Comparative approaches to genome evolution in *Blastocystis* and *Entamoeba*

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Ross Stuart Low

November 2017

Acknowledgements

Firstly, I have to thank my supervisors Andrew Jackson and Neil Hall for their support during the project. I am particularly grateful to Andrew for driving the project and supporting me not only during my time at Liverpool but with thought for the future as well.

I'd also like to thank the staff at the CGR for their role in the sequencing aspects of this thesis (which is quite a lot) and Richard Gregory for his help with installing and running the sizeable suite of programmes on the computing clusters. I probable still owe him some biscuits.

The bulk of this thesis would not have been possible without contribution from Graham Clark at the London School of Hygiene and Tropical Medicine, who provided *Proteromonas* cultures, information and advice whenever I needed them. I would also like to thank Anastasios Tsaousis (Tasos) at the University of Kent for his input on the cellular and mitochondrial metabolism. This project was funded by a grant from the Leverhulme Trust (RPG-2014-005).

I must also give special mention to Sara Silva Pereira who put up with my incessant questions in the office (even when we were no longer desk buddies) and to Catherine Hartley who put up with my incessant questions in the lab, both of whom are fonts of absolute knowledge.

My thanks go to everyone in IC2 who always had time for a cup of tea and a biscuit; it's good to know that caffeine addiction can be a social event. Special thanks go to all the Jitsuka in the region, but especially to those in Liverpool, for offering up their bodies to be living punching bags for hours at a time. You guys kept me sane.

Finally, I must give my wholehearted thanks to my family, particularly my parents whose support right from the beginning made this endeavour possible. My brother, who is always there to stop my head from getting too big and Rebecca who did more than she knew. Suffice to say that I owe her a huge debt and possibly a puppy.

Comparative approaches to genome evolution in *Blastocystis* and *Entamoeba*

Ross Stuart Low

Abstract

Parasitism has arisen independently in numerous lineages of eukaryotes. Investigating the origins of parasitism is a core question in evolutionary biology and allows identification of parasite-specific factors that aid in diagnosis and treatment. Comparative genomic studies have often been applied within clades of parasites, which allows their ancestral state to be imagined, but cannot elucidate the processes that surrounded the emergence of parasitism. This question must be approached by comparison with a free-living out-group, to reconstruct the ancestral non-parasitic state. In this thesis, I examine free-living relatives of two intestinal protists of global importance, *Blastocystis* sp. and *Entamoeba histolytica*, to explore their evolution.

A draft genome sequence for *Proteromonas lacertae*, the non-pathogenic sister-taxon of *Blastocystis*, is presented along with a transcriptome for *Cafeteria roenbergensis*, a free-living out-group to the *Blastocystis-Proteromonas* clade. Together with the published *Blastocystis* sp. genome sequences, the *P. lacertae* genome and the *C. roenbergensis* transcriptome were used in a comparative genomic analysis. This revealed that the *Blastocystis* genomes are genuinely small, compared to other Stramenopiles and that this reduction is genome-wide as well as with respect to specific cellular apparatus, such as the flagellum and other motility-associated genes, which have been totally lost from the ancestor of *Blastocystis*. Rather than observe the same loss of function from metabolic capability, this reduction was associated with loss of gene complexity and is indicative of genomic streamlining. This is coupled with gene family expansion of Ig-like domain-containing proteins, potentially bestowing adhesive qualities to the cell surface.

A transcriptome for *Mastigamoeba* sp., a free-living out-group to the *Entamoeba* genus, is also presented. The *Mastigamoeba* sp. transcriptome was used in a comparative analysis of the *E. histolytica* genome. This analysis revealed large-scale expansion of Ras-family proteins in the ancestor of *Entamoeba,* which may be linked to motility and phagocytosis required for pathogenesis. Analysis of cathepsins revealed processes of genomic reduction and expansion occurring within the same gene family indicating genomic streamlining and subsequent specialisation in the parasite.

I have shown how we might revisit crucial questions in evolutionary biology using the latest genome sequencing technology. By generating new genomic resources for free-living protists, this thesis exposes the mechanism by which two common intestinal parasites of humans and animals evolved. It makes substantial contribution to our understanding of the origins of parasite genomes, and of microbial biodiversity, while revealing numerous parasite-specific features that will sustain future research.

Table of contents

Table of figures

Table of tables

Abbreviations

AIC	Akaike information criterion
ALA	Amoebic liver abscesses
ATCC	American Type Culture Collection
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks substitution matrix of amino acid substitution
BUSCO	Benchmarking Universal Single-Copy Orthologues
CCD	Charge-coupled device
CCS	Circular consensus sequence
CDS	Coding sequence
CGR	Centre for Genomic Research
CHEF	Contour-clamped homogeneous electric field
CLR	Continuous long reads
CP	Cysteine protease
E-value	Exponent value (in NCBI BLAST search)
EF	Elongation factor
EGF	Epidermal growth factor
EhCP	*Entamoeba histolytica* cysteine protease
EM	Electron microscopy
ETC	Electron transport chain
F	Frequency correction
G	Gamma correction
GMM	Gaussian mixture model
GO	Gene ontology
GS	Genome size
HGAP	Hierarchical Genome Assembly Process
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
HSP	Heat shock protein
IBD	Inflammatory Bowel Disease
IBS	Irritable Bowel Syndrome
Ig	Immunoglobulin
IL	Interleukin
lnL	Log likelihood
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG ontology

LTR	Long terminal repeat
MFS	Major facilitator superfamily
MRO	Mitochondria related organelle
MSRL	Minimum seed read length
Nc	Codon usage
NCBI	National Centre for Biotechnology Information
NGS	Next generation sequencing
ORF	Open reading frame
P-value	Probability value
PCA	Principal Components Analysis
PKD	Polycystic kidney disease
RPKM	Reads per kilobase per million
SBS	Sequencing by synthesis
SMRT	Single-Molecule Real-Time
SMS	Smart Model Selection
SSR	Sum of squared residues
SSU	Small subunits
ST	Subtype
TGS	Third generation sequencing
VSG	Variable surface glycoprotein
VT	Variable time model of amino acid substitution
WAG	Whelan and Goldman model of amino acid substitution
X^2	Chi squared
YE	Yeast extract
ZMO	Zero-mode waveguide

Chapter 1. Introduction

1.1 The vertebrate gut

Along with the respiratory and urogenital tracts, the vertebrate gut is one of three major mucosal surfaces that interact with the environment. It is anaerobic, eutrophic and maintains temperature within a narrow margin. Food particles move through the lumen propelled by polarised movement of structures known as villi in a process called peristalsis. The villi provide a large surface area for the efficient diffusion of nutrients across the epithelial membrane and into the densely packed blood vessels surrounding it. The epithelium is coated in mucus, a layer of secreted enzymes, mucins, immunoglobulins and inorganic salts produced by goblet cells in the crypts between the villi. This mucosal layer is a thick protective barrier that defends against invasion of the epithelium by any of the diverse microbes which inhabit the gut and co-evolved with it [1].

In recent years there have been attempts to understand and catalogue the diversity of bacterial, archaeal, viral and eukaryotic components of a healthy intestinal microbiome, such that we might better understand its far-reaching implications for the vertebrate gut. The microbiome has multiple roles in the normal function of the gut in aspects such as secretion of cellulases by *Clostridium, Citrobacter* and *Leptotrichia* in herbivorous vertebrates and secretion of proteases by *Halomonas* in carnivorous vertebrates, which aid the digestion of food [2], priming both the adaptive and innate immune systems and influencing the host metabolic capabilities [3]–[6]. The association between the host and the microbiome is dynamic and enduring, though the scale of this association has yet to fully appreciated. Because of the importance of understanding this relationship, the majority of studies have been aimed at mammalian microbiomes with special focus on humans in order to explore how this relationship might be exploited to improve health and to diagnose and prevent illness.

The microbiome plays important roles in allowing access to otherwise inaccessible nutrients such as within plant matter [7], [8], thereby allowing vitamin acquisition from food stuffs, and it contributes to normal immune development and function [3], [5], [6], [9], [10]. However, dysregulation of the intestinal flora

(dysbiosis) is associated with a number of health conditions including obesity [9], [11], malnutrition [12], inflammatory bowel disease [13], [14], neurological disorders [15] and cancer [16]. Some of the individual species that make up the microbiota may be opportunistic and, while incapable of causing disease themselves, they may be able to complicate or exacerbate existing diseases.

1.1.1 Bacteria in the vertebrate gut

The majority of the microbial flora, in terms of biomass and biodiversity, and the component that has been most studied, is bacterial. Only with the advent of sequencing technologies and metagenomic techniques has a concerted effort been made to determine a true measure of the bacterial diversity inhabiting the intestinal tract without the need to rely on culture methods, which are often inadequate at capturing the scale of diversity, especially of anaerobes [17], [18]. These sequence-based estimates have predicted around 200 prevalent bacterial species and over 1000 uncommon species [19]. Initial studies aimed to define a 'core' microbiome representing a subset of organisms ubiquitous across the human population. However, molecular techniques have shown that high variation is seen within human populations and even within the same individual sampled over time. This variation is heavily influenced by host factors such as age, gender, genetic factors and immune disorders as well as environmental ones such as diet, use of antibiotics, country of residence and proximity to animals. Therefore, a 'core' microbiome cannot be described using a species level description of the microbiome, however, this problem is solved at higher taxonomic levels, for example in adults, most species arise from a few phyla such as Bacteriodetes and Firmicutes, with Actinobacteria, Proteobacteria and Verrucomicrobia contributing smaller constituents [20]. Each of these components will share a particular relationship with the host, which may range between commensal, mutualistic, opportunistic and parasitic. Here, I will use the term 'symbiont' to define any organism in obligate association with a host, regardless of whether that relationship is beneficial or antagonistic.

Alongside the bacterial aspect of the microbiome there are data that suggest that there are also archaeal [21], eukaryotic and viral components. Typically, these have been less studied and are less diverse than their bacterial counterparts. Eukaryotic gut symbionts are mostly *Blastocystis* and fungi [22], [23] and their diversity is often much lower than bacterial taxa; fewer than 10 eukaryotic

phylotypes may be recovered per individual, though as sampling of more diverse human populations increases, this figure may rise. Nonetheless, these components add more dimensions to the already complex community of interactions present in the intestine by introducing further competition [24], [25], mutualism between symbionts [26] and also predation on bacteria [27].

1.1.2 Eukaryotes in the vertebrate microbiome

The majority of the studies into intestinal eukaryotes has centred around disease and they are generally perceived as pathogenic [23]. This has resulted in focused studies that often do not appreciate the complexity of the setting. Recently, eukaryotic metagenomics has become a field in its own right, though less advanced than bacterial metagenomics [28]. The data that has been collected thus far, has shown that there are some parallels between eukaryotes and bacteria in the microbiome. Much like the bacteria, there appears to be no 'core' eukaryotic microbiome at the species level, but representatives from the same genus are often represented across diverse populations [11]. Secondly, despite relatively few genera being associated with symbiosis, there are a large number of individual species and strains within each genus, often capable of inhabiting a broad range of hosts. *Blastocystis* and *Entamoeba* spp. have a diverse range of hosts including mammals, birds, reptiles and amphibians (Figure 1.2.2) [29]. This indicates that while few transition events have occurred, once the hurdle is overcome, diversification and speciation is widespread. This may be due to the fact that the host intestine can be considered an extreme environment as it differs appreciably from a free-living one, primarily because of host immunity and microbial density [19].

Many organisms that have been investigated show evidence of pathology that reflects flexible phenotypes, for example, *Entamoeba histolytica* only causes symptoms in around 10% of cases [30]. There are also examples where normally absent or commensal organisms become virulent when abundant or when colonising the immunocompromised, for example *Blastocystis* spp. [31]. Manifestation of pathology may be dependent on multiple factors including the host, co-infection with multiple symbionts or composition of the bacterial microbiome. Comprehensive insight into the nature of these factors and their interplay will be instrumental in categorising, treating and preventing symptoms from eukaryotic symbionts.

In this thesis, I will focus on the evolution and adaptation of two eukaryotic components of the human microbiome; *Blastocystis* sp. STs and *Entamoeba histolytica* as these are two of the most common taxa associated with the human gut and both have been linked to disease.

1.2 *Blastocystis*

1.2.1 Discovery

The *Blastocystis* genus was first named in the literature in 1911 by Alexeieff who classified the cells as intestinal yeast after differentiating them from *Trichomonas intestinalis* (now *Pentatrichomonas hominis* [32]) [33]. Eventually, the name *Blastocystis hominis* was proposed to refer to these cysts found in human stool samples [34]. However, this trend continued and species continued to be named for the hosts they were isolated from. Unfortunately, there is considerable overlap in host range between species of *Blastocystis* and it became increasingly necessary to cross-reference between nomenclatures in the literature until 2006 [35] when a broad consensus was established whereby each species name of human infective *Blastocystis*, regardless of what host is was isolated from, was replaced by '*Blastocystis* sp.' followed by the subtype (ST) designation (Figure 1.2.2). This nomenclature will apply throughout this thesis.

Detailed morphological examination was not conducted until 1967 when Zierdt *et al.* [36] classified *Blastocystis* as protozoan rather than yeast. This assignation was based on the morphological features of the cell such as the nucleus, nucleolus, Golgi apparatus, smooth and rough endoplasmic reticulum, mitochondria with tubular cristae and the lack of a cell wall. Different morphological forms of *Blastocystis* were also identified which prompted a renewed interest in the organism.

Since then *Blastocystis* sp. STs have been found in almost every country that has been searched. Certainly the distribution is worldwide [37], [38] but reports on prevalence vary, even within the same country, due to differences in identification technique, sensitivity and sampling of distinct populations within a country. It is considered the most common eukaryote found in human stool samples

and recent estimates suggest that there may be as many as 1 billion individuals carrying *Blastocystis* sp. STs [39]–[49].

1.2.2 Morphology of *Blastocystis*

Blastocystis sp. STs can be collected from the stool of many vertebrates, including mammalian and avian hosts, and while isolates from different species may not be morphologically identical, they are similar enough to make differentiation based on morphology alone very difficult [50]. So far, four morphological forms have been confidently described: vacuolar, granular, amoeboid and cyst [36], [50]–[54]. Other forms have been observed, though many of these are unconfirmed and may be artefactual due to oxygen stress [52]–[55]. The vacuolar form is the most common *in vitro* and divides by binary fission, though there is considerable size variation (2 – 200 µm). They consist of a large central vacuole with the nuclei and mitochondria related organelles (MROs) located in the thin layer of cytoplasm at the periphery of the cell. Granular forms gradually become more prevalent as a culture ages and are categorised by the appearance of granules within the central vacuole, the function of which remains unclear [50]. Cysts are the smallest form of *Blastocystis* (2 – 5 µm) and are difficult to detect in stool samples, which is why they are the most recent form to be attributed to *Blastocystis*. They possess a multi-layered cell wall and sometimes a loose outer coat. There is evidence to suggest the cysts are more resistant to oxygen, temperature and dehydration stresses than the other forms [50] with some studies showing survival for up to a month in water at 25˚C and two months at 4˚C [56]. While there is some discrepancy in the literature as to their particular tolerances, these may be due to strain variations between isolates in each study [50], [56], [57]. The final form is the amoeboid, which is the least common. Typically they are 2 – 15 µm and may show features consistent with the vacuolar form with the exception of one or two pseudopodia, though there are conflicting reports possibly due to strain variation [50], [58], [59]. While this form is less frequently observed, it appears to be able to phagocytose bacteria suggesting a biological relevance *in vivo*.

1.2.3 Taxonomy

Conflicting reports in the literature reflect confusion about the morphological forms of *Blastocystis* sp. STs, its size, features and potential artefacts, which was compounded by inconsistency in the nomenclature. This confusion and uncertainty

resulted in *Blastocystis* being incorrectly assigned to a number of taxonomic positions including yeast, fungi, Sporozoa (now Apicomplexa), Amoebozoa and the cysts of a flagellate [51], [60], [61]. In 1989 Johnson *et al.* [62] made an attempt to use small subunit (SSU) rRNA sequence data to overcome the morphological discrepancies and definitively place *Blastocystis* taxonomically, however this attempt was unsuccessful and failed to group *Blastocystis* into an existing clade primarily because of their use of a restricted sample, seeking to place *Blastocystis* with either the Apicomplexa or yeast (*Saccharomyces*). In 1996 Silberman *et al.* [63] used the SSU rRNA sequence data to systematically test the placement of *Blastocystis* with each major eukaryotic lineage. Silberman *et al.* found that it robustly clustered on a deep branch within the Stramenopiles (or Heterokonts) (Figure 1.2.1). Since then, multigene analyses have been used to confirm the phylogenetic placement within the Stramenopiles [64], [65].

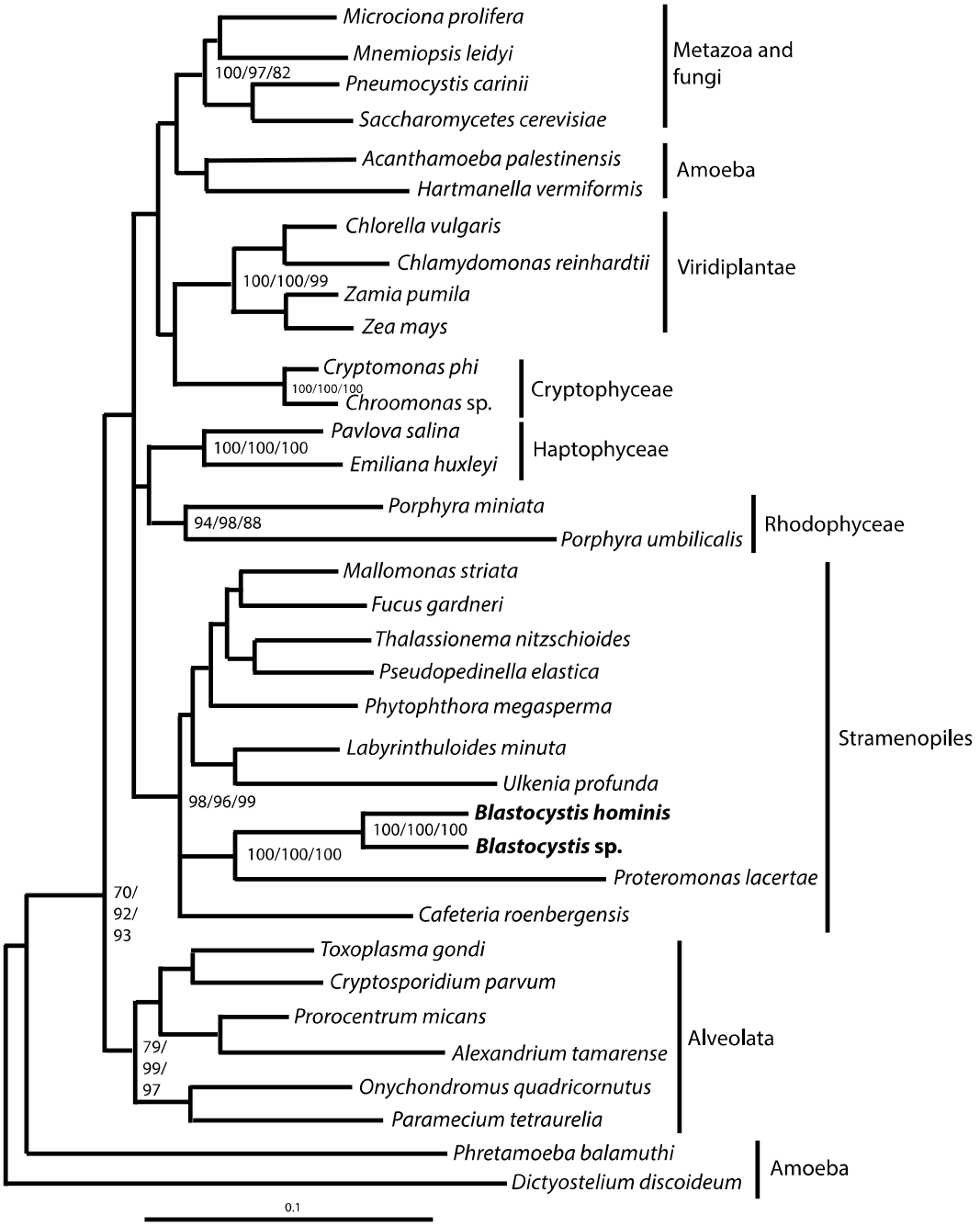


Figure 1.2.1 Relationship of* Blastocystis *(ATCC 50177 and 50578) among Stramenopiles as determine by phylogenetic analyses of SSU rRNA gene sequences. *Redrawn from Silberman* et al. *1996* [63] *using Maximum-likelihood, Fitch-Margoliash, neighbour-joining and parsimony analyses show that* Blastocystis *is embedded in the Stramenopile lineage (only values from 100 bootstrap replicates of the latter three analyses supporting the major eukaryotic groups are shown).*

The Stramenopiles are an extremely diverse group of organisms belonging to the SAR clade (Stramenopiles; Apicomplexa; Rhizaria) and are categorised by tubular cristae in the mitochondria and tripartite hairs on one of two flagellar [66].

