	DUB-1
GL50803_14460	Ubiquitin carboxyl-terminal hydrolase 4
GL50803_8189	Ubiquitin carboxyl-terminal hydrolase 14
GL50803_17386	Ubiquitin-protein ligase E3A
Others	
GL50803_7533	Angio-associated migratory cell protein
GL50803_88438	ATPase
GL50803_9528	Methyltransferase like 2
GL50803_15531	Periodic tryptophan protein 1, putative
GL50803_113143	Lipopolysaccharide-responsive and beige-like anchor protein
GL50803_10822	WD-40 repeat protein family
GL50803_17502	Mannosyltransferase
GL50803_9382	Prenyltransferase
GL50803_27747	Flavohemoprotein B5+B5R
GL50803_7287	Small glutamine-rich tetratricopeptide repeat-containing protein
GL50803_17294	Degreening related gene dee76 protein
GL50803_7030	Prefoldin subunit 3, putative
GL50803_6835	Brix domain containing protein
GL50803_13616	Glycine-rich protein
GL50803_94653	Periodic tryptophan protein 2-like protein
GL50803_11204	Plant adhesion molecule 1
GL50803_8819	Protein required for cell viability
GL50803_10755	Splicing factor 3A subunit 2
GL50803_3993	Polyadenylate-binding protein, putative
GL50803_14174	Glutamate-rich WD-repeat protein
GL50803_15487	WD-40 repeat protein
GL50803_16920	WD-containing protein
GL50803_16202	Axoneme central apparatus protein
GL50803_16572	N-terminal acetyltransferase complex ARD1 subunit, putative
GL50803_17353	G beta-like protein GBL

GL50803_17013	Isoprenylcysteine carboxyl methyltransferase
GL50803_16264	WD-40 repeat protein
GL50803_21512	S-adenosylmethionine-dependent methyltransferase, putative
GL50803_16957	WD-40 repeat protein family
GL50803_33762	WD-40 repeat protein
GL50803_16863	Caffeine-induced death protein 1-like protein
GL50803_27310	Stress-induced-phosphoprotein 1
GL50803_88581	Synaptic glycoprotein SC2

Hypothetical proteins

GL50803_103074	Hypothetical proteins
GL50803_137754	Hypothetical proteins
GL50803_16734	Hypothetical proteins
GL50803_17068	Hypothetical proteins
GL50803_16805	Hypothetical proteins
GL50803_32531	Hypothetical proteins
GL50803_33022	Hypothetical proteins
GL50803_22338	Hypothetical proteins
GL50803_15280	Hypothetical proteins
GL50803_112258	Hypothetical proteins

S2.3 List of 37 *Giardia* proteins which are conserved in all organisms

GL50803_4365	26S protease regulatory subunit 6A
GL50803_7950	26S protease regulatory subunit 6B
GL50803_21331	26S protease regulatory subunit 7
GL50803_86683	26S protease regulatory subunit 7
GL50803_17106	26S protease regulatory subunit 8
GL50803_113554	26S proteasome ATPase subunit S4, putative
GL50803_16867	AAA family ATPase
GL50803_137726	ABC transporter ABCA.1, putative
GL50803_16592	ABC transporter family protein
GL50803_87446	ABC transporter family protein
GL50803_16605	ABC transporter family protein
GL50803_3470	ABC transporter family protein
GL50803_16575	ABC transporter family protein
GL50803_21411	ABC transporter, ATP-binding protein
GL50803_9741	ABC transporter, ATP-binding protein
GL50803_113876	ABC transporter, ATP-binding protein, putative
GL50803_112692	ABC-type multidrug transport system, ATPase component
GL50803_94478	ABC-type transport system ATP-binding chain, putative
GL50803_96460	Alanyl-tRNA synthetase
GL50803_8227	ATP-binding cassette protein 5
GL50803_13777	Cell division control protein 48
GL50803_16200	Developmentally regulated GTP-binding protein 1
GL50803_114246	GTP-binding protein, putative
GL50803_16065	Hypothetical protein
GL50803_15368	Katanin
GL50803_17132	MRP-like ABC transporter
GL50803_17315	Multidrug resistance ABC transporter ATP-binding and permease protein
GL50803_28379	Multidrug resistance-associated protein 1 (most likely to be an ABC transporter)
GL50803_115052	Multidrug resistance-associated protein 1 (most likely to be an ABC transporter)

GL50803_114776	NSF
GL50803_112681	NSF
GL50803_8389	P60 katanin
GL50803_10361	P60 katanin
GL50803_15469	SKD1 protein
GL50803_101906	SKD1 protein
GL50803_16795	Topoisomerase II
GL50803_42442	Transitional endoplasmic reticulum ATPase

S2.4 List of 44 *Escherichia* proteins which are conserved in all bacteria and not found in archaea

gi|16130515|ref|NP_417085.1	23S rRNA pseudouridine	
gi|1788946|gb|AAC75643.1	23S rRNA pseudouridine	
gi|16129049|ref|NP_415604.1	23S rRNA pseudouridylate	
gi|1787327|gb|AAC74170.1	23S rRNA pseudouridylate	
gi|16131057|ref|NP_417634.1	30S ribosomal subunit protein	
gi|1789556|gb|AAC76199.1	30S ribosomal subunit protein	
gi|16130527|ref|NP_417097.1	50S ribosomal subunit protein	
gi|1788958|gb|AAC75655.1	50S ribosomal subunit protein	
gi|90111095|ref|NP_414717.2	CDP-diglyceride	
gi|87081696|gb|AAC73286.2	CDP-diglyceride	
gi|16131993|ref|NP_418592.1	delta(2)-isopentenylpyrophosphate	tRNA-adenosine
gi|1790613|gb|AAC77128.1	delta(2)-isopentenylpyrophosphate	tRNA-adenosine
gi|16128177|ref|NP_414726.1	DNA polymerase III alpha	
gi|1786381|gb|AAC73295.1	DNA polymerase III alpha	
gi|16131569|ref|NP_418156.1	DNA polymerase III, beta	
gi|1790136|gb|AAC76724.1	DNA polymerase III, beta	
gi|16131972|ref|NP_418571.1	Elongation factor	
gi|1790590|gb|AAC77107.1	Elongation factor	
gi|16130793|ref|NP_417367.1	peptide chain release factor	
gi|16129174|ref|NP_415729.1	peptide chain release factor	
gi|2367172|gb|AAC75929.1	peptide chain release factor	
gi|1787462|gb|AAC74295.1	peptide chain release factor	
gi|16129167|ref|NP_415722.1	peptidyl-tRNA	
gi|1787455|gb|AAC74288.1	peptidyl-tRNA	
gi|90111399|ref|NP_416676.4	predicted elongtion	
gi|87082061|gb|AAC75232.2	predicted elongtion	
gi|16128091|ref|NP_414640.1	preprotein translocase subunit,	
gi|1786287|gb|AAC73209.1	preprotein translocase subunit,	

gi|16128052|ref|NP_414600.1	pseudouridine synthase for 23S rRNA (position 746) and tRNAphe(position 32)
gi|1786244|gb|AAC73169.1	pseudouridine synthase for 23S rRNA (position 746) and tRNAphe(position 32)
gi|16131878|ref|NP_418476.1	replicative DNA
gi|1790486|gb|AAC77022.1	replicative DNA
gi|16128075|ref|NP_414624.1	S-adenosyl-dependent methyltransferase activity on membrane-located
gi|1786270|gb|AAC73193.1	S-adenosyl-dependent methyltransferase activity on membrane-located
gi|16131061|ref|NP_417638.1	transcription termination/antitermination L
gi|1789560|gb|AAC76203.1	transcription termination/antitermination L
gi|16130539|ref|NP_417110.1	trans-translation
gi|1788973|gb|AAC75669.1	trans-translation
gi|16130528|ref|NP_417098.1	tRNA m(1)G37 methyltransferase,
gi|1788959|gb|AAC75656.1	tRNA m(1)G37 methyltransferase,
gi|16130698|ref|NP_417271.1	tRNA U65 pseudouridine
gi|1789155|gb|AAC75833.1	tRNA U65 pseudouridine
gi|16129595|ref|NP_416154.1	tyrosyl-tRNA
gi|1787925|gb|AAC74709.1	tyrosyl-tRNA

S2.5 Poster

ESP results were presented as a poster during Genome Informatics Workshop (GIW) 2008 conference, held in Gold Coast, Australia. GIW is the longest running international bioinformatics conference, with a high impact factor for its proceedings. The results on the poster are slightly different from those presented in this chapter, because the data from the poster were the primary results, and more organisms and more up to date databases were used for this chapter.

Eukaryotic Signature Proteins: Guides to modern eukaryotic parasites

Jian Han
Supervisors: Lesley Collins, Patrick Biggs, David Penny
Allan Wilson Centre for Molecular Ecology and Evolution,
Massey University, Palmerston North, New Zealand

Eukaryotic signature proteins (ESPs) are proteins that are found in every eukaryotic proteome but have no significant homology to proteins in Archaea and Bacteria. The protozoans *Giardia lamblia*, *Plasmodium falciparum* and *Trichomonas vaginalis* are all obligate intracellular parasites of humans, causing infections that are amongst leading causes of morbidity and mortality worldwide. We have now calculated ESP datasets for the three parasites and human (the host). These ESP datasets and their databases form the ground work for future research about the parasites in how they maintain their proteomes, and understanding mechanisms of protein loss. Future research will look for essential proteins involved in the growth of the parasites, leading to the investigation of potential drug targets.

Calculation of *Giardia* ESP dataset

From all 6500 annotated *Giardia* proteins, first we removed those that have homologues in any of the 16 bacterial and 9 archaeal species, then we removed proteins that do not have homologues in *Drosophila melanogaster*, *Caenorhabditis elegans* (animals), *Arabidopsis thaliana*, *Oryza sativa* (plants), *Saccharomyces cerevisiae*, *Eremothecium gossypii* and *Yarrowia lipolytica* (fungi). Lastly we screened against human and mouse, and removed proteins which do not have homologues in human and mouse. BLAST searches were performed under default parameters and hits with e-value <1e-4 were evaluated as homologs. This yielded 267 ESPs, these include 262 distinctive proteins (four ESPs possess multiple gene copies in the genome). These 267 ESPs were divided into seven groups according to their function based on their description.

Protein category	Sub category	Number of proteins	
Cytoskeleton	Tubulin	9	43
	Tubulin-associated	25	
	proteasome associated	2	
	microtubule related	1	
	actin	5	
	signal peptidase	1	
membrane	ER and Golgi	10	36
	cytochrome	1	
	SNARE	3	
	vacuole	7	
	calthrin related	10	
	calthrin	1	
	lipid attachments	3	
	WD-repeat	1	
	Endocytosis	1	
nucleus	topoisomerase	1	47
	Zinc finger	11	
	transcriptional factors	4	
	RNA enzymes	9	
	histone-associated	3	
	Helicases	8	
	LIM related	4	
	nucleolus	1	
	DNA polymerase	1	
	ribonucleoprotein	4	
	RNA enzyme	1	
protein synthesis and breakdown	mitochondrial ribosomal protein	1	11
	large ribosomal protein	3	
	ribosome biogenesis protein	2	
	small ribosomal protein	4	
	translation factors	1	
signaling system	ubiquitin conjugation enzymes	12	63
	14-3-3 protein	1	
	calmodulin	5	
	ubiquitin protease	6	
	GTP-binding proteins	8	
	Phosphatidylinositol	6	
	kinases and phosphatases	14	
	kinases and phosphotases	1	
	cell cycle	10	
hypothetical protein	hypothetical protein	26	26
unknown			40

Giardia ESP data listed by protein group

Comparison with previous research

Hartman et al. have previously generated a set of ESPs containing 347 proteins. We performed BLAST search with our set of ESPs versus Hartman's dataset. The results showed that 208 out of our 267 ESPs had homologues in Hartman's set, and 59 did not. The main cause of this variation is the difference in the methods by which ESPs are calculated. Hartman et al. performed their BLAST searches with the yeast proteome as the starting point because *Giardia* genome was very poorly annotated at that time; with recent *Giardia* data we used the more straightforward approach and started our BLAST searches with *Giardia* proteins.

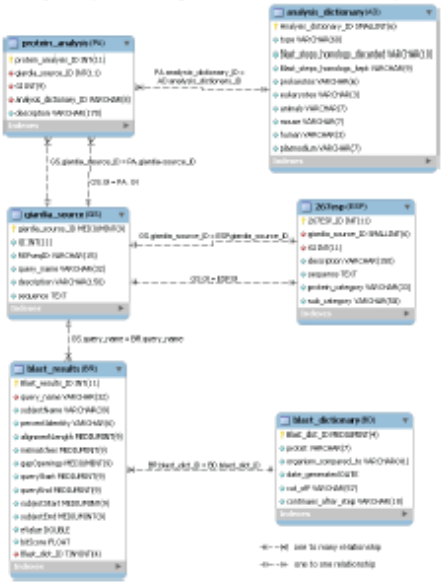
Protein category	Number of ESP with homologues to Hartman's dataset	Number of ESP without homologues to Hartman's dataset
Cytoskeleton	40	3
Membrane	26	11
Nucleus	30	8
Protein synthesis and breakdown	7	4
Signaling system	57	6
Hypothetical protein	13	13
Unknown	26	14
Total	208	59

Future research

Dataset calculation is the preliminary work. We are currently performing a *Giardia* small RNA Solexa run and this data will be integrated into the ESP analysis. Metabolic pathways are presently being chosen for detailed investigations. These investigations will enable close analysis of protein and protein-protein interactions and possible differences between host and parasite proteomes. In future, our research will now focus on parasitic metabolism, observing the trend of their loss and gain of proteins, with the ultimate aim of enabling drug target selection.

MySQL databases containing ESP datasets

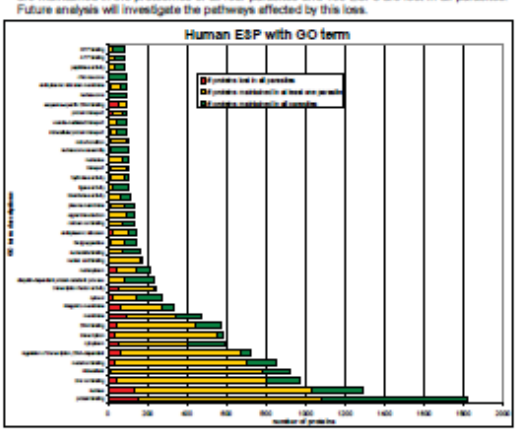
MySQL databases were used for the storage of our ESP data. The advantage of using such databases is that they allow fast and organised information retrieval, and easier updating when newer parasitic genomes/proteomes become available. In addition the relational database management system allows large volumes of information to be efficiently stored.



The *Giardia* database (created using MySQL workbench).

Human ESP dataset

Human ESPs were calculated and 3532 ESPs were derived. Gene Ontology terms have also been linked to the ESPs. We BLASTed the ESPs against the proteomes of *Giardia*, *Plasmodium*, *Trichomonas* and *Cryptosporidium* respectively. We found that 1004 ESPs are maintained in the proteomes of all four parasites and 400 ESPs are lost in all parasites. Future analysis will investigate the pathways affected by this loss.



Top 40 GO term of 3532 human ESPs

Reference

Hartman, H. and A. Fedorov, The origin of the eukaryotic cell: A genomic investigation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(3): p. 1420-1425, 2002.

Human and parasites genome reference: NCBI Genome project home www.ncbi.nlm.nih.gov

Acknowledgements

Health Research Council of New Zealand.
Allan Wilson Centre
Institute of Molecular BioSciences
Massey University

Chapter 3: Phylogenetic analysis using ESPs

3.1 Introduction

3.1.1 Overview

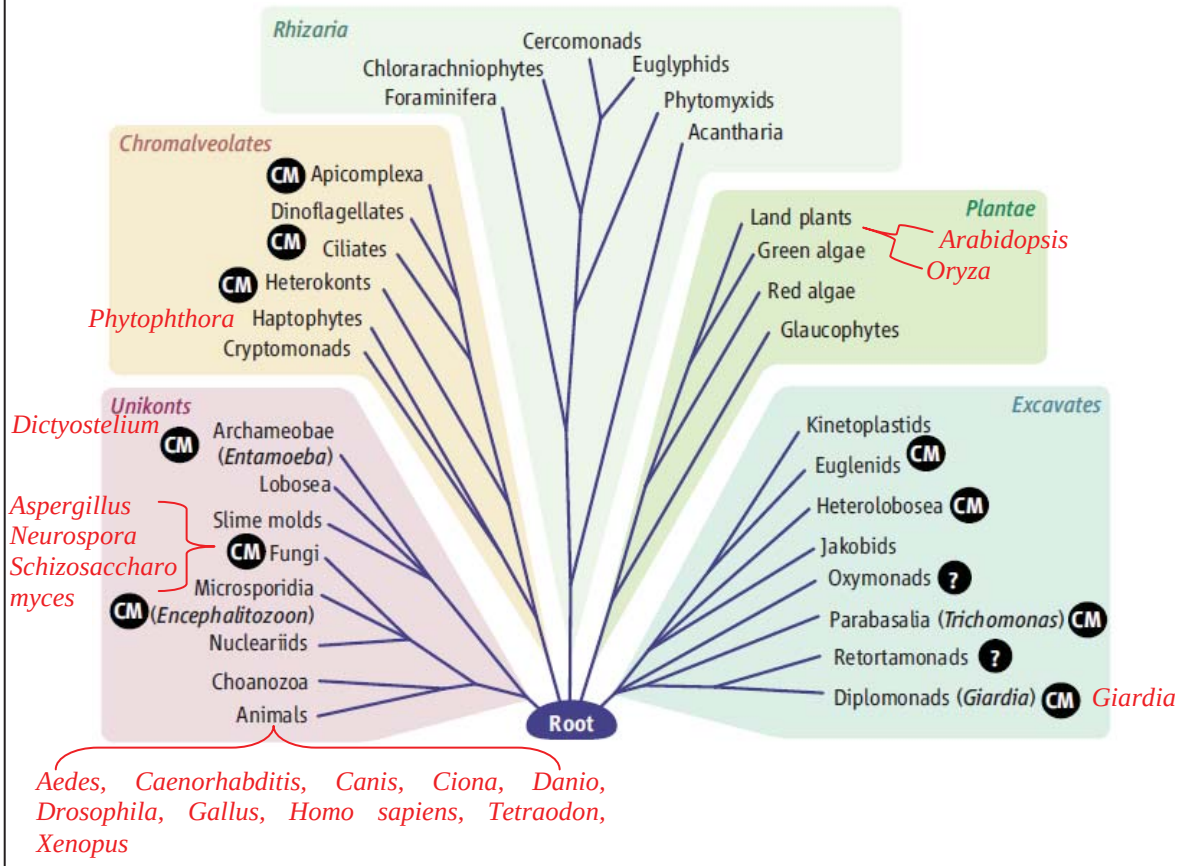
ESPs are conserved in all eukaryotes, and are "considered" to be ancient proteins with a slow and consistent evolving rate. These properties have made them theoretically good candidates for analysing the phylogenetic relationships of eukaryotic species. This chapter examines the possibilities of using ESPs as a special group of proteins in deep phylogenetic analysis. This multi-gene based analysis has used a variety of methods, including building tree networks and concatenation of sequences. A eukaryotic phylogenetic tree was generated here using the longest concatenated sequences to date. The phylogenetic relationship of 18 eukaryotic species, including some divergent species such as *Giardia*, *Dictyostelium* and *Phytophthora*, were analysed using ESPs. In addition, the phylogenetic relationship of 15 mammalian species was also briefly analysed to look at how ESPs perform in less deeper phylogenetic analysis. From the analysis, a proportion of ESPs are found to be good candidates for phylogenetic analysis.

3.1.2 The current phylogenetic system

Due to technological advances, recent years have seen the taxonomy of eukaryotes change rapidly, primarily through the application of phylogenetics. The current system classifies eukaryotes into five supergroups based on molecular and morphological/cell-biological evidence (Keeling *et al.* 2005; Keeling 2007); these are Unikonta (note: some literatures divide this supergroup into Opisthokonta and Amoebozoa (Simpson 2003)), Plantae (aka Archaeplastida), Rhizaria, Chromalveolata, and Excavata.

A recent eukaryotic phylogenetic tree by Keeling *et al.* (Figure 1) illustrates these five supergroups. The earliest eukaryotic divergences (i.e. the root of the tree) are unresolved at present. The positions of the 18 species used in this study are also shown on this tree.

Figure 1. Phylogenetic position of eukaryotic organisms chosen for this project



The positions of organisms chosen are indicated on the eukaryotic tree. CM indicates the presence of cryptic mitochondria (hydrogenosomes or mitosomes). A question mark indicates that no organelle has yet been found. This eukaryotic tree is from Keeling (Keeling 2007) with extra annotations added here.

Giardia along with other diplomonads are placed into the "Excavata", which is a supergroup composed predominately of free-living heterotrophic flagellates. The Excavata supergroup was originally proposed based on the basis of shared morphological characters - a ventral feeding groove and associated cytoskeletal structures (Simpson *et al.* 1999; Simpson 2003), with some additional taxa (parabasalids, euglenids, and oxymonads) linked to the group primarily through molecular studies. The Excavata supergroup includes diplomonads, parabasalids, euglenozoa, heterolobose, jakobids, and several other protists. Molecular phylogeny has not provided clear evidence that Excavata (and Chromalveolata) is monophyletic (i.e. all members are derived from a unique common ancestor) (Hampl *et al.* 2009).

The monophyly of excavates has been challenged in several cases (Simpson *et al.* 2002; Simpson *et al.* 2006; Luo *et al.* 2009). These studies used diverse approaches, ranging

from single gene, multi-gene based phylogenies, to analysis of small nucleolar RNAs (snoRNAs). There are also studies that support Excavata supergroup's monophyly. Hampl *et al.* (Hampl *et al.* 2009) performed analysis by concatenating 143 gene sequences from 48 species including 19 excavates, from which they have concluded that Excavata forms a monophyletic supra-kingdom-level group. However, even with the removal of some fast evolving gene sequences (i.e. remove individual genes which have accumulated large number of changes), they could only obtain an unconvincing bootstrap value of 54%. With long branch taxa removed they did have the support going up to 90%, but *Giardia* was one of the taxa that they removed, thus there were no diplomonads in their tree.

3.1.3 How deep phylogenetic analysis was done in the past

For decades, molecular phylogeneticists have attempted to infer the deepest relationships within the eukaryotic domain of the tree of life. Phylogenetic relationships between distant species are usually performed based on single ubiquitous genes such as 18s rRNA, elongation factors and tubulins (Hashimoto *et al.* 1994). The 18S rRNA is one of the main markers used for *Giardia* diagnostics. Using 18s rRNA has the advantage of easy to amplify with PCR due to highly conserved flanking regions allowing for the use of universal primers (Meyer *et al.* 2010). However, the disadvantage of using 18s rRNA, is that accuracy can suffer from factors such as mutational saturation, unequal mutation rates and rapid evolutionary radiation (Philippe *et al.* 1998). It cannot resolve nodes at all taxonomic levels and its efficacy varies considerably among clades (Abouheif *et al.* 1998), We can end up with a lack of resolution (stochastic error) because of a low number of informative sites and systematic error in tree estimation caused by model violations, and problems related to long branch attraction (LBA, a phenomenon when highly divergent lineages are grouped together, regardless of their true evolutionary relationships) (Felsenstein 1978; Hendy *et al.* 1989; Philippe 2000; Lockhart *et al.* 2005).

Due to an increasing number of genomes available to the public, it is now possible to compute gene trees for many different genes. Now researchers can attempt to obtain a more reliable species tree by building a consensus tree from a set of gene trees (Huson *et al.* 2006), and the approach of constructing a consensus tree from multiple genes is well represented in literature (Holland *et al.* 2003). The other approach of using

multiple genes to confer a species tree is by concatenating sequences (Hampl *et al.* 2009). This approach can eliminate stochastic error, but it is complicated and somewhat controversial, because theoretically, different evolutionary models should be used in different parts of the concatenated alignment. Partitioning analysis can help but, this can make analysis very difficult and time consuming because of the large number of genes present. Other issues include that a large number of model tests would also be required, and the tree building process takes a very long time when all these factors are taken into account.

3.1.4 The ESP approach

The selected genes chosen in previous studies are more or less a random selection of available genes as long as the gene sequence was available in species of interest (e.g. (Hampl *et al.* 2009)). ESPs are conserved throughout eukaryotes, and it is possible that ESPs can outperform other random selections of proteins to determine the phylogenetic relationship of eukaryotes. ESPs could also outperform proteins present in all domains of life, because some species, such as *Giardia*, have a large number of genes which are more similar in sequence to bacterial genes which could bias the position of *Giardia* in eukaryotic phylogenetic studies (Nixon *et al.* 2002; Andersson *et al.* 2003; Morrison *et al.* 2007).

To evaluate the usefulness of ESPs in phylogenetics, I present here an analysis of phylogenies of 18 eukaryotic species with ESPs, using approaches of generating consensus networks and concatenating sequences. The consensus network approach includes data from every distinct ESP, 267 proteins altogether; and the approach of concatenating sequences was performed based on 140 genes and the entire concatenated alignment contained 139,625 sites. My analysis used longer sequences than the previously analysis by Hampl *et al.*, who performed their analysis based on 143 genes with their entire concatenated alignment contained 35,584 sites, but also had a large amount of missing data (averaging 44% per taxon) (Hampl *et al.* 2009). In theory by increasing the length of the dataset, more robust trees can be built. The study will also give more indication of phylogenetic position of *Giardia*, which has long been questioned along with all other long branched lineages.

3.2 Method

3.2.1 Phylogenetic software

Clustal W and Clustal X

The oldest and most widely used multiple sequence alignment (MSA) program that estimates trees as it aligns multiple sequences, is ClustalW (Higgins *et al.* 1988; Thompson *et al.* 1994). ClustalX is an integrated graphical-user interface (GUI) version of the ClustalW multiple sequence alignment program (Thompson *et al.* 1997). It provides an easy-to use work environment for performing MSA and pattern analyses. The main advantage of ClustalX 2.0 is that it provides various formats of output that is needed for other applications. The new guided-tree implementation, compared with the older version, enables larger, faster computations. The latest version of ClustalX (version 2.0) (Larkin *et al.* 2007) was used to align sequences in this study. Other alignment software was considered (such as T-Coffee), but Clustal was chosen for speed and simplicity of operation for this analysis. All alignments in this chapter used ClustalX default parameters (GONNET protein matrix, gap opening cost = 10, gap extension cost = 0.2).

Model testers

Various models have been developed to estimate the total number of substitutions between sequences based on their present states, such as amino acid substitution matrices and gamma distribution models. A model test helps to pick out the best available model for the analysis. ProtTest version 2.4 (Abascal *et al.* 2005) was used for model testing in this study. ProtTest is a bioinformatics/phylogenetic tool for the selection of the most appropriate model of protein evolution (among the set of candidate models) for the data at hand. The software makes its selection is by finding the model with the best likelihood score, or the model with minimum Akaike Information Criterion (AIC), which is a measure of the goodness of fit of a statistical model (Akaike 1974). In this study, the model with best AIC score was used.

Geneious Pro

Geneious Pro is a bioinformatics software platform that allows the user to search, organize and analyse genomic and protein information via a single desktop environment. The platform contains many plug-ins that allow to perform basic and

complex bioinformatic tasks such as aligning sequences and constructing phylogenetic trees (http://www.geneious.com). The PHYML (Guindon *et al.* 2003) plug-in was used for all ML analyses, and Mr Bayes (Huelsenbeck *et al.* 2001) plug-in was used for all Bayesian analyses.

SplitsTree

SplitsTree (Huson 1998; Huson *et al.* 2006) is able to generate a consensus tree network, which attempts to represent all phylogenetic signals present in the given set of gene trees simultaneously up to a given level of complexity (Holland *et al.* 2003). In practice, for a given set of taxa of interest, it is often the case that some of the genes under consideration are not present in all genomes, in which case a super network is able to address this problem (Huson *et al.* 2004). This is because a super network is able to take a collection of partial trees defined on subsets of full taxa set and produces as output a phylogenetic network representing all phylogenetic signals present in the input partial trees. For our study with ESPs we do have protein sequences from all taxa so this feature of the super-network was not required, but the program offered important visualisation aids for further analysis.

3.2.2 Phylogenetic methods

There are two general types of phylogenetic algorithms that were used in this study. The first, Distance method (e.g. neighbor-joining) is a scoring matrix based method. It is very fast but may not give reliable estimates of pairwise distances of divergent sequences, therefore neighbor-joining was used only for some primary analyses. Tree searching methods are better at solving this problem than simple neighbour-joining methods. The maximum parsimony is one tree searching method that tries to find the minimum number of mutations that could possibly reproduce the data. The drawback of this method is that the score of a tree is simply the minimum number of mutations and it does not account for the mechanism or the site on which the mutations occurred, e.g. multiple mutational events at the same site are not considered. Thus, parsimony is very susceptible to long branch attraction – the tendency of highly divergent sequences to group together in a tree regardless of their true relationships (Holder *et al.* 2003). My research covers a range of eukaryotic species including those typically having long branches in phylogenetic trees, therefore maximum parsimony was considered to be unsuitable for my analysis.

Accurately reconstructing the relationships between sequences that have been separated for a long time, or are evolving rapidly, requires a method that corrects for multiple mutational events at the same site. By using the maximum likelihood (ML) method, all possible mutational pathways that are compatible with the data are considered. Bootstrap analysis is often performed to assess the confidence of each branch. Bootstraps are done by taking subsamples and testing if each particular branch occurs in the resulting tree (Holder *et al.* 2003). PHYML is one software used to perform ML analysis (Guindon *et al.* 2003), and as it runs with reasonable speed and reliability, this method is used during majority of this chapter.

Bayesian inference is relatively new, as it was first proposed in 1996 (Rannala *et al.* 1996). It has several advantages over other phylogenetic inference, including easier interpretation of results, and the ability to incorporate prior information. Bayesian inference uses Markov Chain Monte Carlo (MCMC) to approximate the posterior probabilities (Huelsenbeck *et al.* 2001). A Geneious software plug-in Mr Bayes(Huelsenbeck *et al.* 2001) was also used for this project. It has been suggested that Bayesian method is less time consuming than ML (Huelsenbeck *et al.* 2001). However this was not the case in my experiments, where one tree took up to a week to construct by using default settings in Mr Bayes (MCMC chain length: 1,100,000, sample frequency: 200, burn-in length 100,000, substitution matrix: Poisson). Because my analysis involved building hundreds of trees, Bayesian inference was confined to generating trees for only a few specially chosen alignments.

3.2.3 Analysis procedure

For each *Giardia* ESP, sequences of its homologues were obtained using the MySQL *Giardia* database which contained all the BLAST results used for ESP calculation. By use of a Perl script (see supplementary material S3.2), the highest scored homologue from each of 17 eukaryotic organisms was recovered. The original sequences from *Giardia* as well as its homologues from the 17 eukaryotic organisms were then aligned using ClustalX version 2.0.11 (Larkin *et al.* 2007). The procedure was performed on all 267 *Giardia* ESPs. All alignments were imported into software Geneious Pro version 5.0.4 (http://www.geneious.com) for further phylogenetic analyses.

Bayesian trees were built for a three alignments of various lengths in order to test the method. Analyses were performed using default settings in Geneious plug-in Mr Bayes

(Huelsenbeck *et al.* 2001) (MCMC chain length: 1,100,000, sample frequency: 200, burn-in length 100,000, substitution matrix: Poisson). Later, different chain lengths and burn-in length were used in order for the process to take less time. However Mr Bayes still took a long time (about four days) to build trees, and therefore this process was abandoned early on in favour of PHYML (see below). Densitree Version 1.45 was used to display tree samples (Bouckaert 2010).

The Geneious plug-in PHYML (Guindon *et al.* 2003) was used to draw maximum likelihood trees, with one tree was built for each alignment. Ten bootstraps were performed for each tree. This number was relatively low but it was a compromise for the time taken to build 267 trees (using this method a ~500 residue alignment of the 18 species takes about five minutes), these trees are only the primary analyses and a high number of bootstrap was not essential.

The 267 trees were converted into NEXUS format with Geneious. By using SplitsTree4 (version 4.11.3), neighbor-net trees were built for each tree. These trees give some information on how suitable the alignments are to build trees. A consensus network was also built using data collected from all 267 trees.

The 267 trees were then manually examined to determine the best functional group for phylogenetic studies. By using prior phylogenetic knowledge (e.g. the animals should be grouped together, *Giardia* should not be found grouped with animals, fungi or plants, the trees were divided into three groups: Group A (excellent), Group B (good) and Group C (bad) based on the definitions below.

Group A trees have all animals (*Aedes*, *Caenorhabditis*, *Canis*, *Ciona*, *Danio*, *Drosophila*, *Gallus*, *Homo sapiens*, *Tetraodon*, *Xenopus*) in one clade, all plants (*Arabdopsis* and *Oryza*) in one clade and all fungi (*Aspergillus*, *Neurospora* and *Schizosaccharomyces*) in one clade, the bootstrap value of any of these three clades has to be no less than 70 percent. Other organisms *Phytophthora*, *Dictyostelium* and *Giardia* are of less concern as long as they do not show up inside the three clades mentioned above (an assumption for this study). This is because the phylogenetic ordering of these three longer-branching organisms is less clear. Group A trees were considered of excellent quality, because the animals, fungi and plants are expected to be monophyletic.

Group B trees have only one or two species being misplaced within the major clades, or if there are low bootstrapping values for the three clades mentioned above (even if the

topology of the tree fulfils all the requirement of Group A).

Group C trees contain so called “star” trees, in which all major branches originates from a single point, and implies that all of these branches are unresolved. This group also contains trees have more than two clearly misplaced species (e.g. animal species grouped with fungi), likely due to incorrect paralogue used. Any tree that displayed properties not falling into the Group A and Group B was placed in this group.

After this filtering procedure, a new consensus network tree was built using the Group A (excellent) and Group B (good) trees only. Also the Group A and B protein sequences of the same organism were concatenated, and a new PHYML tree was built on concatenated ESP sequences. A model test was first performed before this analysis, and out of the four substitution models (Dayhoff (Dayhoff *et al.* 1978), Mitochondrial Adachi and Hasegawa (Adachi *et al.* 1996), Jones-Taylor-Thornton (Jones *et al.* 1992), and Whelan and Goldman (Whelan *et al.* 2001)) that was available in the Geneious plug-in PHYML, the best model was Whelan and Goldman (WAG) substitution matrix with the proportion of invariable sites being 0.037 and with a gamma shape (4 rate categories) of 1.164 (WAG+Γ4+I). This model showed the best likelihood score, as well as the best AIC (Akaike Information Criterion) score, indicating this is the most accurate and the most fitting model to this data.