This group includes organisms which contain an example of an independent acquisition of chloroplast giving rise to the brown algae and diatoms [67], a clade of plant, animal and fish parasites, known as the Oomycetes, single celled flagellates and a group intestinal anaerobes which includes *Blastocystis.* While the definitive morphological features are relatively well conserved across the large phylogenetic distances within this taxonomic group, molecular data has recently included members with more derived features, including *Blastocystis* which has MROs as a result of its adaptation to an anaerobic environment [68] and has yet to be observed with a flagellum.

1.2.4 Subtypes

Blastocystis can be isolated from the faeces of several birds and mammals, including humans, though there have been accounts of it appearing in the stool of reptiles, amphibians and even insects. It is unlikely that isolates from the latter hosts overlap much with *Blastocystis* species isolated from mammals and birds and they possess their own species names and, from what has been sampled to date, seem to represent distant species within the *Blastocystis* genus [54]. There are currently 17 subtypes from mammalian and avian hosts each with their own distinct SSU rDNA sequence and each of these subtypes has its own zoonotic preference, for example only subtypes 1 – 9 are found in humans, though only 1 – 4 are common [69]. A breakdown of subtype host repertoires is shown in Figure 1.2.2 [70].

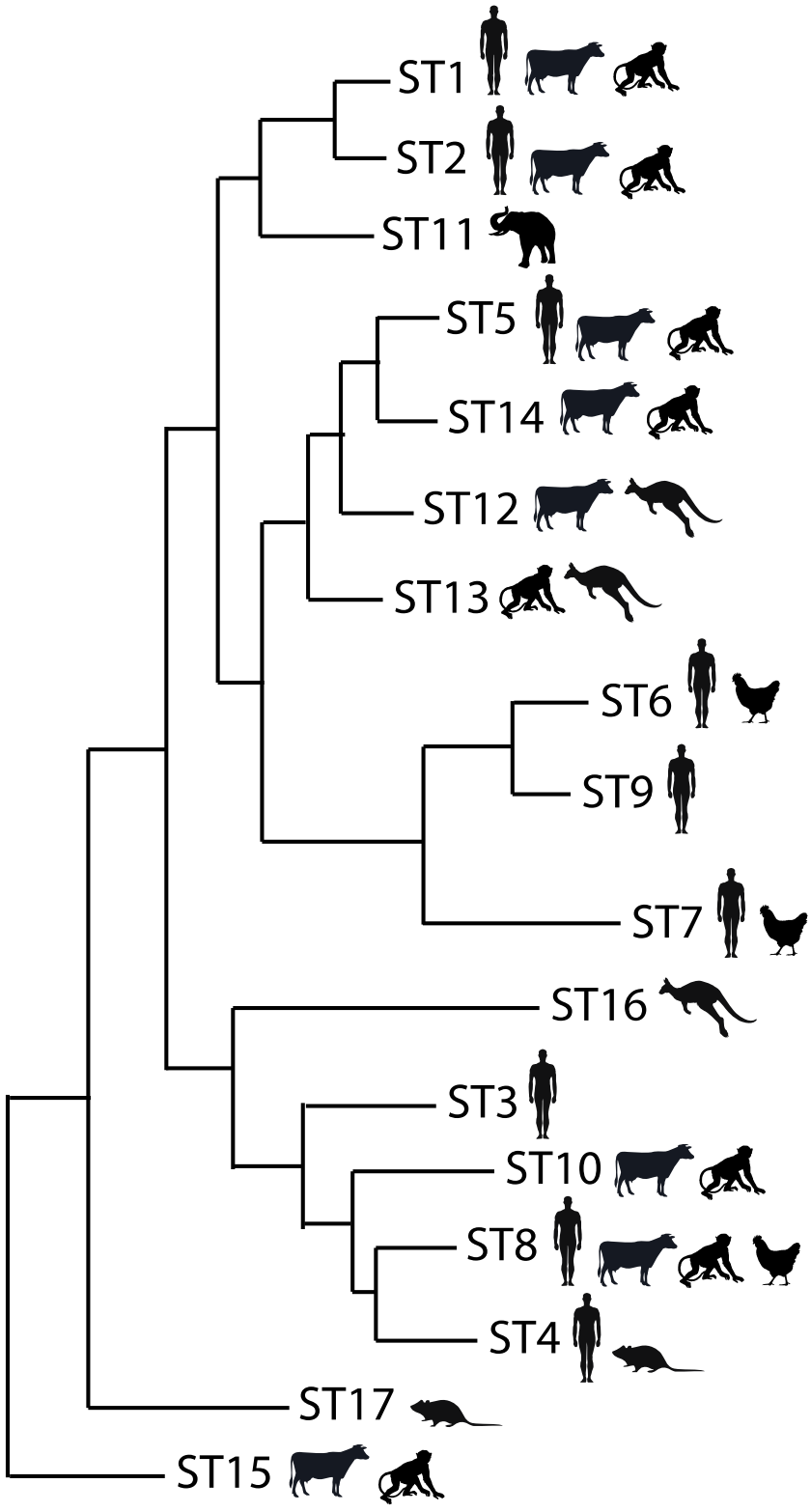


Figure 1.2.2 Host range of <u>Blastocystis</u> subtypes in relation to subtype phylogeny. *The phylogeny was redrawn from a maximum likelihood SSU rDNA phylogeny Alfellani* et al*. 2013* [71] *and the host range is based on Tan 2008* [50]*. Subtype host range indicated by symbols representing human, Artiodactyla (cow), Proboscidea (elephant), nonhuman primates (macaque), aves (chicken), rodents (rat) and marsupials (kangaroo).*

1.2.5 Life cycle

As *Blastocystis* sp. STs are capable of forming a cyst and are isolated from faecal samples, they are most likely transmitted via the faecal-oral route similar to other intestinal protozoa such as *E. histolytica, Giardia lamblia, Trichomonas intestinalis* and *Cryptosporidium* [51], [52], [61], [72]–[79]. This has yet to be confirmed experimentally which is likely due to the absence of a suitable animal model though several life cycles have been proposed with varying degrees of confidence [38], [51]–[53], [79], [80]. Adding to the confusion around the myriad morphological forms of *Blastocystis* are the proposed mechanisms of reproduction which have included binary fission, endodyogeny, schizogeny, plasmotomy and budding [79], [81], [82]. The current consensus on life-cycle (Figure 1.2.3) takes into account the unknown roles of various confirmed morphological forms and the uncertainty that surrounds the mode of transition between these forms. A few attempts have been made to examine the transition between forms using transmission electron microscopy, which showed that cyst forms can transition into vacuolar or granular forms [83], [84]. There is also some evidence that vacuolar forms are able to transition into amoeboid forms, though it is unknown whether this is a permanent or intermediary state [85], [86].

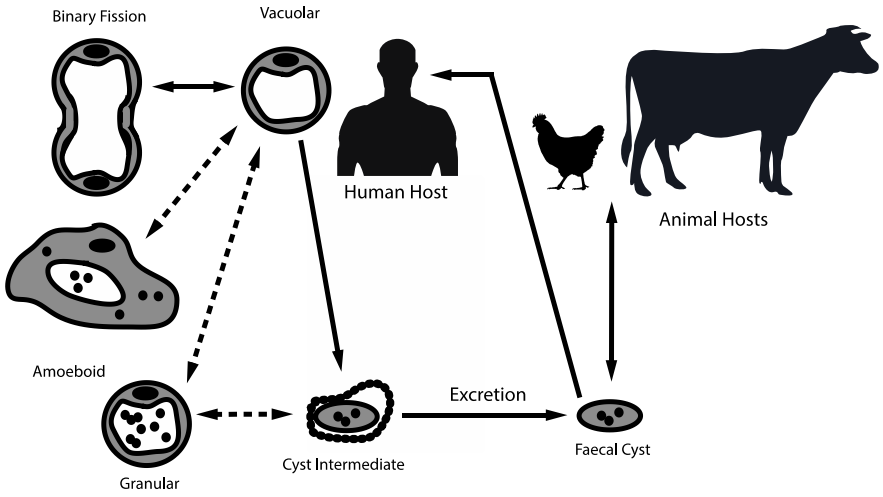


***Figure 1.2.3 Proposed life cycle for* Blastocystis.** *Infection in humans and animals is initiated when the faecal cyst is ingested. Faecal cysts develop into vacuolar forms, which subsequently reproduce by binary fission. Some vacuolar forms encyst and these faecal cysts lose their outer fibrillar layer as it matures. The environmental faecal cyst is transmitted to human and animal hosts via the faecal oral route and the cycle is repeated. The transition of other forms with respect to the vacuolar form is less well understood and is represented with dashed lines.*

Due to the lack of an effective infection model, little is known about how *Blastocystis* interacts with the host and the rest of the microflora. While the amoeboid form exhibits phagotrophic behaviour the other forms may be biotrophic as they have been found in close association with digested food debris or with the host mucosa [87], [88]. Despite recent efforts to sequence the genomes of *Blastocystis* subtypes, little is also known about the molecular biology such as ploidy or whether it reproduces sexually [50], [89]. The karyotype is likewise unconfirmed; between 9 and 13 chromosomes have been observed in an analysis of 15 strains using a contour-clamped homogeneous electric-field (CHEF) system ranging between 2200 kbp and 260 kbp [90], however, pulsed-field electrophoresis of *Blastocystis* sp. ST7 characterised 15 chromosomic bands [91]. It is therefore likely that this feature is also variable based on subtype.

1.2.6 Disease

The role of *Blastocystis* in disease is still widely debated in the literature. Some authors consider *Blastocystis* to be an emerging pathogen [92]–[95], while others doubt the role of *Blastocystis* in disease [96], [97]. In the majority of cases, individuals 'infected' with *Blastocystis* exhibit asymptomatic carriage with only a small proportion developing symptoms. Recent metagenomics studies have shown that *Blastocystis* can have high prevalence in healthy individuals and so may be part of a healthy intestinal flora [98]–[100]. The tenuousness of the link with disease arises from the fact that no *Blastocystis* outbreaks have been verified and that there is only anecdotal evidence that treatment of *Blastocystis* results in improvement of clinical symptoms [39].

Having acknowledged that the evidence for *Blastocystis* causing disease is inconclusive, *Blastocystis* sp. STs have nevertheless been implicated in a range of non-specific gastrointestinal symptoms (occasionally known as blastocystosis), including diarrhoea, abdominal pain, flatulence, nausea, vomiting, constipation, weight loss and fatigue [50], [53], [101], [102] but is particularly associated with irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD) [102]–[104]. Some studies suggest that *Blastocystis* persists at higher concentration in individuals with IBS [105]–[107], however, *Blastocystis* is not implicated as the cause of the symptoms and other studies have found insufficient statistical evidence of this association [108]. *Blastocystis* has also been implicated in the development

of non-intestinal symptoms including cutaneous legions and urticaria. All symptoms, including intestinal ones, can range from acute disease to chronic infection [50], [70], [109]–[112].

One hypothesis for the differences in outcome to infections was thought to be genetic variation between subtypes [50], [102], [104], [113], for example ST1 [114], ST2 [109], [115] and ST3 [111] have all been found, independently, at higher prevalence in symptomatic patients. ST4 has been implicated in intestinal disease in European patients [116]–[118] while the amoeboid form of ST3 was implicated in the development of urticaria [119] by adhering to the host mucosa and causing inflammation followed by lesions [120]. Conversely, other studies either found the subtypes to be more associated with carriage [121], [122], or found no correlation between subtype and disease outcome [115], [122]–[127].

The confusion surrounding the pathogenicity of *Blastocystis* most likely results from the lack of knowledge about its biology and lack of an animal model. This has hindered work investigating pathogenesis, though several experimental models have been described and found to be unsuitable [50], [102].

1.3 *Entamoeba*

1.3.1 Discovery

The first appearance of *Entamoeba histolytica* in the literature can be difficult to pinpoint. It may have appeared in the literature under a number of pseudonyms before its differentiation from *Entamoeba coli* by Schaudinn in 1903. Schaudinn posited that it be called '*histolytica'* because of its ability to lyse tissues [128]. Later it was proposed that a second, morphologically similar species existed as well, *Entamoeba dispar*, though it took until 1978 until this proposition was taken seriously and until 1993 before *E. dispar* received its own classification [129]. *E. histolytica* is isolated from the intestinal tract of humans and has a proven association with intestinal disease.

1.3.2 Taxonomy

Entamoeba spp. are located on a deep branch within the Archamoebae, which is the major anaerobic group of protists within the Amoebozoa. The Archamoebae consists primarily of amoeboid flagellates and amoebae and was created to combine the Pleobionts and the Entamoebae [130]. Currently, the Archamoebae consists of less than 500 individual species organised into five families: Entamoebidae, Pelomyxidae, Mastigamoebidae, Tricholimacidae, and Rhizomastixidae [131]. The majority of the species described in these groups are free-living, though there are distinct symbiont lineages including *Entamoeba* which falls into the Entamoebidae family (Figure 1.3.1) [132], [133].

Despite numerous attempts, the relationships between the branches within the Archamoebae were unresolved until recently. Morphological examination was inconclusive even between genera and there was confusion as to which characters were informative [134], [135]. SSU rDNA and actin trees were unable to robustly resolve the relationships between the families, despite monophyly of each family being confirmed, until a multigene analysis was conducted, made possible due to advances in sequencing technologies and sampling [131], [133], [135]–[137]. This analysis showed that Entamoebidae branches more deeply than the other families which contain the flagellate Archamoebae and are known collectively as the Pelobiontida [131].

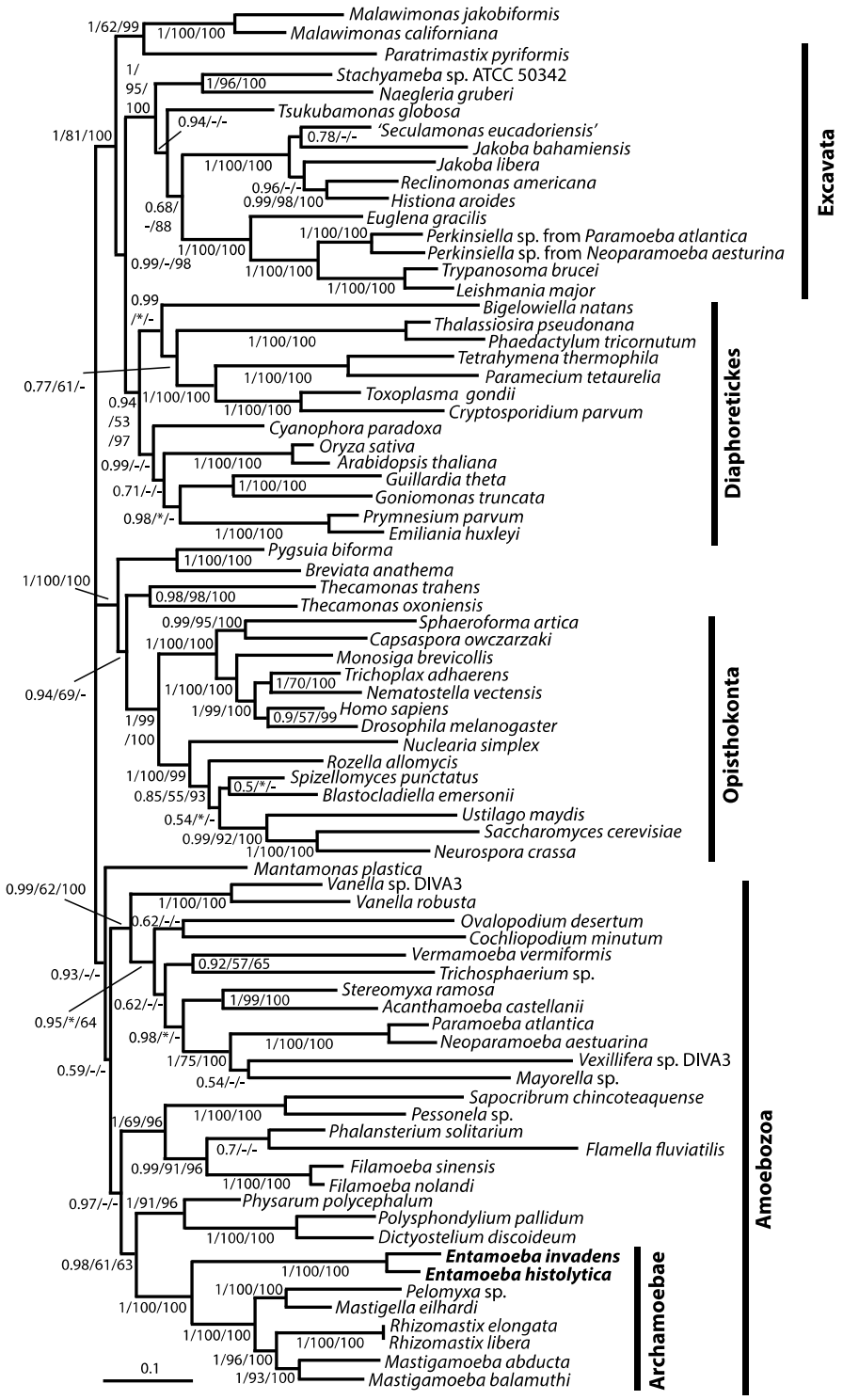


Figure 1.3.1 Eukaryotic phylogeny showing the robust placement of* <u>Entamoeba histolytica</u> *among the Archamoebae. *The phylogeny was redrawn from a concatenation of seven protein-coding genes from Pánek* et al*. 2016* [131] *(actin, a-tubulin, b-tubulin, EF1a, EF2, HSP70, HSP90). The topology was constructed using CAT Poisson model. The values at nodes represent posterior probabilities, non-parametric bootstraps, and maximum likelihood bootstrap support. The values lower than 50% or 0.5 are marked by "*". Branches that were missing in the best maximum likelihood tree topology are marked by "–".*

Several species of *Entamoeba* are known and have been classified primarily based on molecular data as morphological distinction is difficult and large generic variation exists between morphologically identical organisms. The classic example is *E. histolytica* and *E. dispar,* which are morphologically indistinguishable and yet have distinct phenotypes *in vivo; E. histolytica* is pathogenic while *E. dispar* is avirulent. Other *Entamoeba* species are associated with disease with varying degrees of severity, though as identification is based on the host, it is unclear to what extent taxa are synonymous [132]. Despite much molecular work focusing on *Entamoeba* much of its biology remains a mystery including chromosome number, ploidy and whether they can sexually reproduce [73].

1.3.3 Life cycle and distribution

E. histolytica is acquired through ingestion of food or water contaminated with the infective cyst form of the amoeba [138]–[140]. Once in the ileum, the cyst develops into the potentially pathogenic trophozoite stage [141]. In this form it is often able to persist asymptomatically, dividing by binary fission and phagocytosing the surrounding bacteria [142], [143]. However, in a minority of cases it is able to cause the characteristic symptoms of infection including abdominal pain, dysentery and extra-intestinal abscesses. It is estimated that there are approximately 500 million people infected with *E. histolytica* worldwide, though the majority of these are asymptomatic [144]. 10% of those infected, however, suffer from invasive amoebiasis which results in between 40,000 and 100,000 deaths per year [144]–[146]. Due to the route of infection, areas with poor sanitation and sewage infrastructure are at higher risk of being endemic, for example, an annual incidence of children living in an urban slum in Bangladesh was as high as 40% [147]. There are also more affluent populations which are endemic, for example among those who engage in oral or anal sex, commonly homosexual men, whereby transmission is sexual [148]–[150].

1.3.4 Invasive amoebiasis

In asymptomatic carriers, *E. histolytica* produces no symptoms presumably because the symbiont is not attacking the host tissues, however, invasive amoebiasis occurs when the trophozoite stage begins to degrade the host mucosa allowing adherence directly to the epithelium [151], [152]. This not only allows *E.*

histolytica to invade the epithelial layer but also allows other members of the microbiota to cause secondary infections [153]. Adherence to the unprotected epithelium causes an inflammatory response of cascading cytokines and trogocytosis triggers apoptosis [154]–[156]. The resulting invasion causes development of flask-shaped ulcers characteristic of the disease which leads to disruption of the absorptive qualities of the intestine and, eventually, dysentery [157].

Trophozoites that penetrate the epithelium may then be disseminated to other organs, commonly the liver, whereby they form abscesses (Amoebic Liver Abscesses, ALA). A fatal outcome to infection is often due to destruction of the liver function caused by lytic enzyme damage to parenchymal cells. The lytic enzymes concentrate around the abscess and are released by trophozoite-induced apoptosis of hepatic immune cells [158]. It is still unclear which features dictate the outcome of disease, though undoubtedly it will involve a complex interplay of host, symbiont and environmental factors. Investigation into the evolutionary processes that have facilitated the ability of *Entamoeba* to cause disease may help to shed light on some of the colonisation factors involved in determining the outcomes of disease.

1.4 Genome evolution

1.4.1 Mechanisms of evolution

Mechanisms of evolution can be broken down into two components: mutation, describing changes in the frequency of genotypes, and selection, describing changes in the frequency of phenotypes. Mutation is the process by which the genome sequence changes, which may or may not lead to phenotypic change. The smallest occurrence of mutation in the genome is a point mutation. This involves a substitution or the insertion or deletion (indel) of one nucleotide of nucleic acid. This will consequently involves a change to the complementary strand as well. Combinations of three nucleotides code for amino acids and therefore changes that involve a different amino acid being coded for are termed nonsynonymous and may result in a phenotypic change. If the phenotype is affected then the mutation is subject to selection. However, the genetic code is highly redundant and therefore a change to a nucleotide may not involve a change in amino acid, these are termed synonymous mutations. Indels have the additional

effect of imposing a frame shift. This means they do not only alter the nucleotides coding for a particular amino acid, but also all subsequent amino acids in a cistron. This assumes, however, that the mutation is occurring in a transcribed part of the genome. In fact, mutations occur continuously across the genome though where DNA is functional these are more likely to be subject to selective pressure. While non-coding DNA contains regulatory regions necessary for expression of a gene or genes, these non-coding regions are often under less selective pressure than coding regions.

Large changes to the genome may occur by accumulation of point mutations but also by mutations on a larger scale. Some mutations produce multiple copies of genes [159], [160]. Gene duplication is the most important mechanism for evolution [161] and is usually categorised by the size of the duplication event [162], [163]. This serves to relieve selection on one copy of the gene, which is then allowed to accumulate mutations. This may result in a number of different outcomes such as neofunctionalisation, subfunctionalisation or pseudogenisation [164]. The most common outcome is pseudogenisation, whereby genes accumulate mutations, become non-functional and may be eventually lost. However, while less common it is possible for a copy of the gene to acquire a slightly different function (subfunctionalisation), such as the *pax6* genes in zebrafish which preferentially act on different tissues [165]. Finally, it is possible for a gene to acquire a new function altogether, as was the case in the Antarctic zoarcid fish which developed a type III antifreeze protein from a copy of sialic acid synthase [166], though this type of event is rare [167], [168].

Segmental duplications give rise to tandem arrays and are most likely the result of ‘unequal crossing over’ during homologous recombination between paralogous sequences. The position of segmental duplications can be used to identify regions undergoing rapid evolution where diversity of function confers an advantageous [169], [170]. In primates segmental duplications may be responsible for large genetic variation and susceptibility to diseases [171]. Recombination is also responsible for deleting, inverting and rearranging the exons of existing genes, producing mosaic genes with varied functions [160], [172]. On a larger scale, whole chromosomes or even the entire genome may be duplicated [173], [174]. This would represent a large opportunity to contribute to the complexity of the organism but is a rare event, at least in sexually reproducing organisms, because of the necessity to

conserve the number of chromosomes and the proportions of dose dependent gene families [175].

Another way of obtaining functional genes is to acquire them via Horizontal Gene Transfer (HGT, sometimes also known as lateral gene transfer) [176]. HGT denotes a process whereby an organism obtains a gene or set of genes from another microorganism, often a bacterium that bestows a novel function to the recipient. The newly acquired genes may then be integrated into the recipient's genome and be subsequently influenced by the processes discussed previously. The best examples of this are antibiotic resistance genes which may be transferred between pathogens [177] and metabolism-associated genes which confer selective advantages especially in anaerobic environments [168].

Selection acts on mutations that give rise to phenotypic changes and may also be broken down into three major components: positive, balancing and negative. Negative, or purifying, selection is the selective removal of deleterious variants from a population. It operates by eliminating the individual mutations that generate detrimental phenotypes and acts to preserve the adaptive characteristics of an organism. Typically, purifying selection will affect most of the coding sequences in a genome and is greatest in magnitude within those genes that produce the most essential proteins.

Positive selection is the process by which an advantageous mutation becomes fixed within a population by conferring survival or reproductive advantage to an organism. Selectively neutral variants may also become fixed in a population by genetic drift, which complicates the detection of positive selection and also has implications for interpretation of fixed genotypic features, which may not be the results of adaptive evolution.

Overall, selection tends to act on genotypic variants and regulates the phenotypes of individuals. However, selection does not just act on the genotypic variants. High throughput methods have allowed insight into the role of epigenetics in genome evolution [178]. Epigenetics refers to heritable changes in gene function that cannot be attributed to changes in the nucleotide sequence [179]. This principally refers to histone variation and DNA methylation which affect gene expression [180], [181]. DNA methylation in particular has been associated with

repression of genes, variants are often lethal [182] and thereby subject to purifying selection.

There may be many more mechanisms affecting genome evolution that have yet to be described or that are specific to individual clades or genera. An example of this is poly-A termination in *Blastocystis,* whereby one or both final characters in the UAA stop codon are provided by the poly-A tail of the mRNA [183]–[185]. This may relieve selective pressure on part of the gene and have consequences for the evolution of parts of the genome. If individual groups of organisms possess their own mechanisms for genome evolution, then these groups may be subject to differential rates of evolution.

1.4.2 Gene gain and loss

Gene gain and loss through whichever mechanism determines the gene content of a genome. Investigation of gene gain and loss within a gene family is a popular method for studying genome evolution. A gene family is described as a group of homologues that includes orthologues (genes related by speciation) and paralogues (genes related by duplication) [186]. Due to the conservation of function, orthologues often retain common functionality better than paralogues [187] and are therefore commonly used to transfer gene annotation [188], [189].

Once a gene family has been identified, its evolutionary history may be inferred from patterns of gene gain and loss [190]–[194]. The role of duplications in copy number variation and gene family amplification has been well documented [195]–[201]. Methods for investigating this evolutionary history relies on the discrepancy between the gene phylogeny and the species phylogeny [202], [203]. Differences in how the genes and species have diverged mean that reconciliation is a powerful tool in explaining the processes that produced the differences [202]. This method may be used to examine gene gain and loss across multiple genomes [190], [203]–[205]. By considering the function(s) of a gene family it may be possible to infer the causes that have led to the observed changes in phenotype of gene content between two genomes.

If genotypic differences explain phenotypic differences, then where a gene family is expanded, it may be possible to conclude that components of the family are important to a phenotype of a particular organism. In the case of Trypanosomes,

they have expanded families of Variant Surface Glycoproteins (VSG) when compared with the free-living out-group *Bodo saltans* [206]. This is because the VSG are used to escape detection by the immune system. *B. saltans* does not need this function as it does not inhabit an environment with an immune system [206]. Expansion is also commonly associated with diversification of gene function (subfunctionalisation). Again, the VSG can be taken as an example, as monoexpression of a single variant followed by switching to another variant is the mechanism by which immune evasion is achieved. This is only possible because the VSG are non-identical and each variant instigates a different antibody response. Gene families that are expanded in parasitic genomes compared to free-living ones may signify that the function(s) provided by that family are be important in explaining the differences seen between the two organisms' life strategy.

In opposition to gene family expansion is gene family reduction. This occurs when the function of a particular gene or genes becomes non-essential to the survival or reproduction of the cell. Selection pressure is reduced and the gene is allowed to accumulate mutations, become non-functional and is eventually lost. Parasites were initially considered to be examples of genomic devolution, becoming reduced and less complex. However, genomic reduction is not seen in all parasite lineages, including nematodes [207], [208] and Oomycetes [209], [210] and parasite genomes that are reduced also show innovation and expansion suggestive of streamlining rather than simplification [206], [211], [212]. Gene loss is still an important feature of some symbiont lineages and is associated with the redundancy of function. An example would be the loss of components of the Electron Transport Chain (ETC) in the MROs of *Blastocystis* sp. STs. *Blastocystis* sp. STs are anaerobic and therefore have little use for a mechanism which uses O_2 as the final electron acceptor. The ECT is therefore redundant and subsequently, elements have been lost [213].

1.4.3 Genomic adaptations

Expansion of a gene family is generally thought to reflect the functional necessity for the function of that family, while reduction is associated with the redundancy of function. The vertebrate intestinal tract is a highly specific environment, as discussed in Section 1.1.2, and therefore it may be expected that specific adaptations are necessary for survival therein. It was long thought that adaptation to a host environment is associated with reduction in genomic complexity

especially in the case of obligate parasites. The most extreme example of this is microsporidia, which is possibly the most host-dependent parasite identified [214] and shows not only reduction in gene repertoire but also contraction of the genome size [215] though examples of this can be seen across many parasitic lineages including Trypanosomes [216], Giardia [217] and the Apicomplexa [218]. However, there are also many examples of parasite genomes where there is no genome-wide reduction such as nematodes [219], [220], Schistosomes [221] and the Oomycetes [209], [210].

This inconsistency among studied symbiont genomes would suggest that genomic reduction is not a hallmark of symbiosis that applies across all lineages. However, there does appear to be a pattern of genomic reduction when you consider the life strategy of these symbionts. Obligate intracellular parasites, such as Microsporidia, are the most extreme examples of genomic reduction, whereas nematodes show no evidence of reduction but part of their life cycle requires that they are free-living. Thus, they still require the machinery to persist outside of the host. This pattern can be seen across different symbiotic clades and suggests that it is not a symbiotic life-strategy *per se* that determines the genomic trend, but the level of dependence on the host throughout the life cycle, for example, in the Apicomplexa gene loss appears to be the strongest, though not the only evolutionary force at work [211], [222], [223]. The specific gene families that have been lost are dependent on the symbiont in question and, to a certain extent, the host it inhabits. One common theme, is a loss of metabolism-associated genes [224]. This is because one of the advantages of having a host is that nutrients are often readily available and easily accessible. Many symbionts lack the capacity to generate purines which are fundamental components of both nucleic acid and energy metabolism and must obtain these from the host throughout their life cycles [225]–[228].

While reliance on the host shapes some aspects of genome loss, the host environment also represents novel challenges to the organisms that require genomic innovation. The host immune system necessitates the expansion of certain gene families that enable symbionts to persist and even invade. These adaptations are often novel and are typically lineage-specific and evolve independently but display some similarities that reflect their common target (i.e. immunity). Moreover, their functions may give insights into the mechanisms of disease and invasion. Common examples include expansion of variant protein families expressed at the

cell surface that enables various symbionts to successfully evade the immune response during chronic colonisation such as Trypanosomes [229]–[231], Apicomplexa [232], Diplomonads [198] and the Oomycetes [209], [233]. Expansion of proteases is also common and enables symbionts to degrade host proteins either in defence, such as cleaving bound antibodies in trematodes [234], [235], or attack, such as degrading the host mucosa allowing direct contact to the epithelium beneath such as in *E. histolytica* and Apicomplexa [236]–[238]. Given that the pattern of gene family amplification is indicative of those important for colonisation and persistence, it is often possible to detect important families by examining copy number and divergence between symbionts and free-living organisms.

In this thesis, I'll be considering the contributions of gene gain and loss to the molecular evolution of the *Blastocystis* and *Entamoeba* genomes.

1.5 Comparative genomics

The comparative method is the principal method by which evolutionary origins and mechanisms are investigated. Since the advent of genome sequencing, it has become possible to examine the evolution of symbiont adaptations at the level of individual genes and even nucleotides. Hence, comparative genomics is the method through which two or more genomes are compared in order to infer the genotypic changes that explain observed phenotypic differences. Choice of phylogenetic distance between comparators is determined by the nature of the question being asked. Thus, by comparing closely related organisms it is possible to create a detailed reconstruction of the last common ancestor and investigate 'recent' evolutionary events by looking at preservation and divergence in contemporary genome sequences [239]. Large phylogenetic distances can be used to give clues about essential processes that are present in all forms of life. By using a combination of phylogenetic distances in an analysis it is possible to build detailed records of the evolutionary events in the ancestors of organisms of interest and allow interrogation of a particular moment associated with a phenotypic change, such as the origin of parasitism, and speculate about the processes that drove them. In order to gain meaningful insights into particular evolutionary events, it is important to carefully consider phylogenetic distances and sampling of different events.

1.5.1 Sampling and phylogenetics

The main issue with attempting to infer the evolutionary changes from comparison of the contemporary genotypes of related organisms is the sampling of the genomes used. By using an incomplete sample, it is possible that the outcomes of the investigation may be artefactual or misleading. This is akin to the choice of treatment in an experimental setting. Meaningful results can only be collected where sensible treatment choices allow questions to be asked of the data.

As with phylogenetic studies, sampling can have an important effect on the accuracy of comparative genomics [240]–[246]. While phylogenetic studies are subject to sampling error at both that the individual gene level and the species level [247], [248], comparative genomics utilises multiple genes from across the genome. This relieves some of the sampling error at the gene level and can help to overcome error introduced by differential rates of evolution between genes, however, sampling at the species level can still have a measurable effect on the outcomes of the inference [249].

1.5.2 Free-living out-groups

Comparison of symbionts, both morphological and molecular, has a long history [212]. However, these typically compare symbionts with each other, which allow identification of species differences and for reconstruction of the ancestral, symbiotic state but often cannot provide data on the evolutionary transition to a symbiotic life strategy. Reconstruction of this ancestor requires an out-group that is entirely free-living but as closely related to the symbiont of interest as possible. Inclusion of this out-group makes it possible to determine the features that are involved in symbiosis from those that are features of the organisms generally [206]. In the case of symbionts, this out-group is very important because it may give clues about how an organism evolves to become symbiotic from free-living [250]. As both *Blastocystis* sp. STs and *E. histolytica* are predicted to have small genomes compared with other eukaryotes, our inclusion of the free-living out-groups will show whether this feature is indicative of the symbionts or whether this is a feature of the organisms generally.

There are many examples in the literature of studies that have used comparative genomics to investigate the genomes of symbiont lineages. This has

included comparisons over multiple phylogenetic differences including between strains, such as Jerlström-Hultqvist *et al.* 2010 [217], who showed that, despite a set of well conserved 'core' *Giardia* genes, there was significant genome variation between isolates and Ankarklev *et al.* 2015 [251], who suggested that these differences may play key roles in determining the host range and virulence of each isolate. Longer distances such as those between species like Mazurie *et al.* 2013 [74] and Cai *et al.* 2013 [252] who investigated species differences in *Cryptosporidium* and *Plasmodium* respectively. These studies focused on genotypic differences that awarded interesting phenotypes such as drug resistance, host range, infectivity and pathogenicity. Similarly, within *Trypanosomes* [253] and *Leishmania* [254] comparative analyses have identified 'core' genes shared between species as well as species-specific families. This allowed detection of important parasitic factors, diagnostic and potential drug targets. However, even larger distances have been studies including comparisons between species belonging to different genera in the Kinetoplastids and Apicomplexa. These have provided more detail and data on the biology of the parasites and the host-parasite interactions, including identification of key parasite molecules that elucidate the mechanisms of these complex diseases [255], [256]. Studies have investigated both gene expansion and reduction between parasite species [214], [257] but while these studies can provide details about the individual parasites studied, they do not sample widely enough to ask questions about the origin of parasitism in each of these lineages. In order to investigate this moment in evolutionary history, reconstruction of the parasitic ancestor is necessary and this requires a free-living out-group. This allows identification of factors specific to parasites and investigation into how these parasite factors have evolved [206], [233], [258]. It allows differentiation between adaptations that are necessary for parasitism and those that are features of the organisms in question, regardless of their life strategy [259]–[261].

In this thesis, I will use free-living relatives of *Blastocystis* and *E. histolytica* as out-groups in a comparative genomics analysis, to investigate adaptations to symbiosis and disease.

1.6 Genome sequencing

To create genome and transcriptome sequences for free-living relatives of *Blastocystis* and *Entamoeba*, I have utilised two types of sequencing technology in order to obtain genome information about the organisms of interest; Illumina and PacBio. These technologies are based on different sequencing techniques and both have advantages and disadvantages associated with them.

1.6.1 History

Genome sequencing has been a popular scientific tool for 20 years but in that time, has followed a classic example of Moore's law [262] becoming faster, more efficient and cheaper with each generation. The first generation was developed in 1977 by Sanger and is based on chain-termination method (Sanger Sequencing). Because of its ease to use and the lack of radioactive reagents this was the first sequencing technology to become commercially available [263]. Shortly after this an automated sequencing machine was developed by Applied Biosystems which utilised capillary electrophoresis improving speed and accuracy so that by 1995 it was possible to sequence 2.88 million bases per day with read lengths up to 900 bases. The use of these technologies in the Human Genome Project helped to drive the search for faster, more accurate methods that were cheaper and less manually intensive [264]. Thus, the era of Next Generation Sequencing (NGS) began. These technologies differ from Sanger sequencing in their reduced cost and parallel analyses, allowing for much higher throughput [263].

1.6.2 Illumina Sequencing

Illumina sequencing is an example of NGS and works using a method of Sequencing By Synthesis (SBS). A library is generated using single stranded, short DNA sequences (or cDNA sequences in the case of RNA-sequencing [265]) which are ligated to adaptors and fixed to a flow cell in the presence of oligonucleotides which complement the adaptors at the 'free' end. This causes the (c)DNA to form a bridge, which has an incomplete complementary strand. Clonal DNA fragments are then produced in a process known as bridge amplification [266]–[268]. The actual sequencing step incorporates four nucleotides (ddATP, ddTTP, ddGTP and ddCTP) with a fluorescent dye and a cleavable blocking group. As each nucleotide is

incorporated the dye is excited by a laser and is detected by a Charge-Coupled Device (CCD).

Illumina HiSeq is able to produce approximately 650GB per run, which takes about 4 days to complete, depending on which system is employed. Illumina MiSeq is even faster using smaller flow cells and reduced imaging times, though this is achieved by producing shorter read lengths [269]. Overall, Illumina produces the biggest output and the lowest running cost compared to competing NGS technologies [263], [266], [268]–[270].

1.6.3 PacBio sequencing

The short read-lengths generated by NGS technologies mean that assembly software must be used to piece together genomes from very short length reads. Progressively innovative algorithms have been produced in an attempt to compensate for this. However, there has been little attempt to compensate for the bias introduced during the amplification step which is affected by factors such as high or low GC content [271].

Single-Molecule Real-Time (SMRT) sequencing, herby referred to as 'PacBio sequencing', is able to produce much longer reads (from hundreds of bases to tens of thousands of bases per read) and sequences from a single molecule of DNA. The lack of an amplification step has led to this technology being referred to as 'Third Generation Sequencing' (TGS) [272]. This is achieved by capturing the sequence data during the replication process of the target DNA. A template is produced using the target dsDNA to which hairpin adaptors are ligated at both ends, creating a dumbbell-shaped template or 'SMRTbell' [273]. A SMRTbell sample is then loaded onto a chip called a SMRT cell where it diffuses into a sequencing unit called a Zero-Mode Waveguide (ZMO) [274]. A ZMO is a unit with a single polymerase molecule fixed at the bottom and contains the minimum volume that allows for light detection. The SMRTbell enters the ZMO, binds to the polymerase via the adaptor and is allowed to begin the replication process [275]. Fluorescent nucleotides are added to the SMRT cell and as each one is added to the SMRTbell a pulse of light at a specific wavelength is emitted and detected by a camera. Depending on the reagents the SMRT bell can be sequenced multiple times (passes) in a Continuous Long Read (CLR). In this instance the adaptors are trimmed, producing sub-reads, and a Circular Consensus Sequence (CCS) is

produced by mapping multiple reads with high accuracy [276]. If the SMRT bell is too long or there are insufficient passes then no CLR is produced and the sub-reads are output instead.

The main advantage of PacBio sequencing is the long read-lengths. At the time of writing, the PacBio RSII system using C4 chemistry is able to produce average read lengths of 10kb [277], compared to Illumina which is typically paired end ~250bp [278]. However, due to physical restraints on the SMRT cells, such as loss of the polymerase or multiple SMRTbells per ZMO, the throughput of PacBio is much less than that of Illumina [276]. It also has a high error rate (11%-15%) for the CLR [279], though this is random and therefore repeated sequencing improves CCS accuracy up to >99% with 15 passes [275]. This means that the read length and read accuracy are a trade-off, as CLR length is determined by the durability of the polymerase [280]. While PacBio is faster than NGS it is also more costly [276].

The long reads produced by Pac Bio mean that it is well suited to the task of *de novo* assembly. The long reads are able to overcome some of the limitation of NGS in producing large scaffolds which are able to cut down computational load during assembly, span repetitive regions and close gaps in existing genome assemblies [281], [282].

1.7 Aims of the Thesis

We have learned a great deal about obligate symbiont genomes and their evolution in the past 10 years. However, these have often lacked appropriate free-living comparators. This thesis will investigate the evolutionary origins of two neglected unicellular symbionts, *Blastocystis* sp. and *Entamoeba histolytica,* utilising free-living counterparts to infer the ancestral state of each genome and explore the adaptations exhibited by these genomes during their transition into the host intestinal environment.