3.3 Results

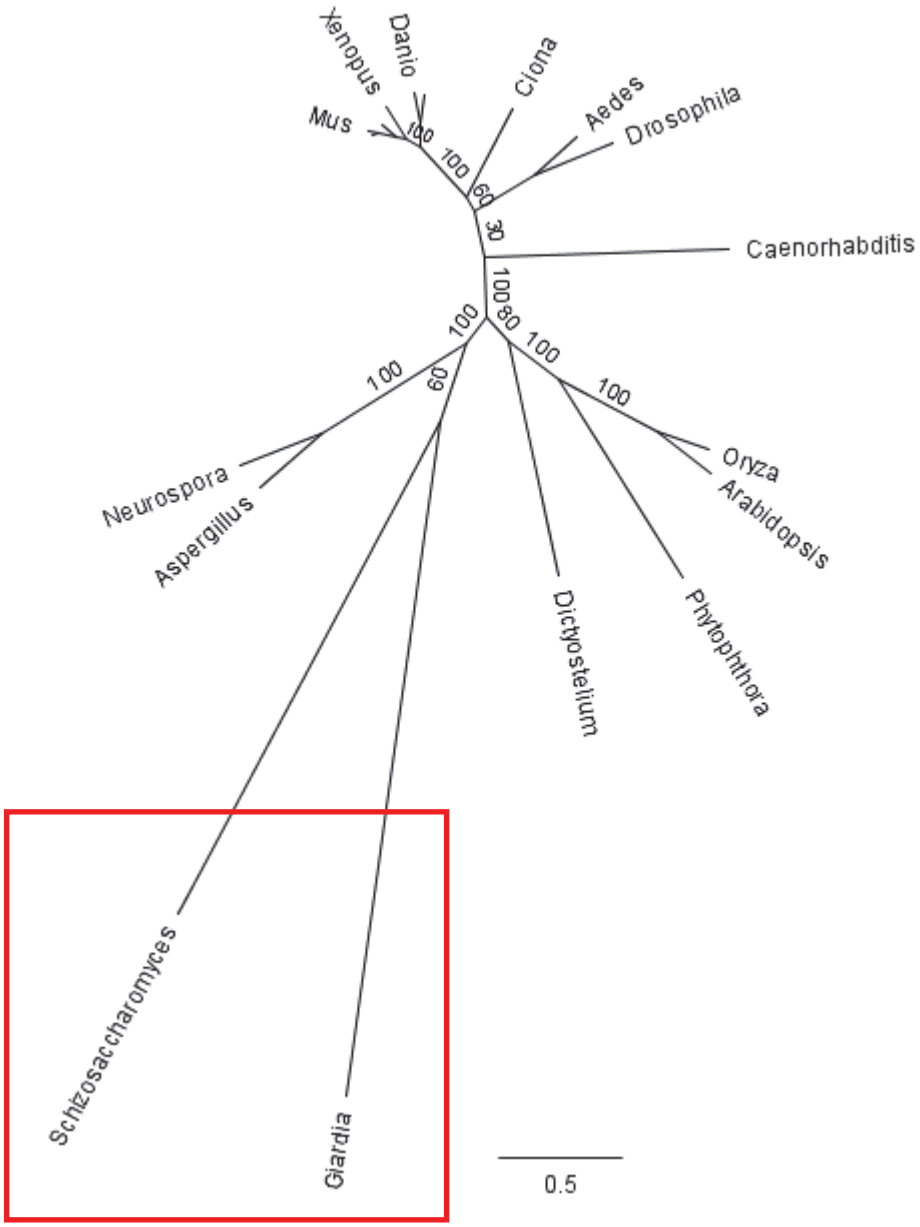
3.3.1 ML trees of ESP

A total of 267 groups of sequences were aligned with ClustalX, the ML trees built using the PHYML plug-in of Geneious. One tree was built for each ESP and a total of 267 trees were built. For ESPs with multiple identical copies (e.g. Alpha tubulin, Beta tubulin, Histone H2A, Histone H2B and Histone H4), only one copy was used. Because of the large dataset, the number of bootstraps was limited to 10, due to time constraints.

An example of a PHYML tree is show in Figure 2. Note that we see the long *Giardia* branch which was expected, and the animals forming a branch of its own, as do the three fungi and plant species. From this tree we can see the animals are grouped together in the top half of the tree, the two land plants (*Oryza* and *Arabidopsis*) are grouped together, as well as the two fungi (*Neurospora* and *Aspergillus*), the other fungi *Schizosaccharomyces* has been obviously mis-grouped with *Giardia*. This is likely to have been caused by a wrong paralogue from *Schizosaccharomyces* used to build the

tree (see Section 3.3.3 for more explanation of this scenario). The bootstrap values for all three aforementioned clades were 100%, indicating the robustness of the grouping. The divergent species (*Giardia*, *Dictyostelium* and *Phytophthora*) formed long branches. The majority of ESPs produced similar trees, with more or less misplaced species, but about 40% of ESPs produced trees of lower qualities, such as uninformative "star" trees or had misplaced a large number of species.

Figure 2. Unrooted ML tree of protein GL50803_93275 (Translational activator GCN1) from different species

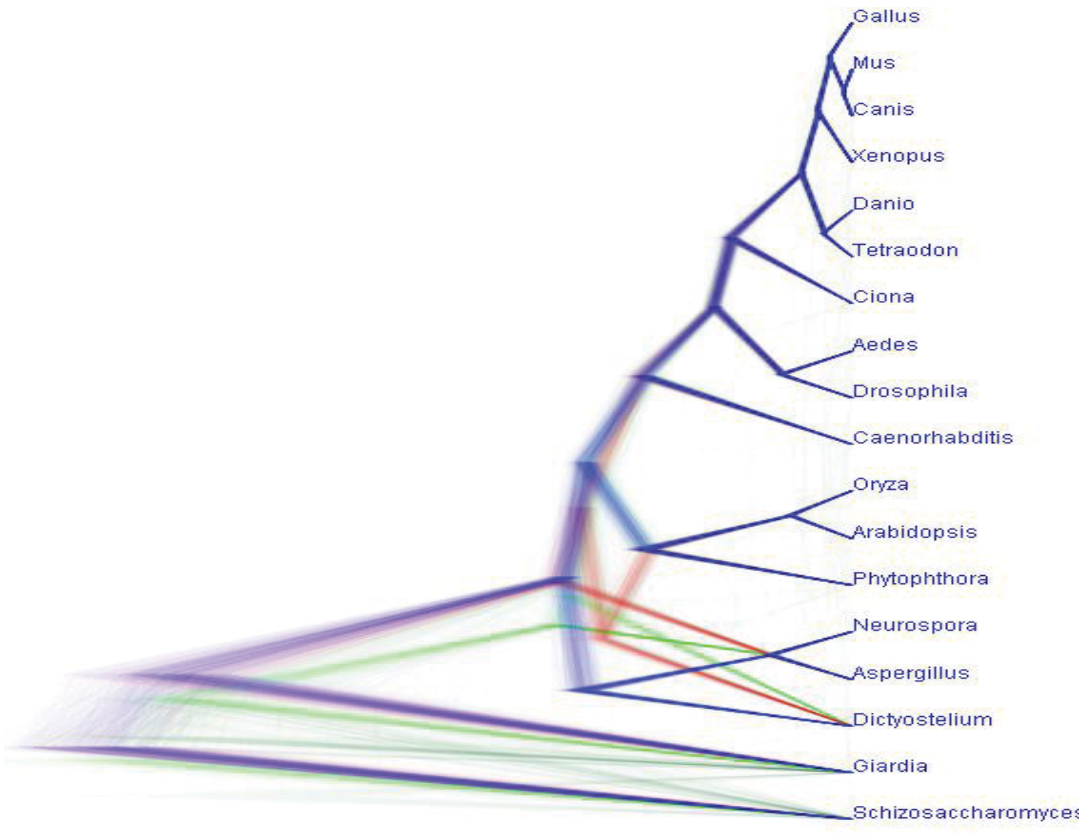


Bootstrapping values are shown on branches, the scale bar is number of substitutions per site. Note the mis-grouping of *Schizosaccharomyces* with *Giardia* (red box), this is likely to be caused by a wrong paralogue from *Schizosaccharomyces* was used to build the tree.

3.3.2 Bayesian analysis

Bayesian analysis was performed on only a few alignments due to its high time consumption. The analysis was performed using default settings (MCMC chain length: 1,100,000, sample frequency: 200, burn-in length 100,000, substitution matrix: Poisson) either with "unconstrained branch lengths" or with "uniform branch lengths" as a prior. The prior did not make any difference in the topology of the few trees generated but made some differences in the posterior probability.

Figure 3. DensiTree output of Bayesian analysis of protein GL50803_93275



DensiTree is capable of showing the all the trees that the heuristic search went through in MCMC chain. It is very useful to show the disagreements of the tree. The blue lines are the branches which agree with the final tree; lines of other colours (red and green) are other signals. A molecular clock was assumed in this analysis.

DensiTree can displays all trees sampled in the MCMC chain simultaneously. If the "uniform branch lengths" setting was used during tree searching, this will result in a diagram that displays the degree of uncertainty very well. An example of DensiTree diagram is shown in Figure 3. The alignment used is same as that of Figure 2, generated

using ML. The main difference between the two trees is the position of *Dictyostelium*: with ML, *Dictyostelium* is closer to the plants; whereas with Bayesian inference *Dictyostelium* is closer to fungi, but with large amount of disagreement with between trees. It should also be noted if the long branches of *Giardia* and the "problem taxon" *Schizosaccharomyces* (probably result of a wrong homologous used, see next section) are deleted, both trees would be very similar to the established tree in Figure 1. This indicates both methods are reasonable for the tree building.

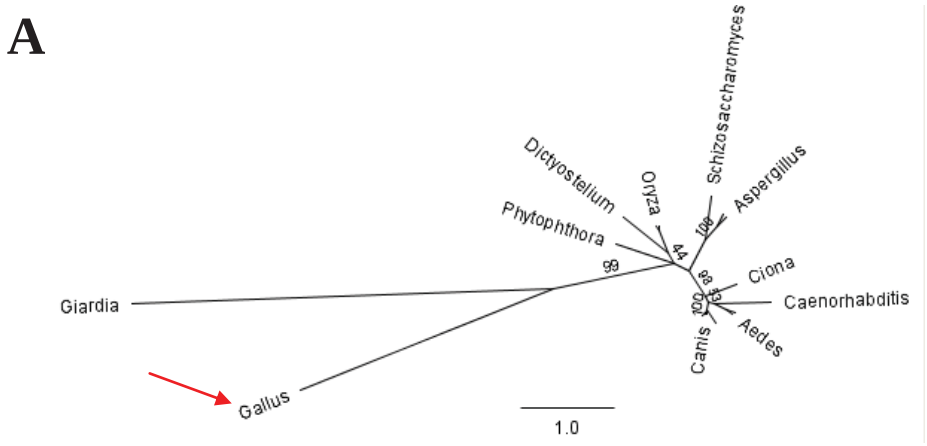
The issue with time consumption of Mr Bayes has prevented this analysis from being performed extensively. Due to the large number of trees required, ML was used to generate nearly all trees in this chapter.

3.3.3 Unexpected tree shapes

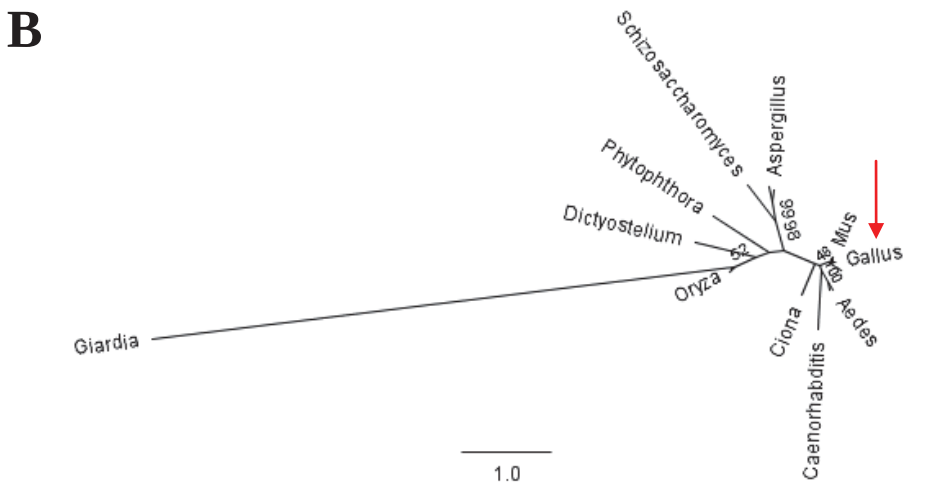
Trees with unexpected shapes can be formed if a wrong paralogue (i.e. after a gene duplication, one copy of the gene may change function as it accumulates mutations) was used for tree construction. If the right paralogue (i.e. the original copy of the gene that has retained the original function) is used then the tree should display the true phylogenetic relationship. One of the ESPs showing this misplaced paralogue effect was the 26S proteasome non-ATPase regulatory subunit 7 (GL50803_7896). The ML tree was generated using the default procedure (see section 3.2.3) and the following tree was produced (Figure 4). Note this tree (and also Figure 5) was chosen to demonstrate the outcome of using incorrect paralogue, not the presentation of phylogeny.

Clearly *Gallus* has been misplaced into the same clade with *Giardia*, with 99% bootstrap value (Figure 4A). I found that this protein has many paralogues in *Gallus*, the best match being ENSGALP00000008530 (Protein A) with bit-score of 65.5; the other match was ENSGALP00000000999 (Protein B) with slightly less bit-score of 65.1. In the default tree generating procedure, Protein A was used as the *Gallus* protein because of its higher bit-score. When Protein B was used as the *Gallus* protein, a different tree was generated, and *Gallus* was placed back to the animal clade where it clearly belongs (see Figure 4B).

Figure 4. Unrooted ML tree of orthologues for GL50803_7896 from different species showing effect of including an incorrect gene paralogue



ENSGALP00000008530 (Protein A) was used as *Gallus* orthologue. Bootstrapping values are shown on branches, the scale bar is number of substitutions per site. *Gallus* (indicated by the arrow) has been grouped with *Giardia* when it should be grouped with other metazoans.

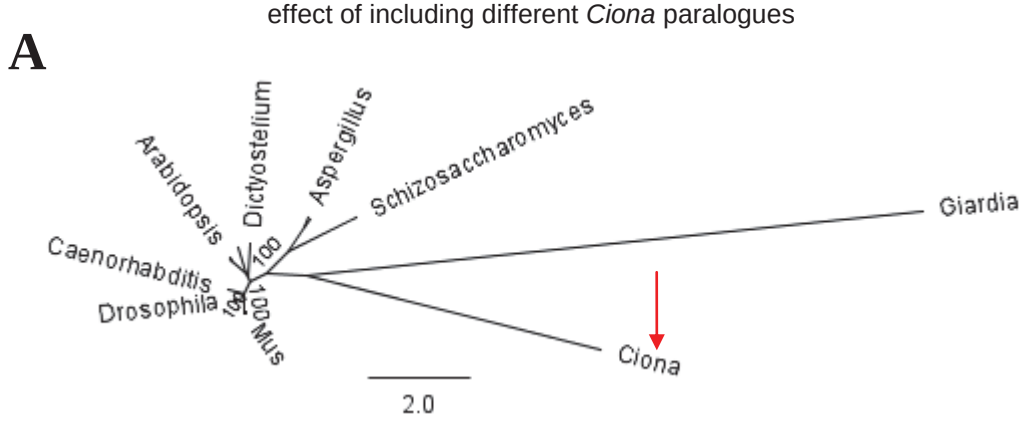


ENSGALP00000000999 (Protein B) was used as *Gallus* orthologue. Bootstrapping values were shown on branches, the scale bar is number of substitutions per site. Using the correct homologue Protein B has placed *Gallus* back in the animal clade and displays a greater distance between the *Giardia* protein and its homologues (*Gallus* branch is indicated by the arrow).

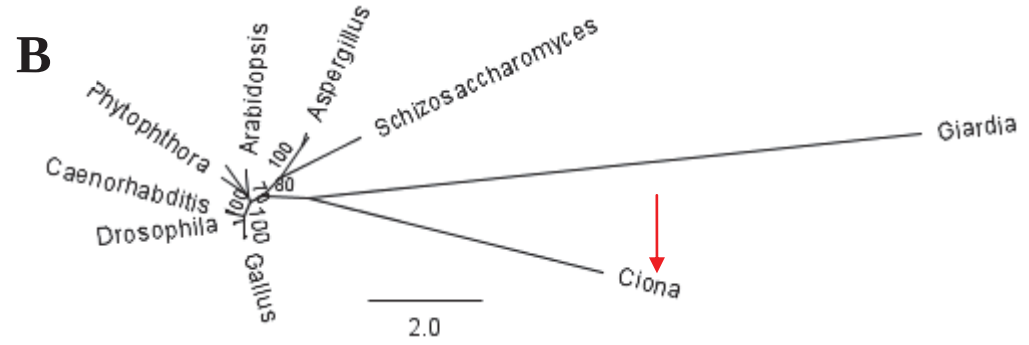
However, for some trees, the obvious misplacing mistakes could not be fixed by using a different paralogue, e.g. GL50803_15339 (Adaptor protein complex large chain subunit BetaA). No matter which paralogue was used, *Ciona* remains as a long branch (Figure 5). This may indicate that the paralogues may both have evolved rapidly. The problem of using the wrong paralogues was not resolved for all ESPs, as it would take a long

time, thus I simply carried on the analysis using the alignments and trees already obtained.

Figure 5. Unrooted ML tree of orthologues of GL50803_15339 from different species showing effect of including different *Ciona* paralogues



ENSCINP00000018012 was used as *Ciona* orthologue. Bootstrapping values are shown on branches, the scale bar is number of substitutions per site. *Ciona* (indicated by the arrow) has been grouped with *Giardia* when it should be grouped with other metazoans.



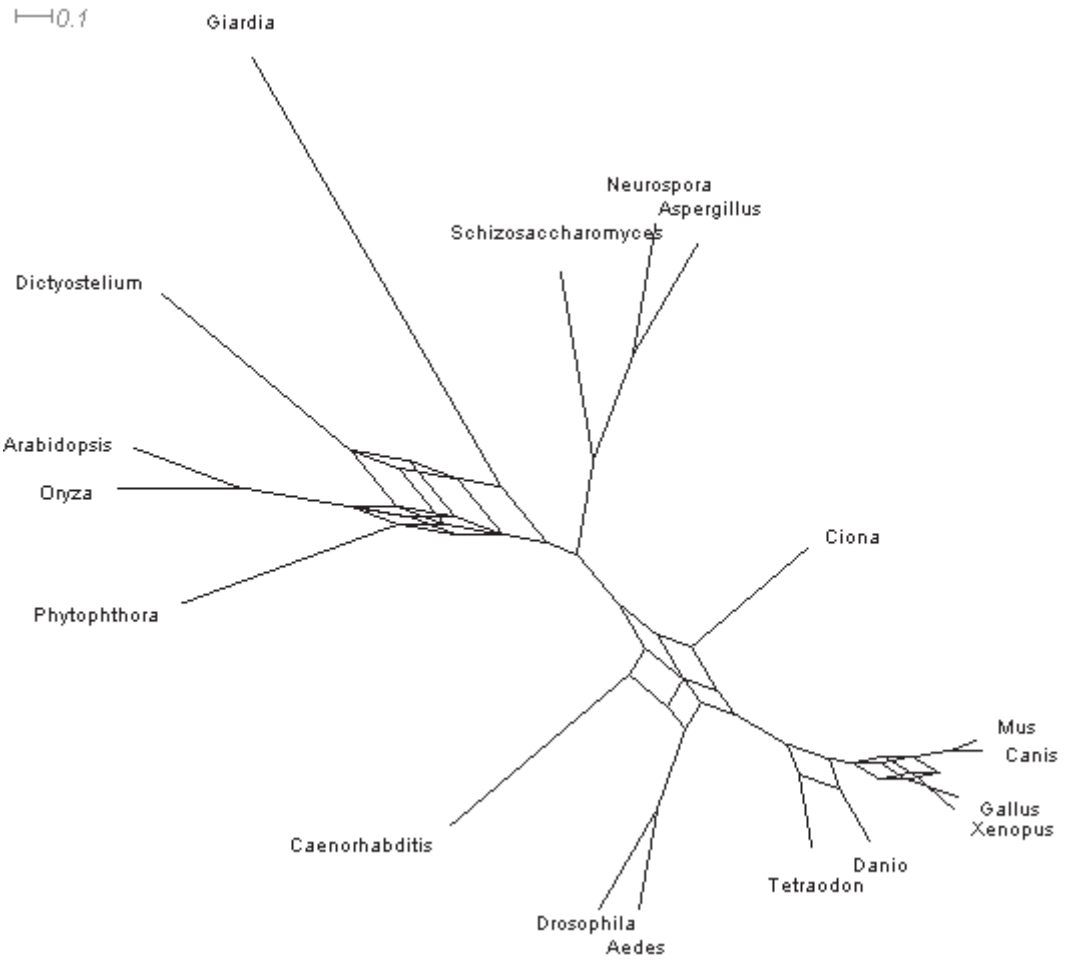
ENSCINP00000018007 was used as *Ciona* orthologue. Bootstrapping values were shown on branches, the scale bar is number of substitutions per site. Using this homologue *Ciona* is still approximately at the same position (*Ciona* branch is indicated by the arrow).

3.3.4 Consensus tree

One good way of visualising the conflicting signals is to build a consensus tree network. Consensus tree networks were built using SplitsTree 4 (version 4.11.3), using the 267 ML trees generated for each ESP as input (Figure 6). The way to visualising the conflicting signals in a consensus tree network is by "splits". The edges lengths are proportional to the average branch length from gene trees in which a particular edge occurs, and this method allows the branch lengths to be visualised (refer to

supplementary material S3.1 for further explanation). The splits threshold was set to be 0.1, which means only splits occurring in at least 10% of all trees are displayed. Conflicting signals were displayed as splits, and this splits threshold value made tree easier to visualise. If all the noise from the data were shown, the consensus tree would be very messy (containing many box-like format), whereas setting a high splits threshold can result in a "star tree" incapable of resolving some taxa.

Figure 6. Unrooted consensus tree built using 267 ML trees



The edges lengths are proportional to the average branch length from gene trees in which a particular edge occurs, and the splits threshold was set to be 0.1. The scale bar is the number of substitutions per site.

The consensus tree (Figure 6) has integrated information from all 267 trees to a single network, showing the phylogenetic relationship between the 18 organisms. The metazoa clade (containing ten metazoan species), fungal clade (containing *Aspergillus*,

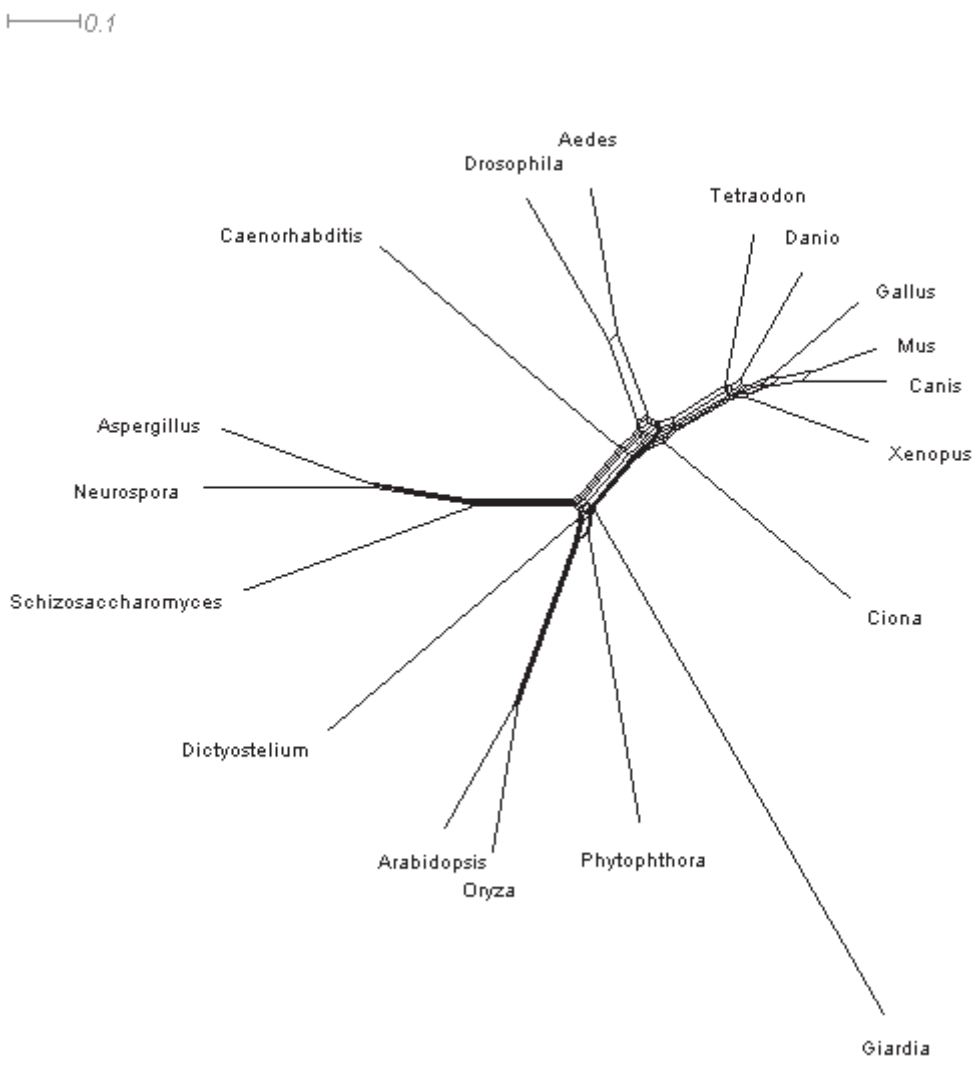
Neurospora and *Schizosaccharomyces*) and the plant clade (containing *Arabidopsis* and *Oryza*) can be clearly identified from the network. The network shows *Giardia* is relatively close to *Dictyostelium*, but this result comes with large amount of noise; in addition, these two species formed the longest branches out of all taxa, and result could be due to long branch attraction. *Dictyostelium* was the sole Amoebozoa representative in this study. Keeling *et al.* has placed *Dictyostelium* and other Amoebozoa species along side Opisthokonta species (containing animals and fungi) into group "Unikonts". From the consensus network, there is no evidence that Opisthokonta and Amoebozoa are monophyletic.

Ciona is in the Chordata (vertebrate) phylum, the same phylum as Mouse (*Mus*), Dog (*Canis*) and Chicken (*Gallus*) etc, but this result showed *Ciona* and other chordates being polyphyletic and *Ciona* is on a branch on its own. The suspected reason is that the wrong paralogues were often included from this organism (other possible reason: *Ciona* may have undergone recent genome duplication or mass gene duplication which altered the rate of some proteins). For this very reason, another consensus tree was built only using selected ESPs which were considered to be more informative on the phylogenetic relationships (see Section 3.3.6). From that consensus tree, *Ciona* did appear to be monophyletic with other chordates.

Although this consensus tree suggested that there are conflict signals around the central eukaryotic node, it had a greater resolution around this node than previous results.

An average consensus tree of 18 eukaryotic species was built using SplitsTree (Figure 7). Average consensus trees have the advantage of taking both the number of trees with a particular split and the branch lengths into account. Bootstrapping values however could not be implemented into building of the consensus tree, because SplitsTree could only take the final tree generated from each protein to consideration. This network is very similar to the consensus tree, suggesting the closest relative to *Giardia* is *Dictyostelium*, but this result also comes with large amount of noise.

Figure 7. Unrooted average consensus tree built using 267 ML trees



Average consensus tree take account of both branch length and the number of trees with this split. The scale bar is the number of substitutions per site.

3.3.5 Divide trees based on topology comparisons with expected tree

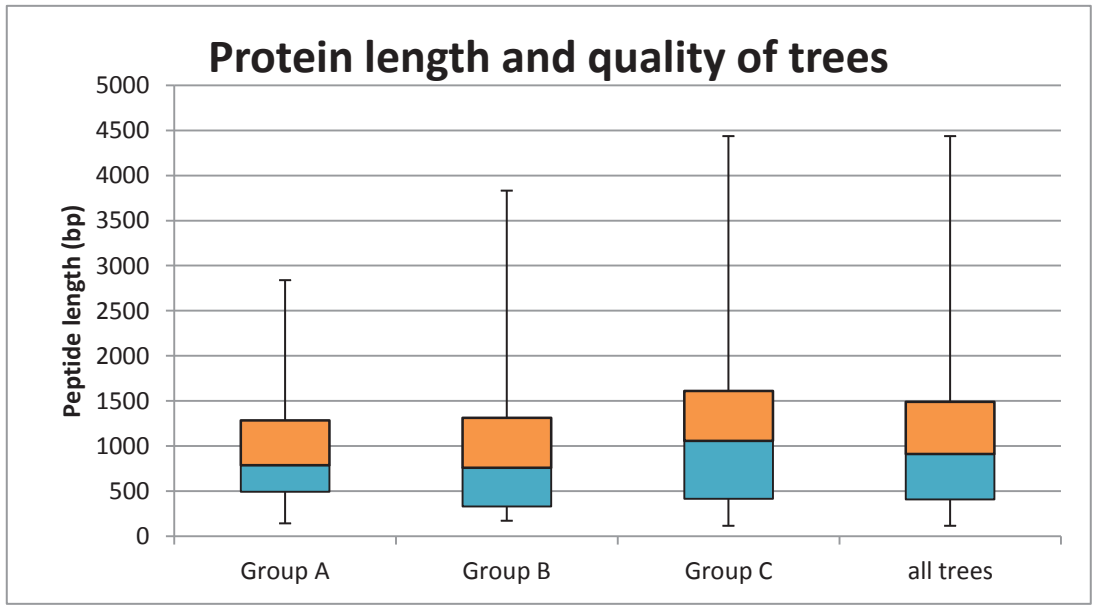
The 267 trees were divided into three groups (see method section for detailed explanation of the three groups) based on the topology and bootstrap support of the tree:

- Group A contained what considered to be 50 excellent trees, each tree has all animals, fungi and plants into three separate clades shows these three are clear monophyletic with bootstrapping value no less than 70 percent.
- Group B contained 90 trees each having only one or two species being misplaced, or with low bootstrapping values for the three clades mentioned above even the topology of the tree is good.

- Group C contained 127 trees, which are considered trees not very useful for phylogenetic study.

The reasons that Group C trees are less informative include the following: the sequences can be too short for analysis to give any meaningful phylogenetic signal; the genes have many paralogues, and inclusion of the wrong paralogues from species caused incorrect phylogenetic relationship being portrayed; and events such as horizontal gene transfer may have taken place in these genes. In addition, the log likelihood values of trees had no bearing on which group the trees were put into, as it gives no indication of how accurate the tree is compared with the true phylogenetic relationship of the species. Even a tree with a excellent log likelihood value could be very messy and have many misplaced species. This is because the log likelihood value is solely based on the sequence lengths and number of gaps and substitutions in the sequences.

Figure 8. Box plot of gene length distribution.



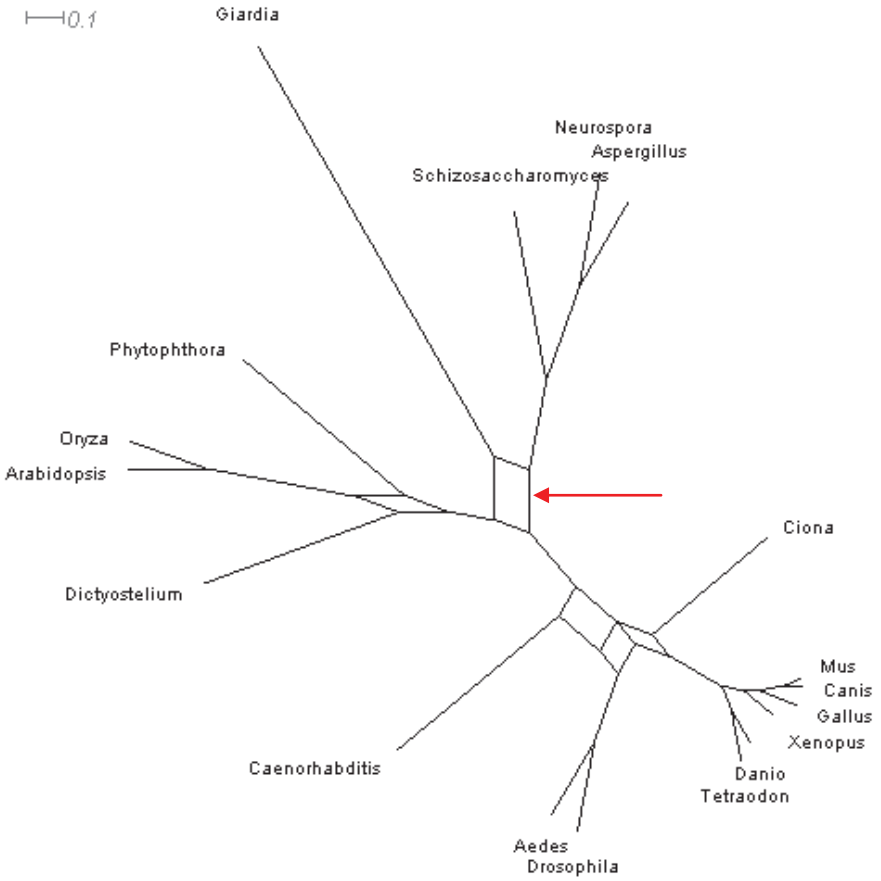
The relationship between the grouping of the trees and sequence lengths was then investigated (Figure 8). The above box plot suggests that proteins between 500-1000 amino acid in length are generally suitable for phylogenetic analyses in this situation. Short sequences may not have enough substitutions to dictate a meaningful phylogenetic relationship. Genes with long sequences might contain multiple domains, the event of gaining or losing a domain in one clade species but not others will magnify the evolutionary distance between them; moreover the different segments of the query

Giardia protein can match to a different proteins paralogue in each organism, resulting orthologue ambiguity.

3.3.6 Consensus tree with split tree, software results can be deceptive

A splits tree of the consensus tree was generated combining the excellent trees (Group A) and good trees (Group B) (Figure 9). If all the noise from the data were shown the consensus tree would be very messy. Hence the split threshold was set to 0.15 (i.e. only include splits if they occur in at least 15% of all trees), to make the tree easier to visualise. Setting a too high splits threshold can result in a "star tree".