In Chapter Two the genome and transcriptome of *Proteromonas lacertae* strain LA is sequenced, assembled and annotated providing the first draft reference genome and transcriptome for this organism. These resources are then analysed in terms of their contiguousness and completeness in an effort to judge their suitability

as a robust representation of this lineage and therefore their usefulness in a comparative analysis conducted in Chapter Four.

In Chapter Three the transcriptome of the free-living *Cafeteria roenbergensis* is sequenced providing the first transcriptomic dataset for this organism. The transcriptome is then analysed for its usefulness as an out-group to both *Blastocystis* and *Proteromonas* to be utilised in a comparative analysis conducted in Chapter Four.

Chapter Four makes use of the newly generated *P. lacerate* genome and the *C. roenbergensis* transcriptome in a comparative analysis with three previously published genomes for *Blastocystis* sp. ST1 [283], ST4 WR1 [284] and ST7 [91] as well as five other published Stramenopile genomes representing diversity across the Stramenopile clade. This allows the identification of gene families and genomic features absent or specific to *Blastocystis*.

In Chapter Five the transcriptome of free-living *Mastigamoeba* sp. is sequenced and analysed for its completeness. It is then included in a comparative analysis with *E. histolytica* [285], *Mastigamoeba balamuthi* [286] and *Dictyostelium discoideum* [287] in order to investigate adaptations in *Entamoeba* that may be linked to its life strategy and its ability to cause disease.

Chapter 2. Sequence and analysis of the *Proteromonas lacertae* strain LA genome

In order to investigate evolutionary events in the ancestor of *Blastocystis*, a genomic resource for a non-pathogenic outgroup is required. As no such resource exists, it was necessary to produce a *de novo* genome assembly for *P. lacertae* that allows for a comparative analysis with *Blastocystis*. The genome was sequenced using PacBio RSII technology to take advantage of the long read lengths produced by this technique. This data was then assembled and assessed for its completeness and contiguousness. The draft genome produced here is a useful representation of an intestinal Stramenopile and is a good comparator to *Blastocystis* sp. STs.

2.1 Introduction

The intestinal tract of all vertebrates is a melting pot of bacterial and eukaryotic flora each contributing to the environment. Organisms that colonise this environment lie on the spectrum between commensal, parasitic, opportunistic and mutualistic. Eukaryotes associated with the host mucosa, such as *Blastocystis*, are often assumed to be pathogenic [23]. In order to investigate the genomes of *Blastocystis* an appropriate out-group is required. While there has been no attempt to establish pathogenicity of species within the *Proteromonas* genus, no pathogenic effects have been observed in the hosts these protists have been isolated from [288], [289]. In order to include *Proteromonas lacertae* in a comparative genomic analysis and identify changes associated with the *Blastocystis* genomes, a genomic resource for this organism is required. In this chapter, I present the first draft genome and transcriptome for *P. lacertae* and assess its usefulness as an out-group to *Blastocystis*.

2.1.1 What is *Proteromonas lacertae*?

P. lacertae is an anaerobic protist isolated from the intestinal tract of lizards, urodelans and rodents [290], [291]. It is a single-celled protist belonging to the Stramenopiles and thought to be closely related to *Blastocystis* based on morphological and molecular comparisons [65], [291]–[294]. However, unlike *Blastocystis*, it possesses some of the characteristic features used to define the Stramenopiles, in particular it possesses two flagella The anterior flagellum is covered with tripartite hairs (mastigonemes) and the posterior flagellum is slightly

shorter and smooth [294]. Notable features inside include an anterior cell surface which is corrugated, with each fold being supported by a single microtubule and a posterior cell surface covered in tubular hairs (somatonemes) [295]. Sub cellular features include a single, large mitochondrion with tubular cristae often found in proximity to the nucleus [294]–[297] (Figure 2.1.1).

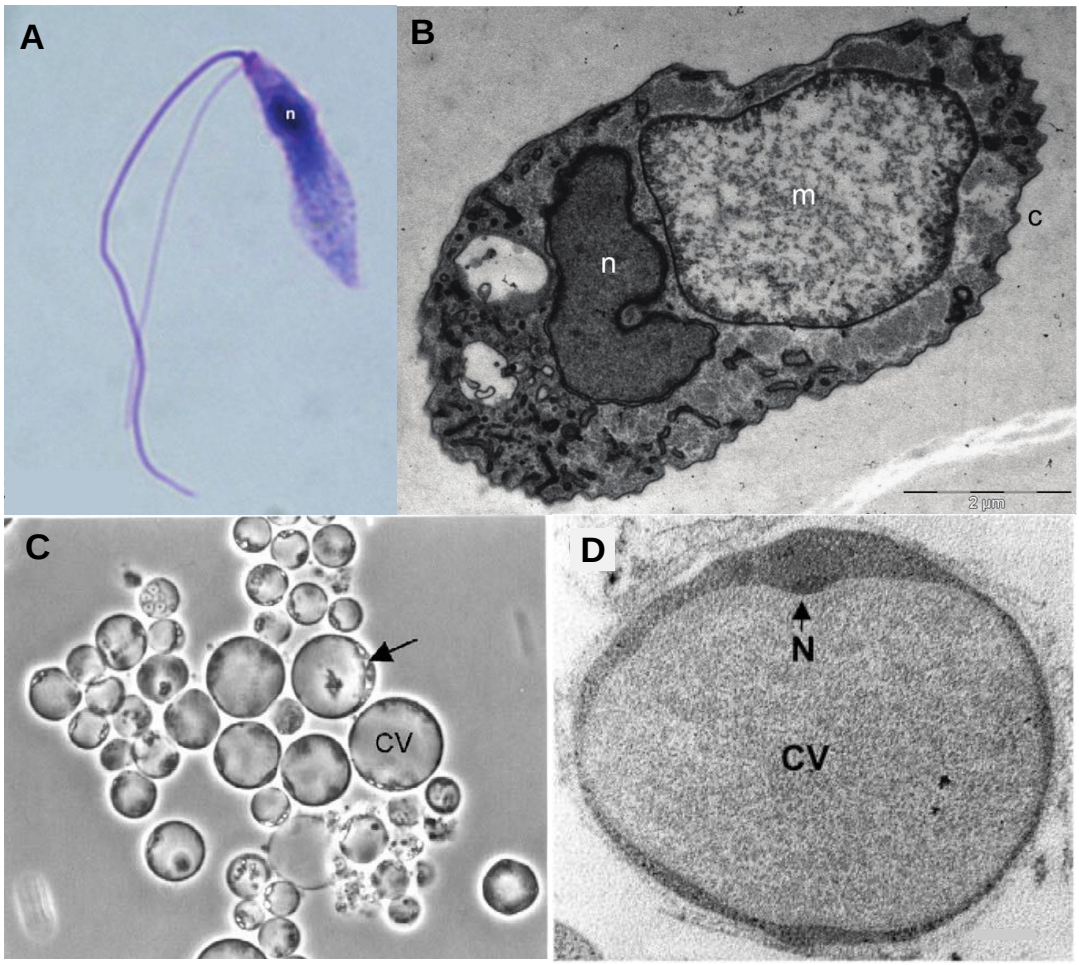


Figure 2.1.1 Light and electron micrographs of* P. lacertae *and* Blastocystis *sp. showing differences in morphology. A. *Light micrograph of a Giemsa-stained cell ~13 x 3 µm from Pérez-Brocal* et al*. 2010* [294]*. The anterior of the cell bears two flagella, one thicker and longer than the other. The single nucleus (n) is visible at the anterior pole of the cell.* ***B.*** *Electron micrograph of* P. lacertae showing the *nucleus (n) is in proximity to the single large mitochondrion (m) and the corrugation of the cell surface (c) is clearly shown.* ***C.*** *Light micrograph of* Blastocystis *from Tan 2004* [38]*. The vacuolar form is spherical with a large central vacuole (CV) and a resultant thin rim of peripheral cytoplasm, where organelles such as the nucleus and mitochondria-like structures reside (arrow).* ***D.*** *Electron micrograph of healthy* Blastocystis *from Yin* et al. *2010* [298] *in vacuolar form with large central vacuole (CV) and crescent shaped chromatin in the nucleus (N).*

In vitro the flagellated cells are highly motile and must be grown in rich medium or in the presence of a bacterial food source, suggesting that *P. lacertae* may utilise a phagotrophic and/or pinocytic life strategies *in vivo*. The cells are also capable of encysting, most likely in order to transmit between hosts via the faecal-oral route [56], [61].

2.1.2 The phylogenetic position of *P. lacertae*

The taxonomic placement of *P. lacertae* has been relatively straightforward, aided by the creation of an axenic culture. SSU rRNA and rDNA sequence data place *P. lacertae* within the Stramenopiles and more specifically as a sister clade to the Bicosoecida and *Blastocystis* within the Slopalinida [63], [65], [291]–[293], [299], [300] (Figure 2.1.2). Its position was used to place *Blastocystis* within the Stramenopiles despite the highly derived morphological characteristics of the latter, such as a large central vacuole and lack of observable flagella *in vitro*. In this chapter, I will attempt to confirm the phylogenetic position of *P. lacertae* through a multigene phylogenomics analysis.

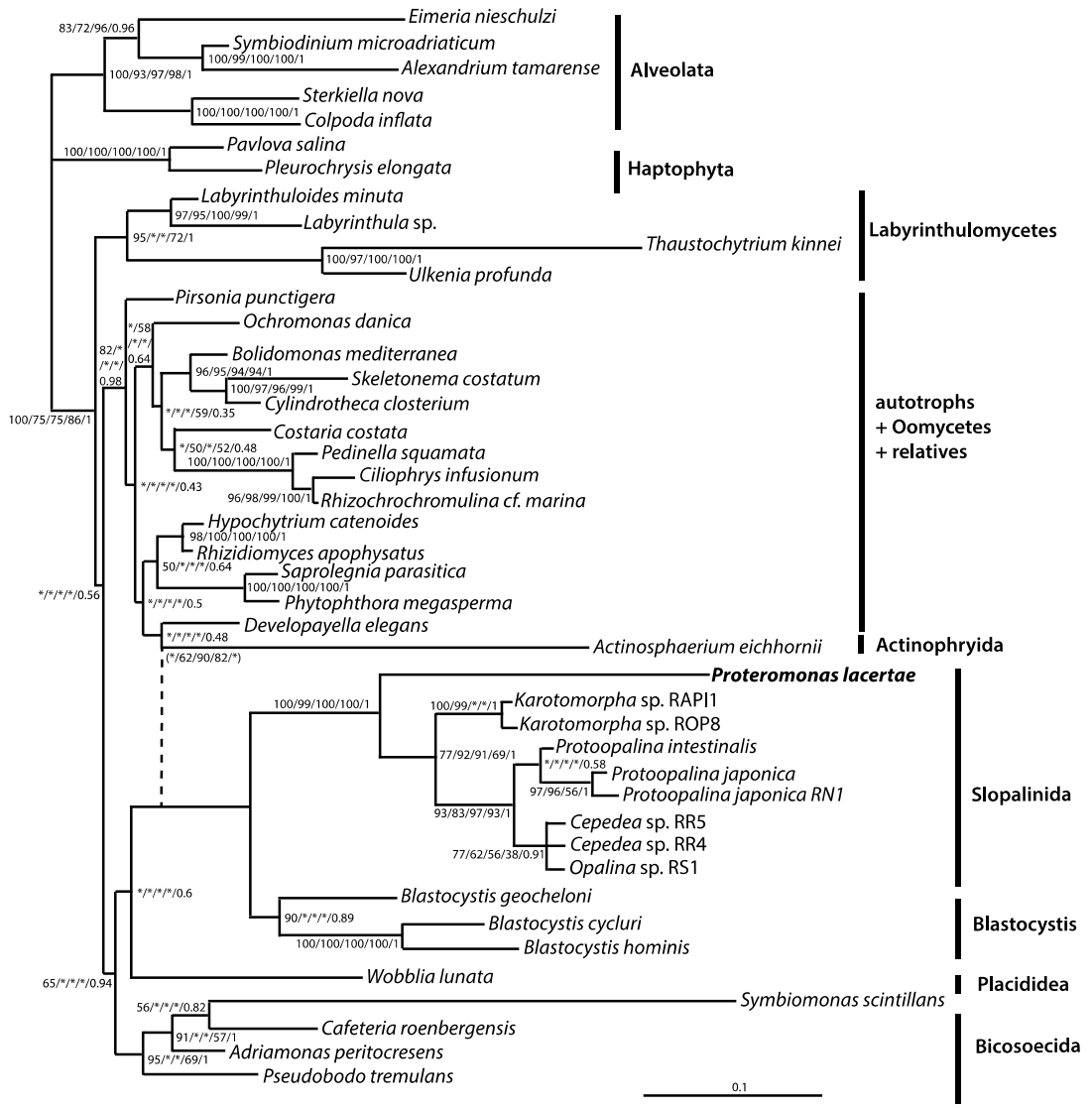


Figure 2.1.2 Bayesian phylogeny of SSU rDNA from 36 Stramenopile taxa and seven Alveolate and haptophyte out-groups. P*hylogeny redrawn from Kostka et al. 2007* [291]***. Bootstrap values from maximum likelihood (100 replicates), maximum parsimony (1000 replicates), Fitch-Margoliash method with Log Det distances (1000 replicates), maximum likelihood distances (1000 replicates) and Bayesian posterior probabilities are shown at the nodes, respectively. Asterisks (*) represent bootstrap value lower than 50%.* A. eichhornii *was shown by maximum Likelihood and Bayesian analysis to be the sister group to* Developayella*, but other methods resolved it as a sister group to Slopalinida and* Blastocystis*. Statistical support for both positions is shown.*

2.1.3 Comparison between *P. lacertae* and *Blastocystis*

The Stramenopiles have been the subject of many investigations as they contain an example of independent chloroplast acquisition [299] and agriculturally

important plant pathogens, the Oomycetes [209], [210], [301]–[306]. However, the intestinal Stramenopiles have not received the same attention. As a result, there are only three annotated genomes for *Blastocystis* sp. and none for closely related genera. *P. lacertae* is a useful organism to investigate as an out-group to *Blastocystis* for a number of reasons. Firstly, it is one of the closest known relatives to *Blastocystis* in phylogenetic analyses [65], [291], [293]. It also inhabits a similar intestinal environment but is not associated with disease and it is easy to culture axenically, which makes it easy to obtain abundant genetic material. For these reasons *P. lacertae* has been used previously to examine the structure and function of the unusual *Blastocystis* MROs. Pérez-Brocal *et al.* (2010) [294] sequenced the mitochondrial genome of *P. lacertae* and found that it was linear, with a unique central region flanked by two inverted repeats. Comparison with the *Blastocystis* MRO sequence showed that there were surprising similarities between the encoded machinery despite the structure and organisation being markedly different. This suggests that despite the divergence between *Blastocystis* and *P. lacertae*, they remain close enough to clues about their common ancestor and therefore shed light on *Blastocystis*-specific adaptations.

2.1.4 'PacBio' sequencing

When producing genome sequences *de novo*, i.e. without existing reference sequences, it is important to maximise sequence contiguity and minimise gaps. There is, therefore, desirable to maximise read length, increase the overlap between tiled reads, and so reduce assembly artefacts. PacBio sequencing is able to produce longer reads than the older technologies (from hundreds of bases to tens of thousands of bases per read) and by sequencing from a single molecule of DNA, rather than requiring an amplification step, it avoids bias in the sequencing process [272]. Despite the increased cost and lower throughput [276], I have chosen to sequence the *P. lacertae* genome using this technology for two reasons. The Long read lengths (up to 10 kb) will mean that a highly contiguous assembly is more likely and less computationally intensive to produce. This will also reduce the likelihood of collapsing repetitive regions and help to sequence closer to regions with unusual composition such as the sub telomeres. The second reason is that, while the molecular composition of the organism was unknown, I wanted to avoid issues arising from bias in the sequencing of regions such as those with low GC content. It is accepted that, while PacBio sequencing has a high error rate relative to other methods, it lacks the systematic bias with respect to specific nucleotide formations

(e.g. purine strings) that affects other methods. The unbiased error rate may be solved by massive parallel sequencing, but the systematic bias of other methods is inherent [271], [275].

2.1.4 Aims and objectives

In this chapter, the aim is to produce a genome and transcriptome for *P. lacertae*, suitable for use in comparative genomic analyses with *Blastocystis* (Chapter 4). There are seven specific objectives:

1. Culture *P. lacertae* cells successfully and obtain electron microscopy images of the ultrastructure
2. Prepare high molecular weight DNA and RNA preps from cultures
3. Sequence the DNA with PacBio and sequence the mRNA with Illumina
4. Optimise the assembly of the genome and transcriptome into contiguous drafts
5. Annotate the genome and provide a measure of completeness for the gene set and transcriptome
6. Survey the general features of the *P. lacertae* genome
7. Build a multigene phylogeny to confirm the position of *P. lacertae* with respect to *Blastocystis*

2.2 Materials and methods

2.2.1 Culture maintenance and origin

P. lacertae was isolated from the rectal contents of a Eurasian sand lizard (*Lacerta agillis*) [290]; the initial isolate contained bacterial and yeast contamination from the intestinal environment. Axenic cultures were established by Prof. Graham Clark (London School of Hygiene and Tropical Medicine) though treatment with antibiotics and repeated migration protocols [290]. Axenic cultures were grown in LYI-S-2 medium as previously described [307]. Cultures were kept upright at room temperature inside a switched-off incubator in sealed glass screw-capped tubes containing approximately 15 ml of medium. Complete medium contained: 0.5% neutralised liver digest, 2.5% yeast extract, 10.0 g glucose, 1.0 g L-cysteine hydrochloride, 2.0 g sodium chloride, 0.2 g ascorbic acid, 1.0 ml of 22.9 µg/ml ferric ammonium citrate, 1.0 g potassium phosphate, dibasic and 0.6 g potassium phosphate, monobasic. Medium was adjusted to pH6.8, autoclaved and supplemented with 10% heat inactivated adult bovine serum and a vitamin mixture.

2.2.2 Preparation of cells for electron microscopy

In order to collect data on the ultrastructure of *P. lacertae*, 1.0 x 10^8 free-swimming flagellated cells at log phase were pelleted and washed in 0.1M phosphate buffer (PB) pH 7.4 to remove growth medium and fixed in 0.1M PB pH 7.4 containing 2.5% electron microscope grade glutaraldehyde. The samples were then passed to Alison Beckett, Department of Cellular and Molecular Physiology at the University of Liverpool. These were washed in 0.1M PB pH 7.4 and fixed with 3% agarose and iced for 30 minutes prior to being cut into pieces. Samples were then stained with 2% OsO_4 and 1.5% potassium ferrocyanide in ddH_2O, 1% TCH in ddH_2O and 2% OsO_4 in ddH_2O. After each stain step samples were heated in a Biowave 20s on, 20s off, 20s on, 20s off; 20Hg and washed 3 x 5 minutes in ddH_2O. The final stain was 1% UA in ddH_2O overnight at 4°C followed with a final wash as described above. Samples were dehydrated in graded acetone dilutions for 8 minutes each on ice (30%, 50%, 70%) and 100% for 2 x 10 minutes. Infiltration was in graded TAAB hard in 100% ethanol (1:1 v/v) and 2 x 100% resin. Samples were then embedded in TAAB hard in mould and cured for 48 hours at 60^0C. Images were taken on an FEI 120kV Tecnai G2 Spirit BioTWIN transmission electron microscope.

2.2.3 DNA preparation, sequencing and QC

To produce genomic information for *P. lacertae* 4.0 x10^7 free-swimming flagellated cells were pelleted at 3000 x g for 10 minutes and genomic DNA was extracted using DNeasy mini kit (Qiagen) according to manufacturer's protocol. Quantification was carried out on a Tecan bio-analyser using Magellan software. This method uses the QuantiFlor dsDNA System that employs fluorescent dsDNA-binding dyes, which reduces interference from ssDNA and RNA; this is then plotted onto a 5-point control curve to determine the concentration in a given sample. Using this method, 20 µg of DNA of high molecular weight was submitted for sequencing in 200 µl total volume.

The single-molecule, real-time (SMRT) sequencing by Pacific BioScience was carried out by the University of Liverpool's Centre for Genomic Research (CGR) [276]. Libraries were prepared by shearing DNA to approximately 10kb fragments and sequencing was done on primarily on a single SMRT cell (P5 / C3 chemistry). The output from the single SMRT cell was assembled using default parameters and the assembly metrics was then plotted (Figure 2.2.1). Further SMRT Cells were sequenced and added to each assembly until the assembly metrics plateaued. Adding SMRT Cell data to each assembly decreased the total number of contigs and increased both the N50 contig length and overall size of the genome smoothly. The increase of maximum contig length was more jagged, however, the addition of the ninth and final SMRT Cell failed to improve the overall genome metrics. Therefore, a total of nine SMRT Cells were used to build the final assembly.

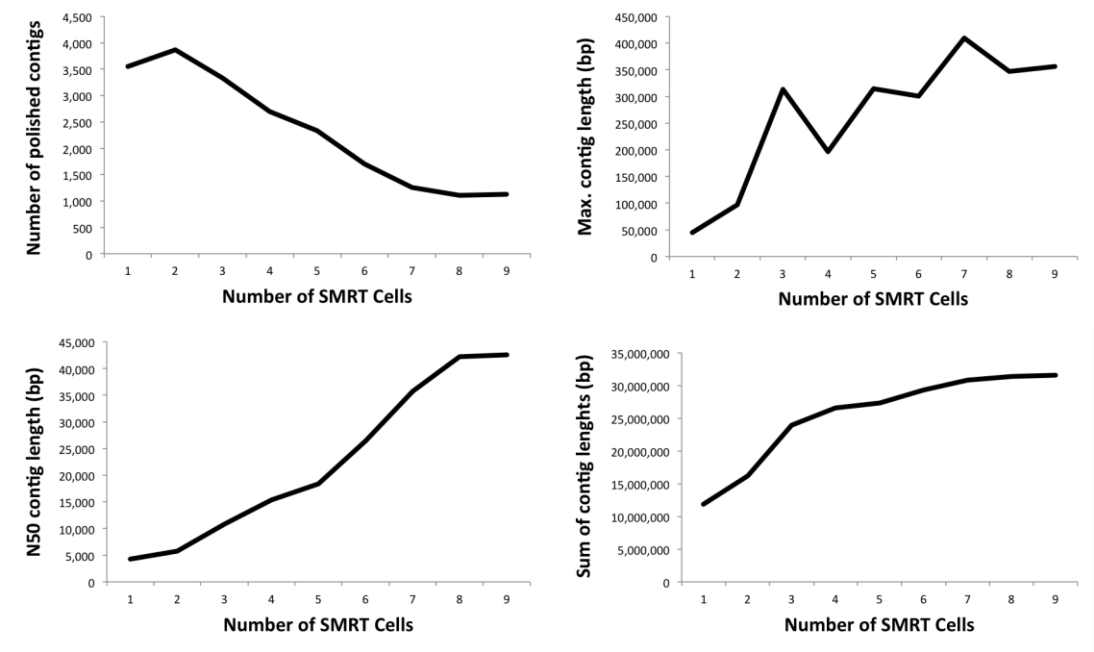


Figure 2.2.1 Assembly metrics for <u>P. lacertae</u> genome assembly when successive SMRT Cells were added. *Assembly metrics including the total number of contigs, the maximum contig length, the N50 contig length and the total size of the genome were used to determine at which point the addition of more SMRT Cells failed to improve the quality of the assembly.*

<u>2.2.4 RNA preparation, sequencing and QC</u>

As no gene expression data exists for *P. lacertae*, whole RNA was isolated to produce a transcriptome. 4.0 x10^7 free-swimming flagellated cells were pelleted at 3000 x g for 10 minutes and whole RNA was extracted using RNeasy kit (Qiagen) according to manufacturer's protocol. Quantification was carried out on a Tecan bio-analyser using Magellan software.

RNA samples from four cultures were pooled and processed by the University of Liverpool's Centre for Genomic Research (CGR) using poly-A selection according to the manufacturer's protocol. These were used to produce three Illumina RNASeq libraries from enriched RNA using the strand-specific ScriptSeq kit. Paired-end sequencing (2x125bp) was carried out on one lane using Illumina HiSeq platform generating in excess of 180M clusters. Post sequencing qc (fastqc) and trimming of adapters resulted in a final dataset of 196,300,790-paired reads.

2.2.5 Assembly of DNA

To assess the potential size of the assembly, kmer frequencies were estimated using Jellyfish v2.2.6 [308] and used to estimate genome size from the filtered reads [309]. This works based on the principle that for a sequence of length l, and a kmer size of k, the total kmers is given by (l - k) +1. This is prone to high error for very short sequences but this decreases exponentially for longer sequences. For genome length sequences, this approach can provide a very accurate estimation of the genome size. Kmers of 19, 21 and 23 were used here; this is a trade-off between accuracy (longer kmers are more accurate) and computational intensity. A range was chosen to ensure consistency. Kmer frequencies were counted and plotted against total number of kmers for a given frequency to calculate the mean coverage. Genome size is then estimated from N = total number of kmers / mean coverage.

In order to produce the best assembly possible, assembly of the raw reads was performed *de novo* by SMRT Portal software (HGAP 3 algorithm) using default parameters except expected genome size which was altered to 35,000,000 to reflect the findings of the Kmer analysis and minimum seed read length which was increased from 6,000 to 10,000. Canu v1.5 was also used to assemble the genome *de novo* with default parameters, except genome size, which was kept consistent between the two assemblers as an independent measure of the quality of the assembly. Canu produced a more fragmented genome assembly consisting of shorter contigs.

In order to obtain an *in-silico* prediction of the ploidy of *P. lacertae* the filtered reads were mapped back to the genome in the SMRT Portal. The resulting BAM file was tested against ideal histograms of base frequencies for diploidy: N(0.5, 0.5), triploidy: N(0.33, 0.04) + N(0.67, 0.04) and tetraploidy: N(0.25, 0.04) + N(0.5, 0.5) + N(0.75, 0.04) using two methods in nQuire [310]. Firstly, this programme performs linear regression on the y-values of the empirical and ideal histograms. Secondly, it utilises a Gaussian Mixture Model (GMM) to describe the data as a mixture of distribution that represents the probability distribution of observations across the whole genome. The likelihood of certain assumptions based on this model given the empirical data is maximised using an Expectation-Maximisation algorithm.

2.2.6 Assembly of RNA

Reads were mapped to both the assembled genome sequence and, initially, to an individual contig using TopHat [311] in order to aid identification of protein coding regions on the genome sequence. In order to assess the completeness of the genome, the transcriptome was also assembled *de novo* by Trinity v2.1.1 [312] using a Kmer size of 25 and default parameters into 27,952 transcripts. A Kmer of 25 was used as per the recommendation of Trinity.

2.2.7 Gene finding and sequence annotation

In order to produce a gene set for *P. lacertae*, two independent gene prediction software packages were used. AUGUSTUS [313] and SNAP [314] both require a 'training set' of protein sequences from the target organism, or related species, in order to customise the parameters to a particular genome in terms of start and stop codons and splice sites. To generate this training set the largest contig from the default SMRT assembly was taken and annotated manually in Artemis v16.0.0 [315] using open reading frames (ORFs) over 100 bp and in conjunction with transcript data mapped to the contig using TopHat. Of the 188 gene models, 78 contained introns. Exons were annotated if there was corroborating transcript evidence for expression, including RNA read pairs spanning the intron, and presence of a splice site. Both AUGUSTUS and SNAP packages were provided with the resulting training set of 188 gene models. Each programme offers slightly different interpretations of gene models leading to large overall differences in the number and structure of the predicted genes. To compensate for this, each gene model was inspected manually to provide consensus across both prediction outputs additionally aided with mapped transcript data (above) and correlation scores. Correlation scores are provided by Artemis and show the correlation between amino acid composition and globular proteins in TREMBLE.

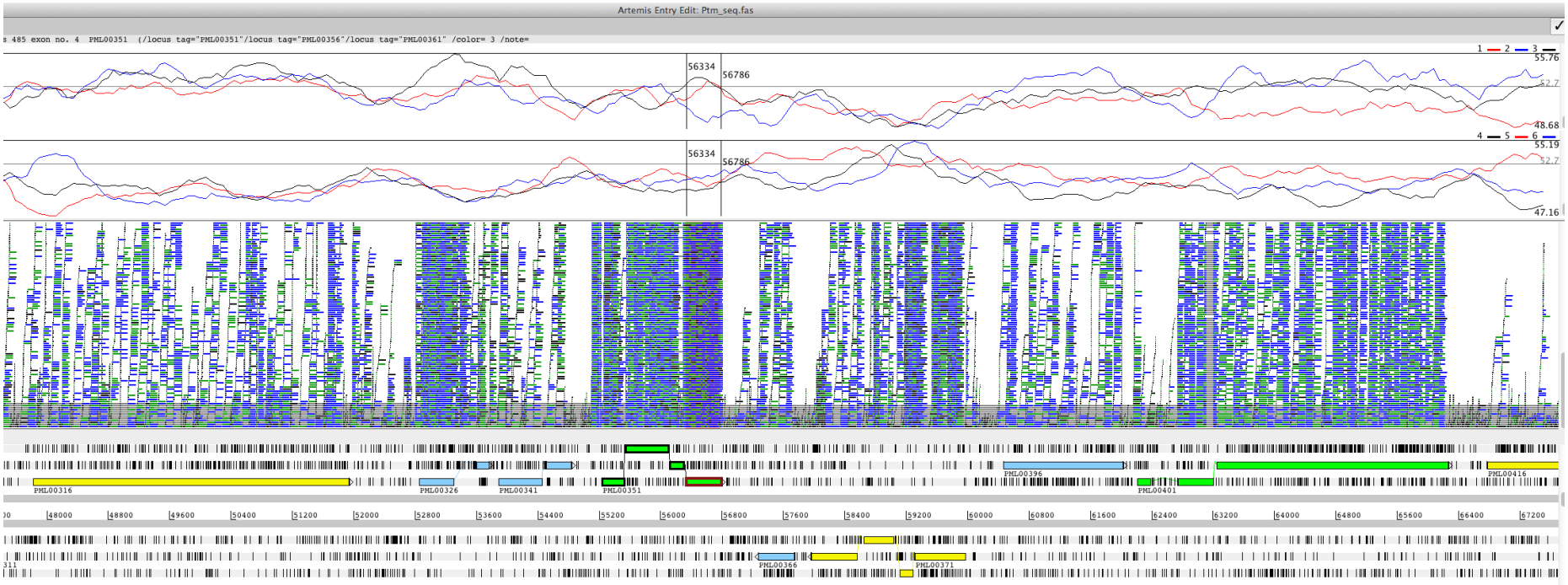


Figure 2.2.2 Snapshot of the integration of data used to aid gene calling in Artemis. *Gene model predictions were provided by AUGUSTUS and SNAP and manually curated using transcript data and correlation scores to provide consensus where the gene model predicted by each package differed.*

Initial annotation of the curated gene models was applied by homology searching using BLAST2GO v4.1 [316] with default parameters. This was done using a BLASTX search against a non-redundant protein database.

To gain more detailed information about gene models the coding sequences (CDSs) were subjected to a suite of independent programmes. InterProScan v5.21-60.0 [317] uses a number of domain databases to assign domain annotations to sequences based on a signature rather than sequence similarity. TMHMM [318] develops a Hidden Markov Model (HMM) in order to identify transmembrane helices in proteins and can distinguish between soluble and membrane proteins. SignalP v4.1 [319] was used to predict cleavage sites and presence and location of signal peptides. ModPred [320] was used to predict post-translational modification sites including GPI-anchor amidation sites, glycosylation sites, phosphorylation and SUMOylation among others. Functional predictions were used to manually refine the gene models for many genes. Other software was applied to better characterise non-protein coding sequences. RfamScan v1.1.1 [321] was used to search a database of various nuclear-encoded RNAs using a HMM for each sequence. RepeatModeler v1.0.4 [322] is a programme which runs two complimentary *de novo* repeats finding programmes (RECON and RepeatScout). Differences in repeat content between assemblies was assessed using a Chi^2 test in R. LTRfinder v1.0.5 [323] identifies full-length LTR retrotranspsons. All software was used with default parameters unless otherwise stated.

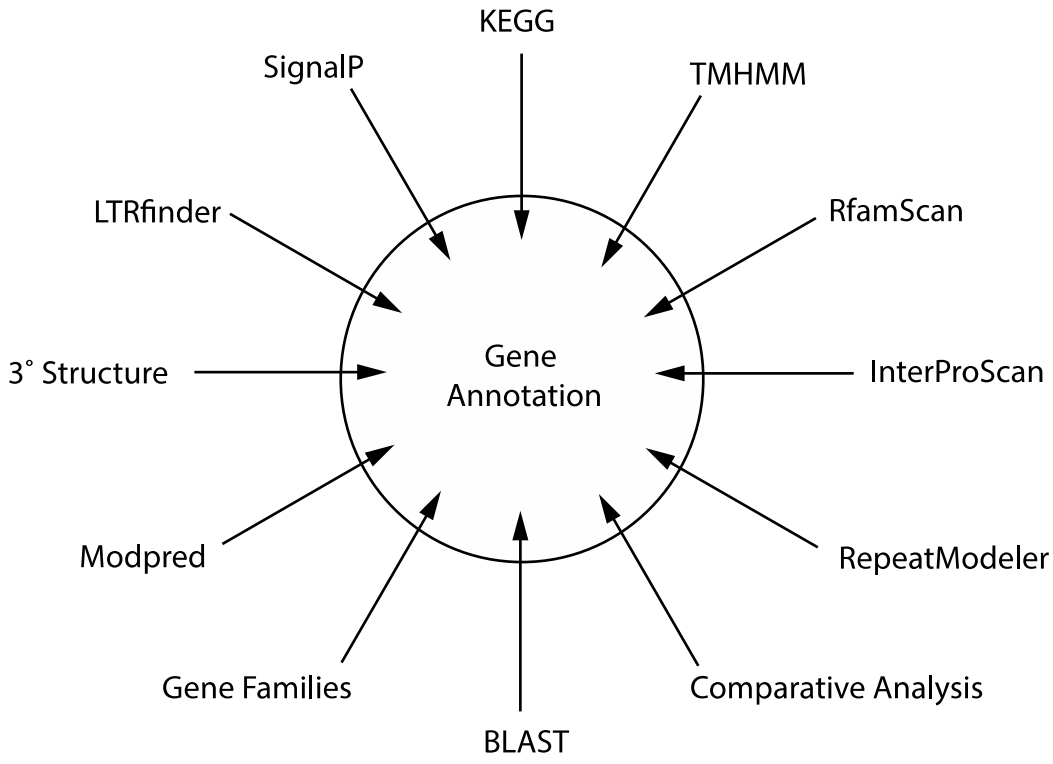


Figure 2.2.3 Sources and information that contributed to gene annotation.

Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations were assigned to protein sequences by GhostKoala [324] and BLAST2GO and were mapped to KEGG pathways using KEGG Mapper – Reconstruct Pathway [325].

2.2.8 Calculating completeness

In order to gain an objective description of how complete the genome was it was given to BUSCO v1.1b1 [326] (Benchmark Universal Single-Copy Orthologs), a programme that assigns a score to an assembled genome based on its BUSCO content. The rationale behind this approach is that all eukaryotes should contain orthologs of a set of 429 core genes. The score is assigned based on how many of these are present in the annotation.

As a validation method for the genome annotation the gene set was tested with a reciprocal BLAST hit to the transcriptome using BLAST. This was to ensure that gene models annotated on the genome reflected assembled transcripts. Genes with no reciprocal hit were examined and splice variants in the transcriptome were included in the dataset. Transcripts with no BLAST result were compared to the

contigs and, where applicable, gene models were modified to reflect transcriptomic data.

2.2.9 Phylogenomics

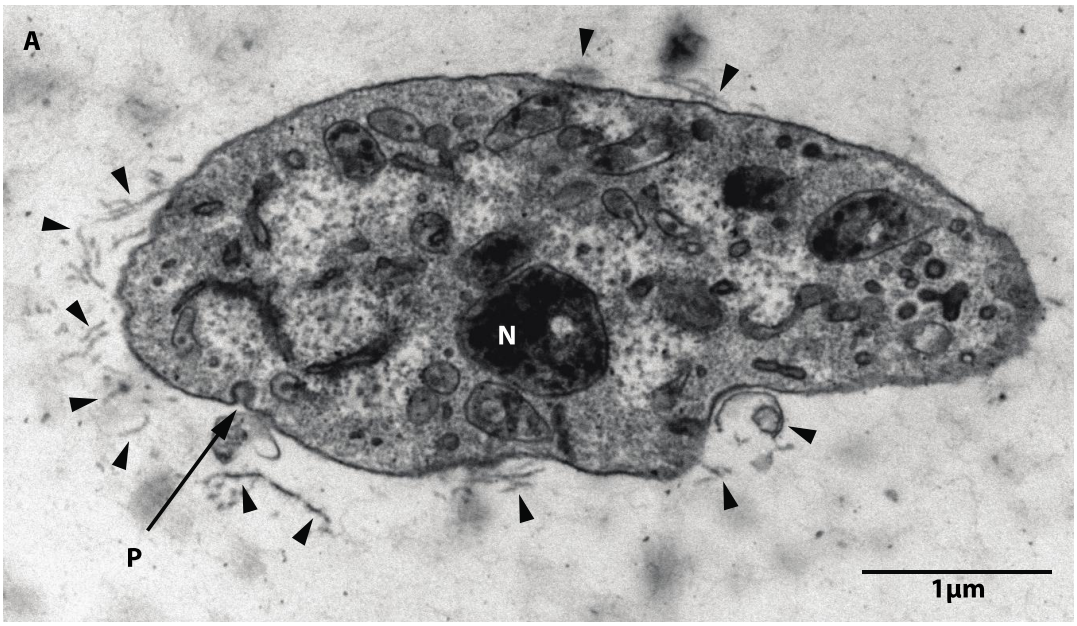
MUSCLE [327] was used to align sequences from 891 orthogroups (Chapter 4) containing orthologues from at least 8 Stramenopile genomes, including *P. lacertae* and *Blastocystis* sp.. Each alignment was passed to Gblocks [328] which curates the alignment by removing segments which are poorly aligned and concatenated into a single alignment containing 83,410 characters. PhyML v3.0 [245] generated maximum likelihood phylograms using smart model selection (SMS) [329] and provided maximum likelihood bootstrap support. SMS uses Akaike information criterion (AIC) [330] to evaluate the quality of statistical models. This is needed to select the substitution matrix and model for rates across sites. Rather than test all combinations, SMS uses a heuristic approach comparable to other model estimation software [329]. MEGA v7.0.14 was used to provide maximum parsimony and neighbour joining bootstraps.

2.3 Results

2.3.1 Electron microscopy

The ultrastructure of *P. lacertae* has been previously reported and its morphology was used to place it with the Stramenopiles prior to the availability of sequence data [294], [297], [331]. I have both confirmed previous findings and report previously unobserved features for *P. lacertae.*

I have confirmed the observation that there is a single, large mitochondrion with tubular cristae and that this is often seen in close proximity to the nucleus, which often contains darkly stained chromatin (Figure 2.3.1). I also observed the Golgi apparatus although no rhizoplast, which has been shown to pass through the Golgi. Extracellular features include the flagella, but no evidence for the kinetosome and the corrugation of the cell surface, which is clearly visible only in certain planes and sections of the cell.