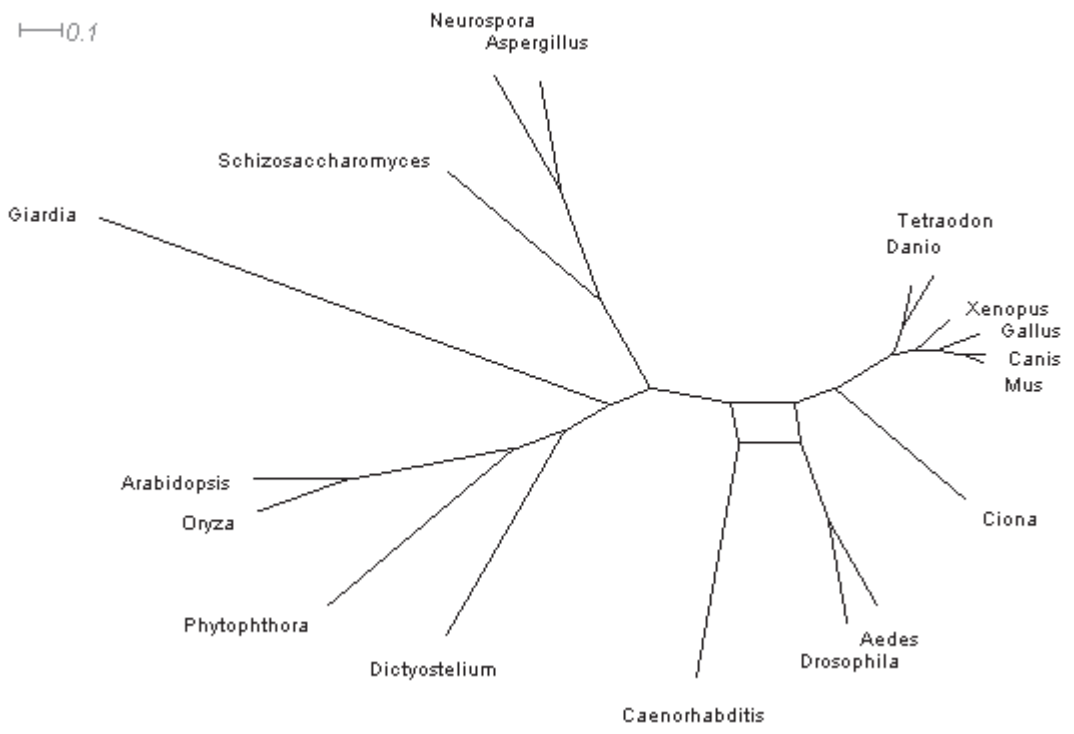
Figure 9. Consensus network Type 1



The splits threshold was 0.15 (splits occur less than 15% of all trees were not shown). The lengths of the edges are proportional to the average branch length from gene trees in which a particular split occurs; thus the branching pointed to by the red arrow does not suggest there were more trees that placed *Giardia* with fungi, but rather that the trees that placed it there had a longer branch length at that point (refer to Supplementary material for further explanation). The scale bar is the number of substitutions per site.

When the splits threshold was increased to 0.18 (i.e. splits occur less than 18% of all trees were not shown), we almost have a tree without any disagreement, the only exception was that the node separating *Caenorhabditis* and the insects was unresolved (Figure 10). Even the central root of the eukaryote has been resolved in this tree. The result could be easily misinterpreted as this is a network tree with solid support, and that all branches (including *Giardia*) can be placed where it is with little doubt. This however, appeared too good to be true. The 0.18 split threshold means that a split is shown if it occurred in more than 18% of all trees; splits associate with the same organism could have appeared in several different places in up to 82% of other trees, but the frequency the split occurs in each of these other place is less than 18%, hence they were not shown. Users should be aware about this in the future.

Figure 10. Consensus network Type 2

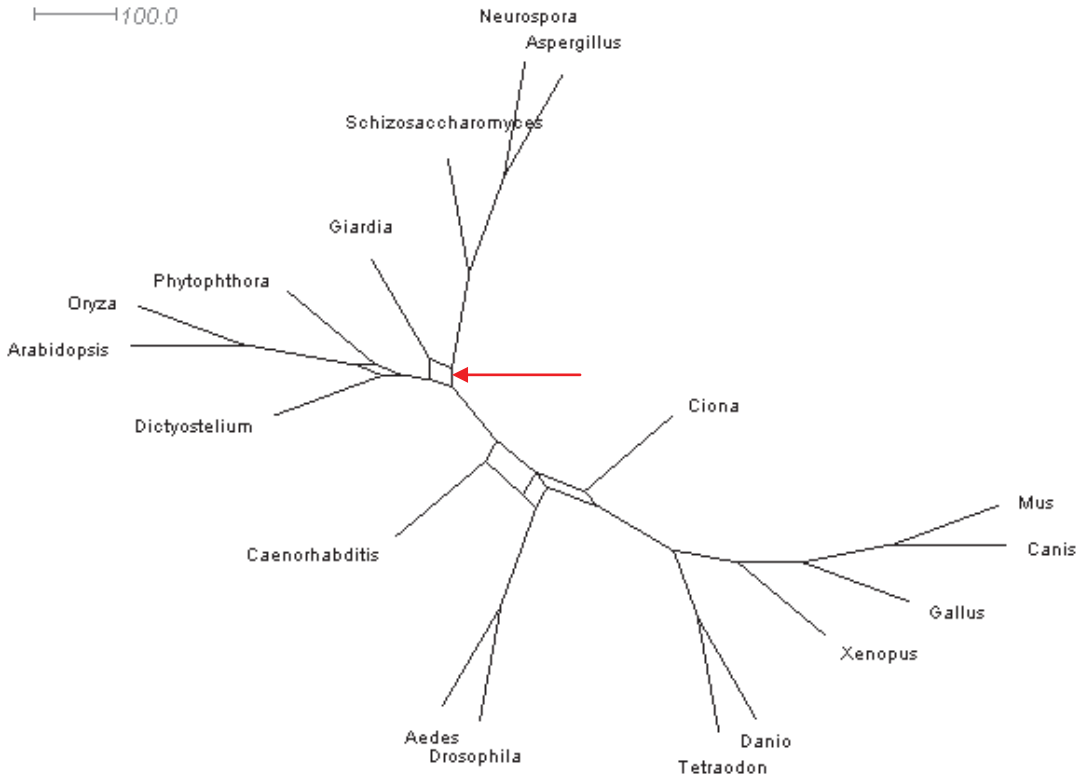


The splits threshold was 0.18 (splits occur less than 18% of all trees were not shown). Note *Giardia* is now grouped with plants, *Phytophthora* and *Dictyostelium*. The scale bar is the number of substitutions per site.

Another consensus tree was built differently (Figure 11), with the edge lengths drawn proportional to the number of trees the splits occur. The actual branch lengths from each tree are ignored. Therefore even though *Giardia* should form a long branch, the length

of *Giardia* in this network is the same as all other species. This visualisation method focuses on the number of trees with the same topology rather than branch lengths.

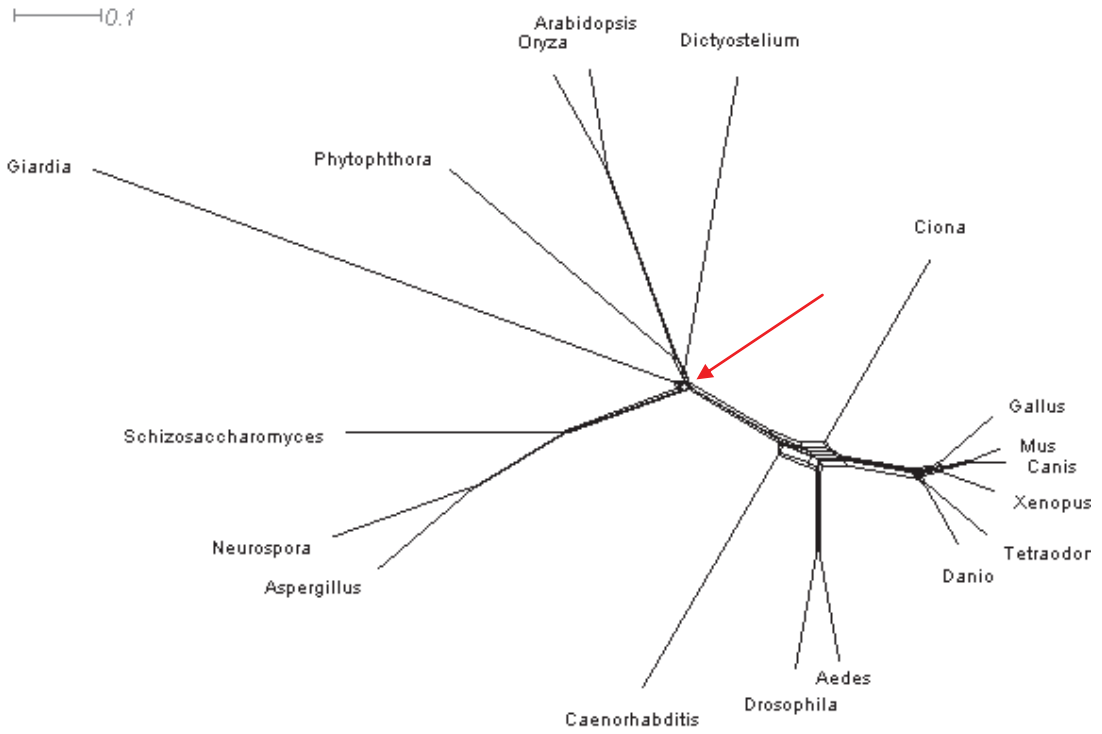
Figure 11. Consensus network Type 3



The splits threshold was 0.15 (splits occur less than 15% of all trees were not shown). The lengths of the edges are proportional to the number of trees the splits occur; thus the red arrow suggest there were more trees placed *Giardia* with plants-*Dictyostelium* and *Phytophthora* than fungi (c.f. Figure 9, refer to Supplementary material for further explanation). The scale bar is the number of trees count.

An average consensus network was then constructed using SplitsTree (Figure 12). This network has the advantage of taking both number of trees with a particular split and the branch lengths into account. This representation is more tree-like, and the splits are not seen as clearly as in consensus networks. This network clearly has well defined animal, fungal and plant clades, but the branches separating the central node are still very much unresolved, shown by the large amount of noise indicated by the red arrow. In addition, the splits between *Ciona*, insects and Chordates have a large amount of disagreement.

Figure 12. Average consensus



Average consensus generated with 50 Group A and 90 Group B trees. The red arrow indicated the large amount of conflicting signals at the central eukaryotic node. The scale bar is the number of substitutions per site.

When using average branch lengths for presentation (Figure 9), the network suggested that *Giardia* is placed in same clade as fungi, but by using the number of trees in the representation, the result suggested otherwise (Figure 11). Consensus networks also have the drawback of taking all trees equally, despite the variation of sequence lengths of alignments the trees were built from.

Overall, even using ESPs the phylogenetic relationship of *Giardia* was inconclusive by constructing networks. There are many disagreements between trees generated from each ESP, because the sequences in each tree are relatively short, and thus, there will be stochastic error because of the low number of informative sites. In addition, more species in the Chromalveolata and Excavata supergroup would be useful for future phylogenetic analyses to break the long branch lengths and permit better resolution.

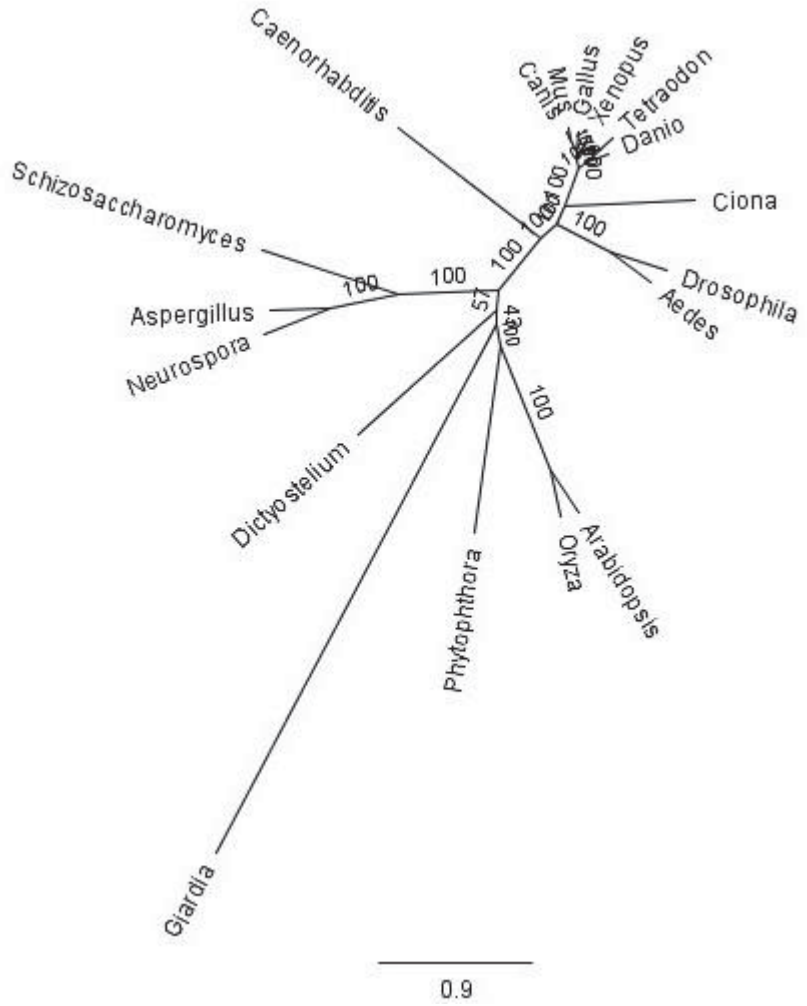
3.3.7 Tree building by concatenating sequences

Another method for analysing phylogeny with multi-genes approach is by concatenating sequences. Using a Perl script, sequences of the same species from group A and B alignments were concatenated. This analysis was based on a total of 140 genes and the entire concatenated alignment contained 139,625 sites. Previously Hampl *et al.* performed a similar analysis based on 143 genes and their entire concatenated alignment contained 35,584 sites, but this alignment suffered from a large amount of missing data (averaging 44% per taxon) (Hampl *et al.* 2009). A model test was first performed before tree-building. The best model was Whelan and Goldman (WAG) substitution matrix with the proportion of invariable sites being 0.037 and with a gamma shape (4 rate categories) of 1.164 (WAG+Γ4+I). Therefore the WAG+Γ4+I model was chosen to be the amino acid substitution model for the tree-building. The tree was built and was bootstrapped 100 times (Figure 13).

The tree built in this analysis was much easier to interpret than the trees using consensus networks. The animals, fungi and plants all formed monophyletic groups of with 100% bootstrap support. The bootstrap for supergroup Opisthokonta (containing animals and fungi) was moderate (57%), considering previous studies strongly supported that opisthokonts form a monophyletic group (Parfrey *et al.* 2006; Steenkamp *et al.* 2006).

There are only four species that do not belong to supergroup Unikonta, which are *Giardia, Photophthora, Arabidopsis* and *Oryza*. The support for supergroup Unikonta (containing Opisthokonta and Amoebozoa) was low (43%). There are very few recent studies on the monophyly of unikonts (this supergroup was originally proposed on that unikonts ancestrally had a single flagellum and single basal body (Cavalier-Smith 2002)). This grouping is however unlikely, since flagellated opisthokonts, as well as some flagellated Amoebozoa actually have two basal bodies, as in typical 'bikonts'. From our analysis, it is inconclusive whether Unikonta is monophyletic, due to the low bootstrap support.

Figure 13. Unrooted tree generated using the WAG+Γ4+I model



This tree was built with 50 Group A and 90 Group B sequences concatenated, using WAG substitution model with estimated proportion of invariable sites (+I) and 4 substitution rate categories and estimated gamma distribution (+G). Bootstrapping values were shown on branches. The scale bar is number of substitutions per site.

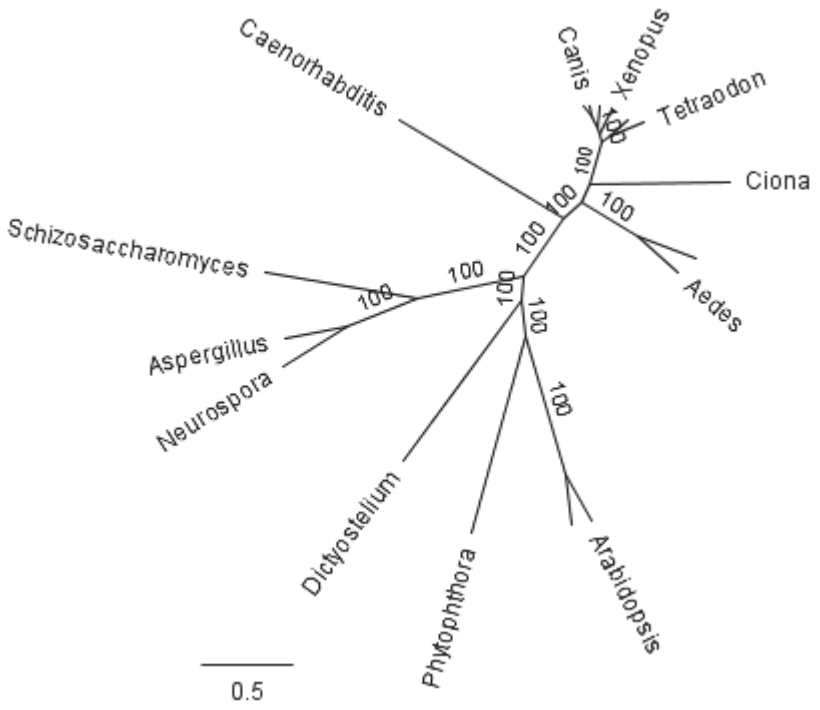
Phytophthora, the only Chromalveolata representative branched closer to plants than species from any other supergoups, with 100% bootstrap support. This is different to the tree by Keeling *et al.* (Keeling 2007) which indicated Chromalveolata and Plantae were two independent supergroups. This result agreed with Hampl *et al.*'s papers (Hampl *et al.* 2009) which suggested that Chromalveolata and Archaeplastida are paraphyletic (i.e. all members along with some other unmentioned species are derived from a unique common ancestor). Hampl *et al.*'s papers have also included some rhizaria species inside the clades which included Chromalveolata and land plants. The different results

between Keeling *et al.* and Hampl *et al.*'s papers may have due to difference in their methods, Hampl *et al.* used sequence concatenation method similar to ours, whereas Keeling *et al.* paper was a review article and the eukaryotic tree was hypothesised using various molecular and morphological data.

Giardia formed the longest branch out of all taxa, this was also observed in the consensus networks constructed. Reason for this can be that the parasitic life style of *Giardia* caused its rate of mutation becoming higher than that of other taxa analysed (i.e. heterotachy).

Another tree was built using the same model, but with the *Giardia* removed to study how the presence of a species with a very long branch affected the entire tree. With *Giardia* taken out, the bootstrap support for every branch went up to 100 (Figure 14).

Figure 14. Unrooted tree generated with *Giardia* removed

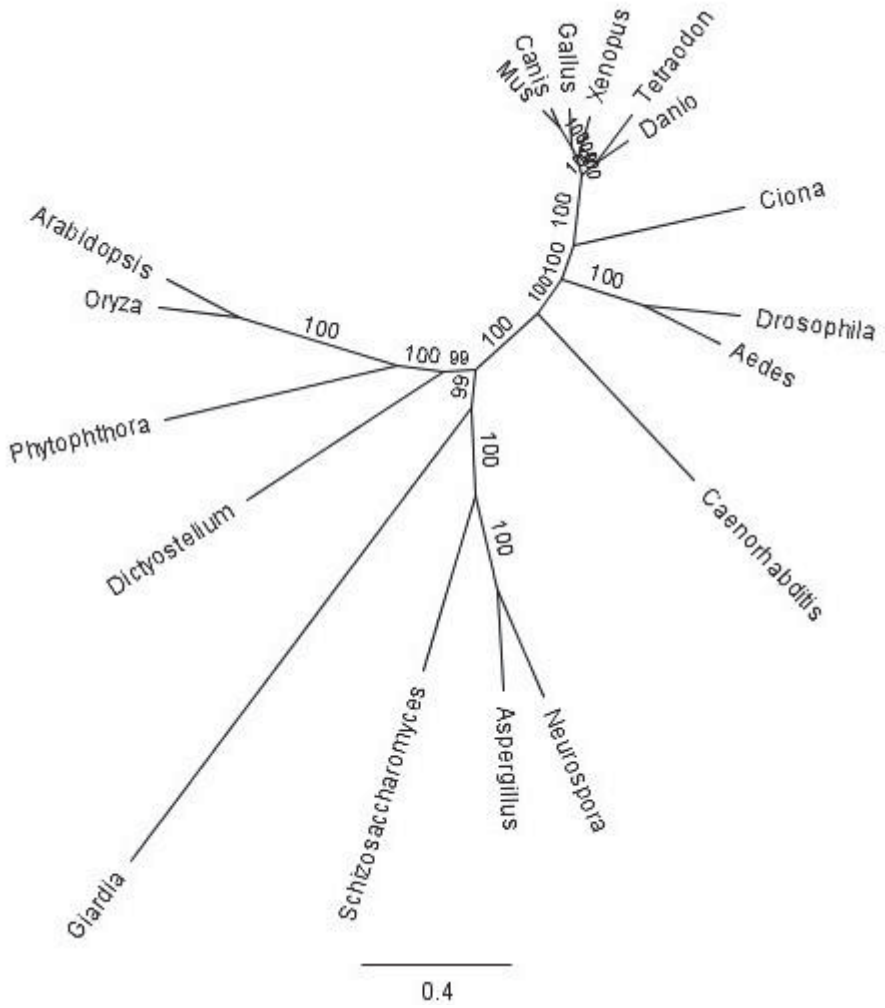


The tree was built with 50 Group A and 90 Group B sequences concatenated using WAG substitution model with estimated proportion of invariable sites (+I) and 4 substitution rate categories and estimated gamma distribution (+G). Bootstrapping values were shown on branches. The scale bar is number of substitutions per site.

3.3.8 Tree built with different model

For comparison, a PHYML tree was built using the Dayhoff substitution model and the tree was bootstrapped 100 times (Figure 15).

Figure 15. Unrooted tree generated using the Dayhoff model



The tree was built with 50 Group A and 90 Group B sequences concatenated using Dayhoff model with fixed proportion of invariable sites of 0 and only 1 substitution rate category. Bootstrapping values were shown on branches. The scale bar is number of substitutions per site.

Using the Dayhoff model, the topology of the tree is exactly same as the WAG+Γ4+I model, except that the position of *Giardia*. *Giardia* has been placed inside the "Opisthokonta", which consists of animal and fungi and a few others, and there is strong support here that this super-kingdom is monophyletic (the bootstrap support was

also deceivingly high at 99%). Thus it seems the result is likely due to long branch attraction. Another tree built (not shown because the topology extremely close to Figure 15) using WAG model but this time using fixed proportion of invariable sites (0 and 1 substitution rate category), built a tree with the same topology as above, but the bootstrap support for the *Giardia*/fungi clade was as low as 26%. This indicates that although the bootstrap value is very high for this tree, it does not necessarily reflect the true phylogenetic relationship. For an accurate phylogenetic result, using the right model is vital.

3.3.9 Relationship between protein function and its phylogenetic usefulness

Moving away from the mechanics of tree-building, this section relates protein functions to the "quality" of tree generated. Previously *Giardia* ESPs were divided into several categories according to the protein function (section 2.3.1). For each recorded category, the numbers of proteins belonging to Groups A, B and C described in section 3.3.5 are shown in Table 1.

All four ribosome biogenesis proteins have generated Group A (considered excellent quality) trees; membrane proteins have also generated a high proportion of (14/34) Group A trees, including proteins found in cellular signature structures (CSSs) such as vacuole, ER and Golgi. Proteins from these two groups have given more consistant results in phylogenetic analysis.

Proteins of the signalling system gave less consistent trees (more Group C trees and less Group A trees). This group contains a variety of signal transduction proteins and enzymes (noticeably kinases and phosphatases) which can evolve at different rates due to speciation., These proteins play different roles in different species and appear to quickly evolve to adapt to a new role (protein engineering and directed evolution demonstrated new enzymes can arise quickly (Quin *et al.* 2011)). The other group which gave a high proportion of Group C trees were the cytoskeletal proteins. This might be due to the fact that actins and tubulins have many paralogues (genes arose from duplication, e.g. alpha and beta tubulin are paralogues), and a wrong paralogue may be annotated in a species (e.g. as with the case described in 3.3.3). Having multiple copies of the same gene could have an effect on the evolving rate of these proteins in some species.

Table 1. Function and phylogenetic utility

Protein category	Sub category	Group A trees		Group B trees		Group C trees	
Cytoskeleton	actin	6	2	9	1	21	1
	microtubule related		1		0		0
	proteasome associated		0		1		1
	tubulin		1		1		3
	tubulin-associated		2		6		16
Hypothetical protein	hypothetical protein	4	4	4	4	2	2
Membrane	cell adhesion	14	1	11	0	9	1
	clathrin related		5		4		2
	endocytosis		0		0		1
	ER and Golgi		4		1		4
	lipid attachments		0		3		1
	vacuole		4		3		0
Nucleus	DNA polymerase	2	0	20	1	19	0
	histones		0		2		5
	histone-associated		0		2		2
	LIM related		1		2		1
	ribonucleoprotein		0		2		0
	RNA enzymes		0		5		4
	topoisomerase		0		0		1
	transcriptional factors		1		3		1
	transcriptional transactivators		0		0		2
	zinc finger		0		3		3
Protein synthesis and breakdown	ribosome biogenesis protein	7	4	5	0	3	0
	large ribosomal protein		0		4		0
	small ribosomal protein		2		0		1
	translation factors		1		1		2
Signalling system	14-3-3 protein	13	0	21	0	63	1
	calmodulin		0		0		5
	cell cycle		2		3		3
	GTP-binding proteins		3		5		12
	kinases and phosphatases		5		7		23
	phosphatidylinositol		2		1		4
	ubiquitin		0		0		2
	ubiquitin conjugation enzymes		1		4		9
	ubiquitin protease		0		1		4
Unknown	unknown	6	6	18	18	10	10

Nuclear ESPs have also produced poor results in our tree building. This is quite surprising since this group comprises RNA enzymes, histones and transcription factors, which are all robust ESPs. Possible reasons are that short sequence lengths (in case of

histones) or multiple paralogue in the genomes (RNA enzymes and transcription factors) may be having an effect here.

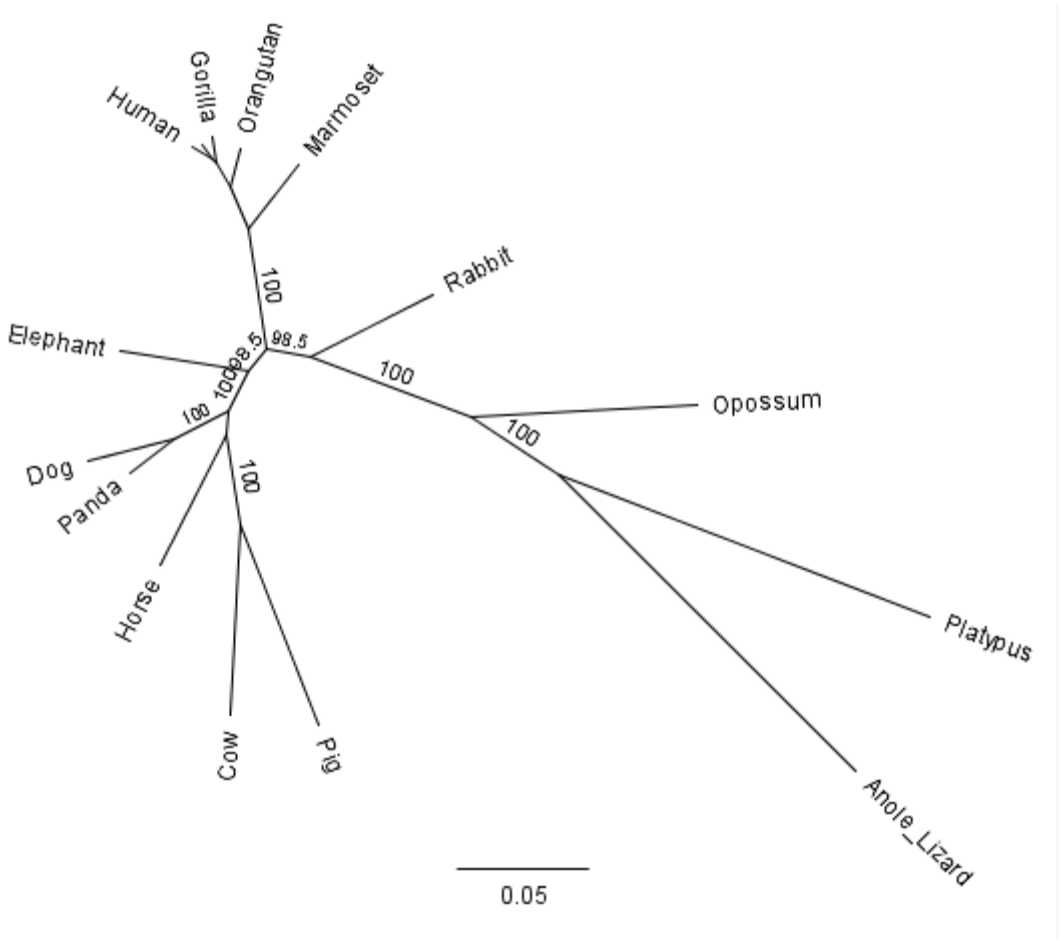
3.3.10 Phylogenetic analysis of mammal species using ESP

Building a phylogenetic tree containing organisms from many supergroups is a difficult task, and the true phylogenetic relationship between distant species is often debatable. Therefore a simpler phylogenetic analysis was performed on mammalian species. Mammals first appeared ~225 million years ago (Kielan-Jaworowska 2007) and since there is good fossil evidence for mammalian evolution there are publications with which the ESP results can be compared. This analysis here can investigate if ESPs are good candidates for phylogenetic analysis over a shorter evolutionary distance.

Fifteen mammalian species from Ensembl, all with at least 6X genome coverage were used. The species are: *Ailuropoda melanoleuca* (Panda), *Bos taurus* (Cow), *Callithrix jacchus* (Marmoset), *Canis familiaris* (Dog), *Equus caballus* (Horse), *Gorilla gorilla* (Gorilla), *Homo sapiens* (Human), *Loxodonta africana* (Elephant), *Monodelphis domestica* (Opossum), *Mus musculus* (Mouse), *Ornithorhynchus anatinus* (Platypus), *Oryctolagus cuniculus* (Rabbit), *Pan troglodytes* (Chimpanzee), *Pongo pygmaeus* (Orangutan), *Sus scrofa* (Pig). *Anolis carolinensis* (Anole Lizard) was also downloaded to serve as the outgroup for this study.

Each of the 50 Group A and 90 Group B (described in section 3.3.5) ESPs were BLASTed against each mammal, and the hit with highest bit-score was recorded from each organism. The annotated transcript sequences were used for this analysis because the mammals are closely related, and comparing nucleotides should obtaining better results. The 140 cDNA sequences were concatenated for each organism, and a phylogenetic tree was built using PHYML (Figure 16). The resulting tree is almost identical to previously published and highly regarded mammalian trees (e.g. (Campbell *et al.* ; Prasad *et al.* 2008; Asher *et al.* 2009)), indicating that the ESPs are very good for phylogenetic analysis of species with around 200 million years of divergence.

Figure 16. Phylogenetic tree of mammalian species



Tree was built using 140 cDNAs concatenated from each species. HKY85 was the substitution matrix. The tree was bootstrapped 200 times and bootstrap values are shown on the tree.

3.4 Discussion

3.4.1 ESPs as candidates for evolutionary studies

ESPs as a group, hold interesting potential for further evolutionary studies, due to their presence in all eukaryotic genomes. Not all ESPs are useful, with some ESPs being less informative for tree-building, because:

1. Sequences can be too short for analysis to give any meaningful phylogenetic signal. This problem can be solved if more ESPs were used (by means of concatenating sequences).

2. Though ESPs are considered to be very slow evolving proteins, different evolving rates can still occur for individual ESPs. Partitioning analysis can potentially solve this problem but has a computational barrier (i.e. will take a long time for model testing).

3. Some genes have many paralogues, and inclusion of the wrong paralogue can cause an incorrect phylogenetic relationship to be portrayed. This problem can be solved by manually adding the right paralogues into alignments to replace the wrong ones (this process was not performed on many proteins due to time constraints).

4. Rare events such as horizontal gene transfer may have taken place for some genes. There is no simple solution of this problem apart from not to use any ESPs which may have this problem.

My results show some conflict across the eukaryotic tree as expected. The concatenated sequence analysis did however, separate the main supergroups of eukaryotes, although there were some low bootstrap values from the concatenated sequences, and it is unknown but expected that by adding more species to break long branches would increase the bootstrap support. On the positive side the bootstrap support was 100% for all branches in animal and fungi clades in the concatenated analysis. ESPs were well performed to resolved phylogenies of closer species, such as phylogenies among mammals.

Resolving the central root in the eukaryotic tree (see Figure 1) is a difficult task for any researcher, and this study definitely did not expect to solve this problem. However, some trees had potential to point the way and with the addition of ESPs from more species to break up the long branches may lead towards a clearer answer.

3.4.2 Limitations

Alignments

The problem of using wrong paralogue (see section 3.3.4 for more details) was not solved for all ESPs, as it would have taken a long time for the manual corrections required. This was not a large problem for the overall analysis since if there were a large number of wrong paralogues in one particular tree, then the ESP was simply discarded for the subsequent analysis as it would naturally fall into the Group C tree category.

Limitations of generating consensus network

In constructing the consensus trees, the bootstrap values of each tree that it was built from were ignored. Therefore there is no information about how strongly the data supports each branch. The other drawback of using consensus networks is that each tree was taken equally despite variation on their sequence lengths (this issue can be addressed using the sequence concatenation method).

When using average branch lengths to a construct a network, the splits can sometimes be misleading, because they show only the average branch length of trees with the split, but can completely ignore how many trees has this particular split. By tree counts, the other method, this problem is solved but this is unable to show branch lengths.

Limitations of concatenating sequences

In general, this method is being considered a better method than generating consensus networks, due to its easier methodology and easier interpretation of results as the uncertainties are displayed as bootstrap values instead of complicated geometrical boxes.

Individual ESPs are expected to have evolved at different rates, and theoretically different substitution models should be used for the different ESPs. Partitioning analysis could be performed with a better understanding of the protein set, but this would be very computationally intensive. Even trees built without partitioning can take up to two weeks per tree, thus partitioning analysis was not performed due to time constraints. Generally speaking, all ESPs have arisen around the time that first eukaryotic cells were formed, thus the rates of evolution over the long period until today are expected to be quite similar. The aim of this research was not focused on the actual phylogenetic relationship between the supergroups of eukaryotes, but instead to investigate how good ESPs are as candidates for phylogenetic research. Concatenating ESP sequences in general can be useful but their quality status (Group A, B or C) should be taken into account before deciding to include a protein in the concatenation.

Long branch attraction

Long branch attraction (LBA) is the tendency of distant sequences to group together in a tree regardless of their true relationships. In general, LBA is more likely cause problem when fast evolving sequences or highly divergent sequences are used for analysis. In this study, the effect of LBA has been attempted to be minimised by using slow evolving, ancestral proteins. Protein sequences are less susceptible than DNA sequences to LBA, since there are 20 amino acids and only four nucleotides. However, LBA could is still inevitable since the organisms in the study were so divergent and there were not many species available which will break the long branches. The problem has been seen when *Giardia* was grouped with fungi when the Dayhoff model was used (see Figure 15).

LBA can occur purely because the grouping of longer branches is more statistically supported (Hendy *et al.* 1989), or it can be the outcome if different rate of evolution occur in taxa (i.e. heterotachy) (Lockhart *et al.* 2005). The case of heterotachy can occur in two scenarios, first is that one species have faster mutational rate across all sites (Felsenstein 1978), and second the faster evolving species possesses more variable sites than others (called mosaic evolution) (Simon *et al.* 1996). It seems that *Giardia* forms a significantly longer branch than other species, meant it possibly have a higher rate of evolution -a case of heterotachy. However we do not yet know which type of heterotachy it is. The LBA seen in Dayhoff model might be caused by heterotachy, and not merely a statistical matter. Using the WAG+Γ4+I model, *Giardia* was not grouped with fungi, LBA did not occur.

When more basal eukaryotic organisms are available, adding taxa from the Excavata supergroup would definitely help resolved the long branch of *Giardia*. If the long branch of *Giardia* was taken out (see Figure 14), the bootstrap value of the tree went up, this again consistent with given a good model, ESPs are very good candidates for phylogenetic analyses.

3.4.3 Conclusion and Future work

To conclude, the ESP dataset could be a powerful tool to study eukaryotic phylogeny. I have demonstrated phylogenetic analysis using ESPs worked very well over short evolutionary distances as shown by the mammal trees. When analysing taxa with longer evolutionary distances such as phylogenetic relationship between the supergroups of eukaryotes, there were some promising results, although the long branching of some taxa (most notably *Giardia*) was problematic, and it was placed differently with different models. Adding close related taxa will help resolve long branches, and with more genomes becoming available, especially those of highly divergent organisms, the ESP approach is expected to produce even better results.

Not all ESPs are equal in their ability to produce good trees. I have grouped ESPs according to the quality of trees they produced (Groups A, B and C). Future research might consider discarding Group C ESPs as they can produce biased results. Group A and B ESPs are excellent candidates for tree-building especially using the sequence concatenation method of tree-building. The other method (consensus network) also did

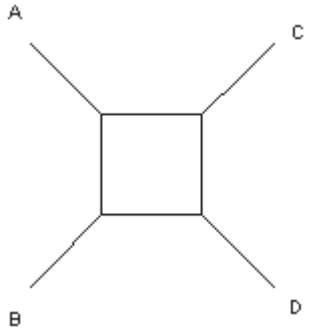
not perform too badly though the results are harder to interpret, and this method may not necessarily be discarded.

For future work, it would be interesting to see if ESPs can produce good results when analysing phylogeny of other well understood clades of eukaryotes. This will further consolidate that ESPs are valuable for phylogenetic analyses. A number of animal clades with unclear phylogenetic relationship can also be analysed using the ESP approach, e.g. the relationship between insects and other groups of arthropods (555mya of divergence) (Strausfeld *et al.* 2011). Furthermore, phylogenetics of deep branching eukaryote taxa can take place using the ESP approach for a second time, when more high coverage genomes of other early diverging and evolutionarily important eukaryotes become available (e.g. *Naegleria gruberi* has recently been sequenced (Fritz-Laylin et al. 2010) and could be a useful species to include in the study). Inclusion of more basal eukaryotic organisms, such as excavates, chromalveolates or Rhizaria species in the analysis should answer many more questions about the relationship between the supergroups, and decipher the monophyly of unresolved groups of eukaryotes.

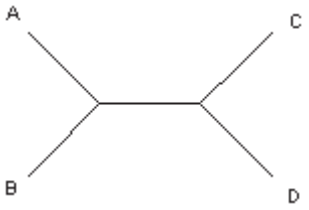
Supplementary material for Chapter 3

S3.1 SplitsTree consensus network explanation

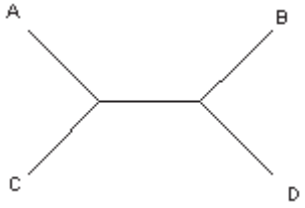
The consensus network method uses splits to indicate evidence for possible relationships. For example, the graph below:



Indicates the relationships shown in the two tree topologies:

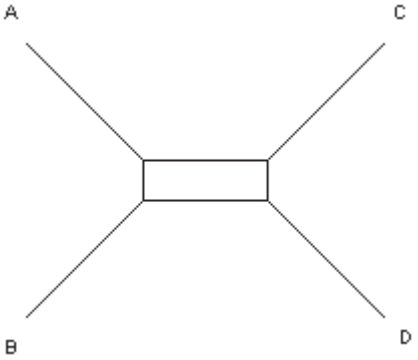


and



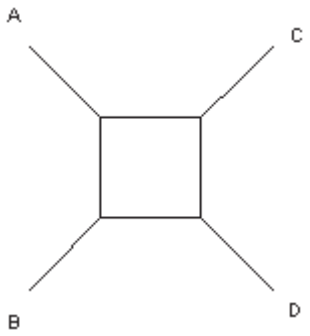
Topology A Topology B

Now suppose our dataset contains three trees of topology A (75%), and one tree of topology B (25%), with the branch lengths drawn to scale as above. when generating a consensus network, we have two options: either we can set the edge lengths to be proportional to the number of trees the splits occur, or we can set the edge lengths to be proportional to the average branch length when the splits occur. If we use the former, the consensus network generated from our dataset is shown below:



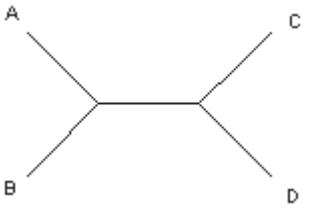
Notice the lengths of the edges are 3:1 in ratio favouring topology A, this ratio corresponds to the tree count in the dataset – three topology A trees and one topology B tree. In addition (not illustrated), this method also ignores branch lengths completely, even if one species forms a very long branch, it would still be shown as same branch length as all other species.

Now the second option is to set the edge lengths proportional to the average branch length, the consensus network created is shown below:



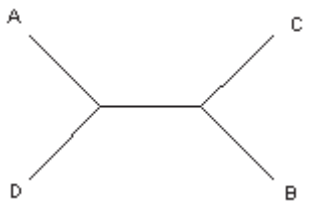
Notice now the edge lengths are equal, this is because the average branch length of the three trees of topology A is still same as the tree of topology B, despite more topology A trees in the dataset.

There is another option in SplitsTree, the splits threshold, which dictates at what percentage of disagreement which would be ignored. If we set this value to 0.3, the consensus network would be look like:



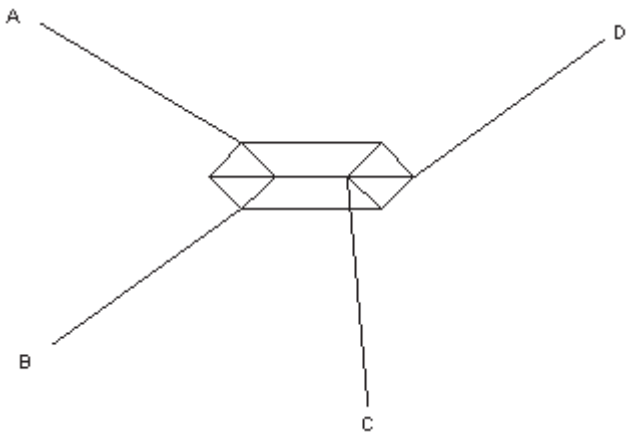
Note the consensus network is exactly same as topology A, this is because there are only one topology B tree in the dataset, 25%, which is below the 0.3 threshold, hence the disagreement is not shown in dataset.

Now if we add another topology to the dataset:



Topology C

We have three trees of topology A, and one tree of each topology B and C, the consensus network is going to be look like:



The conflicting signals are now shown three-dimensional.

S3.2 Perl script used in this chapter

Using this Perl script, the user can connect with the MySQL database (table blast_results in this study) and fetch *Giardia* ESP's homologues with best bit-score from each of the 17 eukaryotic organisms. The program puts all homologue sequences and the original *Giardia* ESP sequences in a text file in FASTA format.

```
#!/usr/bin/perl -w
#Step 1, use mysql to track down all homologues exist for an ESP.
use DBI;
use strict;
use warnings;
my $giardiaseq = ""; #set $giardiaseq to the accession number of ESPs
my $dbh = DBI->connect('dbi:mysql:giardia8909','username','password')
or die "Connection Error: $DBI::errstr\n"; #put in the actual username
and password.
my $sql = "SELECT * FROM blast_results where GL =
'gb|GL50803_$giardiaseq' order by blast_dict_ID, bitscore desc";
my $sth = $dbh->prepare($sql);
$sth->execute
or die "SQL Error: $DBI::errstr\n";
my $output1 = "homologues.txt";
open (OUTPUT1, ">$output1") or die "output file not opened";
my @row;
```

```
while (@row = $sth->fetchrow_array) {
print OUTPUT1 "@row\n";
}
close OUTPUT1;
#Step 2, select best hit for each organism.
open (INPUT1, "$output1") or die "Input file not opened";
my $output2 = "besthits.txt";
open (OUTPUT2, ">$output2") or die "output2 file not opened";
my @allhits;
while (<INPUT1>) {
my $rows1 = $_;
chomp $rows1;
push @allhits, $rows1;
}
for (my $ii = 0; $ii < @allhits - 1; ++$ii) {
      my @f = split / /, $allhits[$ii];
      my @g = split / /, $allhits[$ii + 1];
      if ($ii == 0) {
      print OUTPUT2 "@f\n";
      }
      if ($g[12] ne $f[12]){
      print OUTPUT2 "@g\n";
      }
}
close INPUT1;
close OUTPUT2;
#Step 3, fetch homologue sequences from fasta files. Need to create a
folder d:\\allorganisms, which #contains database files of each
organism, match name of database file to the organism's blast_dict_ID.
open (INPUT2, "$output2") or die "Input file not opened";
my $final = "$giardiaseq homologs.txt";
open (FINAL, ">$final") or die "final file not opened";
#delete the following lines if don't want giardia sequence to appear:
my $giardiadb = "d:\\allorganisms\\giardia.fasta";
open (GIARDIADB, "$giardiadb") or die "GIARDIA DB didnot open";
            print "giardia sequence gb|GL50803_$giardiaseq was
added\n";
            my $found = 0;
```

```
            while (<GIARDIADB>) {
                  if ($_ =~ /^\>/) {
                        if ($_ =~ /GL50803_$giardiaseq/){
                        $found = 1;
                        }
                        else {$found = 0;}
                  }
                  if ($found == 1) {
                  print FINAL "$_";
                  }
            }
close GIARDIADB;
my @goodhit;
my $fastafile;
my $xx;
while (<INPUT2>){
my $rows2 = $_;
chomp $rows2;
push @goodhit, $rows2;
}
my $zz=51;
for ($xx = 51; $xx < 68; ++$xx) {
      my $stop = 0;
      my @h = split / /, $goodhit[$xx - $zz];
      if ($h[12] != $xx) {
            my $diff = 52 - $zz;
            print  "a  sequence  was  skipped,  input  line  $diff,
blastdict_id $h[12]\n";
            $zz--;
            $xx--;
            $stop = 1;
      }
      if ($stop == 0){
            my $fastafile = "d:\\allorganisms\\$xx.fasta";
            open (FASTAFILE, "$fastafile") or die "fasta file did not
open, print $xx";
            my $thisSeq = $h[1];
            if ($thisSeq =~ /^gi.(\d*)/) {
```

```
            $thisSeq = $1;
            }
            print "$thisSeq was added\n";
            my $found = 0;
            while (<FASTAFILE>) {
                  if ($_ =~ /^\>/) {
                        if ($_ =~ /$thisSeq/){
                        $found = 1;
                        }
                        else {$found = 0;}
                  }
                  if ($found == 1) {
                  print FINAL "$_";
                  }
            }
      }
}
close FASTAFILE;
close FINAL;
#Step4, Double check how many sequences were found
open (FINAL, "$final") or die "final file not opened";
my $seqcount = 0;
while (<FINAL>) {
      if ($_ =~ /^\>/){
      $seqcount ++;
      }
}
print "total of $seqcount sequences in $final";
```

Chapter 4: Reconstruction of metabolic pathways in *Giardia*

4.1 Introduction

Examining genomic differences between a host and its parasite heavily relies on the annotation that has been given by the larger well established databases. The human genome is heavily annotated with connections to functions, metabolic pathways and other information. However most of *Giardia* and other non-model organisms' annotation have been performed automatically using variants of the BLAST algorithm. Most has not been manually checked and is not undergoing any detailed annotation in the near future (personal communication with GiardiaDB). This situation is similar when we look for metabolic information. KEGG (Kyoto Encyclopaedia of Genes and Genomes, http://www.genome.jp/kegg) is considered one of the most important resources for understanding higher-order functional utilities of organisms from genomic information (Kanehisa *et al.* 2006; Morrison *et al.* 2007). However, KEGG does not yet contain many enzymes from *Giardia*, partly due to the incomplete annotation of *Giardia* proteins.

In this chapter I have developed a method for analysing metabolic pathways from an organism, even with less defined annotation, and allow a fast scan of a pathway's presence or absence from an organism. The method was especially developed to work with *Giardia* information but could in theory be adapted for use with other species. The sugar pathways (e.g. glycolysis, TCA cycle and electron transport chain) were used for testing, and demonstrated success of the method. This method also indicates what cluster an enzyme belongs to, whether it is similar or different from that of the host's, which would be a very important factor when identifying drug targets.

Finding new methods to identify drug targets is one of the objectives of this study. It is especially relevant to *Giardia*. *Giardia* as explained in the introduction, is a major cause of human waterborne diarrheal disease, infecting an estimated 10% of the world's population during their lifetime (Huang *et al.* 2006). Infection is by faecal-oral transmission and is initiated by ingestion of infectious cysts in contaminated water or through person-to-person contact. After excystation, flagellated trophozoites colonise

the upper small intestine where they attach to the epithelial lining but do not invade the mucosa. Around 50% of *Giardia* infections are asymptomatic, in others the major symptoms of *Giardia* infection include diarrhoea, with malabsorption, dehydration, weight loss, cognitive impairment in children, and chronic fatigue in adults as well as other symptoms (Dunn *et al.* 2010).

One of the main drugs for treating *Giardia* infection is metronidazole (Mz), a synthetic 5-nitroimidazole (NI) derivative which is also active against *Trichomonas vaginalis* and *Entamoeba histolytica* (Harris *et al.* 2001; Valdez *et al.* 2009). Metronidazole is activated when its 5-nitro group is reduced by ferredoxin that has in turn been reduced by pyruvate:ferredoxin oxidoreductase (PFOR), generating toxic free radicals, and it is these short-lived free radicals that cause lethal damage to the parasite. PFOR is a good drug target because humans have an alternative pathway to PFOR, the pyruvate dehydrogenase complex. However, Mz treatment fails in about 20% of patients (Upcroft *et al.* 2001) and there are other issues including developing resistance to 5-NI compounds from *Giardia* (Dunn *et al.* 2010), and that Mz is inactive against *Giardia* cysts (Adam 2001).

The discovery and development of new therapeutics is important to expand the arsenal for controlling parasitic infection. Typically a drug target is a key molecule involved in a metabolic or signalling pathway that is specific to a disease condition or pathology, or to the infectivity or survival of a microbial pathogen (Rao *et al.* 2011). Since *Giardia* is a parasite with limited metabolic diversity, a better understanding of its metabolic pathways is important to the discovery of new drug targets. Although it has been described as having some bacteria-like metabolism (Adam 2001), *Giardia* displays typical eukaryotic features (e.g. cellular structure and ncRNAs such as a spliceosome (Chen *et al.* 2008), snoRNAs (Chen *et al.* 2008), and RNAi (Chen *et al.* 2009). However, given the large evolutionary distance between *Giardia* and other eukaryotes, and expected genome reduction due to its parasitic lifestyle, it is no surprise that the metabolism and these eukaryotic characteristics are slightly different in *Giardia*. It is these differences that can be highly effective as drug targets.

To date, only a few metabolic pathways in *Giardia* have been described, which include some carbohydrate metabolic pathways. However, these pathways were suggested using comparative analysis and are yet to be confirmed (Adam 2001; Morrison *et al.* 2007). An issue is that only a few enzymes have been biologically verified for *Giardia*, so the

presence/absence of a pathway component is heavily reliant on the annotation. This annotation once generated is seldom questioned even if the majority of the annotations are "hypothetical proteins". We expected that since ESPs are found throughout eukaryotes, they will also be found in key metabolic pathways (although this was later proven not the case as the *Giardia* has a set of prokaryote-like enzymes).

I employed a bioinformatics approach to develop a metabolic pathway analysis procedure to look at the sugar metabolism of *Giardia*. The results have been compared with the information from KEGG. This resulted in an overview of its metabolic repertoire, and predicts candidates for enzymes in these pathways. The three sugar-related metabolic pathways investigated in this chapter were glycolysis (including gluconeogenesis and glycogen synthesis), the tricarboxylic acid cycle, (TCA cycle, also known as citric acid cycle or Krebs cycle) and oxidative phosphorylation.

The glycolysis pathway is the most basic sugar metabolic pathway and it occurs with variations in nearly all organisms (Romano *et al.* 1996). During glycolysis, a glucose molecule is catabolised into two pyruvate molecules. Two adenosine triphosphate (ATP) and two NADH molecules can be gained from one glucose molecule, which provides energy to the organism (Voet *et al.* 2004). The reverse pathway of glycolysis is gluconeogenesis, which utilises the majority of the same enzymes as glycolysis. This process enables organisms to store energy in the form of glucose.

The product of glycolysis, pyruvate can be converted to acetyl CoA, which is the input molecule for the TCA cycle, which is the second pathway investigated. Acetyl-CoA and oxaloacetate combines to form citrate to begin the cycle, which then goes through a series of reactions until oxaloacetate is reformed to repeat the cycle. This enables carbohydrates to be fully oxidised into carbon dioxide and water, and generates two further ATPs and six reduced co-enzymes NADH per glucose molecule (Voet *et al.* 2004). In eukaryotes this process typically occurs inside the mitochondrial matrix and in prokaryotes it occurs in the cytosol. *Giardia* is an aerotolerant anaerobe with no mitochondria, but instead has closely related organelles called mitosomes (Dolezal *et al.* 2005; Jedelsky *et al.* 2011). These are a reduced form of mitochondria, but whether they actually participate in ATP synthesis is currently unknown (Emelyanov *et al.* 2011). The biosynthesis of FeS clusters, which plays an important role in oxidation/reduction reactions during the electron transport chain, has been said to be the only mitochondrial function retained by mitosomes in *Giardia* (Jedelsky *et al.* 2011). The TCA cycle

provides precursors for many compounds including some amino acids, therefore some part of the cycle may be functional in *Giardia*.

The third pathway, oxidative phosphorylation, uses energy released during glycolysis and the TCA cycle to produce ATP. Electrons are transferred from electron donors (such as NADH generated from the TCA cycle) to electron acceptors (such as H_2O) to the transfer of H^+ ions (protons) across a membrane. The resulting electrochemical proton gradient is used to generate chemical energy in the form of ATP (Voet *et al.* 2004). In eukaryotes, these redox reactions are carried out by five main protein complexes within mitochondria, named Complex I to Complex IV and ATP synthase. Some bacterial species can carry out the electron transport chain quite differently by using different electron donors, acceptors, and different enzymes. The chain may contain three proton pumps like those found in mitochondria (Complex I, III and IV), or it may contain only one or two pumps. Because *Giardia* is anaerobic, the majority of enzymes in the mitochondrial electron transport chain are expected to be absent, but further assessments could aid in determining if there are any bacteria-like electron transport chains in *Giardia*.

In this chapter, candidates for key enzymes in these three pathways have been identified by similarity searching against all annotated enzymes from KEGG. I identified good candidates for several enzymes that were not recognised by KEGG, including phosphoglucomutase, phosphofructokinase and enzymes for ethanol fermentation. I have also determined which of the *Giardia* enzymes are more bacteria-like and which are more eukaryote-like. I use terms such as “bacteria-like” or “eukaryote-like” here to refer to enzyme amino acid sequence similarity to bacteria or eukaryotic enzymes respectively, and not to infer any phylogenetic relationship. I have identified a number of enzymes which show differences between *Giardia* and its hosts, thus making them as potential targets for drug discovery. This chapter lays the groundwork for metabolic comparisons using KEGG to enable further work towards identifying treatments for *Giardia*.

4.2 Materials and Methods

Because of the high standard of curation for genes in the KEGG database (http://www.genome.jp/kegg), this methodology was simple but precise enough to

permit an ‘overview’ look at metabolic pathways without years of proteomics and metabolomics. The KEGG database contains networks represented by wiring diagrams of protein and other gene products responsible for various cellular processes, such as metabolism (examples see Figure S1 and S2).

In KEGG, enzymes that catalyze the same reaction typically have the same enzyme commission (EC) numbers in the major databases, regardless of their homology. Enzymes with the same EC number may show significant sequence and structural similarity. However, in some cases enzymes with the same activity (i.e. same EC number) can be associated with different phylogenetic lineages and have different catalytic mechanisms with little structural similarity. I used EC numbers during this study because enzymes can have variation on their names, but the EC numbers will remain the same (e.g. “phosphohexose isomerase” can also be referred to as “phosophglucose isomerase” or “Glucose-6-phosphate isomerase”, but the EC number will always be “EC: 5.3.1.9”).

The “genes.pep” file was downloaded from KEGG (http://www.genome.jp/kegg, accessed January 2011), containing all the sequences in the KEGG database in FASTA format (a total of 5,338,631 sequences, with the EC number included in the annotation of each protein). This file made it possible to pull out all sequences belonging to the same EC number from different organisms). For each EC class, enzymes from all species were collated into a single FASTA file by the use of a Perl script, the *Giardia* protein database was then BLASTed (Altschul *et al.* 1990) against this FASTA file, and the proteins with the highest bit-score were recorded. A metabolic map of the pathway was generated according to bit-scores (the log scaled score given to alignments by BLAST, higher numbers correspond to higher similarity) of the hits for each enzyme. Hits with bit-scores higher than 300 were considered to be high-quality candidates for the enzyme, and hits with scores between 100 and 300 were considered as lower quality candidates. This procedure was repeated for each enzyme in the glycolysis, TCA cycle and oxidative phosphorylation pathways. EC numbers for all enzymes detected in *Giardia* in these pathways are given in tables S1-S3.

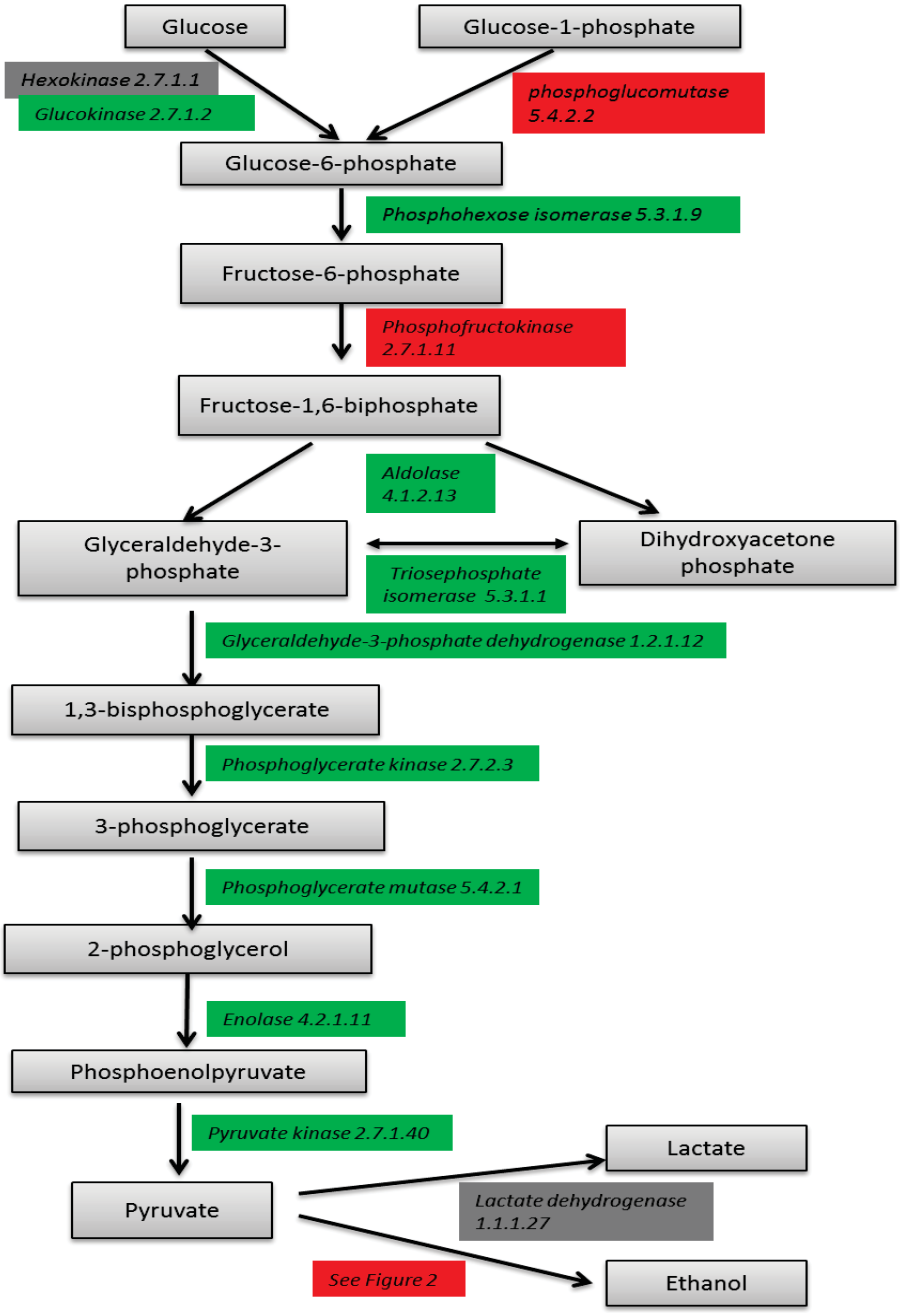
4.3 Results

4.3.1 Glycolysis and Gluconeogenesis

As expected, because glycolytic proteins are highly conserved in eukaryotes, the major enzymes in the backbone of the glycolysis pathway were found in the *Giardia* genome (see Figure 1). However, an unexpected feature is that most of these enzymes showed greater similarity to their bacterial orthologues than their eukaryotic orthologues (refer to supplementary data); some of these bacteria-like enzymes (e.g. phosphofructokinase, EC: 2.7.1.11) are also found in other eukaryotic protists (*Toxoplasma*, *Tetrahymena*, *Trypanosoma*, *Plasmodium* and *Trichomonas*). There were a few eukaryote-like enzymes from the glycolysis pathway detected, including phosphoglucomutase (EC: 5.4.2.2), phosphoglycerate kinase (EC: 2.7.2.3), dihydrolipoyllysine-residue acetyltransferase (EC: 2.3.1.12), and enolase (EC: 4.2.1.11). These results are explained in more detail below.

Phosphoglucomutase (PGM, EC: 5.4.2.2) facilitates the inter-conversion of glucose-1-phosphate and glucose-6-phosphate. *Giardia* protein GL50803_17254 showed high similarity to the PGM from eukaryotes (bit-score of 310). Experimental evidence from Mitra *et al.* (Mitra *et al.* 2009) indicates that this protein has phosphoglucomutase activity, validating in part the potential of this method for finding new enzymes. Another glycolytic enzyme not yet in KEGG to *Giardia*, phosphofructokinase was also recovered, with the *Giardia* protein GL50803_14993 showing high similarity to phosphofructokinases from many bacterial species. KEGG have assigned the EC number “EC: 2.7.1.90” to this protein, suggesting that this is a pyrophosphate based phosphofructokinase.

Figure 1. Glycolysis in *Giardia*



The diagram indicates which enzymes have been directly identified by KEGG (green), which have been identified during this chapter (red) and which are not present (grey). As can be seen, most glycolytic enzymes are present in *Giardia*. A more technical representation of this image is present in Figure S1. Key: The metabolites are labelled and grey boxes, the enzyme which catalyse reactions from one metabolite to another are shown in rectangles with their EC number indicated, and are coloured according to their homology to enzymes of other species: green: enzymatic function registered in KEGG; red: found in *Giardia* with bit-score >300, these are enzyme candidates with fairly high degrees of certainty; grey: found in *Giardia* with bit-score <100; there was no enzymes found in *Giardia* with bit-scores between 100-300 in this pathway.

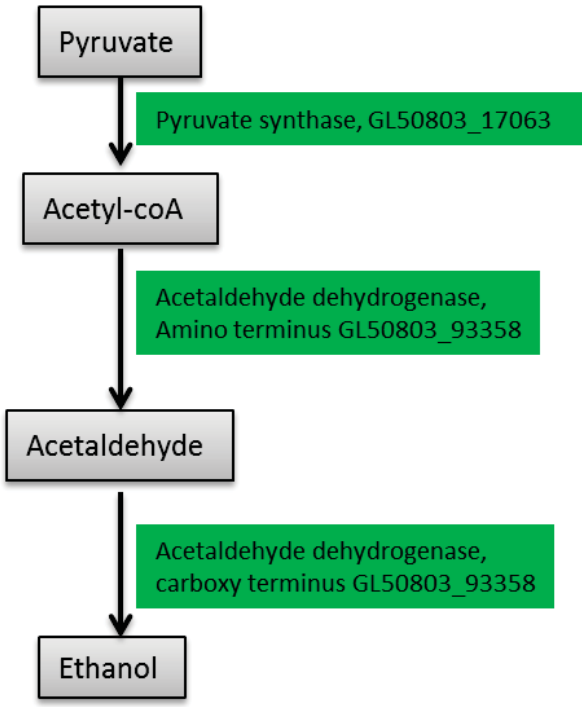
Two enzymes are responsible for the inter-conversion of glucose to glucose-6-phosphate, namely hexokinase (EC: 2.7.1.1) and its isozyme glucokinase (EC: 2.7.1.2). The difference is that glucokinase has a lower affinity for glucose than hexokinase. A *Giardia* glucokinase has been described by KEGG (GL50803_8826), with this protein showing more similarity (i.e. higher bit-score) to the bacterial (cyanobacteria) glucokinase than eukaryotic glucokinase. *Tetrahymena thermophila* (a free living protozoan species) also has a similar type of glucokinase. As yet it is unknown why *Giardia* has a lower affinity enzyme, but one possible reason is that the trophozoite living environment (intestines of animals) provides a generous glucose supply so a high-affinity enzyme is not required. I did not identify any hexokinase from *Giardia* although one fungal protein (uma:UM03093.1, "uma" indicates the species: *Ustilago maydis* and "UM03093.1" is the accession number of the protein) was mislabelled as a hexokinase by KEGG, and returned a *Giardia* protein with high similarity. Upon further analysis of uma:UM03093.1 by using genomic resource databases (http://www.ncbi.nlm.nih.gov) and homology search, I determined that the fungal and *Giardia* proteins were in fact false positives for hexokinase. In light of this it appears all conversion of glucose to glucose-6P is carried out by glucokinase exclusively in *Giardia*.

The conversion of pyruvate to lactate is catalysed by lactate dehydrogenase (EC: 1.1.1.27), and the coupled reaction also oxidises coenzyme NADH to NAD^+. This reaction occurs in lactate fermenting bacteria and in eukaryotes such as humans in the absence of oxygen, to provide a constant supply of oxidised form of coenzyme NAD^+ for glycolysis. The only *Giardia* protein with high bit-score to any known lactate dehydrogenase is GL50803_17325 with a bit-score of 161 against one (and only one) lactate dehydrogenase from *Toxoplasma gondii* (also a parasitic protozoan). However, further analysis indicated that tgo:TGME49_060600 may have been incorrectly assigned by KEGG (in the same manner as described above for hexokinase). The results here suggest that *Giardia* lacks lactate dehydrogenase, and that lactic acid fermentation does not take place in *Giardia*. Instead the re-oxidation of coenzyme NADH to NAD^+ is performed by ethanol fermentation (discussed later).

Pyruvate synthase (EC: 1.2.7.1) is also known as pyruvate:ferredoxin oxidoreductase (PFOR). This is an alternative enzyme to the pyruvate dehydrogenase complex (formed together by EC: 1.2.4.1, EC: 2.3.1.12 and EC: 1.8.1.4) found in mammals. PFOR is able

to oxidise pyruvate to acetyl-CoA, but utilizes ferredoxin rather than NAD^+ as the electron acceptor. The PFOR of *Giardia* is GL50803_17063 and is the main target for the drug Metronidazole (Mz) (Valdez *et al.* 2009). The selective toxicity of Mz is achieved because the parasite has PFOR only.

Figure 2. A possible ethanol fermenting pathway in *Giardia*



Pyruvate is metabolised into acetyl-CoA, and acetyl-CoA is subsequently converted to acetaldehyde and alcohol by a bi-functional acetaldehyde-CoA and alcohol dehydrogenase.

Giardia performs ethanol fermentation to maintain a constant supply of NAD^+, but this pathway is different from that found in some bacteria and yeast in that it converts pyruvate into acetaldehyde and then into ethanol. *Giardia* seems to be unable to convert pyruvate directly to acetaldehyde because pyruvate decarboxylase (EC: 4.1.1.1) is noticeably absent. *Giardia* performs ethanol fermentation by first converting pyruvate to acetyl-CoA (by pyruvate synthase, EC: 1.2.7.1), then to acetaldehyde and finally to ethanol (see Figure 2). It has been reported (Sanchez 1998; Dan *et al.* 2000) that a *Giardia* enzyme has acetaldehyde dehydrogenase (EC:1.2.1.10) activity in the amino-terminus which catalyses the conversion of acetyl-CoA to acetaldehyde, and alcohol dehydrogenase (EC: 1.1.1.1) activity in the carboxy-terminus which catalyses the conversion of acetaldehyde to ethanol, but the paper did not include the accession number used to identify the protein. I identified the aforementioned protein, as

GL50803_93358 which scored high bit-scores (870) for both alcohol dehydrogenase and acetaldehyde dehydrogenase. The closest homologue of this protein, tel:tlr0227, is also incidentally a bi-functional acetaldehyde-CoA and alcohol dehydrogenase from the cyanobacterium *Thermosynechococcus elongates*.

I also found acetyl-CoA synthetase (EC: 6.2.1.13) in *Giardia* (GL50803_13608), indicating that pyruvate can also be converted to acetyl-CoA and then to acetate. Experimental evidence suggests that the metabolism of trophozoites is markedly affected by small changes in oxygen concentration (Paget *et al.* 1993). Under anaerobic conditions, ethanol is the major product of carbohydrate metabolism, and under aerobic conditions, alanine and acetate are the predominant products of energy metabolism. Thus, the pyruvate metabolism pathway appears to be flexible for dealing with different aerobic/anaerobic environments (Paget *et al.* 1993).

I have identified enzymes which are significantly different from that of the host, as many enzymes in the glycolytic pathway are more closely related to those from archaea and bacteria, and thus different from those of eukaryotes. The *Giardia* enzymes that are different from eukaryotic enzymes and hence are possibilities for future drug discovery are listed in Table 1. Prokaryotic looking enzymes were known from Morrison *et al.* paper, but what there will be key eukaryotic pathways such as glycolysis, we expect *Giardia* to have eukaryotic glycolysis, because *Giardia* is a eukaryote. This raised some interesting evolutionary questions, as to whether this is a case of convergent evolution between *Giardia* and bacterial enzymes, or the last common ancestor of *Giardia* and bacteria had the same enzymes (discuss later section 4.4). It is also noted that none of the glycolytic enzymes are ESPs, because all of them have bacterial homologues, there are however, six enzymes that are conserved in all eukaryotes (discuss later in section 4.4, see Table 2).