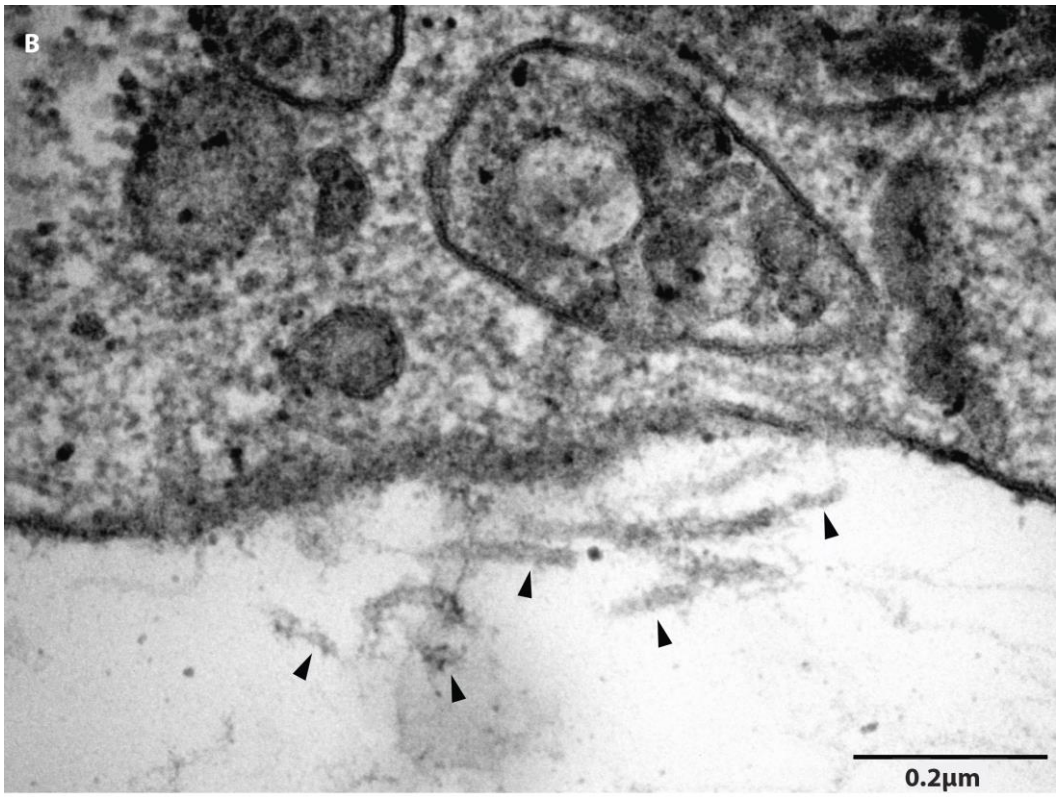
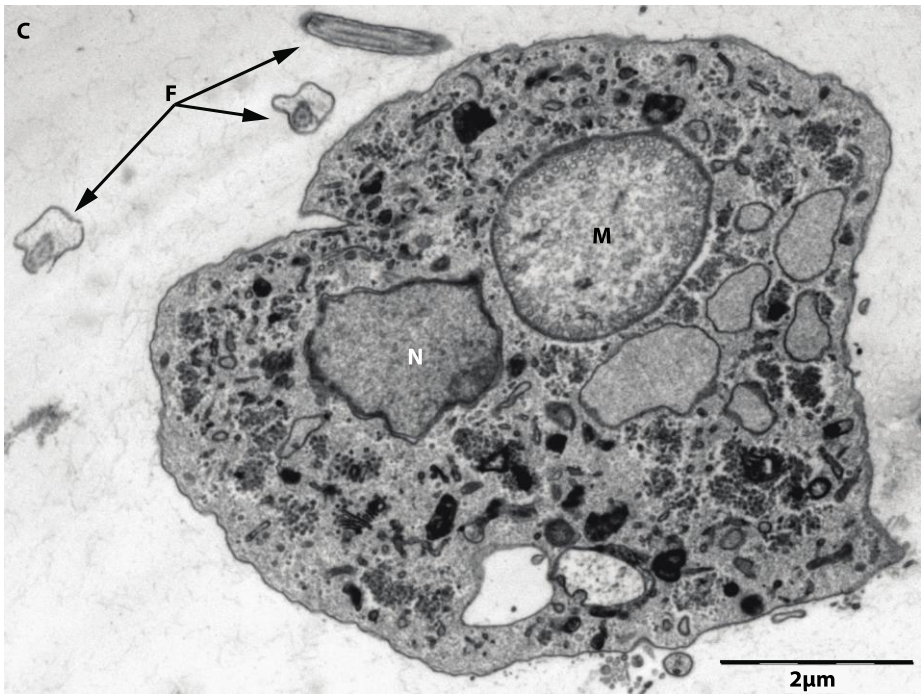


Figure 2.3.1 Transmission electron micrographs of* <u>P. lacertae</u>**. *1.0 x 10^8 cells were fixed in 3% agarose, stained with 1% UA overnight and embedded in TAAB hard mould. Images were taken* an *FEI 120kV Tecnai G2 Spirit BioTWIN transmission electron microscope.* ***A*. Cell with an electron-dense nucleus (N) and showing parts of the surface coat, which have detached (arrowheads) and potential pinocytosis (P).* ***B****. A close-up of the surface coat, which has clearly disconnected from the body of the cell (arrowheads).*



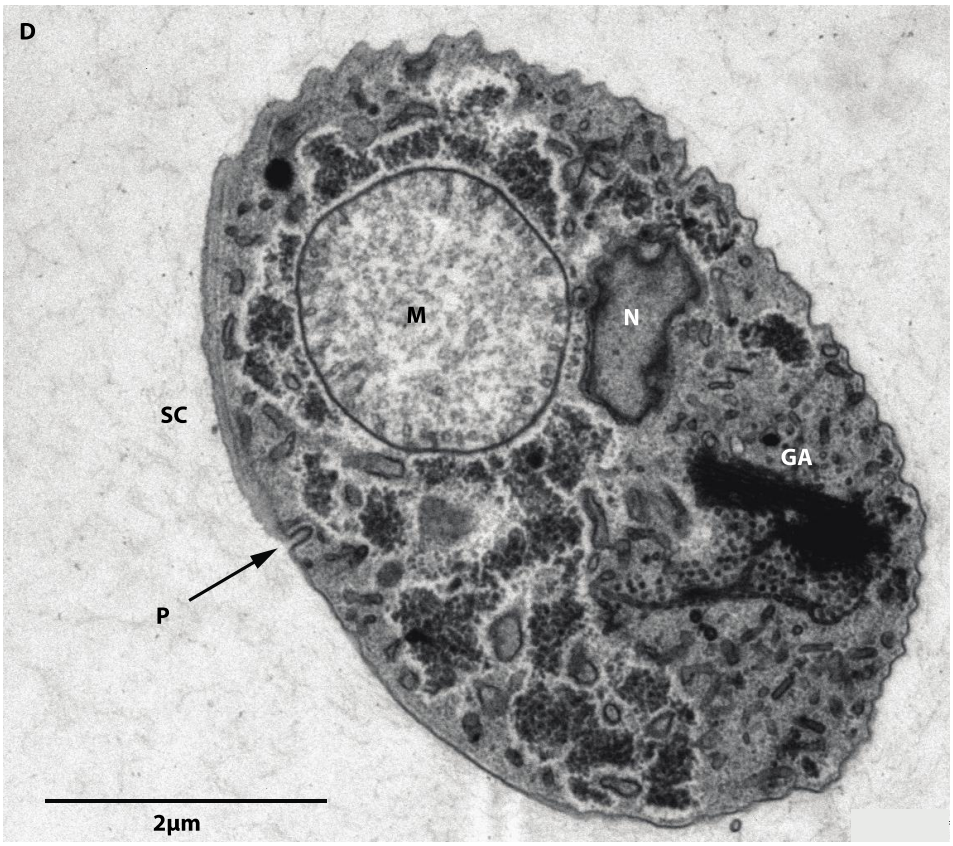


Figure 2.3.1 Transmission electron micrographs of* <u>P. lacertae</u>**. ***C. *Cell with the large mitochondrion (M) in proximity to the nucleus (N) with chromatin. The flagella are also clearly seen (F).* ***D***. *Single cell showing the large mitochondrion (M) in proximity to the nucleus (N) and the very electron dense Golgi apparatus (GA). The surface coat (SC) is seen here still attached to the cell membrane and there is another incidence of pinocytosis (P).*

While the complexity of *P. lacertae* ultrastructure is markedly different from that of *Blastocystis*, there are some interesting similarities with *Blastocystis*, some of which have not yet been reported. There are multiple examples (Figure 2.3.1 A and C) of small, pinocytic pits in the membrane, which are distinct from cell surface corrugation because they are more electron dense than the surrounding membrane and cytoplasm. These may be examples of clathrin-mediated endocytosis, whereby the cell is able to take up nutrients from the environment in the absence of a bacterial food source.

It also appears as though *P. lacertae* may possess a surface coat similar to that exhibited by *Blastocystis,* sometimes referred to as the fibrillar layer, slime layer or capsule (Figure 2.3.1 A and C). The surface coat of *P. lacertae* is homogeneous but may be being continuously shed into the environment by an as yet unidentified mechanism (Figure 2.3.1 A and B). This behaviour has been observed in *Blastocystis* but a function has not yet been confirmed [332].

2.3.2 Genome assembly

In order to produce a genome sequence a single 10 kb library was sequenced on a PacBio RS II machine in 9 SMRT cells. This generated 278,301 paired end reads where the mean read length was 9,642 bp and the N50 was 14,691 bp, of these, 267,828 were mapped.

Genome assembly was performed using SMRT Portal. The reads were initially assembled using default parameters in order to provide a benchmark for subsequent analysis (Figure 2.3.2). In order to optimise the assembly, the genome size was estimated from the filtered reads using a kmer frequency analysis. Kmer frequencies were counted and plotted against total number of kmers for a given frequency to calculate the mean coverage. Genome size is then estimated from N = total number of kmers / mean coverage. Kmer frequency estimation suggested a genome size of between 35 and 37 Mb for kmer sizes of 19, 20 and 21 bp. The predicted genome size was used along with a range of other genome sizes to sequentially investigate the best assembly. Other parameters that were also altered include minimum seed read length and maximum divergence. The best metrics were then combined in an assembly using an estimated genome size of 35 Mb and a minimum seed read length of 10,000, which yielded 1,449 contigs where the maximum contig length was 864,525 bp, the N50 contig length was 92,586 bp. It is

worth noting that although the predicted genome size used in the assembly was 35 Mb, the assembly produced totalled 52 Mb. Multiple iterations of this process failed to reconcile these values and increasing the predicted genome size failed to reproduce this increase (Figure 2.3.2).

The contigs were concatenated where the shortest contig was 1003 bp and the average CG content across the genome was 27.1 % (Table 2.3.1). To test whether this assembly was better than one produced using a different approach, the reads were also assembled using Canu. The reads were assembled into a 55.8 Mb genome consisting of 3,068 contigs where the N50 was 31,981 bp and the max contig size was 450,152 (Figure 2.3.2). As this assembly is more fragmented than the HGAP assembly, the HGAP assembly was preferred to the Canu one.

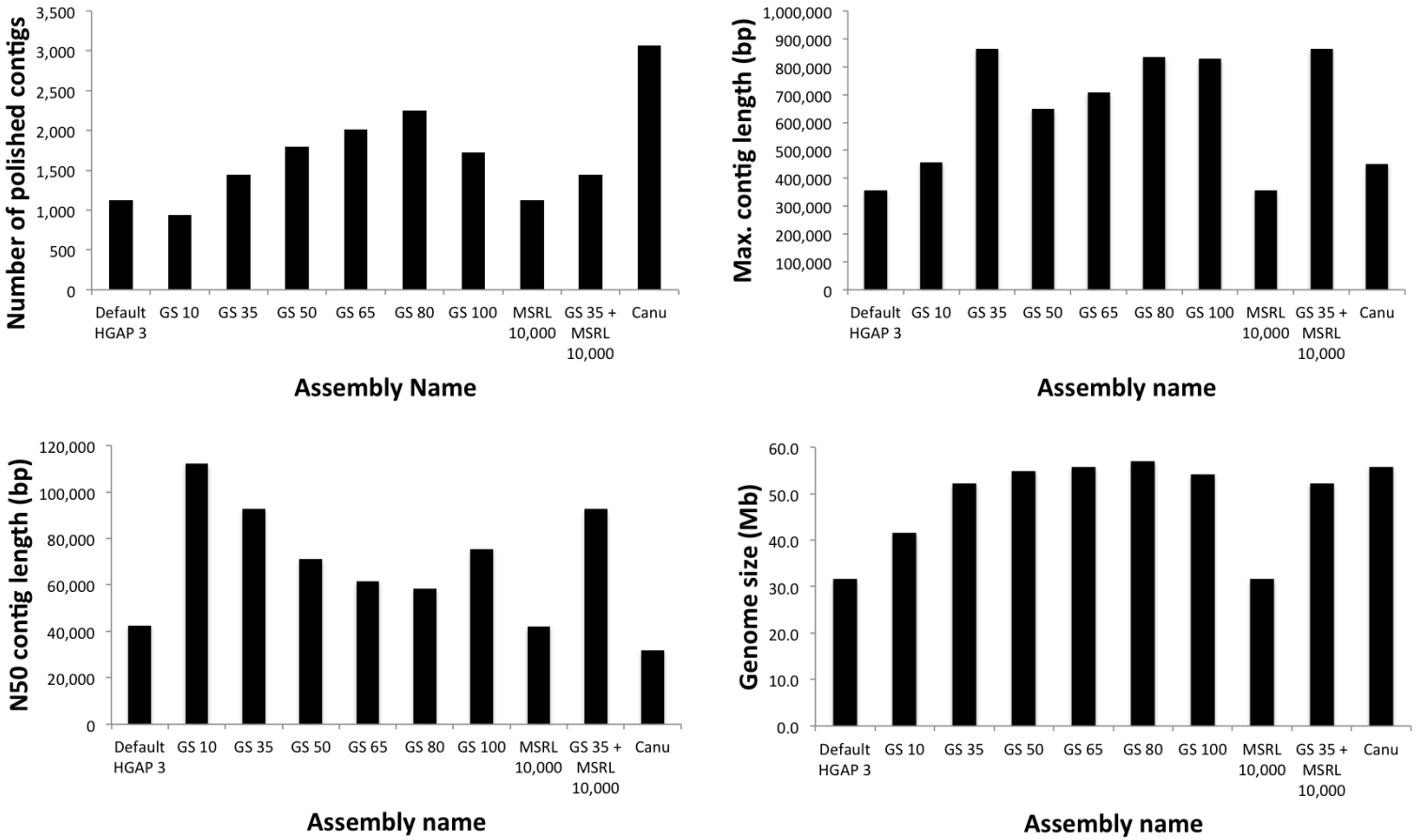


Figure 2.3.2 Assembly statistics for <u>P. lacertae</u> genome assemblies to determine the optimum parameters. *Genome size (GS) and minimum seed read length (MSRL) were altered sequentially and compared to assemblies produces by HGAP 3 and Canu with default parameters.*

Table 2.3.1 Assembly and genome statistics for *P. lacertae, Blastocystis* sp. ST1, ST4 and ST7

	P. lacertae	*Blastocystis* sp. ST1	*Blastocystis* sp. ST4	*Blastocystis* sp. ST7
Scaffolds	-	580	1,301	54
Contigs	1,449	1,092	1,355	155
N50 (bp)	92,586	36,659	29,524	296,810
Genome size (Mb)	52.25	16.4	12.92	18.82
Gene number	35,706	6,544	5,707	6,020
GC (%)	27.1	53	39.7	45.2
Average coverage	54.3	80	300	12.4
BUSCO (%)	85.55	85.08	81.81	78.79
Total gene length (Mb)	33.1	11.6	7.9	7.8
Total gene length (%)	63	70	61	41
Gene density (Genes/Mb)	684	399	442	320
Intergenic length (Mb)	19.1	4.8	5	11

The assembly statistics of the genome are comparable to the published genomes of *Blastocystis* sp. STs, which are the closest relatives with genome data available. The largest difference comes from the size disparity; the *P. lacertae* genome is much larger than the *Blastocystis* sp. ST genomes. As a result, the genome is made up of more contigs, though the N50 value is higher than those for *Blastocystis* sp. ST1 and ST4, so there is confidence that the genome is not more fragmented. The proportion of the genome that is protein coding is similar between the *P. lacertae* genome and *Blastocystis* sp. STs, contributing between 60 – 70% with the exception of *Blastocystis* sp. ST7. Differences are notable for GC content, and gene density, which are lower and higher respectively in *P. lacertae*. The BUSCO score for *P. lacertae* can be broken down into details, the gene set contains 245 complete single-copy BUSCOs, 101 duplicated BUSCOs, 122 fragmented BUSCOs and only 62 BUSCOs missing rendering the overall completeness >85%, again comparable to the existing *Blastocystis* genomes.

As this assembly is composed of long reads it is likely that the approach has managed to sequence over long repetitive regions and therefore is a good estimation of the actual genome size. Interestingly, during the assembly process, the default parameter for genome size estimation was set at 5 Mb (the HGAP algorithm was optimised for bacterial genomes). When assembled under default parameters, the whole assembly was compressed in to 31 Mb (Figure 3.2.2). Increasing the expected genome size to 35 Mb had the effect of almost doubling the size of the assembled genome to 52 Mb and the N50 to 92,586 bp. This was

potentially due to collapsing repetitive regions and discarding reads that were too divergent.

To test this, RepeatModeler predicted repeats across both the 31 Mb assembly (default HGAP) and the 52 Mb assembly (optimised HGAP). It showed that the default assembly was repetitive over 24% of the total length, whereas the optimised assembly was repetitive over 37% of the total length, suggesting that collapse of repetitive reads may have been partially responsible for the reduction in genome size. Distribution of repeat size was also significantly different between the two assemblies, X^2 (20) = 577.79, p < 0.0001, meaning that the repeats in the default assembly were significantly shorter. Increasing the estimated genome size up to 100 Mb was not associated with a similar expansion of genome size.

2.3.3 Gene annotation

There were 188 manually generated gene models used to train AUGUSTUS and SNAP. 78 (41%) of these had multiple exons the rest were single exon. 24 (12.7%) of these genes did not have a BLAST hit to a sequence in the database. Of the remaining gene models, 32 (17.0%) were best matched to proteins annotated as 'hypothetical protein' and the rest had BLAST hits to annotated proteins of various functions. Each gene was >100 bp beginning with a start codon and ending with a stop. All 188 gene models were annotated in conjunction with transcript data mapped to the longest contig in the assembly.

The gene model predictions from AUGUSTUS and SNAP differed in both the number of genes and their positions within the genome. AUGUSTUS predicted 22,869 genes while SNAP predicted 151,750. This difference was resolved by manual curation of each gene model aided by the inclusion of transcript data mapped to the whole genome. This curation of the gene models resulted in 40,042 predicted genes. These genes were given functional annotations by a suite of programmes, which allowed refinement of some of the models. In some cases where adjacent genes received the same annotation or where a single gene received multiple annotations gene models were revised, again with the support of transcript data. This refinement resulted in 35,685 gene models of which 28,067 (79%) had transcriptomic support.

As a validation method for the gene models the *de novo* assembled transcriptome was compared to the gene set by reciprocal BLAST. 13,950 (39%) of the gene models had a reciprocal best hit to a transcript and *vice versa*.

2.3.4 Transcriptome assembly +QC

In a reciprocal BLAST search between the gene set and the transcriptome 13,950 of the 27,952 transcripts (50%) had a reciprocal best hit to a gene model. However, this may have been due to the splice variants in the transcriptome. If there were several isoforms of a gene expressed in the transcript only one would receive a reciprocal BLAST hit. When investigated, it transpired that without the splice variants, the transcriptome consists of 15,101 transcripts of which only 8563 (56%) had a reciprocal best hit to the proteome. This is very similar to the former result and suggests that there is some disagreement between the transcriptome and the predicted proteome, however, 28,067 gene models have transcript support when the reads are mapped to the contigs.

In order to get an independent measure of the assembly quality it was also assessed by BUSCO, which measures how complete the transcriptome is by measuring how many BUSCOs it contains. BUSCO estimated that the *P. lacertae* transcriptome is 75.9% complete.

2.3.5 General features of the *P. lacertae* genome

The ploidy of *P. lacertae* and *Blastocystis* sp. STs are not known, so in an effort to investigate the features of the *P. lacertae* genome and gain insight into the change in genome size between assembly parameters, the genome was used in an *in-silico* analysis of ploidy. Reads were mapped back to the genome and the base frequencies were then tested against assumptions based on predicted frequencies for diploid, triploid and tetraploid genomes. The programme nQuire does this in several ways firstly by calculating the sum of squared residues (SSR) of empirical vs. ideal frequencies. The assembly showed best fit against the tetraploid ideals showing a low SSR, a positive slope and a high r^2 (SSR = 0.0482166, y-y = 0.59435, r^2 = 0.0554398), compared with the fit against the diploid (SSR = 0.102001, y-y = -0.1535, r^2 = 0.01) and triploid (SSR = 0.0754403, y-y = -0.181506, r^2 = 0.0125269) frequencies. The nQuire programme also uses a Gaussian Mixture Model to calculate delta log-likelihoods for each model. Again the *P. lacertae*

genome scored highest when compared to the tetraploid (80860430.427226) model and negatively for both the diploid (-837038.879166) and triploid (-545848.790561) models (Figure 2.3.3).

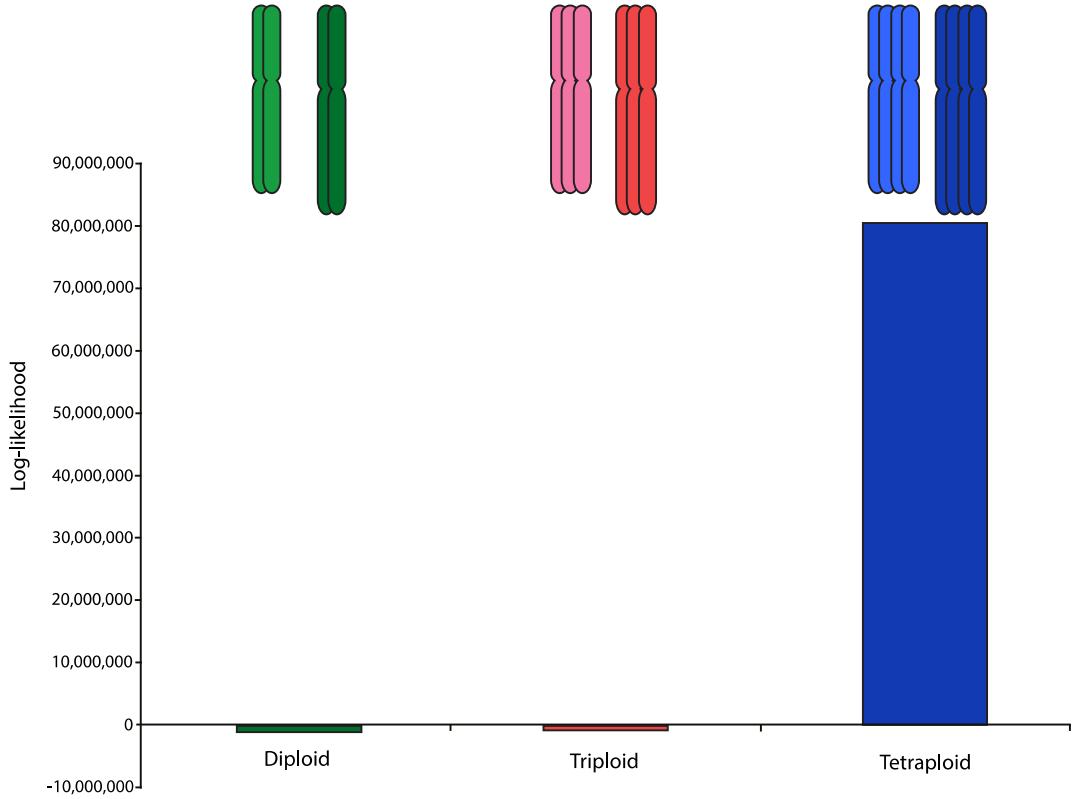


Figure 2.3.3 Graph of delta log-likelihood values for the <u>P. lacertae</u> genome compared against diploid, triploid and tetraploid models.

The genome assembly is comprised of contigs each representing a continuous region of DNA. A search of the contigs using BLASTn revealed the presence of telomeric repeats (CCCTAA) at the ends of 55 contigs. Examination of these contigs revealed that there appear to be no subtelomeric regions; the coding regions end abruptly with the telomere. No contigs were observed with telomeric repeats at both ends, suggesting that there are no contigs representing whole chromosomes.

<u>2.3.6 Protein coding genome</u>

The majority of the proteins coding genes have no significant similarity with databases and are therefore annotated as ‘hypothetical protein’. This accounts for 65% of the predicted proteome. This is comparable to the *Blastocystis* sp. STs

genomes where ‘hypothetical’ or ‘uncharacterised’ proteins account for 94%, 49% and 34% of the genomes for ST7, ST4 and ST1 respectively. For other published Stramenopile genomes (Chapter 4) this proportion ranges from 97% for *Phytophthora sojae* to 28% for *Pythium ultimum*. These proteins are often annotated by software packages with vague domain descriptions that do not provide useful information on the function of the protein, such as repeat domain-containing or transmembrane superfamily, which may refer to a wide array of proteins and functions. The annotations performed by InterProScan assigned 4,015 domains to 17,832 proteins.

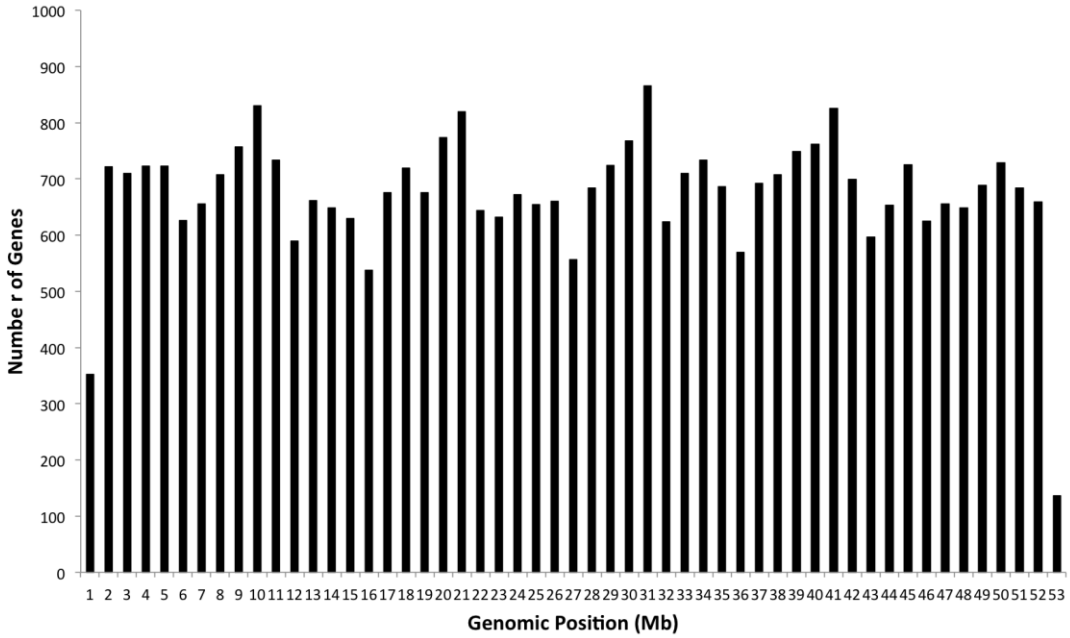


Figure 2.3.4 Gene density plotted across the* <u>P. lacertae</u> *genome. *Number of genes is plotted against 1 Mb bins.*

P. lacertae has an average gene density of 673 genes per Mb. While there are some regions with slightly lower or higher gene density overall it is fairly consistent across the genome (Figure 2.3.4). The gene density does however, appear to drop off at both ‘ends’ of the genome. The contigs at the lowest gene density are in the final (53rd) bin. Because the genome is less than 53 Mb in length, this bin accounts for only 0.25 Mb, which may explain its low density. However, position 1 accounts for a full Mb but only contains half the average number of genes. Examination of contigs 28 and 973 which contribute to this Mb revealed few insights, other than some large proteins 00473 (59 Kb, including introns) and 00351 (41 Kb, including introns) in this region, there is nothing to suggest such a drastic reduction in gene density.

2.3.7 KEGG mapping

Annotation and mapping of the *P. lacertae* genome to KEGG pathways revealed that 4,490 proteins mapped to 328 pathways (Figure 2.3.5). Thus, the *P. lacertae* genome contains representatives from the most common pathways expected for a unicellular protist, such as lipid, amino acid, carbohydrate and vitamin metabolism. Interestingly, there seem to be components from the citrate cycle (TCA) missing, including citrate synthase (K01647, EC:2.3.3.1), aconitate hydratase (K01681, EC:4.2.1.3) and isocitrate dehydrogenase (K00031, EC:1.1.1.42). This might be linked to the anaerobic environment that *P. lacertae* inhabits. The pathways with the most genes mapped to them included 03040 spliceosome (77 genes), 03010 ribosome (73 genes), 03013 RNA transport (62 genes) and 00230 purine metabolism (62 genes). These may suggest that high copy number of genes in these pathways is important for *P. lacertae* persistence.

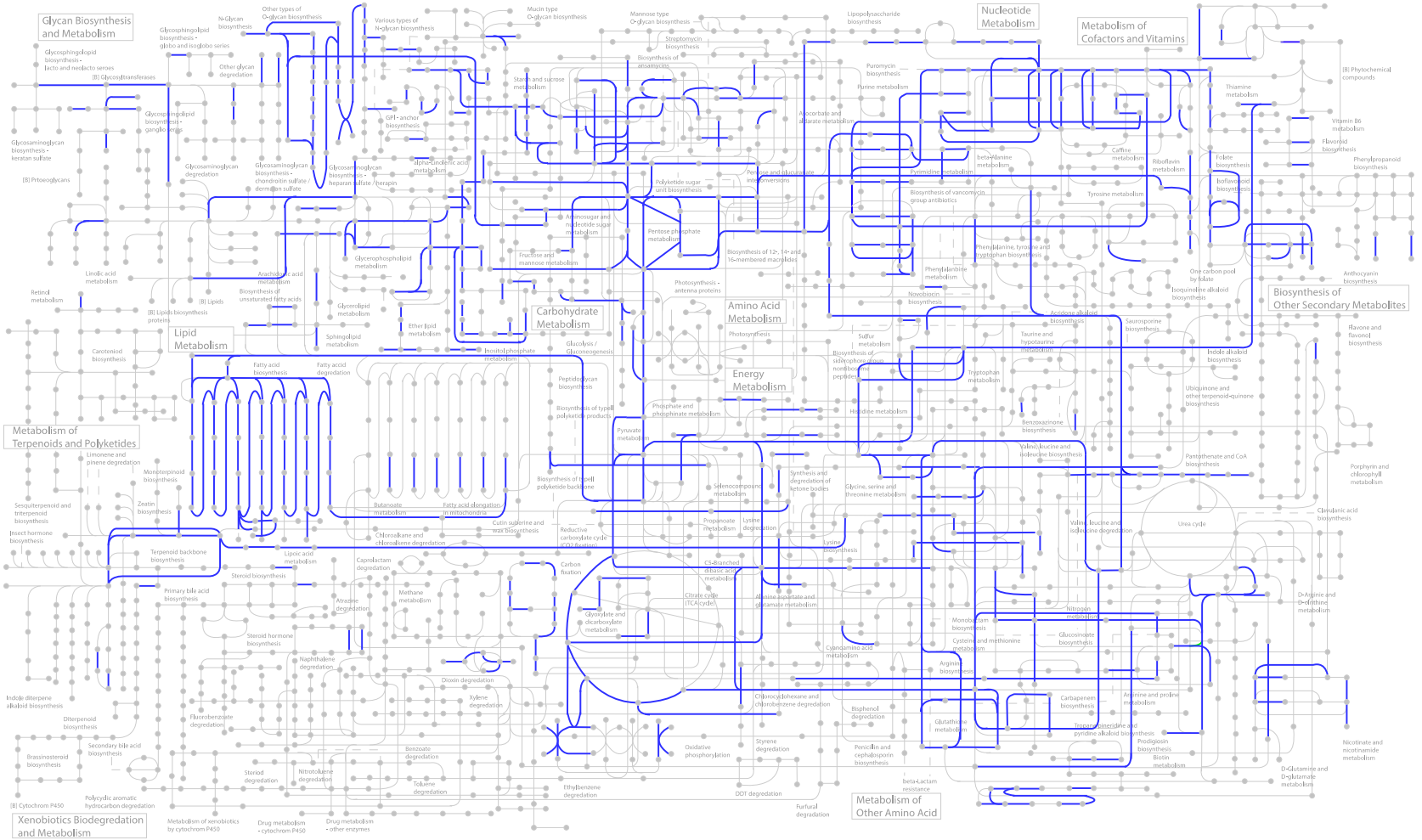


Figure 2.3.5* <u>P. lacertae</u> *protein dataset mapped to KEGG pathways. *Blue lines indicate where proteins with KEGG annotations mapped.*

2.3.8 Phylogenomics

The closest relative of *Proteromonas* is *Blastocystis* however; this genome was produced with the intention to study aspects of the *Blastocystis* genus. Therefore, using *Blastocystis* sp. STs genomes to inform annotation of the *P. lacertae* genome could potentially skew a comparative analysis. There are other published Stramenopile genomes, which may offer comparison, however, these are more distantly related to *Proteromonas* and are therefore less useful. In order to confirm the placement of *P. lacertae* within the Stramenopiles and in proximity to *Blastocystis* I built a multigene phylogeny using conserved orthologues from published Stramenopile genomes (Figure 2.3.6).

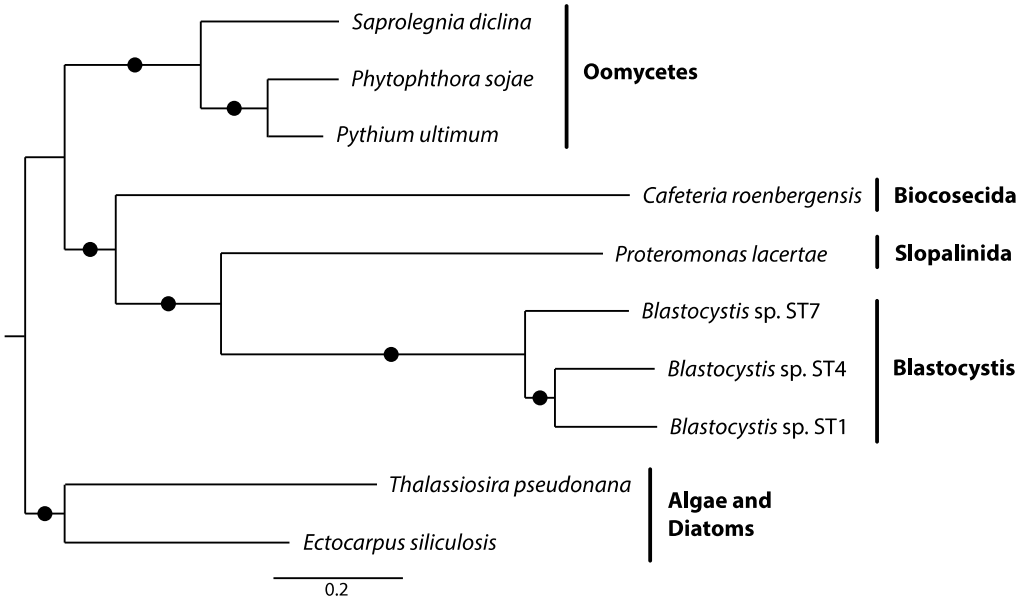


Figure 2.3.6 Maximum likelihood phylogeny of Stramenopile genomes showing the relationship between P. lacertae and other published genomes. *Alignment was generated from representative sequences from 528 conserved orthogroups. The alignment contained 83,410 informative sites. Nodes marked with a point are supported 100/100/100 by bootstraps from maximum likelihood, maximum parsimony and neighbour-joining, 100 replicates each. The tree was rooted with the algae and diatom sequences following Pérez-Brocal* et al. *2010* [294].

The phylogeny robustly places *P. lacertae* as the sister taxon to *Blastocystis,* confirming previous results but with a much-improved character sampling. There is also high bootstrap support for the Oomycete and photosynthetic clades as well as for the proximity of *Cafeteria roenbergensis* to *Blastocystis* and *Proteromonas*.

2.4 Discussion

I have investigated the ultrastructure of *P. lacertae* and made observations about its behaviour *in vitro* consistent with those seen in *Blastocystis*. I have sequenced and assembled the first draft genome for *P. lacertae* and determined that the assembly is contiguous and that the gene set is >85% complete. I have also assessed the relevance of *P. lacertae* as an out-group to *Blastocystis* by confirming their close relation in a multigene phylogeny.

2.4.1 Electron Microscopy

I have confirmed the observation that the ultrastructure of *P. lacertae* is highly different from the derived morphology of *Blastocystis* but I have also shown there are surprising similarities between them. Of particular interest is the occurrence of the surface coat, previously only seen in *Blastocystis* sp. [59], [333]. Further, I have shown that this coat may be extruded into the environment in a similar manner to *Blastocystis* [332], [334]. Zaman *et al.* 1997 [332] were the first to record this phenomenon in *Blastocystis* and postulated that it may be involved in capture and degradation of surrounding bacteria or, less convincingly, involved in resistance to changes in osmotic pressures. It is possible that this mechanism is also used in immune evasion. There are several examples of symbionts that utilise membrane shedding to avoid detection by the hosts immune system including helminths [335], and *Entamoeba* which moves bound immunoglobulins and other bound factors to the uroid region and is extruded as a vesicular cap [336], [337]. The observation of this behaviour in *P. lacertae* may suggest that this machinery was present in the intestinal ancestor of both *Proteromonas* and *Blastocystis* and may therefore play a role in persistence in the intestinal environment.

2.4.2 Evaluation of genome assembly

Two independent software packages were used to assemble the *P. lacertae* genome. The assembly produced by HGAP was preferred because it produced a less fragmented assembly. The biggest difference was between the number of contigs. The Canu assembly was made up of more than twice the number of contigs. This is also reflected in the maximum contig length, which was halved in the Canu assembly, and the N50, which was tripled in the HGAP assembly. While there have been studies which have attempted to compare genome assembly software for

TGS technologies [338], [339], there is no direct comparison between the HGAP algorithm and Canu. Judge *et al*. 2016 [338] compared the assemblies produced by three software tools (PBcR, Canu and miniasm) using MinION sequencing reads and concluded that Canu produced the best result both in terms of low error rate and high contiguity. Liao *et al.* 2015 [339] compared PBcR with HGAP and concluded that both these technologies were able to produce high quality genome assemblies. While this is not a direct comparison, these studies suggest that the Canu assembler might produce higher quality assemblies; however, both of these studies were conducted using bacterial genomes. We have seen that HGAP produces a more contiguous assembly than Canu for the *P. lacertae* genome, which may suggest that the size and complexity of the genome is an important factor.

While the assembly statistics for the *P. lacertae* genome are comparable with the *Blastocystis* ones, the *Blastocystis* sp. STs genomes were sequenced using 454, Illumina or Sanger technologies. As PacBio sequencing is able to produce longer reads than these other methods, it might be expected the assembly for the *P. lacertae* genome would be more contiguous and made up of larger contigs. When compared to the statistics for the *Blastocystis* sp. ST7 assembly, the *P. lacertae* assembly contains more contigs and a lower N50 value (Table 2.3.1). However, it is worth noting that contig number and N50 are not independent of genome size. The *P. lacertae* genome is more than double the size of the *Blastocystis* sp. STs genomes but does not contain double the number of contigs or half the N50. Thus, it may be concluded that the *P. lacertae* genome is not as fragmented as the *Blastocystis* sp. STs genomes.

The *P. lacertae* HGAP assembly compares well to other Stramenopile genome assemblies produced using PacBio sequencing technology: the genome of Oomycete *Phytophthora nicotianae* (unpublished) is 71 Mb but is made up of 5,654 contigs with an N50 of 27,164 bp. Some of the Stramenopile genomes are also larger than *P. lacertae*, Oomycete *Sclerospora graminicola* has a 299 Mb genome (unpublished) made up of 28,799 contigs and has an N50 of 15,902 bp and *Saccharina japinoca* (Phaeophyceae) has a 543 Mb genome assembled into 35,750 contigs with an N50 of 44,573 bp [340], [341]. These statistics show that the *P. lacertae* genome is comparable not only to the *Blastocystis* sp. ST genomes but to other Stramenopile genomes sequenced using the same technology.

The changes in genome size with increase of expected size during the assembly process raises another issue for the assembly of the *P. lacertae* genome. *P. lacertae* is estimated to be tetraploid and therefore the difference in genome size between the two parameters in the assembly may indicate that the software has segregated a number of divergent, heterozygous regions. The smaller estimated genome size of 5 Mb may have forced the programme to compress heterozygous reads into a single sequence (or else discard them), whereas the larger estimated genome size of 35 Mb allowed the programme to segregate and retain these regions. If the genome was tetraploid, though, then increasing the genome size should, conceivably, have increased the overall size by four rather than by two. This notwithstanding, no large-scale duplication of gene content in the annotation is observed, and so the ploidy of the organism has not had an observable impact on the predicted proteome for *P. lacertae*.

2.4.3 Evaluation of gene annotations

More than 65% of the genes predicted for the *P. lacertae* genome have no homologues in the database. While it is expected that there will be some genes specific to this species, and the presence of genes that are not specific but are yet to be investigated, the high proportion of these raises questions about the validity of the gene annotation.

There may be reasons why the gene models did not share homology with sequences in the databases. The databases are biased towards 'model' and well-studied organisms that do not include *P. lacertae* and its Stramenopile relatives. Many of the gene models that did not share homology with sequences in the database were assigned functional annotations based on other methods of comparison such as HMMs.

Despite considerable evolutionary distance, other published Stramenopile genomes show varying degrees of species-specific genes. For example, the Oomycetes, which have been extensively studied, show a much smaller proportion of proteins with no homology in the database compared with more distantly related organisms, which have not been as extensively studied such as the diatom *Thalassiosira pseudonana* [209], [342]. However, the proportion of *P. lacertae* genes without homologs in the database is higher than even *T. pseudonana*.

Nonetheless, I have shown that the genome is >85% complete using BUSCO and it is these genes that will be of use in the comparative analysis. In the event that the *P. lacertae* genome does contain erroneous gene models, these will not influence the results of the comparative analysis with *Blastocystis* (Chapter 4). The comparative analysis will serve to identify conserved genes between *Blastocystis* and the other Stramenopiles and will not use or comment on the species-specific repertoire of each genome. The BUSCO score is comparable to both the *Blastocystis* sp. ST genomes and the genomes of other Stramenopiles which are typically ~90% (Figure 4.1.1). The final 10 - 15% of BUSCO orthologues may be missing through annotation or though sequence divergence. By using a range of Stramenopile genomes along with our assembly for *P. lacertae* it was hoped that the majority of the core Stramenopile genes would be captured.

Only 13,950 gene predictions had reciprocal best hits to *de novo* assembled transcripts. This represents 50% of the overall transcriptome. This implies that the other portion of the transcriptome has no orthologous counterpart in the gene set. I have shown that this is not due to splice variants in the transcriptome. The difference may be due to errors in the transcripts assembly by Trinity, or the gene set may still be incomplete. As the genome contains 28,067 gene models that are associated with transcriptomic evidence, it is unlikely that these are misannotated in the genome. Both the transcriptome and the gene set scored highly when examined by BUSCO, though the gene set scored higher and so will be used as the representative proteome for *P. lacertae*.