Table 1. Bacteria-like *Giardia* enzymes in glycolysis pathway

EC	Enzyme name	Best candidate	Bit score	E-value	Homologous Domain
2.7.1.2‡	glucokinase	GL50803_8826	393	2.00E-110	B, P
2.7.1.11	phosphofructokinase	GL50803_14993	429	9.00E-122	P, B
4.1.2.13‡	aldolase	GL50803_11043	390	3.00E-110	B, P
1.2.1.59	glyceraldehyde-3-phosphate dehydrogenase (NAD(P)+)	GL50803_6687	326	6.00E-92	B
1.2.7.6	glyceraldehyde-3-phosphate dehydrogenase	GL50803_6687	315	5.00E-89	B
5.4.2.1†	phosphoglycerate mutase	GL50803_8822	551	4.00E-142	B
1.2.7.1	pyruvate synthase	GL50803_17063	1008	0.0	B
6.2.1.13	acetyl-CoA synthetase (ADP-forming)	GL50803_13608	507	5.00E-146	A, B, P
1.1.1.1	alcohol dehydrogenase	GL50803_93358	870	0.0	B
eutG	ethanol:NAD+ oxidoreductase	GL50803_93358	717	0.0	B
1.2.1.10	acetaldehyde dehydrogenase	GL50803_93358	870	0.0	B

There is very little literature on the presence of the gluconeogenesis pathway in *Giardia*. It has been suggested that gluconeogenesis may occur during encystation, when glucose uptake decreases substantially and *Giardia* gains its energy by up-taking amino acids (aspartate) followed by gluconeogenesis (Adam 2001); However, another group have also found no evidence of active gluconeogenesis (Ma'ayeh and Brook-Carter, presentation during the IV International *Giardia* and *Cryptosporidium* conference, Wellington, 2012). The gluconeogenesis pathway shares a number of identical enzymes with the glycolysis pathway. There are three subtle differences: first the reaction catalysed by pyruvate kinase (converting phosphoenolpyruvate to pyruvate) is irreversible, but pyruvate carboxylase (EC: 6.4.1.1) and phosphoenolpyruvate carboxykinase (PEPCK, EC: 4.1.1.32) can convert pyruvate into oxaloacetate and then back to phosphoenolpyruvate, which can be used for gluconeogenesis. Homologues of both of these enzymes were found in *Giardia* in this analysis, although pyruvate carboxylase has not yet been described in *Giardia* by KEGG. The second enzyme where gluconeogenesis differs from glycolysis is fructose bisphosphatase (EC: 3.1.3.11) which

converts fructose-1,6-bisphosphate to fructose-6-phosphate in gluconeogenesis, (the reverse of the reaction catalysed by phosphofructokinase in glycolysis). In *Giardia*, only the protein GL50803_17316 has some low similarity (bit-score of 99) to fructose bisphosphatase. Given this low similarity, it is likely this protein does not have fructose bisphosphatase enzymatic activity. Lastly, in the reverse of the reaction that is catalysed by glucokinase, no candidates for glucose-6-phosphatase (EC: 3.1.3.9) were recovered. Overall, this analysis suggests that two key gluconeogenic enzymes are absent, and that *Giardia* does not have the entire set of enzymes required to perform gluconeogenesis. *Giardia* may take up amino acids for energy, but may not convert it all the way back to glucose, and instead the amino acids are possibly converted to pyruvate or oxaloacetate to obtain limited energy (via the likes of the arginine dihydrolase pathway).

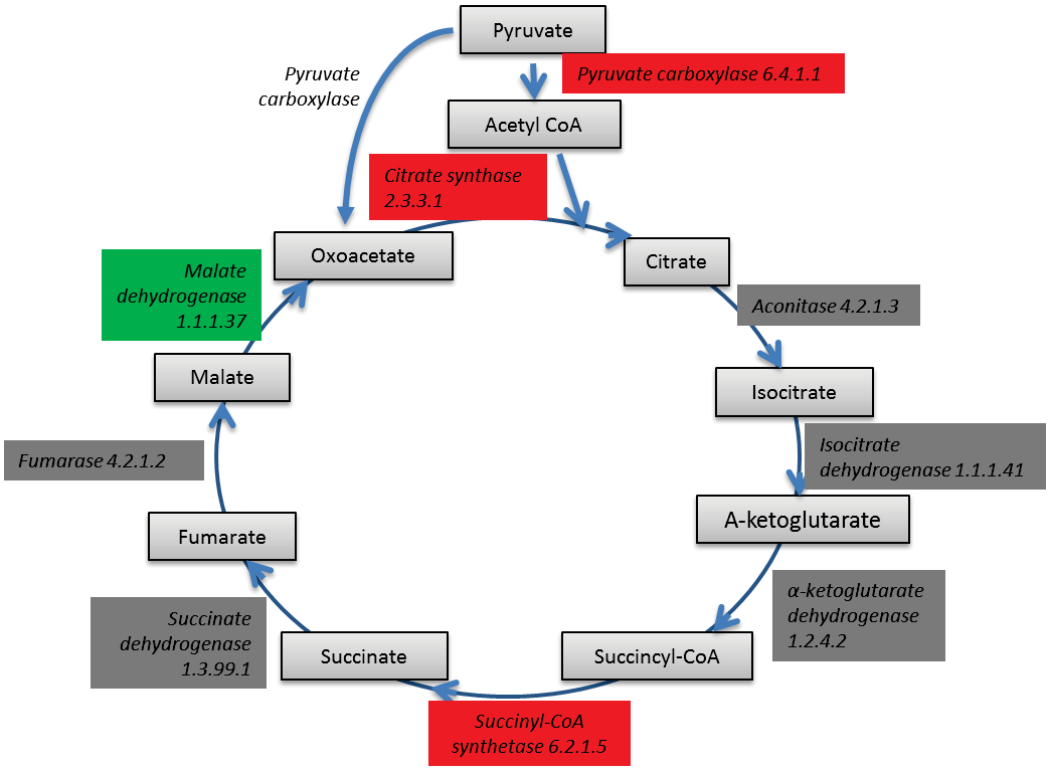
Although unable to regenerate glucose, *Giardia* does appear to have all required enzymes to synthesise glycogen from glucose (glycogenesis): UTP-glucose-1-phosphate uridylyltransferase (EC: 2.7.7.9) and glycogen synthase (EC: 2.4.1.11) have both been noted by KEGG. Combining this result with previous reports that glycogen has been found to be present in trophozoites (Ladeira *et al.* 2005), it appears that *Giardia* is able to generate glycogen from glucose to serve as an energy reserve.

In summary, *Giardia* is able to perform glycolysis, using glycolytic catabolic reactions to provide energy for the organism. *Giardia* is also able to synthesis glycogen from glucose to create an energy reserve. However, it appears that *Giardia* is unable to perform gluconeogenesis to generate glucose from pyruvate.

4.3.2 Tricarboxylic acid cycle

Most of the enzymes in the TCA cycle were not detected in *Giardia* (Figure 3). This was expected because *Giardia* is an anaerobe, and lacks the mitochondria in which the TCA cycle typically operates in other eukaryotes. Those enzymes that are present in *Giardia* are also part of alternative metabolic pathways (pyruvate synthase (EC: 1.2.7.1), pyruvate carboxylase (EC: 6.4.1.1) and PEPCK (EC: 4.1.1.32) are all in the glycolysis pathway). The presence of citrate synthase (EC: 2.3.3.1) and malate dehydrogenase (EC: 1.1.1.37) is expected because citrate and malate are important intermediates involved in the metabolism of highly interconnected cellular metabolites. Thus it is possible that *Giardia* will need pyruvate synthase, pyruvate carboxylase and PEPCK to metabolise malate and citrate.

Figure 3. TCA cycle enzymes in *Giardia*.



Limited candidates were found for enzymes in the TCA cycle. A more technical representation of this image is present in Figure S2. Key as for Figure 1.

The best candidate for succinyl-CoA synthetase (EC: 6.2.1.5) is GL50803_13608, which also has similarity to acetyl-CoA synthetase (EC: 6.2.1.13). The substrates for both enzymes are similar, and it is more likely that this protein is an acetyl-CoA synthetase because of the higher bit-score (507 for acetyl-CoA synthetase vs 435 for succinyl-CoA synthetase) and the fact that GiardiaDB labelled this as acetyl-CoA synthetase. Overall, the lack of the majority of the enzymes in the pathway suggests the absence of the TCA cycle in *Giardia*. The absence of the TCA cycle is likely due to secondary loss when mitochondria were lost from the parasites, but this is not conclusive because of the long evolutionary distance between *Giardia* and other eukaryotes.

4.3.3 Oxidative phosphorylation

The oxidative phosphorylation pathway is harder than the other pathways to analyse because many enzyme subunits with the same EC class, form the multimeric complexes involved in the pathway (e.g. Complex I contains as many as 45 peptides in metazoans). I did not expect *Giardia* to have a typical electron transport chain because they lack mitochondria, and there is likely to be a limited supply of reduced NADH due to the absence of the TCA cycle. However, *Giardia* does have reduced mitosomes and thus some proteins may have remained from the ancestral mitochondria. The main components of a typical oxidative phosphorylation pathway are shown in Figure 4.

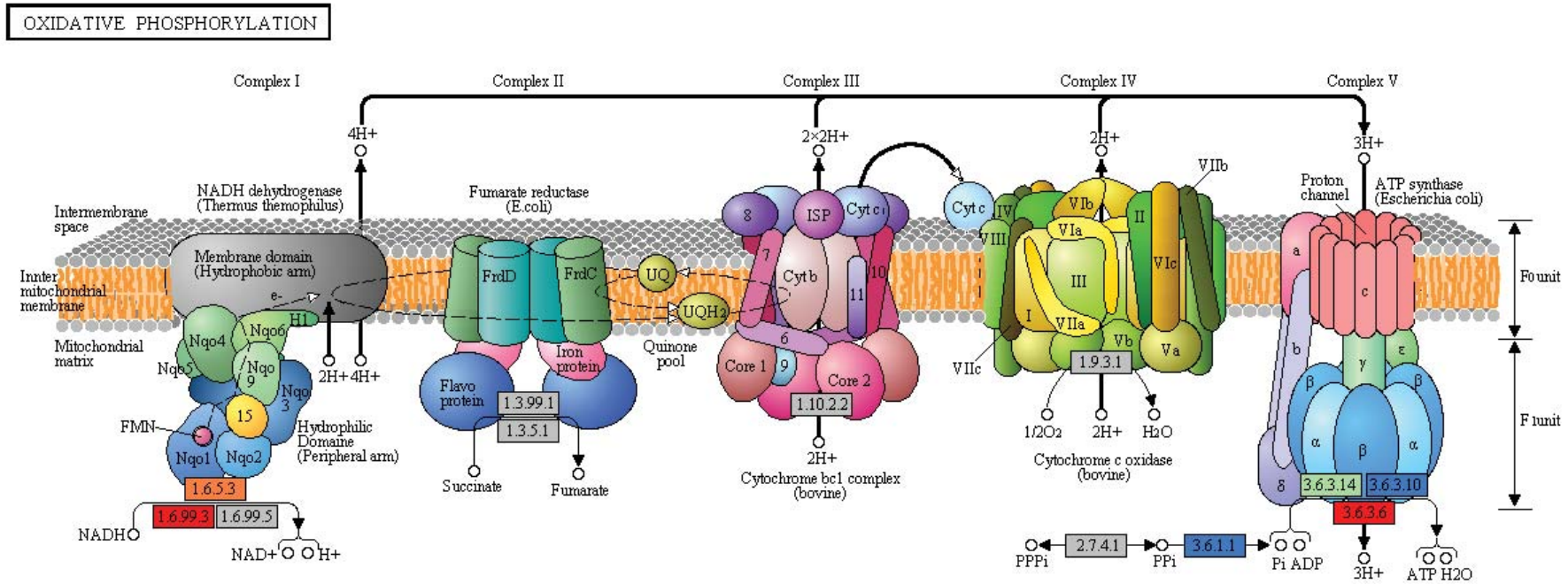
This analysis showed that Complex II (succinate dehydrogenase), Complex III (ubiquinol-cytochrome-c reductase) and Complex IV (cytochrome c oxidase) are clearly absent in *Giardia*. Complex I (NADH dehydrogenase) have two EC classes: EC: 1.6.5.3 and EC: 1.6.99.3, the difference between the two EC classes is the former uses ubiquinone as the electron acceptor, the later does not have a specified electron acceptor. Homologues of Complex I were recovered from *Giardia* (e.g. GL50803_6304 and GL50803_33769). However, Complex I is a polymer containing up to 45 individual peptides, given that very few homologues of these peptides have been found in, it is unlikely that *Giardia* has the entire Complex I.

The ATP synthase (EC: 3.6.3.34, labelled as Complex V in Figure 4) was determined to be present. As many as 14 proteins (GL50803_10530, GL50803_12216, GL50803_13000, GL50803_13603, GL50803_14660, GL50803_14961, GL50803_15598, GL50803_18470, GL50803_30851, GL50803_3678, GL50803_7532, GL50803_8367, GL50803_8559, GL50803_87058) have been assigned to this EC class. These 14 proteins make up the vacuolar (V-type) ATPase (Hilario *et al.* 1998). The V-type is different from F_1F_O (F-type) ATPase, which is present in the plasma membrane of bacteria, the inner membrane of mitochondria, and the thylakoid membranes of chloroplasts. The V-ATPase is present in the endomembrane systems of eukaryotes: vacuoles, Golgi apparatus, and coated vesicles. V-type ATPases build up a H^+ gradient across the membrane via ATP hydrolysis to transport solutes, or to lower the pH inside the endomembrane system, in reverse of reactions catalysed by F-ATPase (Hilario *et al.* 1998). Its function is to generate a proton gradient rather than utilising the proton gradient to harvest ATP. The other two ATP synthase enzymes present in *Giardia* were the H^+/K^+-exchanging ATPase (EC: 3.6.3.10) and H^+-exporting ATPase

(EC: 3.5.3.6). These two enzymes function as transporters rather than ATP generators. No F-type ATPases were recovered from *Giardia*, indicating a possibility that there is no ATP-producing ATP synthase. Overall the lack of Complexes I, II, III and IV suggest that *Giardia* is unable to actively generate the proton gradient that is vital for generation of ATPs by F-ATP synthase.

Some bacterial species can carry out the electron transport chain differently (i.e. during anaerobic respiration), by using different electron donors and acceptors and therefore different enzymes (e.g. *Escherichia coli* can use a large number of electron donor/acceptor pairs such as fumarate/succinate, or pyruvate/lactate (Unden *et al.* 1997)). It cannot be ruled out that *Giardia* might have such a mechanism or an as yet completely novel mechanism, but an F-type ATP synthase is still lacking in *Giardia*, suggesting that *Giardia* is unable to mass produce ATP by using a typical eukaryotic electron transport chain pathway.

Figure 4. The Oxidative phosphorylation pathway in *Giardia*



Only homologues of enzymes in complex I and complex V (ATP synthase) were found in *Giardia*. Key as for Figure 1, except that the blue enzyme boxes indicate the enzyme has registered in KEGG but was not indicated in their map of *Giardia* metabolic pathway. This is because KEGG also have a KEGG orthology (KO) number, which indicate if there are different classes of enzymes with same EC number, some lesser studied enzymes such as those of *Giardia*'s have not been given a KO number, and KEGG was conservative not to show them on the map of *Giardia* pathways. In the above case the enzymes were coloured blue in my figures. The template image was downloaded from KEGG.

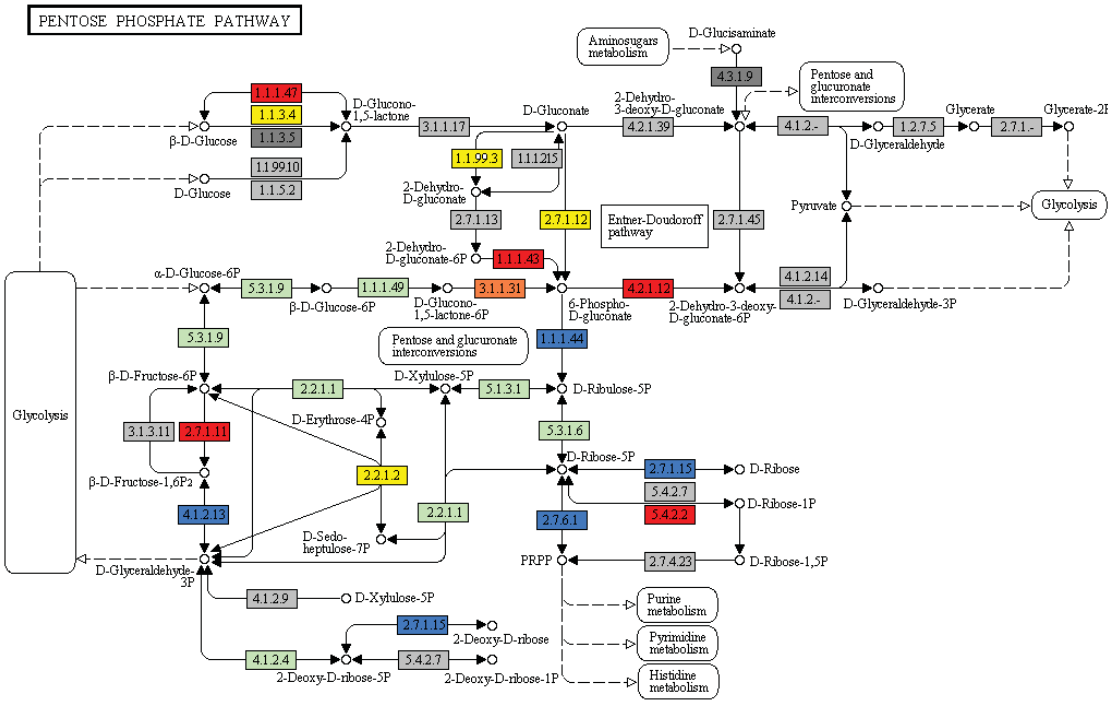
4.3.4 Other metabolic pathways

Using the same procedure, a quick scan of a metabolic pathway to see if it is present in *Giardia* (or any other organisms) can easily be performed. Brief analyses on the *Giardia* pentose pathway, alanine and aspartate metabolism were performed but were not examined in great detail due to time constraints.

Pentose phosphate pathway

The pentose phosphate pathway is an alternative sugar metabolism pathway. The purpose of this pathway is to regenerate NADPH, produce ribose-5-phosphate (R5P), used in the synthesis of nucleotides and nucleic acids, and erythrose-4-phosphate (E4P), used in the synthesis of aromatic amino acids. It is unknown whether *Giardia* can synthesise these sugars *de novo*, and results here only took a brief look at this pathway. The results indicated that the majority of the enzymes are present, and *Giardia* is capable of synthesise ribose and erythrose, as well as regenerate NADPH (Figure 5).

Figure 5. Pentose phosphate pathway in *Giardia*

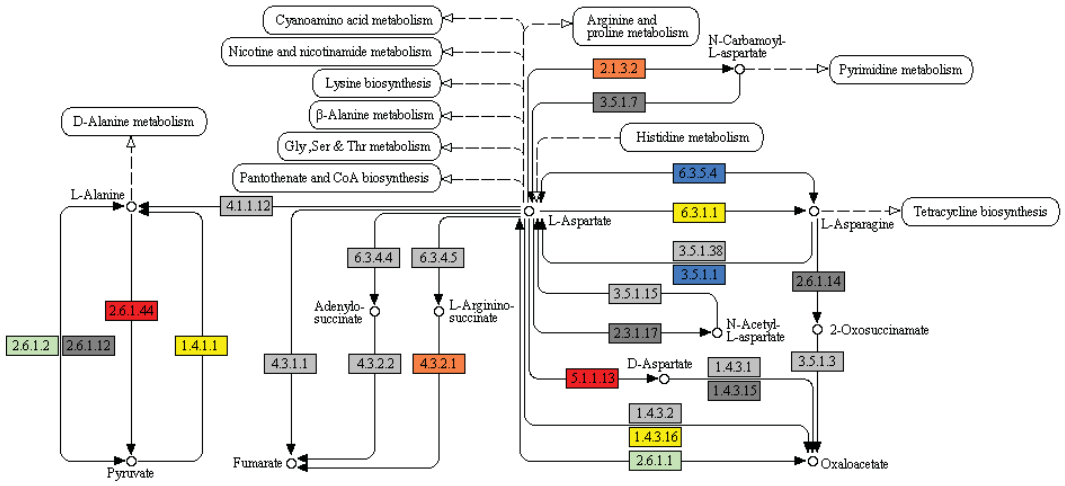


Colour scheme: green: enzymatic function registered in KEGG; blue: EC number registered in KEGG but was not in the KEGG *Giardia* glycolysis pathway; red: found in *Giardia* with score >300; orange: found in *Giardia* with score 200-300; yellow: found in *Giardia* with score 100-200, grey: found in *Giardia* with score <100; dark grey: no enzyme with this EC number registered in KEGG. Template image was downloaded from KEGG.

Alanine and aspartate metabolism

The pathways for alanine and aspartate are quite simple (Figure 6). Alanine is broken down into pyruvate whereas aspartate is broken down into oxaloacetate. *Giardia* also appears unable to inter-convert between aspartate and asparagine. There were no other aspartate metabolism (e.g. to fumarate) pathways found from the results.

Figure 6. Alanine and aspartate in *Giardia*



Colour scheme: green: enzymatic function registered in KEGG; blue: EC number registered in KEGG but was not in the KEGG *Giardia* glycolysis pathway; red: found in *Giardia* with score >300; orange: found in *Giardia* with score 200-300; yellow: found in *Giardia* with score 100-200, grey: found in *Giardia* with score <100; dark grey: no enzyme with this EC number registered in KEGG. Template image was downloaded from KEGG.

4.4 Discussion

Overall, I have been able to reconstruct a number of sugar-related metabolic pathways for *Giardia lamblia*, and highlight notable enzyme absences from these pathways. The glycolytic enzymes from *Giardia* bear a stronger homology with bacterial enzymes, rather than with eukaryotic or archaeal enzymes (except for phosphoglucomutase and phosphoglycerate kinase which are more similar to those found in eukaryotes). The difference between the host's and parasite's pathway can be exploited for medical applications. Only a few enzymes were identified from the TCA cycle and oxidative phosphorylation, indicating the likely absence of these pathways. In addition, because of the similarity to prokaryotic enzymes, glycolytic enzymes in *Giardia* cannot be considered ESPs.

This approach of analysing metabolic pathways could, in theory, be applied to any organism with genome information but limited annotation. The advantage of using this approach is that it is reasonably quick to give an indication of which pathways are likely to be present and which ones are not. There are however some limitations: for a few proteins, KEGG can allocate wrong EC numbers which will result in false positives if users are not familiar with the pathways. False positives can also occur if one EC class is very similar to another EC class (such as in the case of succinyl-CoA synthetase and acetyl-CoA synthetase). KEGG is a database that is still growing and as yet, does not have the enzymes from all species. It is expected that enzyme candidates may not be recovered if they are from a species extremely different from the known enzymes and species. It is expected this issue will decrease with time as enzymatic studies on species such as *Giardia*, add to the improvement of KEGG annotations.

The overall picture of *Giardia* indicates that glucose is absorbed from the host and metabolised into pyruvate through glycolysis, and after that, in order to regenerate the oxidised form of coenzyme NAD^+, pyruvate is reduced to ethanol, alanine or acetate depending on the availability of oxygen. Under aerobic conditions, pyruvate is converted to alanine by a transamination reaction, or to acetate by acetyl-CoA synthetase. Also under anaerobic conditions, pyruvate is metabolised to acetyl-CoA by PFOR, and subsequently into acetaldehyde and ethanol. The TCA cycle and oxidative phosphorylation do not appear to occur. These latter results were not completely unexpected since it is already known that *Giardia* has an anaerobic life style, and has undergone genome reduction (i.e. a smaller genome with fewer unnecessary enzymes will give the parasite advantage when replicating) (Morrison *et al.* 2007).

Giardia shares many metabolic attributes of bacteria, including its fermentative energy metabolism which relies heavily on pyrophosphate rather than adenosine triphosphate. Morrison *et al.* 2007 looked into *Giardia*'s metabolic repertoire briefly when the *Giardia* genome project was completed. Their results indicated that *Giardia*'s sugar metabolic pathways contained a mixture of eukaryote-like (enzymes that appeared more similar in sequence to those enzymes found in eukaryotes) and bacteria-like enzymes (Morrison *et al.* 2007). Morrison *et al.* 2007 indicated that about half of glycolytic enzymes are eukaryote-like (Morrison *et al.* 2007), but they did not distinguish between typical eukaryotic enzymes (i.e. those well studied in mammals, yeasts and plants) and enzymes from eukaryotic protists. This study has considered protists separately from

other eukaryotes, because frequently these eukaryotic protists have prokaryote-like enzymes rather than those from typically studied eukaryotes. Some reasons for *Giardia* having a sizable number of bacteria-like enzymes include the possibilities that mitochondrial genes migrated to the nucleus with the loss of this organelle (Adams *et al.* 2003); lateral gene transfer of bacterial genes (Andersson *et al.* 2003); convergent evolution between bacterial set of enzyme and some of *Giardia* enzymes due to their common anaerobic life style; or that the eukaryotic set of enzymes arose after their divergence from the ancestral eukaryote. There are still many evolutionary questions surrounding *Giardia* and it is expected that the clarification of its somewhat 'atypical' metabolism will aid this research.

The glycolysis pathway occurs, in nearly all organisms with minor variations (Romano *et al.* 1996). So, if the enzymes in the glycolysis pathway are also conserved in all organisms? I also compared the *Giardia* annotated proteins (4889 in total) against 28 bacterial, 12 archaeal species and 17 other eukaryotic species, and identified four groups of proteins according to the conservation of the proteins in the three super kingdoms: Group A (see section 2.3.8) contains 37 *Giardia* proteins that are conserved in all three domains of life; Group B (see section 2.3.8) contains 849 *Giardia* proteins that are found in all eukaryotes; Group C contains 274 eukaryotic signature proteins (ESPs, see section 2.3.1) (Hartman *et al.* 2002; Kurland *et al.* 2006), which are proteins conserved in all eukaryotes, but not found in any archaea or bacteria; and finally Group D contains 278 *Escherichia coli* proteins conserved in all bacteria species (see section 2.3.8).

The candidates of glycolytic enzymes (20 in total) were compared with the above four groups of proteins. None of the glycolytic enzymes matched were matched to Group A (conserved in all three domains), Group C (eukaryotic signature proteins) or Group D (conserved in all bacterial species). However, there were six candidates matched to Group B (conserved in all eukaryotic species, Table 2).

Glycolysis in bacteria occurs in diverse forms. This means that none of the *Giardia*'s bacteria-like glycolytic enzymes are likely to be universal to all bacteria and thus less likely to be found matched to those in Group A or Group D. The eukaryotic glycolytic enzymes are more conserved across eukaryotes, and thus some of *Giardia*'s eukaryote-like glycolytic enzymes were found to be conserved in all eukaryotes; in the contrary, not all enzymes in the glycolytic pathway are maintained in all eukaryotes, this indicate one cannot assume that just because a pathway is conserved throughout eukaryotes then

the individual enzymes are. However, homologues of these enzymes conserved in all eukaryotes are also found in some branches of bacteria, hence they did not show up in Group C (eukaryotic signature proteins). This result is due to the large variety of glycolytic enzymes present in bacteria. Most of the enzymes found from *Giardia* from these key eukaryotic pathways were not classed as ESPs, due to their similarity with prokaryotic proteins.

Table 2. *Giardia* glycolytic enzyme candidates maintained in all eukaryotes

Protein	Enzyme name	EC number
GL50803_11118	enolase	4.2.1.11
GL50803_7260	alcohol dehydrogenase	1.1.1.2
GL50803_7982	aldose 1-epimerase	5.1.3.3
GL50803_90872	phosphoglycerate kinase	2.7.2.3
GL50803_9115	glucose-6-phosphate isomerase	5.3.1.9
GL50803_93938	triosephosphate isomerase	5.3.1.1

Using sugar pathways as examples, this method has been shown to be successful in analysing metabolic pathways from incompletely annotated genomes. More pathways, such as those involved in amino acid metabolism, and the RNA degradation pathway, can be analysed using this method, adding more pieces to the puzzle of *Giardia*'s metabolism. This study also identified *Giardia* candidates for enzymes that had not been recognised before. They bear high homology to known enzymes of their classes, and although the actual functions of these enzymes have not been confirmed, this work gives direction to future experimental confirmation with activity assays, which could then lead to the identification of new drug targets.

Typically a drug target is a key molecule for the infectivity or survival of a microbial pathogen. Selective toxicity would be best achieved if the parasite has a key enzyme that humans do not have or which is remarkably different from the host. For example, PFOR is found in *Giardia*, but the host (human or mammal) uses the pyruvate dehydrogenase complex to perform the same reaction, and thus drugs targeting PFOR such as metronidazole have been designed. From the glycolytic pathway, I have identified enzymes which are significantly different in *Giardia* from those in the host (see Table 2), including glucokinase and phosphofructokinase. Glucokinase has been investigated as a drug target for type 2 diabetes (Matschinsky 2009), and its potential to

be a target for parasite infection is as yet uncertain. Phosphofructokinase has been suggested as a drug target for *Entamoeba histolytica* by Byington *et al.* (Byington *et al.* 1997), and they designed a competitive inhibitor of phosphofructokinase, with the drug inhibiting the growth of the parasite *in vitro*. These enzymes, and especially those that can be compensated in the host by alternative pathways, hold the possibility of new targets for drugs effective against *Giardia*. An even better understanding of this parasite's metabolism will surely provide more ammunition against this worldwide parasitic problem.

Supplementary material for Chapter 4

S1. Enzymes of glycolysis pathway in *Giardia*

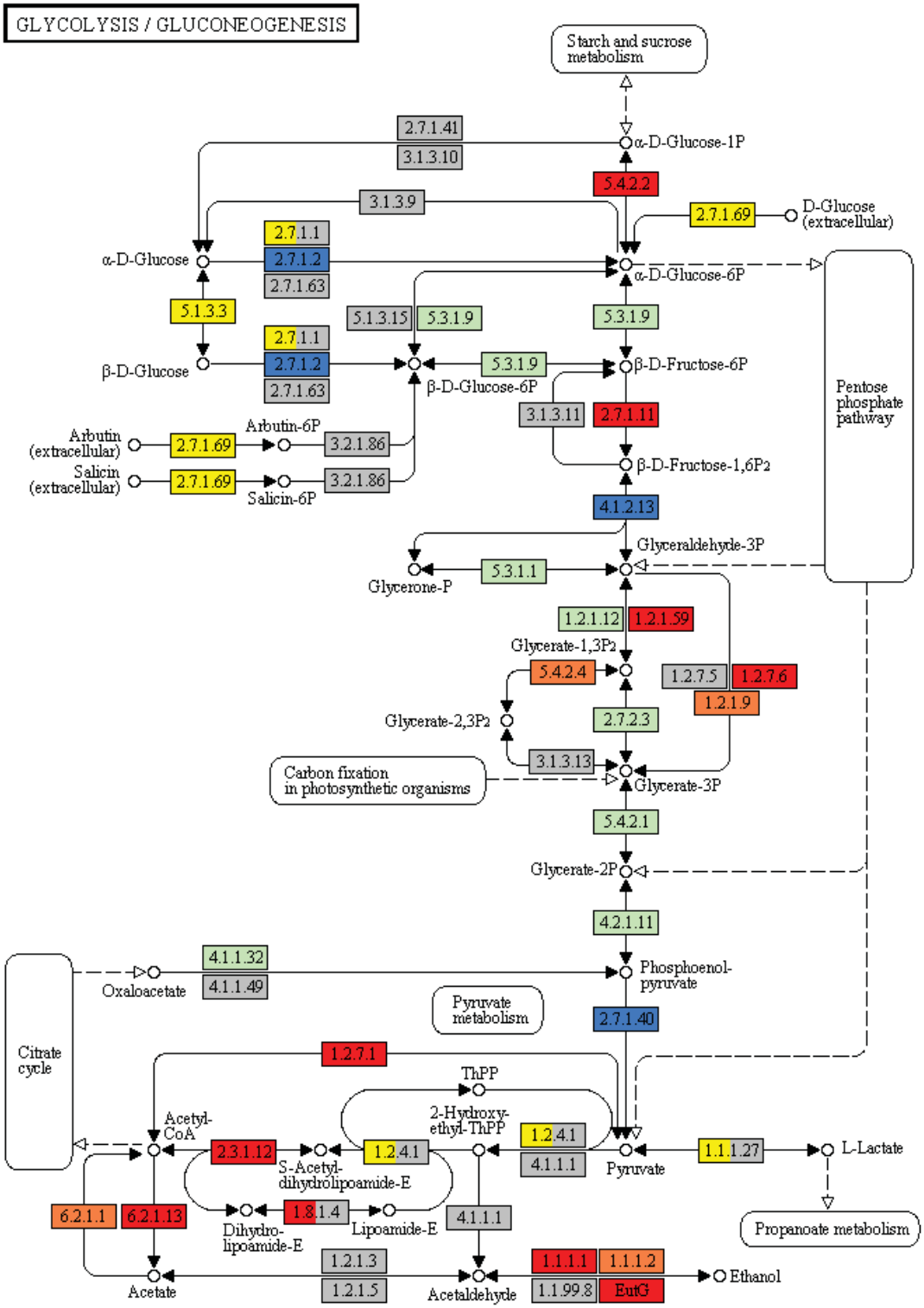
The enzymes are ordered in the approximately direction of the metabolic flux. † indicates this enzyme is already in KEGG *Giardia* glycolysis pathway, ‡ indicates the *Giardia* enzyme has already giving the EC number, but was not in the KEGG *Giardia* glycolysis pathway for unknown reason, ! indicates high possibility of false positive result. The last column indicates whether the *Giardia* enzyme is more homologous to bacterial (B), archaeal (A) or eukaryotic (E) enzymes, P indicates eukaryotic protists, which are considered separately from other eukaryotes.

EC	Name of the enzyme	#sequences in KEGG	Best candidate	Bit-score	E-value	Domain akin to
2.7.1.41	glucose-1-phosphate phosphodismutase	11	GL50803_14038	60.1	6.00E-13	-
3.1.3.10	glucose-1-phosphatase	84	GL50803_7556	31.6	0.003	-
5.4.2.2	phosphoglucomutase	929	GL50803_17254	310	9.00E-86	E, P
3.1.3.9	glucose-6-phosphatase	28	GL50803_5631	30.8	0.004	-
2.7.1.1!	hexokinase	305	GL50803_7260	179	5.00E-47	-
2.7.1.2‡	glucokinase	1110	GL50803_8826	393	2.00E-110	B, P
5.1.3.3	aldose 1-epimerase	736	GL50803_7982	124	3.00E-30	B, E
5.1.3.15	glucose-6-phosphate 1-epimerase	6	GL50803_9115	93.2	2.00E-22	-
5.3.1.9†	Phosphorhexose isomerase	1261	GL50803_9115	394	5.00E-111	P, B, E
2.7.1.69	glucose permease	8977	GL50803_9909	130	6.00E-31	B
3.2.1.86	6-phospho-beta-glucosidase	788	GL50803_35487	38.9	0.004	-
3.1.3.11	fructose-bisphosphatase	1286	GL50803_17316	99	3.00E-22	B
2.7.1.11	phosphofructokinase	1172	GL50803_14993	429	9.00E-122	P, B
4.1.2.13‡	aldolase	1839	GL50803_11043	390	3.00E-110	B, P
5.3.1.1†	triosephosphate	1352	GL50803_93938	348	8.00E-98	B, P, E

	isomerase					
1.2.1.12†	glyceraldehyde-3-phosphate dehydrogenase	1872	GL50803_17043	270	9.00E-74	B, E, P
1.2.1.12†	glyceraldehyde-3-phosphate dehydrogenase	1872	GL50803_6687	459	5.00E-131	E, B, P
1.2.1.59	glyceraldehyde-3-phosphate dehydrogenase (NAD(P)+)	138	GL50803_6687	326	6.00E-92	B
5.4.2.4	bisphosphoglycerate mutase	60	GL50803_8822	278	1.00E-77	B, E
2.7.2.3†	phosphoglycerate kinase	1289	GL50803_90872	453	4.00E-129	E
3.1.3.13	bisphosphoglycerate phosphatase	31	GL50803_135885	45.8	2.00E-08	-
1.2.7.5	aldehyde ferredoxin oxidoreductase	330	GL50803_13616	47.4	4.00E-07	-
1.2.7.6	glyceraldehyde-3-phosphate dehydrogenase	26	GL50803_6687	315	5.00E-89	B
1.2.1.9	glyceraldehyde-3-phosphate dehydrogenase (NADP+)	168	GL50803_6687	209	2.00E-56	B
5.4.2.1†	phosphoglycerate mutase	2987	GL50803_8822	551	4.00E-142	B
4.2.1.11†	enolase	1329	GL50803_11118	455	1.00E-129	P, E
4.1.1.32†	phosphoenolpyruvate carboxykinase (GTP)	322	GL50803_10623	470	2.00E-134	A, E, B
4.1.1.49	phosphoenolpyruvate carboxykinase (ATP)	554	GL50803_10623	41.6	4.00E-05	-
2.7.1.40‡	pyruvate kinase	1575	GL50803_3206	1243	0	
1.1.1.27!	L-lactate dehydrogenase	658	GL50803_17325	161	3.00E-41	-
1.2.7.1	pyruvate synthase	846	GL50803_17063	1008	0	B
1.2.4.1!	pyruvate	2632	GL50803_3281	156	3.00E-39	B

	dehydrogenase (acetyl-transferring)					
4.1.1.1	pyruvate decarboxylase	75	GL50803_9704	40.8	1.00E-05	-
2.3.1.12	dihydrolipoyllysine-residue acetyltransferase	1351	GL50803_113021	647	0	E, P
6.2.1.1	acetyl-CoA synthetase	1718	GL50803_13608	226	4.00E-60	B
6.2.1.13	acetyl-CoA synthetase (ADP-forming)	75	GL50803_13608	507	5.00E-146	A, B, P
1.8.1.4!	dihydrolipoyl dehydrogenase	2009	GL50803_16125	450	2.00E-127	-
1.2.1.3	aldehyde dehydrogenase (NAD+)	1521	GL50803_93358	70.1	4.00E-13	-
1.2.1.5	aldehyde dehydrogenase [NAD(P)+]	90	GL50803_93358	65.5	6.00E-13	-
1.1.1.1	alcohol dehydrogenase	2659	GL50803_93358	870	0	B
1.1.1.2	alcohol dehydrogenase (NADP+)	219	GL50803_7260	240	6.00E-66	E
1.1.99.8	alcohol dehydrogenase (acceptor)	102	GL50803_3861	79.3	2.00E-17	-
eutG	ethanol:NAD+ oxidoreductase	62	GL50803_93358	717	0	B
1.2.1.10	acetaldehyde dehydrogenase	611	GL50803_93358	870	0	B

Figure S1. KEGG diagram of glycolytic enzymes in *Giardia*

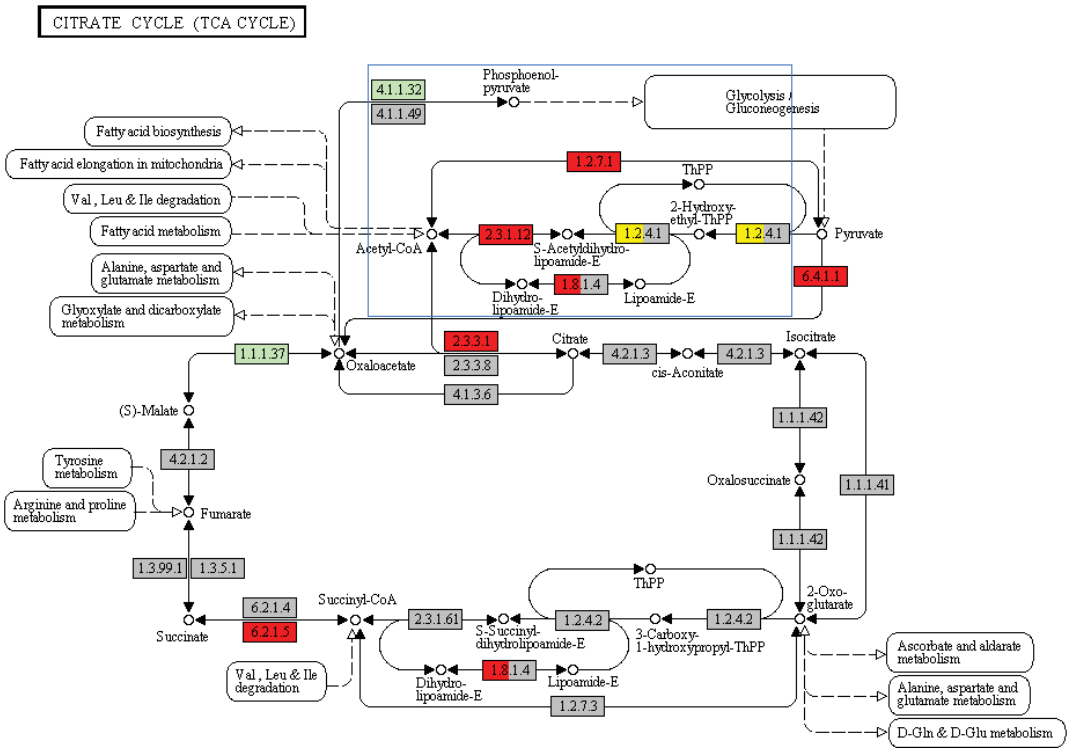


Key: The metabolites are labelled and shown as small circles, the enzyme which catalyse reactions from one metabolite to another are shown in rectangles, with their EC number indicated. The *Giardia* enzymes are coloured according to their homology to enzymes of other species: green: enzymatic function registered in KEGG; blue: EC number registered in KEGG but was not in the KEGG map of *Giardia* metabolic pathways; red: found in *Giardia* with score >300, these are enzyme candidates with fairly high degrees of certainty; orange: found in *Giardia* with score 200-300; yellow: found in *Giardia* with score 100-200, grey: found in *Giardia* with score <100. Half painted grey indicates the result is very likely a false positive. The template image was downloaded from KEGG.

S2. Enzymes of citric acid cycle in *Giardia*

EC	Name of the enzyme	#sequences in KEGG	Best candidate	Bit-score	E-value	Domain akin to
6.4.1.1	pyruvate carboxylase	526	GL50803_113021	644	0	P, B
2.3.3.1	citrate synthase	1374	GL50803_7195	359	2.00E-100	B, A
2.3.3.8	ATP citrate (pro-S)-lyase	166	GL50803_13608	46.6	6.00E-07	-
4.1.3.6	citrate lyase subunit alpha	1028	GL50803_38462	42.7	2.00E-04	-
4.2.1.3	aconitate hydratase 1	1591	GL50803_17063	89	2.00E-18	-
1.1.1.42	isocitrate dehydrogenase	1108	GL50803_14785	43.9	1.00E-05	-
1.1.1.41	isocitrate dehydrogenase (NAD+)	508	GL50803_11230	37.4	4.00E-04	-
1.2.4.2	oxoglutarate dehydrogenase	852	GL50803_33769	52.8	3.00E-08	-
2.3.1.61	2-oxoglutarate dehydrogenase	989	GL50803_33769	51.2	5.00E-08	-
1.2.7.3	2-oxoglutarate synthase	1528	GL50803_22677	114	2.00E-27	-
6.2.1.4	succinyl-CoA synthetase (GDP-forming)	361	GL50803_13608	48.9	1.00E-07	-
6.2.1.5	succinyl-CoA synthetase (ADP-forming)	2107	GL50803_13608	435	2.00E-123	B, A, E
1.3.99.1	succinate dehydrogenase	4118	GL50803_9089	56.6	5.00E-10	-
1.3.5.1	succinate dehydrogenase (ubiquinone)	1137	GL50803_92246	62.8	6.00E-11	-
4.2.1.2	fumarate hydratase	1807	GL50803_14259	84	5.00E-18	-
1.1.1.37†	malate dehydrogenase	1349	GL50803_3331	659	0	B, E

Figure S2. TCA cycle enzymes in *Giardia*



Colouring key is same as Figure S1. The bright blue box indicates this part of the pathway is also part of the glycolysis pathway in Figure S1. The template image was downloaded from KEGG.

S3. Enzymes of oxidative phosphorylation in *Giardia*

EC	Name of the enzyme	#sequences in KEGG	Best candidate	Bit-score	E-value	Domain akin to
Complex I						
1.6.5.3	NADH dehydrogenase	13702	12 proteins	1080	0	B, E, A
1.6.99.3	NADH dehydrogenase	3956	GL50803_33769	385	3.00E-108	B, E, A
1.6.99.5	NADH dehydrogenase (quinone)	3020	GL50803_14058	83.6	1.00E-16	-
Complex II						
1.3.99.1	succinate dehydrogenase	4118	GL50803_39312	42.4	3.00E-04	-
1.3.5.1	succinate dehydrogenase (ubiquinone)	1137	GL50803_9698	61.2	9.00E-12	-
Complex III						
1.10.2.2	ubiquinol-cytochrome c reductase	1838	GL50803_39312	42.4	6.00E-04	-
Complex IV						
1.9.3.1	cytochrome c oxidase	4440	GL50803_103783	38.1	0	-
ATP synthase						
3.6.3.14	F-type H+-transporting ATPase	14587	21 proteins	1800	0	P, E
3.6.3.10	H+/K+-exchanging ATPase	87	GL50803_96670	2665	1.00E-170	E
3.6.3.6	H+-transporting ATPase	196	4 proteins	590	0	E
Others						
3.6.1.1	inorganic pyrophosphatase	1831	3 proteins	1418	5.00E-05	
2.7.4.1	polyphosphate kinase	703	GL50803_8174	42.4	3.00E-04	-

S4 Perl script used

Using this Perl script, the user can fetch sequences from all organisms available in KEGG belonging to a particular EC number. This script requires the “genes.pep” file downloaded from KEGG. Note that future KEGG versions may modify this file and thus the script will also need modification should this occur.

```
use strict;
use warnings;
my $ec = ""; ##add the correct EC number inside the quotation marks.
my $output = "$ec.txt";
open (OUTPUT, ">$output") or die "output file not opened";
my $database = "D:\\Kegg maps\\all genes in kegg\\genes.pep";
open (DATABASE, "$database") or die "DATABASE file not opened";
my $quotedec= quotemeta($ec);
while (<DATABASE>) {
    if ($_ =~ /^\>/) {
          if ($_ =~ /$quotedec\D/){
          $found = 1;
          }
          else {$found = 0;}
    }
    if ($found == 1) {
    print OUTPUT "$_";
    }
    }
close DATABASE;
close OUTPUT;
#double check how many sequences were found
open (OUTPUT, "$output") or die "output file not opened";
my $seqcount = 0;
while (<OUTPUT>) {
    if ($_ =~ /^\>/){
    $seqcount ++;
    }
}
print "total of $seqcount sequences in $output";
```

Chapter 5: Non-coding RNAs of *Giardia* and *Trichomonas* and their relationship to ESPs

5.1 Introduction to small ncRNAs

The occurrence of ESPs and RNAi are both thought to represent ancient mechanisms, and both appear to be present in all main branches of eukaryotes. Have they both existed since the last common ancestor of eukaryotes? This chapter will investigate if there is any correlation between ESP and genes possibly regulated by RNAi.

RNAi is a system by which RNA is used to control the expression of genes. RNAi usually involves two types of non-coding RNA (ncRNA) molecules, micro RNA (miRNA) and small interfering RNA (siRNA) (other types of small ncRNAs such as piRNAs, tasiRNAs will not be discussed in detail here). RNAi is a typical eukaryotic feature, and has been found in most branches of eukaryotes. There are some lineages that have lost their RNAi proteins but some still maintain some form of ncRNA-based regulation. For example, *Saccharomyces cerevisiae* does not have the Dicer-like RNases nor Argonaute, but it does have ncRNAs that act in the regulation of its genes (Drinnenberg *et al.* 2011). Therefore the existence of RNAi appears universal in eukaryotes. Although RNAi (using Dicer and Argonaute proteins) as such does not exist in prokaryotes, bacteria do have a comprehensive small RNA network system which does have some similarities to the eukaryotic RNAi (Collins 2011).

Typical miRNAs are ~21-22 base pair (bp) single stranded RNA (Carrington *et al.* 2003), and siRNAs are about 21-26 bp double stranded RNA (Hamilton *et al.* 2002). The miRNAs and siRNAs are processed in a similar manner (Figure 1). The siRNA precursors (pre-siRNAs) are normally double stranded RNAs that are transcribed by RNA dependent RNA polymerase 2 (RDR2) and RDR6 by using single stranded target RNAs as templates. The pre-miRNAs are produced from the transcription of a genomic locus independent of the target locus by DNA dependent RNA polymerase II (Williams *et al.* 2005), and the precursor molecules usually form double-stranded RNA by intramolecular pairing (Ambros *et al.* 2003). The RNAs precursors are then cleaved into

20 to 25 nt RNAs by Dicer or Drosha, which are RNase II family endonucleases in the cytoplasm (Carrington *et al.* 2003).

Figure 1. Micro RNA and siRNA mechanism of action

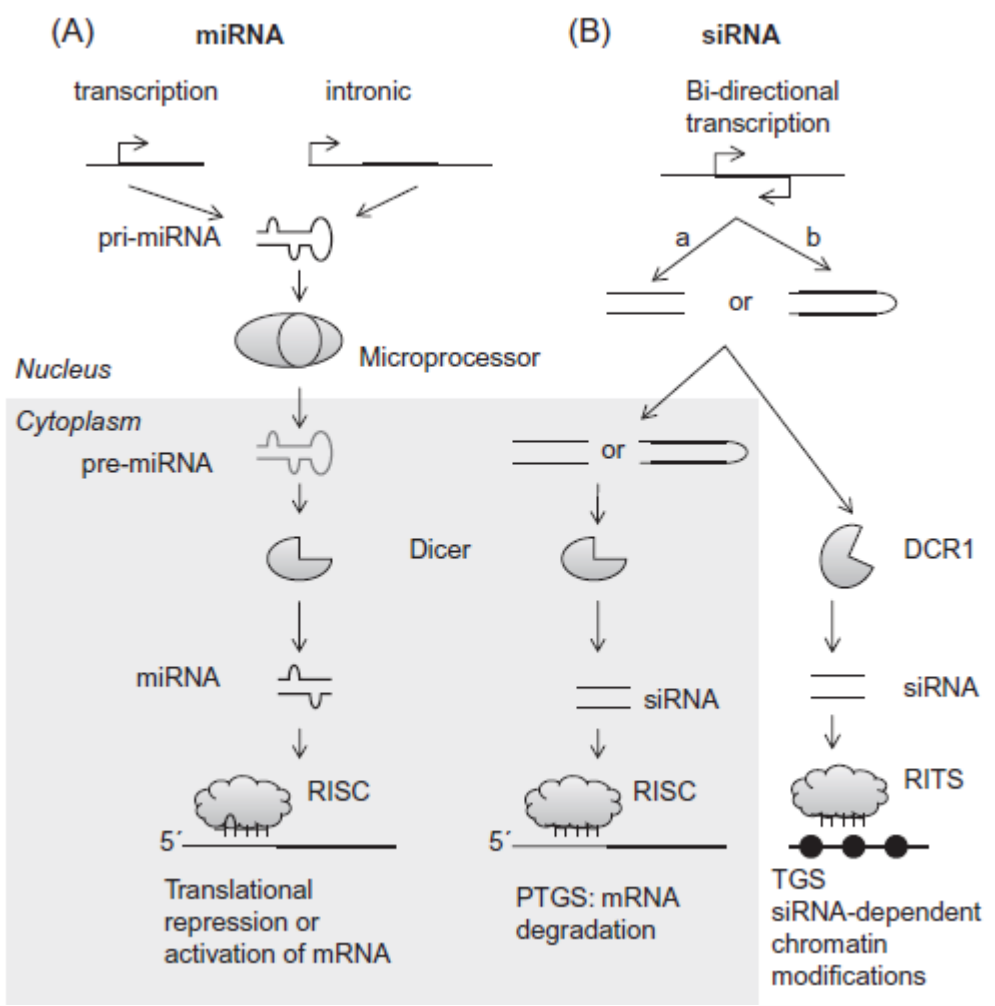


Figure modified with permission from "The Epigenetics of non-coding RNA" Chapter in The Handbook of Epigenetics (Collins *et al.* 2011).

After cleavage by Dicer, the miRNAs and siRNAs are incorporated into ribonucleoprotein particles which assemble to form the RNA-induced silencing complex (RISC). RISC unwinds the RNA duplex, and usually only one strand is active. The single-stranded siRNA or miRNA guides the RISC complex to the target mRNA, and it is strongly bound to the Argonaute protein which then cleaves the target mRNA. The cleaved mRNA is then recognised as aberrant and destroyed (see Figure 1). In metazoans the miRNA target sites are in the 3' untranslated regions (UTR) of the mRNA (Bartel 2009); in plants targets can be located in the 3' UTR but are more often

found in the coding region itself (He *et al.* 2004), and there are also studies suggesting the target sites reside on some promoter regions (Collins 2011). Small interfering RNAs are structurally related to miRNA and act via incomplete complementary base-pair interactions with a target mRNA. The siRNA can also act in RNAi-related pathways, such as in an antiviral mechanism where it binds to foreign DNAs causing cleavage and degradation of these DNAs (Ahlquist 2002).

With the completion of genomes from *Giardia lamblia* and *Trichomonas vaginalis*, we can now use genomics to analyse the RNAi systems in these basal eukaryotes. *Giardia* and *Trichomonas* are both anaerobic eukaryotic parasites (Keeling *et al.* 2005), yet they are separated by a long evolutionary distance (Hampl *et al.* 2009), making them comparable yet distant models to study. *Giardia* and *Trichomonas* both have eukaryotic specific RNAs such as snoRNAs (Yang *et al.* 2005; Chen *et al.* 2007), spliceosomal snRNAs (Chen *et al.* 2008; Simoes-Barbosa *et al.* 2008), and RNase P (Marquez *et al.* 2005). Dicer and Argonaute homologues have also been identified in *Giardia* (Dicer: GL50803_103887; Argonaute: GL50803_2902; neither were ESPs because they have no homologues in some eukaryotes such as *Schizosaccharomyces*) (Saraiya *et al.* 2008) and *Trichomonas* (Dicer: TVAG_491480; Argonaute: TVAG_463390 and TVAG_419780; again neither were ESPs) (Carlton *et al.* 2007). *Giardia* is also well known for its large abundance of antisense RNAs (Ullu *et al.* 2005; Teodorovic *et al.* 2007). Therefore, the presence of other basic small RNAs such as miRNAs and siRNAs in the two parasites is expected. *Giardia* and *Trichomonas* may also possess some different RNA processing components from those found in other eukaryotes (Chen *et al.* 2007). As an example, MacRae *et al.* suggested RNA fragments cleaved by *Giardia* Dicer are slightly longer than the typical miRNA at about 25-27 bp long (MacRae *et al.* 2006).

Investigating the different classes of *Giardia* and *Trichomonas* ncRNAs involves an interesting field to study. Researchers at Massey University, Palmerston North are currently investigating the evolution of ncRNAs by using RNA data from these two organisms (Chen *et al.* 2007; Chen *et al.* 2008; Collins *et al.* 2009). In 2008, Illumina Solexa sequencing and genome wide analysis of small RNAs from *Giardia* and *Trichomonas* were performed. From this sequencing data, Chen *et al.* identified 10 miRNA candidates from *Giardia* and 11 from *Trichomonas* (Chen *et al.* 2009). These candidates were named Gims and Tvms respectively. In addition, Chen *et al.* also

characterised five unusual long tandem repeated double stranded RNAs which were named Girep-1 to Girep-5. The sequence alignment confirmed these five RNAs belong to the same group, and they share high degrees of sequence similarity with a number of variant-specific surface proteins (VSPs). VSP gene expression is crucial for the surface antigenic variation of *Giardia* trophozoites. By displaying different VSPs on the surface, *Giardia* is able to evade the host's immune system (Nash *et al.* 2001). Chen *et al.* suspected Gireps are precursor siRNAs and have strong potential to be involved in regulation of VSP expression. Other research clarified the annotation of RNase P and RNase MRP RNA as well as identifying examples of the H/ACA class of small nucleolar RNA (snoRNAs) (Chen *et al.* 2011).

In this chapter, the same Illumina sequencing data was re-analysed to further investigate the classes of small RNAs and especially how they relate to ESPs. The idea was to undertake preliminary data mining to uncover small RNA groups then for myself and my supervisor to apply that information to our different interests. The method in brief included removing known adaptor sequences from the sequences to yield a dataset of 15-29 nt single stranded RNAs. The sequences were then mapped to the organisms' genomes to identify possible miRNAs and siRNAs. This study involved Perl and various other bioinformatics data mining tools to find potential RNAs targeting sites, and evaluation of these with respect to ESP genes. This idea was then to combine the ESP results with the *Giardia* RNA Illumina results, searching for ncRNAs affecting ESPs. Some correlation between ncRNAs and ESPs was expected, because both are thought to be ancient mechanisms. It is hoped that the results will tell us more about the RNAi in deep-branching eukaryotes, as well potentially some insights on how ESPs are regulated.

5.2 Methods

5.2.1 Sample preparation and sequencing

Giardia was grown and the DNA collected prior to this study by Dr. Sylvia Chen in the following manner: *Giardia lamblia* (WB strain) trophozoites were grown in TY1-S-33 growth media at concentration of $1.4x10^7$ cells/ml and collected by centrifugation. Samples of total RNA were prepared using Trizol (Invitrogen) according to the protocol provided by the manufacturer.

Trichomonas vaginalis was grown in *Trichomonas* broth (Fort Richard) at 37 °C for 3–4 days. The culture was harvested by centrifugation. Growth media was removed and cells were resuspended. An equal volume of phenol:chloroform (5:1, pH 5) was added to the suspension, and the mixture was vortexed for 10 seconds. After that phases were separated by centrifugation, and the upper phase was further extracted twice with phenol:chloroform, then once with chloroform. Finally, total RNA was precipitated by adding LiCl to a final concentration of 0.2 M and 3 volumes of 100% EtOH, and incubated at -80 °C for 1 hour.

For sequencing, 10 μg of total RNAs were separated on a 15% denaturing acrylamide 8 M urea gel and RNAs ranging from 10 to 200 nt were cut out from the gel and prepared according to Illumina's small RNA preparation protocol. 8 pmol of *Giardia* cDNA and 12 pmol of *Trichomonas* cDNA were used for sequencing on an Illumina Genome Analyzer for 35 cycles. The sample preparation and sequencing steps were performed by Dr. Sylvia Chen and technicians of Massey University Genome Service (Chen *et al.* 2009).

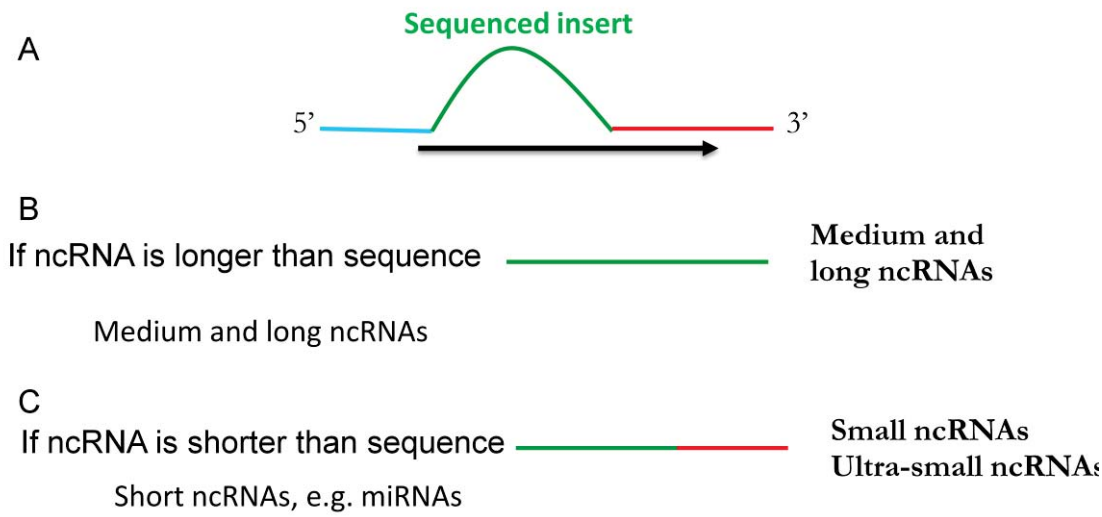
5.2.2 Adaptor trimming and mapping

The collection of our small RNA datasets from *Giardia* and *Trichomonas* is summarised in Figure 3. For each short sequence from the data, adaptor sequences were removed from the short-read sequences using the FastX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/, see Figure 2 for the reason for adaptor sequences to be removed).

Sequences were initially trimmed to 34 nt. Sequences of ≤ 14 nt or ≥ 30 nt were discarded due to possible trimming errors, the remaining 15-29mers were our small ncRNA candidates. To aid in mapping the dataset was collapsed to unique sequences also using the FastX-toolkit. However, many of these sequences were found to be tRNAs and rRNA fragments so mapping to known tRNAs and rRNAs ensured that these sequences were removed. The unique 15-29mers were then mapped to the organism's genome respectively allowing two mismatches for strain differences by mapping software Bowtie (Langmead *et al.* 2009). The mapping steps were performed prior to this study by Dr Lesley Collins. Genome coordinates of mapped sequences have been obtained and a MySQL database was constructed containing small 15-29mer

RNAs. The database also contains sequence information as well as the genome location of RNAs.

Figure 2. Why adaptor trimming was performed



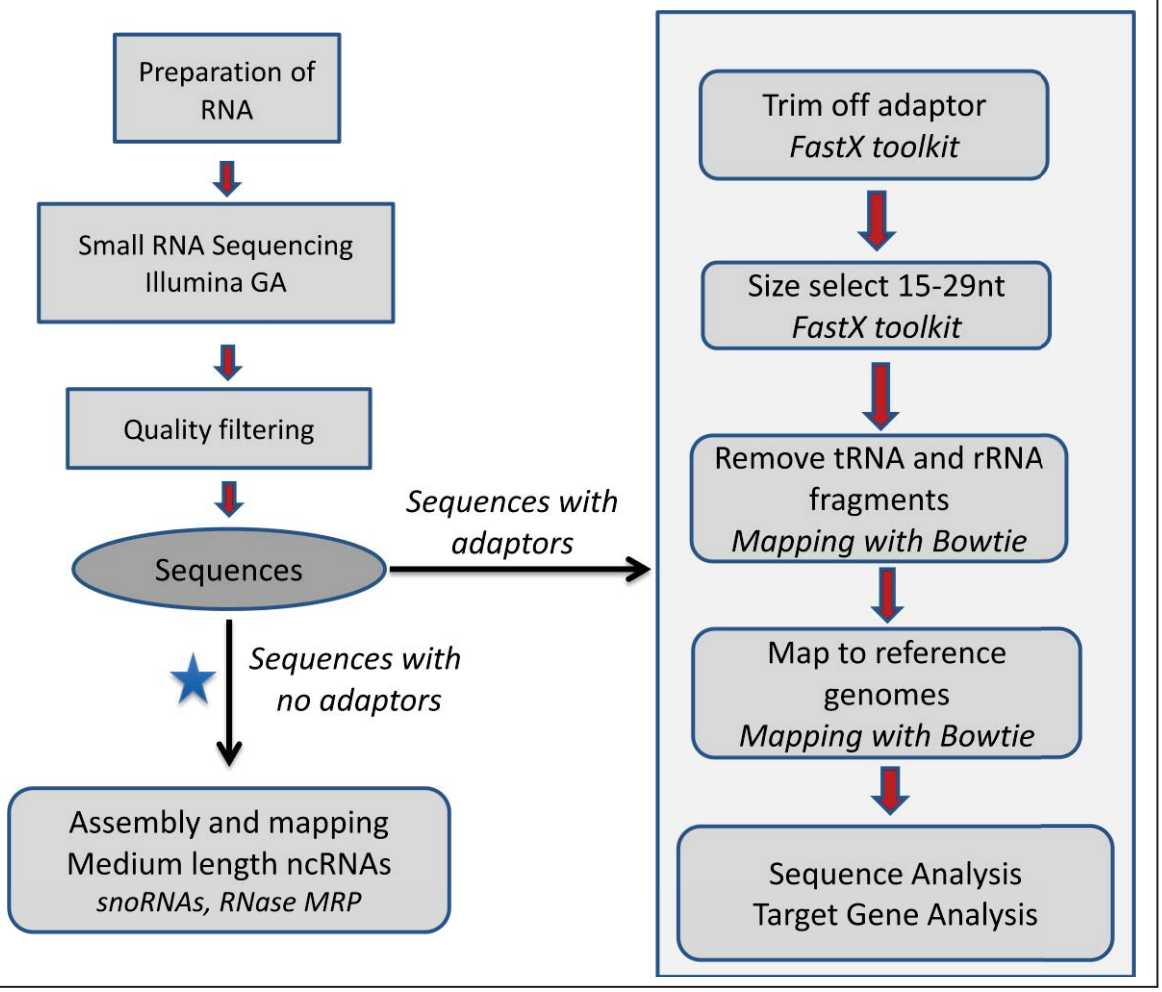
A: A ncRNA sequence (green) with attached 5' adaptor (blue) and 3' adaptor

B: when the ncRNA sequence is longer than the cycles (35 in this case) performed by the sequencer, only the first 35 bases are read.

C: when the ncRNA sequence is shorter than the cycles performed, the sequencer will read through the 3' adaptor. Therefore the adaptor sequences have to be removed to obtain the true sequence of the insert.

For the *Giardia*, the RNAs appeared to form two length peaks. Sequences of the two peaks 15-18 nt (by my supervisor Lesley Collins) and 26-27 nt (by myself) were collected for detailed analysis. Genome sequences and annotation were obtained from GiardiaDB (http://www.giardiadb.org) using *Giardia intestinalis* version 2.3 (Assemblage A, WB strain), and TrichDB (http://www.trichdb.org) using *Trichomonas vaginalis* version 1.2. The 26-27 peak was considered to be those sequences ideally sized for the Giardia Dicer protein so were analysed for potential miRNAs which could regulate gene expression. The 15-18 peak is unusual and is currently under investigation. A manuscript is currently in preparation and will include results from both studies. I did not expect to map 'all' miRNAs or identify in detail any miRNAs but to look as a first pass, at the trend of how these RNAs map against ESPs. A summary of the analysis procedure is presented in Figure 3.

Figure 3. Summary of analysis procedure



5.2.3 Finding mapped RNA targeting sites

As miRNAs can affect other RNAs by complementary pairing, we can find potential target sites by locating ncRNA sequences complementary to genes and gene regulatory regions. The table containing unique RNAs mapped to *Giardia* genome and *Giardia* coding regions table (9747 genes, containing 3846 deprecated genes) were merged into a single table, containing information on genomic coordinates of both RNAs and coding regions. The rearrangement of results by genomic coordinate was performed using MySQL, and the result set was exported to text files. By using Perl scripts, RNAs that are in the vicinity of coding regions were pulled out, and divided into six categories according to strand and position in relation to the gene (Table 1).

The mapping results show a 26mer and 27mer peak for *Giardia* RNAs. These 26mers and 27mers were subjected to the analysis above in order to find RNAs of this length adjacent to or inside of coding regions. The protein products of the coding regions these

RNAs associated with, have also been analysed to see if there is any trend in the proteins' function. This step was not performed for *Trichomonas* RNAs due to the complete lack of a 26mer and 27mer peak.

Table. 1 list of categories for mapped RNAs

Category	Strand	Location in relation to coding region
1	sense	Upstream or partially overlap
2	sense	Inside of the gene
3	sense	Downstream or partially overlap
4	antisense	Upstream or partially overlap
5	antisense	Inside of the gene
6	antisense	Downstream or partially overlap

5.3 Results and Discussion

5.3.1 Summary of number of RNAs yielded after each step

Illumina sequencing of small RNAs of *Giardia* and *Trichomonas* was performed. After the initial trimming of the adaptor sequences, several filtering and mapping steps were performed for the interpretation of results. The yield of RNA sequences after each step are summarised by Table 2.

Table 2. Number of *Giardia* and *Trichomonas* RNAs remained after each step

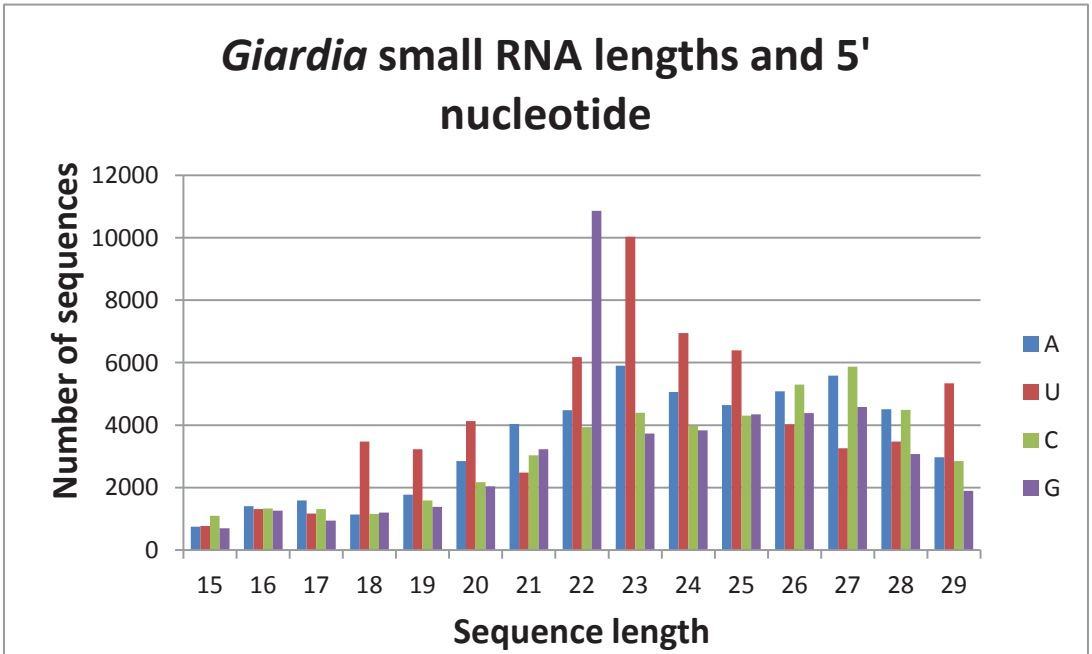
Step	# *Giardia* RNAs	# *Trichomonas* RNAs
Trimming adaptor sequences to 14-34mers	327899	327980
Selecting 15-29mers as candidates	208257	210529
Use only unique results (i.e. delete repeated sequences)	88758	92452
Delete sequences mapped to tRNAs and rRNAs	74647	90384
Sequences mapped to organism's genome	34196	8562
26mers and 27mers	5447/5588	NA

5.3.2 Small RNAs of *Giardia*

The adaptor sequence from raw sequence data was removed, resulting in 327899 RNAs between 14-33 nt long. However, sequences of ≥30 nt may be possible trimming

remnants: because the sequencer only performed 35 cycles, RNAs longer than 33 nt will only have their first 33 nt sequenced, appearing to be 33 nt long (or a little bit shorter). Sequences of 14 nt were also discarded because they were too short for effective mapping (15nt being the shortest effectively used in mapping, personal communication L. Collins). Hence 14 nt or ≥ 30 nt were not used, leaving 15-29mers to be analysed in this study. This length range later did prove to be sufficient for the subsequent analyses. A new database containing 208257 oligonucleotide sequences between 15-29 bases long was constructed. The 5'nucleotide (i.e. the first nucleotide from the 5' end of each RNA) was also analysed, because it may offer hints to the functions of these RNAs and/or how they are processed (Drinnenberg *et al.* 2011). The sequence length and 5' nucleotide distributions were performed for RNAs between 15-29 nt in length (Figure 4).

Figure 4. Length and 5' nucleotide distribution for *Giardia* ncRNA



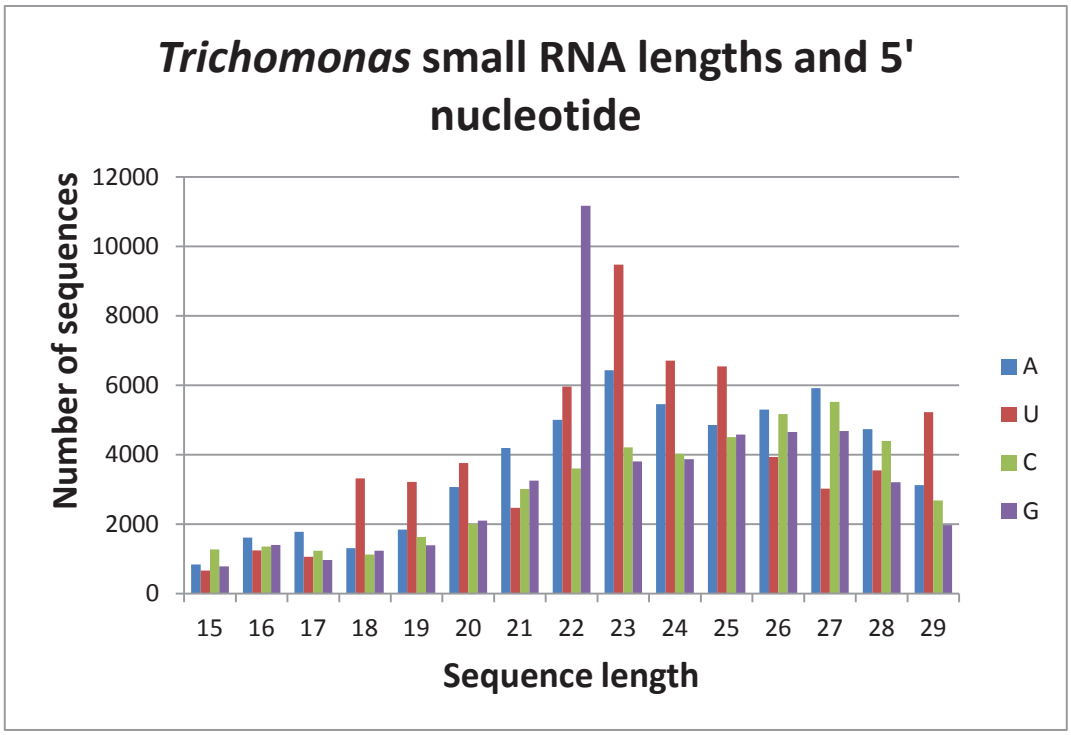
The graph peaked at length of 22-23, which is the expected length of miRNA in metazoan. The most abundant 5' nucleotides are As and Us, which is expected for miRNA and siRNA. There is a large amount of 22 nt long nucleotide with G at 5' end. This is unusual as G is the least abundant 5' nucleotide in small RNAs of other lengths. The small RNA "GTGGAGACCGGGGTTCGACTCC" occurred 7075 times, has contributed to this bias. Another sequence, "TCCGTGATAGTTTAATGGTCAGAA-

TGGGC" and its truncated forms have also occurred >10,000 times. When these sequences were BLASTed at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi), they both matched to nucleotides from a number of yeast species, including *Saccharomyces cerevisiae*. This indicated possible contaminations, which is apparently common with ncRNA sequencing with the earlier kits. By mapping the data to the organism's genome, the contaminant sequences were effectively removed (see sections 5.3.4 and 5.3.5).

5.3.3 Small RNAs of Trichomonas

The adaptor sequence from the *Trichomonas* sequencing data was trimmed off leaving sequences to 14-34 nt, with 15-29 nt RNAs selected for further analysis. The small RNA database contained 210529 oligonucleotide sequences between 15-29 bases long was constructed, and the sequence length and 5' nucleotide distributions were performed (Figure 5).

Figure 5. Length and 5' nucleotide distribution for *Trichomonas* ncRNA



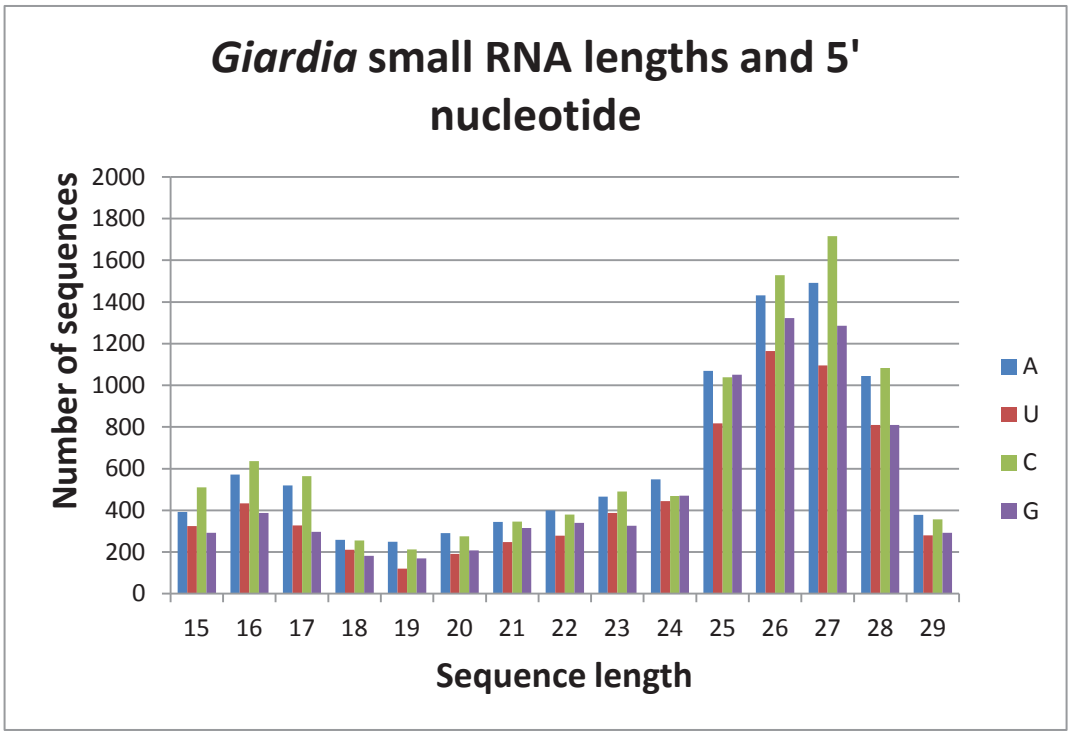
The graph is remarkably similar to that of *Giardia*'s. It peaked at length of 22-23 nt. The most abundant 5' nucleotides are As and Us. Sequences "GTGGAGACCGGGG-

TTCGACTCC" and "TCCGTGATAGTTTAATGGTCAGAATGGGC" again occurred ~10,000 times each indicating similar contamination.

5.3.4 *Giardia* mapping results

For both *Giardia* and *Trichomonas*, sequences that occurred multiple times were collapsed, and only unique results were used from this point onwards. This aids in the mapping process and subsequent interpretation of results. The 88758 unique sequences were mapped to rRNAs and tRNAs first, and 74647 sequences that were not mapped were further to mapped against the *Giardia* genome (Table 2). This step is performed to make sure that the small RNAs were not remnants of rRNA and tRNAs, and after this, the remaining RNAs were mapped to the *Giardia* genome to eliminate any other contaminant sequences.. There were 34196 sequences successfully mapped to the *Giardia* genome, and the sequence length and 5' nucleotide distributions were analysed (Figure 6).

Figure 6. Length and 5' nucleotide distribution for mapped *Giardia* ncRNA



The *Giardia* small RNA distribution graph shows two peaks, one "larger peak" around 26-27 nt in length, and an "ultra small peak" around 15-18 nt in length. The 5' nucleotide distribution has been fairly consistent throughout different sequence lengths,

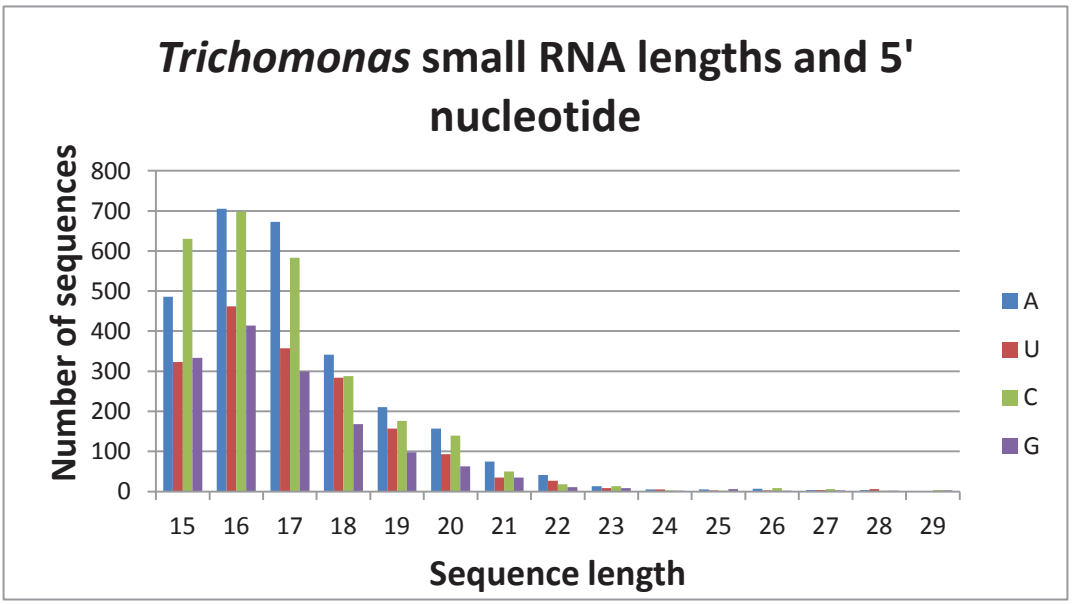
with Cs being the most common, followed by As and Gs; Us are the least common 5' nucleotides. This 5' nucleotide distribution is also quite similar to overall nucleotide distribution of the sequenced RNAs (see section 5.3.6).

The "larger peak" contains 26-27mers, which is with the similar length range to MacRae *et al.*'s finding that RNA fragments cleaved by *Giardia* Dicer are about 25-27 bp long (MacRae *et al.* 2006); the function of the "ultra small peak" is unclear, as it has only been reported previously in human and referred to as "unusually small RNAs (usRNAs)" (Li *et al.* 2009). The human usRNAs could possibly be miRNA degradation-like products or can be non-miRNA-derived. The two peaks are likely to represent two separate RNAi mechanisms and is interesting work for the future.

5.3.5 *Trichomonas* mapping results

The 92452 unique *Trichomonas* sequences were mapped to rRNAs and tRNAs first, then the 90384 sequences that were not mapped were further mapped against the *Trichomonas* genome (Table 2). There were only 8562 sequences successfully mapped to the *Trichomonas* genome, which is significantly less than that of *Giardia*. The sequence length and 5' nucleotide distributions were analysed (Figure 7).

Figure 7. Length and 5' nucleotide distribution for mapped *Trichomonas* ncRNA



The distribution graph shows only the "ultra small peak" 15-18 nt in length. This peak was also observed in *Giardia* RNAs. However the 26-27 nt "larger peak" was not observed from the *Trichomonas* data. The absence of the larger peak suggests that the

"ultra small peak" is not a result of secondary processing of larger miRNA products like human usRNAs were suggested to be (Li *et al.* 2009). The 5' nucleotide distribution is similar to that of *Giardia*, with highest proportion of Cs, followed by As and Gs, Us are the least common 5' nucleotides. Although the *Trichomonas* genome is not completely assembled, most of the unmapped regions are repeat and transposon regions. Whether the genomically-unmapped sequences do in fact lie in these regions and are of different lengths is as yet unknown.

5.3.6 GC content

The total GC content of the mapped RNAs was analysed for any trends (Table 3). The GC content of all annotated transcripts and the entire genome from the organisms were also calculated as references. The results indicate all sequenced small RNAs, including the peaks (*Giardia* and *Trichomonas* 15-18mer peak and *Giardia* 26-27mer peak) have a high GC content.

Table 3. GC content of *Giardia* and *Trichomonas* small RNAs

	A	**T**	**G**	**C**	**GC content**
Giardia					
26-27mers	26.16%	18.63%	30.60%	24.62%	55.21%
15-18mers	23.94%	24.31%	28.83%	22.92%	51.75%
All unique mapped sequences	25.90%	19.31%	30.77%	24.03%	54.80%
All (non-unique) mapped and unmapped sequences	24.25%	19.94%	32.77%	23.04%	55.81%
Annotated transcripts (GiardiaDB1.3)	26.61%	24.25%	24.26%	24.88%	49.14%
Entire genome (GiardiaDB1.3)	25.42%	25.33%	24.62%	24.63%	49.25%
Trichomonas					
15-18mers	24.97%	25.04%	27.09%	22.38%	49.47%
All unique mapped sequences	26.26%	25.09%	27.09%	21.57%	48.66%
All (non-unique) mapped and unmapped sequences	24.43%	19.99%	32.69%	22.89%	55.58%
Annotated transcripts (TrichDB1.1)	36.21%	28.24%	17.70%	17.84%	35.54%
Entire genome (TrichDB1.1)	33.57%	33.60%	16.41%	16.41%	32.83%

5.3.7 Determination of whether the "ultra small peak" of *Giardia* is a result of secondary processing of longer RNAs

Both *Giardia* and *Trichomonas* have the "ultra-small peak" containing RNAs in the size range of 15-18 nt. The aim here was to investigate whether the "ultra-small peak" was associated with the secondary processing of siRNAs (Baulcombe 2007), and thus we would find the presence of shorter sequences (16-17mers) within our longer sequences (include 25-27mers, 25mers were also included because MacRae *et al.*'s finding suggested that RNA fragments cleaved by *Giardia* Dicer are about 25-27 bp long (MacRae *et al.* 2006)) .The analysis were done using the genomic coordinates and with the assistance of Perl scripts. The results did not show significant overlaps between the shorter RNAs and the longer RNAs, indicating there is limited evidence of the "ultra-small peak" being secondary siRNA products (Table 4), thus it appears that the two peaks represent are two separate mechanisms. However, the "ultra small peak" is still being investigated as to whether it contains breakdown fragments from other longer ncRNAs such as snoRNAs.

Table 4. Number of overlapping 16 and 17mers

	16mers	**17mers**
Total mapped sequences of this length	2030	2068
sense overlap sequences	174 (8.6%)	194 (9.6%)
antisense overlap sequences	117 (5.8%)	130 (6.4%)

5.3.8 Possible target sites of *Giardia* 26 and 27mers

The two peaks in the mapped *Giardia* sequence length distribution graph (see Figure 6) might be evidence of two different mechanisms of RNAi, therefore further analysis of their properties is necessary. The 26mers and 27mers of *Giardia* (11035 sequences in total) were subjected to more analysis by me; and the analyses performed on 15-18mers were done by my supervisor Dr. Lesley Collins (and thus not reported in this thesis).

Although the ncRNAs are transcribed elsewhere, the target sites are usually in proximity to the target gene (3'UTR regions are commonly suggested (Bartel 2009; Chen *et al.* 2009)). RNA interference is achieved by means of complimentary binding of ncRNA to the mRNAs. Therefore during mapping, the ncRNAs can be mapped to two places: the loci where they are transcribed and the target sites because they will be

complimentary to the coding strand. A parameter was thus set during the mapping procedure so that if RNAs are mapped to multiple loci, all of these loci will be shown. This analysis will give insights on where the target sites for the 26-27mer RNAs, and what genes are potentially regulated by the RNAs. The number of ESPs with RNAs in proximity were then analysed to look for any connection between RNAi and ESP.

The genome coordinates of 26mers and 27mers were compared with those of protein coding genes in the following manner. RNAs within 100 bp from 5' and 3' end of the gene, as well as RNAs located inside of the gene from both sense and antisense strand have been identified. Table 4 summarises the number of 26mers, 27mers and genes found for each region. The *Giardia* genome database GiardiaDB listed many "deprecated" gene products (3846 deprecated genes within total of 9747 genes), and the functional genes have been recorded as a separate column in Table 5.

Table 5. Loci of 26 and 27mers in relation to genes

Strand	Location in relation to coding region	# RNAs	# all genes	# functional genes	# ESPs
sense	Upstream or partially overlap	635	104	99	3
sense	Inside of the gene	4800	464	409	27
sense	Downstream or partially overlap	758	142	105	4
antisense	Upstream or partially overlap	555	100	62	4
antisense	Inside of the gene	3354	224	152	5
antisense	Downstream or partially overlap	642	119	85	1

In general, there are more RNAs on the sense strand than antisense strand, and possibly there are some mRNA degradation products here. There are large numbers of RNAs located inside the coding region of the gene, but this could be merely a statistical issue, because the average coding region length is 1064 bp, whereas the searched for upstream and downstream RNAs only carried on for 100 bp from the start or end of the coding region. The number of RNAs on the 3' antisense is only slightly larger than on the 5'antisense region. Given that these two sites are the most likely possible target binding sites, the conclusion cannot be made whether RNAs target sites are typically in the 3' UTR of the mRNA (Bartel 2009), coding region itself (He *et al.* 2004) or the promoter region (Collins 2011). There is as yet no data from the *Giardia* genomics community on this issue (personal communication L. Collins).

The functions of genes with small RNAs in proximity were also analysed, and there was no obvious trend in gene functions. This is partly due to the annotation of the *Giardia* genome, as more than half of gene products are listed as "hypothetical proteins". There is no significant higher or lower proportion of ESPs in the genes with 26-27mers in vicinity.

The same Illumina data was also assembled into 'contigs' using a consensus assembly (performed by Dr. Sylvia Chen in 2008) (Chen *et al.* 2009). From the assembled contigs, Dr Sylvia Chen predicted 10 *Giardia* miRNA candidates and *Trichomonas* 11 miRNAs, using a strategy of sequence similarity and searching miRNAs by definition (Chen *et al.* 2009). These miRNA candidates were named Gim1 to Gim10 and Tvm1 to Tvm11, respectively. The target genes of these miRNAs were also predicted by Dr Sylvia Chen. I have compared these putative target genes with the ESP dataset, to find out if any of these genes were ESPs. The results showed that none of the predicted target proteins of the Gims or the Tvms targets were ESPs. So results so far from both analyses suggest that there is no real relation between RNAi and ESP genes, unlike what was hypothesised. However, the work in this area is ongoing.

The experiment to locate RNAs in location of the gene using genome coordinates is a new method. RNAs could map to their actual sites where they are transcribed or the site where it interacts with gene transcripts (mRNAs). The miRNA acts on the gene and can have mismatches in the binding so it is possible that this method here for *Giardia* might not allow enough mismatches for the RNA to be mapped to where it binds to the mRNA depending on whether mismatches are required (as for plants) or not (as for humans).

Overall, this analysis did not show ESPs to have more association with 26-27mers, Gims or Tvms than other proteins do. Searching for eukaryotic signature RNAs (ESRs) can be performed in future, and this will form parallel work to the current ESP analysis. However, more "ribo-genomes" have to be completed before this can take place.

5.4 Conclusion

High throughput sequencing of RNAs <100 nt length for parasites *Giardia lamblia* and *Trichomonas vaginalis* was re-analysed. From the sequencing data RNAs between 15-29 nt long were selected and shown to have a high GC content. Examination of *Giardia* small RNA data uncovered two length peaks: a larger peak around 26-27 nt long,

possibly cleaved by Dicer (MacRae *et al.* 2006); the other an "ultra small peak" containing RNAs 15-18 nt long. Only the "ultra small peak" was present in *Trichomonas*. The two length ranges of RNA are possibly two different groups of ncRNA which are cleaved through different mechanisms. The 26-27 nt sequences could be considered possible miRNAs because this is the range that Dicer has been functionally shown to process (MacRae *et al.* 2006). The 15-18 nt long peak has only been shown in humans, and they could be potential siRNAs or cleavage products from an as yet unknown mechanism (Dr Lesley Collins, personal communications). This study has confirmed the two length types of small RNAs from *Giardia* and the absence of one of these types from *Trichomonas*, which will guide further studies to unravel the actual RNAi mechanisms of the two deep branching eukaryotic parasites.

There appeared to be no connection between ESPs and genes that small RNAs regulate, which indicate the two arose separately, or the RNAi mechanism have evolved to much for the two to have any apparent connections. This study is a part of a larger project currently undergoing in Massey University in RNA systems biology. A manuscript is currently in preparation that will include work from this chapter.

Supplementary material for Chapter 5

S5.1 Abstract for 3rd Next Generation Sequencing Conference

3rd Next Generation Sequencing
Conference
23-24 August 2011
Palmerston North Convention Centre
Palmerston North

Small and Ultra-small RNAs from Parasitic Protists – more needles from the haystack of NGS data

Jian Han[1] and Lesley Collins[2]

[1]*Institute of Molecular BioSciences, and* [2]*Institute of Fundamental Sciences, Massey University, Palmerston North, NZ.*

ncRNAs abound within eukaryotic protists such as *Giardia lamblia* and *Trichomonas vaginalis*, but are not well characterized. From high-throughput sequence data of small RNAs (Illumina small RNA length 36 nt) we have uncovered different ranges of small RNAs attributed to the RNAi mechanism in these organisms. There are similarities and differences between *Giardia* and *Trichomonas* siRNAs which is likely to have a direct reflection on their RNAi-based protein structure. We will also discuss further work involving *Trichomonas* expression data and examination of how small RNAs from these protists may be interacting with their host dsRNA viruses.

Note: Both Jian Han and Lesley Collins presented.

S5.2 Abstract for IV International Giardia and Cryptosporidium Conference

ncRNAs and their evolution in *Giardia lamblia*

Lesley J. Collins[1], Jian Han[2] and David Penny[2]

[1.] Institute of Fundamental Sciences, Massey University, Palmerston North.

[2.] Institute of Molecular BioSciences, Massey University, Palmerston North,

Correspondence to: l.j.collins@massey.ac.nz

ncRNAs include regulatory RNAs such as miRNA and siRNA, but also processing RNAs such as RNase P, RNase MRP and snoRNAs. These ncRNAs do not exist on their own, but interact in complex RNA-protein networks that link transcription, translation, gene regulation and the cell cycle. We collectively call these networks the RNA-infrastructure. Over the last few years we have used high-throughput sequencing of some protists to aid our understanding of how RNAs interact within the cell, and how ncRNAs and their networks evolve throughout eukaryotes. Here we present findings from our work with ncRNAs from the Diplomonad *Giardia lamblia*.

We have used RNA sequencing to investigate some very different classes of ncRNA from *Giardia*. Our study of snoRNAs (C-D box and H/ACA classes), RNase P and RNase MRP, have highlighted defining structural features to aid further classification. We confirm that in *Giardia* we see a peak of sequence lengths 25-27nt believed to be miRNAs (compared to 21-22nt in humans), but we also find a peak of RNAs believed to be siRNAs in the 15-18nt range. These ultra-small RNAs have been previously found in humans, but *Giardia* usRNAs have different features most likely associated to the different domain structure of *Giardia* RNAi proteins Dicer and Argonaute. This leads to interesting questions on how the RNAi system in *Giardia* may have its own unique characteristics.

Overall, we see remarkable differences in the lengths and structure of ncRNAs from *Giardia* to those displayed in 'model' eukaryotes. We also demonstrate that high-throughput sequencing is a valid option for protist ncRNA identification on a genomic scale.

Note: L. Collins presented on behalf of all authors.

Final words

This project is an in-depth study on eukaryotic signature proteins (ESPs), with a focus on *Giardia lamblia*, a single celled eukaryotic anaerobic parasite that causes intestinal disease throughout the world. New Zealand has a higher incidence rate of giardiasis than other developed countries with the annual rate of 44.1 notified cases per 100,000 population. *Giardia* is a parasite that is rather unlike other eukaryotes, which is why it has been difficult to treat infection effectively. New drugs are needed to treat giardiasis, but in order to effectively develop them, an understanding of its metabolism and how *Giardia* has evolved in a very different way to the host is necessary. Thus analysing the phylogeny and metabolism of the organism became very important topics during my work. ESP datasets can guide phylogenetic or metabolic studies and aid our way on the long path towards discovering potential drug targets.

The ESPs datasets for three parasites, *Giardia lamblia*, *Trichomonas vaginalis* and *Plasmodium falciparum*, as well as their host human have all been re-calculated. These new datasets are significantly advanced from previously calculated datasets due to the use of better quality and larger numbers of genomes for comparison. The definition of ESP however is not black or white, as there are issues with distant non-recognisable homologues and possible convergent evolution, and there will be some "fringe" ESPs (e.g. ESPs that do have some very distant homologues in prokaryotes but this detection falls below the cut-offs set in this study). With the new datasets I aimed to find the most precise list of ESPs that delineates eukaryotes from prokaryotes (archaea and bacteria). Through the calculation of the ESP datasets for *Giardia* and other organisms, a detailed protocol was developed. This is important as future databases will be updated and novel organisms' genomes will become available, and these will consolidate and further improve the list of ESPs.

The ESP calculations have already laid the groundwork for other studies. ESPs are an important resource to calculate and clarify ancestral proteins which are important differences between eukaryotes and prokaryotes. There is a potential for drug targeting for prokaryotes using this dataset as the enzymes here delineate metabolism that is different between the two groups (eukaryotes and prokaryotes). The ESP database is still currently being used to look at how ancient proteins and ancient RNAs interact in

networks (Chapter 5 and part of a larger project). Furthermore, the ESP data is also currently being used in a Marsden-funded research project looking at ancestral metabolism.

During my study the differences between the ESP datasets from host and parasites were briefly analysed in a phylogenetic and metabolic manner, where ESPs from the human genome were compared with ESPs from parasites. Gene loss appears to dominate parasitic evolution (Smid *et al.* 2008), and some ESPs were missing from parasites, as indicated by GO annotations. Examples of ESPs likely to be missing are those that are involved in mRNA processing, RNA splicing, transmembrane transport and regulation of transcription. These pathways can hold potential for future drug treatment development, since in order for the parasites to deal with the loss, they must have alternative pathways. Protein interaction data is much needed and better annotations of *Giardia* genome is also required to identify these alternative pathways. Indeed these alternative pathways may be hidden in the mass of "hypothetical proteins" currently annotated. In-depth comparisons between the host and parasites' ESPs can only be complete when the annotation for the parasites becomes more complete, GO terms for parasite proteins and more protein interaction data becomes available.

One interesting aspect of ESPs as a set of proteins is their utility in phylogenetic analysis. There is a long evolutionary distance between different supergroups of eukaryotes, and using ESPs shortens this distance due to their slower rate of evolution. Chapter 3 has demonstrated the use of ESPs to analyse the phylogenetic relationships of eukaryotes. Two methods, consensus network and concatenating sequences, were employed to deal with the large number of discreet protein sequences, with the concatenation method being the more useful approach. The Unikonta eukaryotes formed clades with convincing bootstrap values. Although the two basal species *Giardia* and *Dictyostelium* formed long branches, they largely maintained their positions outside the main eukaryotic groups as expected. The mammalian phylogenetic relationship showed very similar results from published molecular and fossil results, to indicate that ESPs are capable of being good candidates for phylogenetic analysis. In future, ESPs could be used to attack more complex cutting-edge problems such as truly analysing deep phylogenies in detail with many other protists including more members of the controversial supergroups Excavata or Chromalveolata. More completed genomes of the

organisms in these supergroups, however, are still needed to break the long branches the taxa in these groups form at present. Overall, although it was not possible to conduct in-depth phylogenetic analysis with ESPs due to time constraints, the work done here demonstrate that ESPs as a set of proteins could be very useful in future phylogenetic projects.

Basal eukaryotic metabolism is often not well studied and the enzymes are often poorly annotated due to an overall lack of funding, and the fact that annotation of protists is much harder because they are so different from the other eukaryotes we know more about. Chapter 4 has investigated three key sugar metabolism pathways from *Giardia lamblia* by comparing its enzyme sequences to those in the widely used KEGG database. The analysis showed that the glycolysis pathway is present as expected but not the reverse gluconeogenesis pathway. The TCA cycle and the oxidative phosphorylation pathways only have a few enzymes represented in *Giardia*, which are likely to be part of other pathways. ESPs were expected to be present in some of these sugar metabolic pathways because *Giardia* is a eukaryote, and the eukaryotic metabolism is different from that of prokaryotes. However, the results suggest that this was not the case. *Giardia* has a few glycolytic enzymes that are conserved in eukaryotes, but because these enzymes also have clear prokaryotic (mostly bacterial) homologues, they could not be considered ESPs under our definition. *Giardia* has a unique metabolism that has been described as prokaryote-like. Given that many of the enzymes investigated here showed more similarity to prokaryotic enzymes than eukaryotic ones, this description is well earned. This raises questions as to whether these key enzymes hold ancient features in common with prokaryotes (i.e. divergent evolution) or reductive evolution has driven these enzymes to mimic prokaryotic enzymes (i.e. convergent evolution). With such a large evolutionary distance involved it is very hard to decipher which could be more likely. Whatever their origin, these prokaryote-like enzymes may be of important interest as new drug targets due to their dissimilarity with the equivalent host enzyme.

Genomics and proteomics is how drugs are developed nowadays. However, the emphasis in *Giardia* research is on diagnosis and treatment, and not metabolism or proteomic studies. With little money being invested, complete annotation of the *Giardia* genome is likely to be many years away. Until many more enzymatic assay results

become available, annotation is at present the only way to infer function. Chapter 4 has developed a way to deal with poor annotations of parasitic organisms to infer metabolic function. This procedure simply used BLAST to find homology and KEGG as a way to group this homology together. It went further than comparison single proteins and looked at proteins in a group as they would be found in a metabolic pathway. This method resulted in a putative 'map' of a pathway, indicating which enzymes are present or absent, and which ones are different from the host (i.e. prokaryote-like). These maps can only be considered the "best estimate" of a pathway at present, but they do give us a handle on what is actually there in the absence of any hard proteomic assay data. The procedure developed in this study can be used to analyse other complicated pathways (such as amino acid pathways) of basal eukaryotes in future, and could be used to aid in designing which metabolic assays should be performed in future lab work when funding becomes available.

Connecting ESPs and ncRNAs is important because therapeutic applications of ncRNAs has been well documented (e.g. (Zender *et al.* 2003)). The RNA analysis of *Giardia* and *Trichomonas* has yielded interesting results. Through analysing the length distribution of the mapped *Giardia* and *Trichomonas* small RNAs, two distinctive length types of RNAs have been discovered for *Giardia*: the "ultra small peak" of 15-18nt and a "larger peak" of 26-27nt. The "larger peak" matches that the length of reported *Giardia* Dicer product (MacRae *et al.* 2006) and we do expect that Dicer and perhaps the putative Argonaute protein are involved in the *Giardia* RNAi mechanism. Interestingly only the "ultra-small peak" was present in the *Trichomonas* data. The characteristics of small RNAs of this peak are still to be determined, but they could be potential siRNAs or cleavage products from an as yet unknown mechanism. After the analysis of potential target genes for the "larger peak", it appears there is not a significant correlation of these genes with ESPs. Also the predicted targets Gims (potential miRNAs discovered by Chen *et al.* (Chen *et al.* 2009)) also showed no significant correlation with ESPs. These results indicate that although both ESPs and RNAi are ancient and eukaryote specific, either RNAi does not necessarily play a big role in regulation of ESPs, or the mechanism and genes have evolved so that any ancient role is no longer prominent (i.e. the ESP genes no longer stand out from the rest). We do expect ancestral regulation mechanisms to have had a large effect on ancestral proteins. If there is indeed some correlation, then perhaps in future therapies targeting ncRNA regulation of essential

proteins can be designed. So far, any detailed connection between ESPs and ncRNA has not been found, but more study could either find connections, or unveil why connections do not appear between the two.

For future directions, relating the intron splicing mechanism with ESPs may also uncover interesting results. Spliceosomal introns have been demonstrated in both *Giardia* (Nixon *et al.* 2002) and *Trichomonas* (Vanacova *et al.* 2005), and it appears that a RNA splice-site motif shared by these two organisms is also found in yeast and metazoan introns. It is clear that the splicing mechanism (the spliceosome) is present in a common ancestor of *Giardia*, *Trichomonas*, yeast, and metazoans (Lynch *et al.* 2002; Collins *et al.* 2005; Vanacova *et al.* 2005). In addition, it appears that the ancestral mechanism have maintained most of the key components (the small nuclear ribonucleoproteins) across all crown eukaryotes (Collins *et al.* 2005). Any interesting pattern in the splicing of ESPs or analysis of ESPs which are involved in splicing could harvest meaningful insights to the nature of these ancestral proteins.

Throughout my study data management has been crucial. This is because a large amount of information had to be effectively stored in an interactive platform. MySQL has proven to be a very effective data managing system during the project, as data can be easily stored and retrieved from tables and databases. Also the use of specific Perl code allows Perl scripts to communicate with the MySQL databases and perform designated tasks, so that large amounts of data can be processed automatically. Perl programming was also used extensively during the project, because it enables the simple but effective manipulation of data. The Perl scripts used here have been included in supplementary material in each chapter, so that the developed protocols can be used in future research. Using these scripts, updated ESP datasets can be readily re-calculated, and new databases can be constructed using MySQL, which can incorporate new genomes or updated ones when they become available. The database I constructed during my study is presently being used by researchers at Massey looking at ancestral metabolism in eukaryotes. Several manuscripts are currently being prepared for the new ESP datasets and on the RNAi pathway. Also the ESP datasets will be publically available, most likely on GiardiaDB. As co-discoverer of the usRNAs in *Giardia*, I am also working on their characterisation, and their inclusion in the EuPathDB protist databases.

To conclude, ESPs have shown good potential to be good candidates for phylogenetic analysis, for which complex phylogenetic problems can be analysed using the ESP approach in future. *Giardia* metabolic pathways appear to be more similar to those of prokaryotes in places, which means that although some enzymes are essential and are part of a pathway ancestral to eukaryotes, their prokaryotic similarities mean that they cannot be designated as ESPs. Where host and parasites differ in terms of ESPs, are the ones present in less studied pathways (such as RNA splicing). Understanding the parasitic version of these pathways (or their alternative pathways), can guide the way to potential drug targets. The same ESP may be involved in totally different pathways in human and in parasites. So how does a highly conserved protein gain, lose or change functions? How does the domain change affect this? Future studies could be focused on answering these questions and thereby provide more knowledge about these parasites. It is possible that with more accurate ESP datasets, that they could act as even better guides to pinpoint enzymes that could be further analysed. ESPs are essentially the modern equivalents of ancient proteins and not only do they hold clues about our past, but can aid our proteomics push towards drug discovery of the future.

References

Abascal, F., R. Zardoya and D. Posada (2005). "ProtTest: selection of best-fit models of protein evolution." Bioinformatics **21**(9): 2104-2105.

Abouheif, E., R. Zardoya and A. Meyer (1998). "limitations of metazoan 18S rRNA sequence data: implications for reconstructing a phylogeny of the animal kindom and inferring the reality of the cambrian explosion." Journal of Melocular Evolution **47**(4): 394-405.

Adachi, J. and M. Hasegawa (1996). "Model of amino acid substitution in proteins encoded by mitochondrial DNA." Journal of Molecular Evolution **42**(4): 459-468.

Adam, R. D. (2001). "Biology of Giardia lamblia." Clinical Microbiology Reviews **14**(3): 447-475.

Adams, K. L. and J. D. Palmer (2003). "Evolution of mitochondrial gene content: gene loss and transfer to the nucleus." Molecular Phylogenetics and Evolution **29**(3): 380-395.

Ahlquist, P. (2002). "RNA-dependent RNA polymerases, viruses, and RNA silencing." Science **296**(5571): 1270-1273.

Akaike, H. (1974). "New look at statistical-model indentification." Ieee Transactions on Automatic Control **AC19**(6): 716-723.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." Journal of Molecular Biology **215**(3): 403-410.

Ambros, V., B. Bartel, D. P. Bartel, C. B. Burge, J. C. Carrington, X. M. Chen, G. Dreyfuss, S. R. Eddy, S. Griffiths-Jones, M. Marshall, M. Matzke, G. Ruvkun and T. Tuschl (2003). "A uniform system for microRNA annotation." Rna-a Publication of the Rna Society **9**(3): 277-279.

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

Andersson, S. G. E., A. Zomorodipour, J. O. Andersson, T. Sicheritz-Ponten, U. C. M. Alsmark, R. M. Podowski, A. K. Naslund, A. S. Eriksson, H. H. Winkler and C. G. Kurland (1998). "The genome sequence of Rickettsia prowazekii and the origin of mitochondria." Nature **396**(6707): 133-140.

Asher, R. J., N. Bennett and T. Lehmann (2009). "The new framework for understanding placental mammal evolution." Bioessays **31**(8): 853-864.

Aurrecoechea, C., J. Brestelli, B. P. Brunk, J. M. Carlton, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, H. G. Morrison, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, C. J. Stoeckert, Jr., S. Sullivan, C. Treatman and H. Wang (2009). "GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens Giardia lamblia and Trichomonas vaginalis." Nucleic Acids Research **37**: 526-530.

Aurrecoechea, C., J. Brestelli, B. P. Brunk, J. Dommer, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, J. A. Miller, V. Nayak, C. Pennington, D. F. Pinney, D. S. Roos, C. Ross, C. J. Stoeckert, Jr., C. Treatman and H. Wang (2009). "PlasmoDB: a functional genomic database for malaria parasites." Nucleic Acids Research **37**: 539-543.

Baldauf, S. L. (2003). "The deep roots of eukaryotes." Science **300**(5626): 1703-1706.

Bartel, D. P. (2009). "MicroRNAs: Target Recognition and Regulatory Functions." Cell **136**(2): 215-233.

Baulcombe, D. C. (2007). "Amplified silencing." Science **315**(5809): 199-200.

Birky, C. W. (2005). "Sex: Is Giardia doing it in the dark?" Current Biology **15**(2): 56-58.

Bouckaert, R. R. (2010). "DensiTree: making sense of sets of phylogenetic trees." Bioinformatics **26**(10): 1372-1373.

Brinkmann, H. and H. Philippe (2007). The diversity of eukaryotes and the root of the eukaryotic tree. Eukaryotic Membranes and Cytoskeleton: Origins and Evolution. **607:** 20-37.

Brown, D. M., J. A. Upcroft, M. R. Edwards and P. Upcroft (1998). "Anaerobic bacterial metabolism in the ancient eukaryote Giardia duodenalis." International Journal for Parasitology **28**(1): 149-164.

Brown, J. R. and W. F. Doolittle (1997). "Archaea and the prokaryote-to-eukaryote transition." Microbiology and Molecular Biology Reviews **61**(4): 456-502.

Byington, C. L., R. L. Dunbrack, F. G. Whitby, F. E. Cohen and N. Agabian (1997). "Entamoeba histolytica: Computer-assisted modeling of phosphofructokinase for the prediction of broad-spectrum antiparasitic agents." Experimental Parasitology **87**(3): 194-202.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. L. Madden (2009). "BLAST plus : architecture and applications." Bmc Bioinformatics **10**: Article number: 421.

Campbell, V. and F.-J. Lapointe "Retrieving a mitogenomic mammal tree using composite taxa." Molecular Phylogenetics and Evolution **58**(2): 149-156.

Carlton, J. M., R. P. Hirt, J. C. Silva, A. L. Delcher, M. Schatz, Q. Zhao, J. R. Wortman, S. L. Bidwell, U. C. M. Alsmark, S. Besteiro, T. Sicheritz-Ponten, C. J. Noel, J. B. Dacks, P. G. Foster, C. Simillion, Y. Van de Peer, D. Miranda-Saavedra, G. J. Barton, G. D. Westrop, S. Muller, D. Dessi, P. L. Fiori, Q. H. Ren, I. Paulsen, H. B. Zhang, F. D. Bastida-Corcuera, A. Simoes-Barbosa, M. T. Brown, R. D. Hayes, M. Mukherjee, C. Y. Okumura, R. Schneider, A. J. Smith, S. Vanacova, M. Villalvazo, B. J. Haas, M. Pertea, T. V. Feldblyum, T. R. Utterback, C. L. Shu, K. Osoegawa, P. J. de Jong, I. Hrdy, L. Horvathova, Z. Zubacova, P. Dolezal, S. B. Malik, J. M. Logsdon, K. Henze, A. Gupta, C. C. Wang, R. L. Dunne, J. A. Upcroft, P. Upcroft, O. White, S. L. Salzberg, P. Tang, C. H. Chiu, Y. S. Lee, T. M. Embley, G. H. Coombs, J. C. Mottram, J. Tachezy, C. M. Fraser-Liggett and P. J. Johnson (2007). "Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis." Science **315**(5809): 207-212.

Carrington, J. C. and V. Ambros (2003). "Role of microRNAs in plant and animal development." Science **301**(5631): 336-338.

Cavalier-Smith, T. (1987). "Eukaryotes with no mitochondria." Nature **326**(6111): 332-333.

Cavalier-Smith, T. (2002). "The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa." International Journal of Systematic and Evolutionary Microbiology **52**: 297-354.

Chen, X. S., L. J. Collins, P. J. Biggs and D. Penny (paper in preparation). "High throughput genome-wide survey of microRNAs from deep-branching eukaryotes."

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

Chen, X. S., T. S. Rozhdestvensky, L. J. Collins, J. Schmitz and D. Penny (2007). "Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*." Nucleic Acids Research **35**(14): 4619-4628.

Chen, X. S., W. T. White, L. J. Collins and D. Penny (2008). "Computational iIdentification of four splicesomal snRNAs from the deep-branching eukaryote *Giardia intestinalis*." PLos ONE **3**(8): e2106.

Chen, X. W., L. J. Collins, P. J. Biggs and D. Penny (2009). "High Throughput Genome-Wide Survey of Small RNAs from the Parasitic Protists Giardia intestinalis and Trichomonas vaginalis." Genome Biology and Evolution **1**: 165-175.

Collins, L. (2011). RNA Infrastructure and Networks. In

Collins, L. and D. Penny (2005). "Complex spliceosomal organization ancestral to extant eukaryotes." Molecular Biology and Evolution **22**(4): 1053-1066.

Collins, L. J. and X. S. Chen (2009). "Ancestral RNA The RNA biology of the eukaryotic ancestor." Rna Biology **6**(5): 495-502.

Collins, L. J. and D. Penny (2009). "The RNA infrastructure: dark matter of the eukaryotic cell?" Trends in Genetics **25**(3): 120-128.

Collins, L. J., B. Schönfeld and X. C. Chen (2011). The Epigenetics of non-coding RNA. Handbook of Epigenetics - The New Molecular and Medical Genetics. T. Tollefsbol, Academic Press.

Cowman, A. F. and B. S. Crabb (2002). "The Plasmodium falciparum genome - a blueprint for erythrocyte invasion." Science **298**(5591): 126-128.

Cowman, A. F. and B. S. Crabb (2006). "Invasion of red blood cells by malaria parasites." Cell **124**(4): 755-766.

Crouch, A. A., W. K. Seow and Y. H. Thong (1986). "Effect of 23 chemotherapeutic-agents on the adherence and growth of Giardia lamblia in vitro." Transactions of the Royal Society of Tropical Medicine and Hygiene **80**(6): 893-896.

Dacks, J. B., L. A. M. Davis, A. M. Sjogren, J. O. Andersson, A. J. Roger and W. F. Doolittle (2003). "Evidence for Golgi bodies in proposed 'Golgi-lacking' lineages." Proceedings of the Royal Society of London Series B-Biological Sciences **270**: 168-171.

Dan, M. X. and C. C. Wang (2000). "Role of alcohol dehydrogenase E (ADHE) in the energy metabolism of Giardia lamblia." Molecular and Biochemical Parasitology **109**(1): 25-36.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt (1978). A model of Evolutionary Change in Proteins. In Atlas of protein sequence and structure Washington DC, **5**: 345-358

Desai, A. and T. J. Mitchison (1998). "Tubulin and FtsZ structures: functional and therapeutic implications." Bioessays **20**(7): 523-527.

Dolezal, P., O. Smid, P. Rada, Z. Zubacova, D. Bursac, R. Sutak, J. Nebesarova, T. Lithgow and J. Tachezy (2005). "Giardia mitosomes and trichomonad hydrogenosomes share a common mode of protein targeting." Proceedings of the National Academy of Sciences of the United States of America **102**(31): 10924-10929.

Drinnenberg, I. A., G. R. Fink and D. P. Bartel (2011). "Compatibility with Killer Explains the Rise of RNAi-Deficient Fungi." Science **333**(6049): 1592-1592.

Dunn, L. A., A. G. Burgess, K. G. Krauer, L. Eckmann, P. Vanelle, M. D. Crozet, F. D. Gillin, P. Upcroft and J. A. Upcroft (2010). "A new-generation 5-nitroimidazole

can induce highly metronidazole-resistant Giardia lamblia in vitro." International Journal of Antimicrobial Agents **36**(1): 37-42.

Emelyanov, V. and A. Goldberg (2011). "Fermentation enzymes of Giardia lamblia, pyruvate:ferredoxin oxidoreductase and hydrogenase, do not localize to its mitosomes." Microbiology **157**: 1602-1611.

Euzeby, J. P. (1997). "List of bacterial names with standing in nomenclature: A folder available on the Internet." International Journal of Systematic Bacteriology **47**(2): 590-592.

Fanini, F., I. Vannini, D. Amadori and M. Fabbri (2011). "Clinical Implications of MicroRNAs in Lung Cancer." Seminars in Oncology **38**(6): 776-780.

Fedorov, A. and H. Hartman (2004). "What does the microsporidian E-cuniculi tell us about the origin of the eukaryotic cell?" Journal of Molecular Evolution **59**(5): 695-702.

Felsenstein, J. (1978). "Cases in which parsimony or compatibility methods will be positively misleading." Systematic Zoology **27**(4): 401-410.

Ferracin, M., P. Querzoli, G. A. Calin and M. Negrini (2011). "MicroRNAs: Toward the Clinic for Breast Cancer Patients." Seminars in Oncology **38**(6): 764-775.

Fritz-Laylin, L. K., S. E. Prochnik, M. L. Ginger, J. B. Dacks, M. L. Carpenter, M. C. Field, A. Kuo, A. Paredez, J. Chapman, J. Pham, S. Shu, R. Neupane, M. Cipriano, J. Mancuso, H. Tu, A. Salamov, E. Lindquist, H. Shapiro, S. Lucas, I. V. Grigoriev, W. Z. Cande, C. Fulton, D. S. Rokhsar and S. C. Dawson (2010). "The Genome of Naegleria gruberi Illuminates Early Eukaryotic Versatility." Cell **140**(5): 631-642.

Fuchsman, C. A. and G. Rocap (2006). "Whole-genome reciprocal BLAST analysis reveals that Planctomycetes do not share an unusually large number of genes with Eukarya and Archaea." Applied and Environmental Microbiology **72**(10): 6841-6844.

Gardner, M. J., N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. A. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall, J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser and B. Barrell (2002). "Genome sequence of the human malaria parasite Plasmodium falciparum." Nature **419**(6906): 498-511.

Gillin, F. D., D. S. Reiner and J. M. McCaffery (1996). "Cell biology of the primitive eukaryote Giardia lamblia." Annual Review of Microbiology **50**: 679-705.

Glansdorff, N., Y. Xu and B. Labedan (2008). "The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner." Biology Direct **3**: Article number: 29.

Graham, D. E., R. Overbeek, G. J. Olsen and C. R. Woese (2000). "An archaeal genomic signature." Proceedings of the National Academy of Sciences of the United States of America **97**(7): 3304-3308.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Systematic Biology **52**(5): 696-704.

Hamilton, A., O. Voinnet, L. Chappell and D. Baulcombe (2002). "Two classes of short interfering RNA in RNA silencing." Embo Journal **21**(17): 4671-4679.

Hampl, V., L. Hug, J. W. Leigh, J. B. Dacks, B. F. Lang, A. G. B. Simpson and A. J. Roger (2009). "Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups"." Proceedings of the National Academy of Sciences of the United States of America **106**(10): 3859-3864.

Harp, D. F. and I. Chowdhury (2011). "Trichomoniasis: evaluation to execution." European Journal of Obstetrics & Gynecology and Reproductive Biology **157**(1): 3-9.

Harris, J. C., S. Plummer and D. Lloyd (2001). "Antigiardial drugs." Applied Microbiology and Biotechnology **57**(5-6): 614-619.

Hartman, H. and A. Fedorov (2002). "The origin of the eukaryotic cell: A genomic investigation." Proceedings of the National Academy of Sciences of the United States of America **99**(3): 1420-1425.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto and M. Hasegawa (1994). "protein phylogeny gives a robust estimation for early divergences of eukaryotes - phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia." Molecular Biology and Evolution **11**(1): 65-71.

He, L. and G. J. Hannon (2004). "MicroRNAs: Small RNAs with a big role in gene regulation." Nature Reviews Genetics **5**(8): 522-531.

Hendy, M. D. and D. Penny (1989). "A framework for the quantitative study of evolutionary trees." Systematic Zoology **38**(4): 297-309.

Henikoff, S. and J. G. Henikoff (1992). "Amino-Acid Substitution Matrices from Protein Blocks." Proceedings of the National Academy of Sciences of the United States of America **89**(22): 10915-10919.

Higgins, D. G. and P. M. Sharp (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer." Gene **73**(1): 237-244.

Hilario, E. and J. P. Gogarten (1998). "The prokaryote-to-eukaryote transition reflected in the evolution of the V/F/A-ATPase catalytic and proteolipid subunits." Journal of Molecular Evolution **46**(6): 703-715.

Holder, M. and P. O. Lewis (2003). "Phylogeny estimation: Traditional and Bayesian approaches." Nature Reviews Genetics **4**(4): 275-284.

Holland, B. and V. Moulton (2003). "Consensus networks: A method for visualising incompatibilities in collections of trees." Algorithms in Bioinformatics, Proceedings **2812**: 165-176.

Hoyne, G. F., P. F. L. Boreham, P. G. Parsons, C. Ward and B. Biggs (1989). "The effect of drugs on the cell-cycle of Giardia-intestinalis." Parasitology **99**: 333-339.

Huang, D. B. and A. C. White (2006). "An updated review on Cryptosporidium and Giardia." Gastroenterology Clinics of North America **35**(2): 291-314.

Huber, H., M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer and K. O. Stetter (2002). "A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont." Nature **417**(6884): 63-67.

Huelsenbeck, J. P. and F. Ronquist (2001). "MRBAYES: Bayesian inference of phylogenetic trees." Bioinformatics **17**(8): 754-755.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen and J. P. Bollback (2001). "Evolution - Bayesian inference of phylogeny and its impact on evolutionary biology." Science **294**(5550): 2310-2314.

Huson, D. H. (1998). "SplitsTree: analyzing and visualizing evolutionary data." Bioinformatics **14**(1): 68-73.

Huson, D. H. and D. Bryant (2006). "Application of phylogenetic networks in evolutionary studies." Molecular Biology and Evolution **23**(2): 254-267.

Huson, D. H., T. Dezulian, T. Klopper and M. A. Steel (2004). "Phylogenetic super-networks from partial trees." Ieee-Acm Transactions on Computational Biology and Bioinformatics **1**(4): 151-158.

Jedelsky, P. L., P. Dolezal, P. Rada, J. Pyrih, O. Smid, I. Hrdy, M. Sedinova, M. Marcincikova, L. Voleman, A. J. Perry, N. C. Beltran, T. Lithgow and J. Tachezy (2011). "The Minimal Proteome in the Reduced Mitochondrion of the Parasitic Protist Giardia intestinalis." PLos ONE **6**(2): Article No.: e17285.

Jenkins, C., R. Samudrala, I. Anderson, B. P. Hedlund, G. Petroni, N. Michailova, N. Pinel, R. Overbeek, G. Rosati and J. T. Staley (2002). "Genes for the cytoskeletal protein tubulin in the bacterial genus Prosthecobacter." Proceedings of the National Academy of Sciences of the United States of America **99**(26): 17049-17054.

Jimenez, M., A. Martos, M. Vicente and G. Rivas (2011). "Reconstitution and Organization of Escherichia coli Proto-ring Elements (FtsZ and FtsA) inside Giant Unilamellar Vesicles Obtained from Bacterial Inner Membranes." Journal of Biological Chemistry **286**(13): 11236-11241.

Jones, D. T., W. R. Taylor and J. M. Thornton (1992). "The rapid generation of mutation data matrices from protein sequences." Computer Applications in the Biosciences **8**(3): 275-282.

Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki and M. Hirakawa (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Research **34**: D354-D357.

Keeling, P. J. (2007). "Deep questions in the tree of life." Science **317**(5846): 1875-1876.

Keeling, P. J., G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger and M. W. Gray (2005). "The tree of eukaryotes." Trends in Ecology & Evolution **20**(12): 670-676.

Kielan-Jaworowska, Z. (2007). "The beginning of the age of mammals." Nature **446**(7133): 264-265.

Kinsella, R. J., A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey and P. Flicek (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space." Database : the journal of biological databases and curation **2011**: bar030.

Knoll, A. H., E. J. Javaux, D. Hewitt and P. Cohen (2006). "Eukaryotic organisms in Proterozoic oceans." Philosophical Transactions of the Royal Society B-Biological Sciences **361**(1470): 1023-1038.

Korf, I., M. Yandell and J. Bedell (2003). BLAST. In

Kurland, C. G., L. J. Collins and D. Penny (2006). "Genomics and the irreducible nature of eukaryote cells." Science **312**(5776): 1011-1014.

Lacey, E. (1988). "The role of the cytoskeletal protein, tubulin, in the mode of action and mechanism of drug-resistance to benzimidazoles." International Journal for Parasitology **18**(7): 885-936.

Ladeira, R. B., M. A. R. Freitas, E. F. Silva, N. F. Gontijo and M. A. Gomes (2005). "Glycogen as a carbohydrate energy reserve in trophozoites of Giardia lamblia." Parasitology Research **96**(6): 418-421.

Lake, J. A. and M. C. Rivera (1994). "Was the nucleus the 1st endosymbiont." Proceedings of the National Academy of Sciences of the United States of America **91**(8): 2880-2881.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biology **10**(3).

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins (2007). "Clustal W and clustal X version 2.0." Bioinformatics **23**: 2947-2948.

Lee, K., J. Kim, M. Jung, T. Arii, J. Ryu, S. Han, K. E. Lee, J. H. Kim, M. K. Jung, J. S. Ryu and S. S. Han (2009). "Three-dimensional structure of the cytoskeleton in Trichomonas vaginalis revealed new features." Journal of Electron Microscopy **58**(5): 305-313.

Letunic, I. and P. Bork (2007). "Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation." Bioinformatics **23**(1): 127-128.

Li, Z., S. W. Kim, Y. Lin, P. S. Moore, Y. Chang and B. John (2009). "Characterization of Viral and Human RNAs Smaller than Canonical MicroRNAs." Journal of Virology **83**(24): 12751-12758.

Lo, M., M. Reid, M. Brokenshire and M. Lo (2002). "Resistance of Trichomonas vaginalis infections to metronidazole in Auckland sexual health clinics: report of two cases." New Zealand Medical Journal **115**(1160): 147.

Lockhart, P. and M. Steel (2005). "A tale of two processes." Systematic Biology **54**(6): 948-951.

Logsdon, J. M. (2008). "Evolutionary genetics: Sex happens in Giardia." Current Biology **18**(2): 66-68.

Luo, J., M. Teng, G.-P. Zhang, Z.-R. Lun, H. Zhou and L.-H. Qu (2009). "Evaluating the evolution of G-lamblia based on the small nucleolar RNAs identified from Archaea and unicellular eukaryotes." Parasitology Research **104**(6): 1543-1546.

Lynch, M. and A. O. Richardson (2002). "The evolution of spliceosomal introns." Current Opinion in Genetics & Development **12**(6): 701-710.

MacRae, I. J., K. H. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams and J. A. Doudna (2006). "Structural basis for double-stranded RNA processing by dicer." Science **311**(5758): 195-198.

Marquez, S. M., J. K. Harris, S. T. Kelley, J. W. Brown, S. C. Dawson, E. C. Roberts and N. R. Pace (2005). "Structural implications of novel diversity in eucaryal RNase P RNA." Rna-a Publication of the Rna Society **11**(5): 739-751.

Matschinsky, F. M. (2009). "Assessing the potential of glucokinase activators in diabetes therapy." Nature Reviews Drug Discovery **8**(5): 399-416.

Meyer, A., C. Todt, N. T. Mikkelsen and B. Lieb (2010). "Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity." Bmc Evolutionary Biology **10**: Article No.: 70.

Mitra, S., J. Cui, P. W. Robbins and J. Samuelson (2009). "A deeply divergent phosphoglucomutase (PGM) of Giardia lamblia has both PGM and phosphomannomutase activities." Glycobiology **20**(10): 1233-1240.

Morrison, H. G., A. G. McArthur, F. D. Gillin, S. B. Aley, R. D. Adam, G. J. Olsen, A. A. Best, W. Z. Cande, F. Chen, M. J. Cipriano, B. J. Davids, S. C. Dawson, H. G. Elmendorf, A. B. Hehl, M. E. Holder, S. M. Huse, U. U. Kim, E. Lasek-Nesselquist, G. Manning, A. Nigam, J. E. J. Nixon, D. Palm, N. E.

Passamaneck, A. Prabhu, C. I. Reich, D. S. Reiner, J. Samuelson, S. G. Svard and M. L. Sogin (2007). "Genomic minimalism in the early diverging intestinal parasite Giardia lamblia." Science **317**(5846): 1921-1926.

Nanda, N., R. G. Michel, G. Kurdgelashvili and K. A. Wendel (2006). "Trichomoniasis and its treatment." Expert review of anti-infective therapy **4**(1): 125-35.

Nash, T. E., H. T. Lujan, M. R. Mowatt and J. T. Conrad (2001). "Variant-specific surface protein switching in Giardia lamblia." Infection and Immunity **69**(3): 1922-1923.

Neves, S. R., P. T. Ram and R. Iyengar (2002). "G protein pathways." Science **296**(5573): 1636-1639.

Nixon, J. E. J., A. Wang, J. Field, H. G. Morrison, A. G. McArthur, M. L. Sogin, B. J. Loftus and J. Samuelson (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." Eukaryotic Cell **1**(2): 181-190.

Nixon, J. E. J., A. Wang, H. G. Morrison, A. G. McArthur, M. L. Sogin, B. J. Loftus and J. Samuelson (2002). "A spliceosomal intron in Giardia lamblia." Proceedings of the National Academy of Sciences of the United States of America **99**(6): 3701-3705.

Page, R. D. M. (2002). "Visualizing phylogenetic trees using TreeView." Curr Protoc Bioinformatics **Chapter 6**: Unit 6.2.

Paget, T. A., E. L. Jarroll, P. Manning, D. G. Lindmark and D. Lloyd (1989). "Respiration in the cysts and trophozoites of Giardia muris." Journal of General Microbiology **135**: 145-154.

Paget, T. A., M. L. Kelly, E. L. Jarroll, D. G. Lindmark and D. Lloyd (1993). "The effects of oxygen on fermentation in Giardia lamblia." Molecular and Biochemical Parasitology **57**(1): 65-72.

Parfrey, L. W., E. Barbero, E. Lasser, M. Dunthorn, D. Bhattacharya, D. J. Patterson and L. A. Katz (2006). "Evaluating support for the current classification of eukaryotic diversity." Plos Genetics **2**(12): 2062-2073.

Philippe, H. (2000). "Opinion: Long branch attraction and protist phylogeny." Protist **151**(4): 307-316.

Philippe, H. and A. Adoutte (1998). "The molecular phylogeny of Eukaryota: solid facts and uncertainties." Evolutionary Relationships among Protozoa **56**: 25-56.

Prasad, A. B., M. W. Allard, E. D. Green and N. C. S. Program (2008). "Confirming the phylogeny of mammals by use of large comparative sequence data sets." Molecular Biology and Evolution **25**(9): 1795-1808.

Quin, M. B. and C. Schmidt-Dannert (2011). "Engineering of Biocatalysts: from Evolution to Creation." Acs Catalysis **1**(9): 1017-1021.

Rannala, B. and Z. H. Yang (1996). "Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference." Journal of Molecular Evolution **43**(3): 304-311.

Rao, V. S. and K. Srinivas (2011). "Modern drug discovery process: An in silico approach." Journal of Bioinformatics and Sequence Analysis **2**(5): 89-94.

Rivera, M. C. and J. A. Lake (2004). "The ring of life provides evidence for a genome fusion origin of eukaryotes." Nature **431**(7005): 152-155.

Robertson, C. E., J. K. Harris, J. R. Spear and N. R. Pace (2005). "Phylogenetic diversity and ecology of environmental Archaea." Current Opinion in Microbiology **8**(6): 638-642.

Roger, A. J., S. G. Svard, J. Tovar, C. G. Clark, M. W. Smith, F. D. Gillin and M. L. Sogin (1998). "A mitochondrial-like chaperonin 60 gene in Giardia lamblia: Evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria." Proceedings of the National Academy of Sciences of the United States of America **95**(1): 229-234.

Romano, A. H. and T. Conway (1996). "Evolution of carbohydrate metabolic pathways." Research in Microbiology **147**(6-7): 448-455.

Royet, J., T. Bouwmeester and S. M. Cohen (1998). "Notchless encodes a novel WD40-repeat-containing protein that modulates Notch signaling activity." Embo Journal **17**(24): 7351-7360.

Sanchez, L. B. (1998). "Aldehyde dehydrogenase (CoA-acetylating) and the mechanism of ethanol formation in the amitochondriate protist, Giardia lamblia." Archives of Biochemistry and Biophysics **354**(1): 57-64.

Saraiya, A. A. and C. C. Wang (2008). "snoRNA, a Novel Precursor of microRNA in Giardia lamblia." PLoS Pathogens **4**(11): e1000224.

Seema, S., K. Arti, S. Sood and A. Kapil (2008). "An update on Trichomonas vaginalis." Indian Journal of Sexually Transmitted Diseases **29**(1): 7-14.

Sen, C. K. and S. Roy (2007). "MIRNA: Licensed to kill the messenger." DNA and Cell Biology **26**(4): 193-194.

Sgro, F., M. Gai, E. D. Luca and F. D. Cunto (2011). "Microtubule-dependent cytokinesis control by Citron kinase." Febs Journal **278**: 386-386.

Shih, Y.-L. and L. Rothfield (2006). "The bacterial cytoskeleton." Microbiology and Molecular Biology Reviews **70**(3): 729-754.

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

Simon, C., L. Nigro, J. Sullivan, K. Holsinger, A. Martin, A. Grapputo, A. Franke and C. McIntosh (1996). "Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes." Molecular Biology and Evolution **13**(7): 923-932.

Simpson, A. G. B. (2003). "Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota)." International Journal of Systematic and Evolutionary Microbiology **53**: 1759-1777.

Simpson, A. G. B., Y. Inagaki and A. J. Roger (2006). "Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes." Molecular Biology and Evolution **23**(3): 615-625.

Simpson, A. G. B. and D. J. Patterson (1999). "The ultrastructure of Carpediemonas membranifera (Eukaryota) with reference to the "Excavate hypothesis"." European Journal of Protistology **35**(4): 353-370.

Simpson, A. G. B., A. J. Roger, J. D. Silberman, D. D. Leipe, V. P. Edgcomb, L. S. Jermiin, D. J. Patterson and M. L. Sogin (2002). "Evolutionary history of "early-diverging" eukaryotes: The excavate taxon Carpediemonas is a close relative of Giardia." Molecular Biology and Evolution **19**(10): 1782-1791.

Smid, O., A. Matuskova, S. R. Harris, T. Kucera, M. Novotny, L. Horvathova, I. Hrdy, E. Kutejova, R. P. Hirt, T. M. Embley, J. Janata and J. Tachezy (2008). "Reductive Evolution of the Mitochondrial Processing Peptidases of the Unicellular Parasites Trichomonas vaginalis and Giardia intestinalis." PLoS Pathogens **4**(12): e1000243.

Snel, S. J., M. G. Baker and K. Venugopal (2009). "The epidemiology of giardiasis in New Zealand, 1997-2006." The New Zealand medical journal **122**(1290): 62-75.

Spitalny, P. and M. Thomm (2008). "A polymerase III-like reinitiation mechanism is operating in regulation of histone expression in archaea." Molecular Microbiology **67**(5): 958-970.

Staley, J. T., H. Bouzek and C. Jenkins (2005). "Eukaryotic signature proteins of Prosthecobacter dejongeii and Gemmata sp Wa-1 as revealed by in silico analysis." Fems Microbiology Letters **243**(1): 9-14.

Steenkamp, E. T., J. Wright and S. L. Baldauf (2006). "The protistan origins of animals and fungi." Molecular Biology and Evolution **23**(1): 93-106.

Strausfeld, N. J. and D. R. Andrew (2011). "A new view of insect-crustacean relationships I. Inferences from neural cladistics and comparative neuroanatomy." Arthropod Structure & Development **40**(3): 276-288.

Teodorovic, S., C. D. Walls and H. G. Elmendorf (2007). "Bidirectional transcription is an inherent feature of Giardia lamblia promoters and contributes to an abundance of sterile antisense transcripts throughout the genome." Nucleic Acids Research **35**(8): 2544-2553.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins (1997). "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools." Nucleic Acids Research **25**(24): 4876-4882.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Research **22**(22): 4673-4680.

Thompson, R. C. A., J. A. Reynoldson and A. H. W. Mendis (1993). "Giardia and giardiasis." Advances in Parasitology **32**: 71-160.

Ullu, E., H. D. Lujan and C. Tschudi (2005). "Small sense and antisense RNAs derived from a telomeric retroposon family in Giardia intestinalis." Eukaryotic Cell **4**(6): 1155-1157.

Unden, G. and J. Bongaerts (1997). "Alternative respiratory pathways of Escherichia coli: Energetics and transcriptional regulation in response to electron acceptors." Biochimica Et Biophysica Acta-Bioenergetics **1320**(3): 217-234.

Upcroft, P. and J. A. Upcroft (2001). "Drug targets and mechanisms of resistance in the anaerobic protozoa." Clinical Microbiology Reviews **14**(1): 150-164.

Valdez, C. A., J. C. Tripp, Y. Miyamoto, J. Kalisiak, P. Hruz, Y. S. Andersen, S. E. Brown, K. Kangas, L. V. Arzu, B. J. Davids, F. D. Gillin, J. A. Upcroft, P. Upcroft, V. V. Fokin, D. K. Smith, K. B. Sharpless and L. Eckmann (2009). "Synthesis and Electrochemistry of 2-Ethenyl and 2-Ethanyl Derivatives of 5-Nitroimidazole and Antimicrobial Activity against Giardia lamblia." Journal of Medicinal Chemistry **52**(13): 4038-4053.

Vale, R. D. (2003). "The molecular motor toolbox for intracellular transport." Cell **112**(4): 467-480.

van den Ent, F. and J. Lowe (2000). "Crystal structure of the cell division protein FtsA from Thermotoga maritima." Embo Journal **19**(20): 5300-5307.

Vanacova, S., D. R. Liston, J. Tachezy and P. J. Johnson (2003). "Molecular biology of the amitochondriate parasites, Giardia intestinalis, Entamoeba histolytica and Trichomonas vaginalis." International Journal for Parasitology **33**(3): 235-255.

Vanacova, S., W. H. Yan, J. M. Carlton and P. J. Johnson (2005). "Spliceosomal introns in the deep-branching eukaryote Trichomonas vaginalis." Proceedings of the

National Academy of Sciences of the United States of America **102**(12): 4430-4435.

Vasudevan, S., Y. C. Tong and J. A. Steitz (2007). "Switching from repression to activation: MicroRNAs can up-regulate translation." Science **318**(5858): 1931-1934.

Voet, D. and J. G. Voet (2004). Biochemistry. In

Watters, C. (2006). "The bacterial cytoskeleton." CBE life sciences education **5**(4): 306-10.

Whelan, S. and N. Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." Molecular Biology and Evolution **18**(5): 691-699.

White, N. J. (2004). "Antimalarial drug resistance." Journal of Clinical Investigation **113**(8): 1084-1092.

Williams, L., C. C. Carles, K. S. Osmont and J. C. Fletcher (2005). "A database analysis method identifies an endogenous trans-acting short-interfering RNA that targets the Arabidopsis ARF2, ARF3, and ARF4 genes." Proceedings of the National Academy of Sciences of the United States of America **102**(27): 9703-9708.

Woese, C. R., O. Kandler and M. L. Wheelis (1990). "Towards a natural system of organisms - proposal for the domains Archaea, Bacteria, and Eukarya." Proceedings of the National Academy of Sciences of the United States of America **87**(12): 4576-4579.

Yang, C. Y., H. Zhou, J. Luo and L. H. Qu (2005). "Identification of 20 snoRNA-like RNAs from the primitive eukaryote, Giardia lamblia." Biochemical and Biophysical Research Communications **328**(4): 1224-1231.

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

Zender, L., S. Hutker, C. Liedtke, H. L. Tillmann, S. Zender, B. Mundt, M. Waltemathe, T. Gosling, P. Flemming, N. P. Malek, C. Trautwein, M. P. Manns, F. Kuhnel and S. Kubicka (2003). "Caspase 8 small interfering RNA prevents acute liver failure in mice." Proceedings of the National Academy of Sciences of the United States of America **100**(13): 7797-7802.