The *P. lacertae* genome has a predicted proteome of 35,706 proteins, which is more than double the average for published Stramenopile genomes (Figure 4.1.1). *Nannochloropsis gaditana* has the smallest proteome of any Stramenopile [343] containing 3,558 genes in its 34 Mb genome but the largest proteome belongs to *Thalassiosira oceanica* and comprised 34,642 proteins [344]. In order to prevent poor gene models from influencing the subsequent analyses in Chapter 4, the dataset was reduced to contain only genes with transcriptomic support. This reduced the number of proteins to 28,067, which produced a BUSCO score of 75.05%.

2.4.4 Conclusion

In this chapter I have presented the genome assembly and annotation of *P. lacertae*. The assembly is comparable to other Stramenopile genomes and is a valuable taxon to add to the understanding of protist genomes, particularly within the Slopalinida. While the annotation is largely specific it represents a good starting place for the investigation of *Blastocystis* and would benefit from functional characterisation, particularly of novel sequences. Resolution of the number of chromosomes and discovery of the ploidy would also add to the understanding of the behaviour of data during assembly.

Overall, this genome is a useful tool in investigating the biodiversity of the Stramenopiles with particular respect to the evolutionary processes that have led to the development of symbiosis in *Blastocystis.*

Chapter 3. Sequence and analysis of the *Cafeteria roenbergensis* transcriptome

The *P. lacertae* genome described in Chapter 2 is a useful resource with which to compare *Blastocystis*. However, as *P. lacertae* is also an intestinal anaerobe, it is necessary to include a data point from a related, yet entirely free-living organism to distinguish between adaptations for pathogenesis and symbiosis. However, no genetic resource is available for such an organism, therefore a transcriptome was produced for *C. roenbergensis* and assessed for its usefulness in a comparative analysis with *Blastocystis* and *P. lacertae*. The transcriptome produced here was successfully separated from the bacterial component of the cultures and shown to be a useful out-group to the *Blastocystis-Proteromonas* clade.

3.1 Introduction

In Chapter 2 I presented a *P. lacertae* genome sequence. It was clear from the genome annotation that *P. lacertae* possesses a large number of novel, hypothetical proteins. Not all of these will be *P. lacertae*-specific, and so, in order to differentiate conserved hypothetical genes that have been lost from *Blastocystis* and species-specific hypotheticals that are unique to *Proteromonas* (and not pertinent to present aims), an out-group was required to *Blastocystis* and *Proteromonas*, preferably with an entirely free-living life strategy. The majority of Stramenopiles that have an associated genome sequence are not free-living, and those that are free-living, such as the diatoms, are too phylogenetically distant to form a useful comparison.

Marine ecosystems are some of the most diverse and least well-understood food webs. Their ecological importance also has large and rapid impacts on a number of issues such as global warming. *Cafeteria roenbergensis* (ATCC 50561) represents a single link in the complicated diversity of marine life and has been studied to that effect. *C. roenbergensis* is a single-celled marine flagellate that feeds on bacteria. First described by Fenchel and Patterson (1988) [345] it is now considered to be among the most common, ecologically significant species of heterotrophic nanoflagellate in marine biology. Based on the phylogenetic position of *C. roenbergensis*, it forms a useful comparator to *Blastocystis* and *P. lacertae*.

3.1.1 *Cafeteria roenbergensis*: a common marine flagellate

C. roenbergensis was first identified in cultures based on water samples from Limfjord, Denmark in 1981. Initial analysis of its morphology, show that *C. roenbergensis* has two flagella inserted sub-apically; where one extends laterally and carries tubular hairs while the recurrent flagellum anchors the cell to the substrate during feeding *in vitro*. It is possible that *in vivo* cells adhere to debris where they are able to persist in oligotrophic conditions. Cultures were established from an isolation of *Rhizochromulina marina* where it appeared as a contaminant. It was isolated by serial dilution into sterilised natural seawater or enriched seawater medium bacterized with *Klebsiella aerogenes* [346]. Since its identification it has been found in all areas where it has been searched for suggesting that it may be more widespread and potentially of greater ecologically importance than previously though [347]–[352].

3.1.2 The phylogenetic position of *C. roenbergensis*

Morphological features of *C. roenbergensis* are consistent with defining features of the Stramenopiles and along with the lack of observable chloroplast or lorica its initial placement was with the heterokont (Stramenopile) flagellates, specifically within either the Bicosoecid or Chrysomonad flagellates [345]. SSU rRNA sequence data was able to bolster the morphological analysis and place *C. roenbergensis* within the Stramenopiles, however it was often found on very deep-rooted braches and accompanied by little support [67], [299]. General consensus from these trees did show that despite the long branches and deep nodes, it was often proximal to *P. lacertae* when included. Subsequent analyses have added support to these nodes through use of SSU rDNA and the mitochondrial genome sequence data (Figure 1.2.2) [291]–[294], [353]. In Chapter 2 the phylogenetic position of *C. roenbergensis* was confirmed with a multigene analysis (Figure 2.3.3).

3.1.3 *C. roenbergensis* value in comparative genomics

As described above, both *Blastocystis* sp. STs and *P. lacertae* inhabit an intestinal environment and therefore an out-group was required, which is free-living. As a marine flagellate, *C. roenbergensis* is not associated with symbiosis and represents an essential component in identifying adaptations to symbiosis in both *P. lacertae* and *Blastocystis*.

The 43 Kb mitochondrial genome of *C. roenbergensis* can be found online at NCBI (NC_000946.1). The entry describes 34 protein-coding sequences (unpublished). This contrasts with the mitochondrial genomes of *Blastocystis* (28 Kb) and *P. lacertae* (48 Kb) which differ greatly in size and structure between themselves, but which contain similar gene repertoires of 27 protein-coding sequences (Figure 3.1.1) [294], [354]. The difference in gene content between these and *C. roenbergensis* is most likely due to the loss of genes from the anaerobic mitochondria such as the *cox* and *cob* encoded cytochrome complexes that are essential in aerobic *C. roenbergensis*. There is also high variation between all three mitochondrial genomes with respect to their tRNA content. *P. lacertae* contains the most with 25, *C. roenbergensis* is next with 22 and *Blastocystis* has the smallest repertoire with 18 tRNAs. Already this organism has shown insight into the adaptations shared between *Blastocystis* and *P. lacertae.*

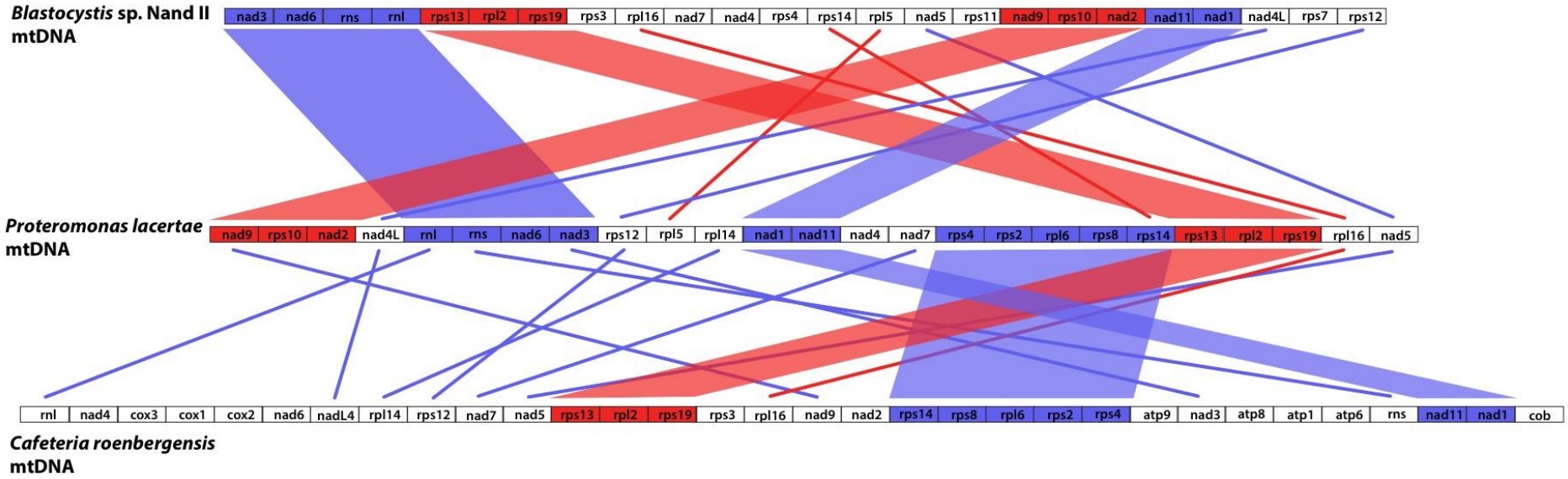


Figure 3.1.1 ACT gene order comparison between the mtDNA of <u>Blastocystis</u> sp. NandII (linearised), <u>P. lacertae</u> (single repeat only) and <u>C. roenbergensis</u> (linearised). *Based on the comparison of Pérez-Brocal et al. 2010* [294]. *Only protein-coding genes have been included for simplicity. Lines indicate synteny of single genes and blocks indicate synteny with conserved order, either forward (red) or reverse (blue) orientation.*

3.1.4 Metagenomics

The xenic nature of the *C. roenbergensis* cultures means that sequence data prepared from them will represent a meta-genomic dataset. There have been a number of methodologies developed that attempt to remove systemic bacterial contamination from sequenced samples and these may be used singularly or in combination. In the first instance, it may be possible to remove much bacterial contamination prior to sequencing. This may be achieved either by using antibiotics to destroy bacterial cells or by enriching for eukaryotic sequences. Eukaryotic mRNA can be selected for using poly-A selection, as bacterial RNA is poorly adenylated [355]–[358]. However, rather than resulting in purification, this method produces a sample enriched for mRNA as non-adenylated sequences may still be carried through [359]. Therefore, bioinformatic approaches have been developed to remove bacterial contamination post sequencing. In order to determine which sequences originated from the eukaryotic aspect of the sample, approaches can be employed based on homology, differential genomic signatures [360], [361] and coverage profiles [362], [363].

Homology is a very common method and may utilise a tool such as BLAST. Differential genomic signatures utilise codon preference, GC content and kmer frequencies, which allow clustering of 'like' sequences. Translational selection is important in bacterial genomes and shapes the most highly expressed genes to favour the most abundant tRNAs. Codon usage and GC content are also linked because optimal codons are those most likely to maintain the GC value of that organism [364]. A Kmer is described as a motif, built up of a number of characters (Kmer distribution analysis is primarily based on tetranucleotides [360], [365]) and that may occur numerous times along a length of nucleic acid. The number of times a kmer appears throughout the genome can be used to group 'like' sequences together (e.g. for GC rich or repetitive regions). Coverage profiles group sequences based on the assumption that certain taxa within the sample will be more or less abundant than others. Especially with respect to bacterial and eukaryotic cultures, components of the bacterial component may be far less numerous than others after poly-A selection and therefore have been sequenced less deeply. Grouping reads based on coverage provides another method for differentiating prokaryotic from eukaryotic sequences (e.g. prokaryotic reads may be sequenced less deeply as they are less abundant after poly-A selection). In this thesis, I will use a combination of methods to remove bacterial contamination from the cultures of *C. roenbergensis*.

3.1.5 Aims of the chapter

In this chapter, the aim is to produce a transcriptome for *C. roenbergensis* to serve as a free-living out-group to the *Blastocystis* – *Proteromonas* clade, which will allow precise identification of gene gains and losses in both genomes during the comparative analyses (Chapter 4). There are five specific objectives:

1. Prepare high integrity RNA preps from cultures
2. Sequence the mRNA with Illumina
3. Assemble the reads into a transcriptome
4. Remove contaminating transcripts from the assembly
5. Annotate the transcriptome and provide a measure of completeness for the protein dataset

3.2 Materials and methods

3.2.1 Culture optimisation

Cafeteria roenbergensis (ATCC 50561) was maintained in artificial seawater for protozoa (ASWP) at 4˚C or at room temperature in light-shielded T-25 culture flasks. Complete medium contained 33.6 g 'Ultramarine Synthetica' sea salts (Waterlife Research Industries Ltd), 0.5 g tricine, 10.0 ml stock solution and 50.0 ml soil extract made up to 1 l with distilled H_2O and adjusted to pH 7.6 – 7.8. Stock solution contained 5.625 g $NaNO_3$, 0.225 g Na_2HPO_4 and 0.188 g K_2HPO_4 (BDH) made up to 1 l with distilled H_20. Soil extract was prepared by taking soil from undisturbed deciduous woodland and sieving through a 1 cm mesh. Soil was spread, air-dried and hand-picked for small invertebrates and roots. This was then sieved through a 2-4 mm mesh. 105 g of this was added to 660 ml distilled H_2O, autoclaved twice and left to settle for one week. Supernatant was decanted, filtered and adjusted to pH 7.0 – 8.0.

The final medium contained a single autoclaved corn kernel (Marks & Spencer) and supplemented with 0.1% yeast extract (BDH). Cell density was established using a haemocytometer under a light microscope.

3.2.2 RNA preparation, sequence and assembly

Five replicate cultures were produced under four different conditions: 1. treated with Penicillin-Streptomycin (150 µg/ml) (Thermo-Scientific) for 24 hours; 2. heat shocked (40˚C) for 10 minutes prior to RNA collection; 3. both antibiotic and heat treatment and 4. left under normal conditions. These steps allowed capture of a more 'complete' expression profile and to reduce the proportion of bacteria in the culture. Cultures were processed at peak density (determined by cell counts) and pelleted at 3000 x g for 10 minutes and RNA was extracted using RNeasy (Qiagen) according to the manufacturers protocol. RNA quantification was assessed with Tecan spectrophotometer and Magellan software according to the manufacturer's protocol. Integrity was assessed on a 1% agarose gel stained with SYBR safe (Invitrogen) run at 100v for 30 minutes.

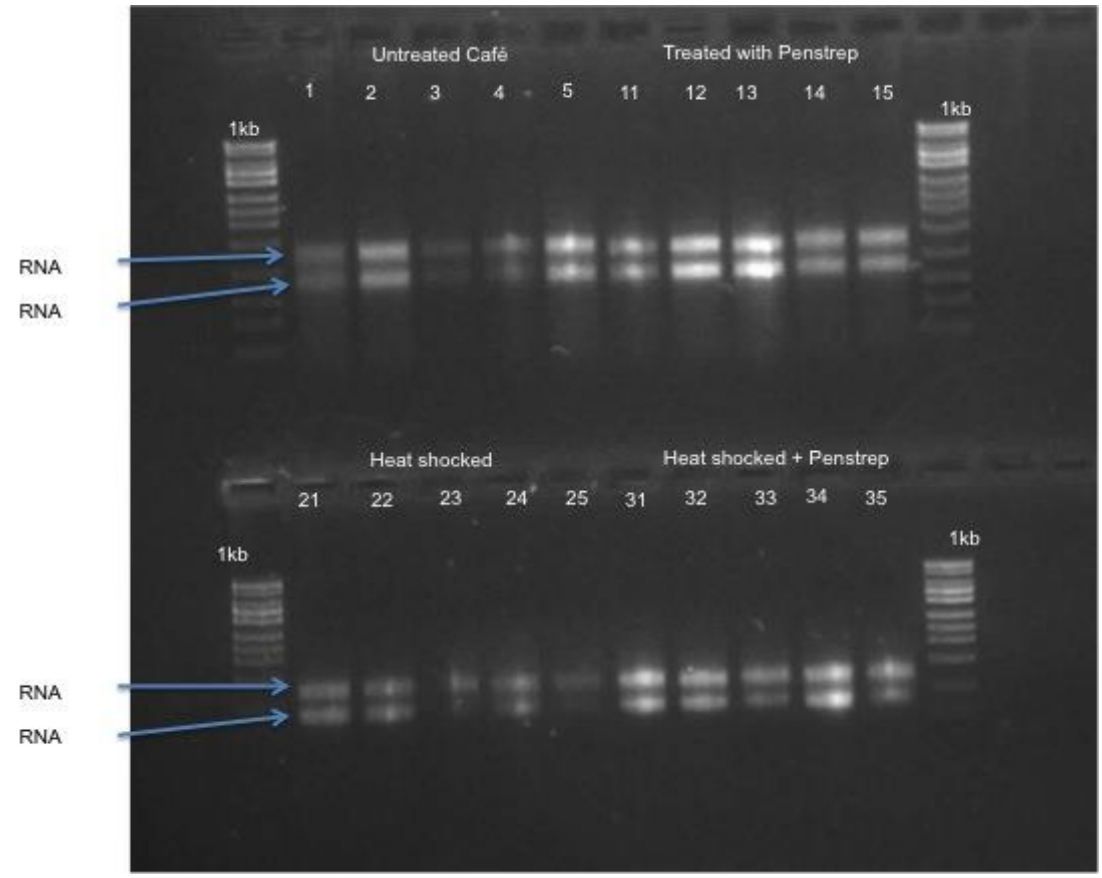


Figure 3.2.1 1% agarose gel showing integrity of RNA collected from five replicates of four treatments.

RNA samples from the four conditions were pooled and processed by the University of Liverpool's Centre for Genomic Research (CGR) using poly-A selection according to the manufacturer's protocol. These were used to produce three Illumina RNASeq libraries from enriched RNA using the strand-specific ScriptSeq kit. Paired-end sequencing (2x125bp) was carried out on one lane using Illumina HiSeq platform generating in excess of 180M clusters. Post sequencing QC (fastqc v0.11.4) and trimming of adapters resulted in a final dataset of 69,013,377-paired reads. Reads were assembled *de novo* by Trinity v2.1.1 [312] using a kmer size of 25 into 40,858 transcripts.

3.2.3 Resolution

In order to remove contaminating bacterial transcripts, those with homology to sequences in the database were used to define bacterial and *C. roenbergensis* control groups, i.e. a 'negative' control group of certain bacterial transcripts (>98% sequence identity to a known bacterium) and a 'positive' control group of high

confidence eukaryotic transcripts (>40% sequence identity to a known eukaryote or a top hit to a known Stramenopile gene). Transcripts with weak or no homology to databases sequence were assigned to either bacterial contamination or *C. roenbergensis* based on their sequence properties. Specifically, the base composition and codon usage were measured and each plotted against read depth. Kmer frequency distribution was also used and related to the control groups, to distinguish between transcripts of bacterial origin and those from *C. roenbergensis*. In order to ensure consistent results, kmer frequencies were estimated for kmers of 3, 4 and 5 nucleotides for each transcript. A principle components analysis (PCA) of kmer values was carried out using R v3.2.4 [366] in R Studio v1.0.136 [367]. Principle component analyses are a statistical method for converting a set of observations into a set of values of linearly uncorrelated variables called principle components.

Transcript metrics were estimated using programmes from the EMBOSS v6.3.1 toolkit [368]. Base composition was calculated per transcript using the EMBOSS infoseq package and codon usage was calculated per transcript using the EMBOSS chips package. Read depth was calculated per transcript by using Bowtie2 [369] to map the trimmed reads to the assembled transcriptome and the samtools v0.1.18-r580 depth programme calculated read depth per position, which was averaged to give a value per transcript. EMBOSS compseq was used to estimate Kmer values.

3.2.4 Transcriptome annotation

Initial annotation was assigned based on homology assigned by BLAST in BLAST2GO [316]. KEGG annotations were assigned to protein sequences by GhostKoala [324] and BLAST2GO and were mapped to KEGG pathways using KEGG Mapper – Reconstruct Pathway [325]. TransDecoder [312] was used to extract peptide information from the transcripts.

In order to gain an objective description of how complete the transcriptome assembly was it was given to BUSCO v1.1b1 [326] (Benchmark Universal Single-Copy Orthologs). This programme asses based on BUSCO content. The rationale being that all eukaryotes should contain orthologs of a set of 429 core genes. The score is assigned based on how many of these are present in the annotation. The *C. roenbergensis* transcriptome scored 70.40% with BUSCO.

3.3 Results

3.3.1 Optimisation of *C. roenbergensis* cultures

C. roenbergensis cultures were grown xenically with bacteria isolated at the time of its collection. Cultures were passaged into fresh medium at peak density. In order to successfully sequence the *C. roenbergensis* transcriptome, effort was made to optimise the cultures, which included increasing the overall proportion of *C. roenbergensis.* Optimal growth conditions were established when supplementing the medium with 0.1% yeast extract and incubating at 4˚C rather than RT (Figure 3.3.1). The addition of yeast extract allowed the bacterial population to increase, which increased the food source for *C. roenbergensis*. Incubation at the lower temperature was sufficient to suppress the growth of the bacteria to ensure they did not overgrow *C. roenbergensis.* The alteration of these conditions produced higher yields of cell number 1 x 10^{6-7} rather than 1 x 10^{3-5} as previously reported [370].

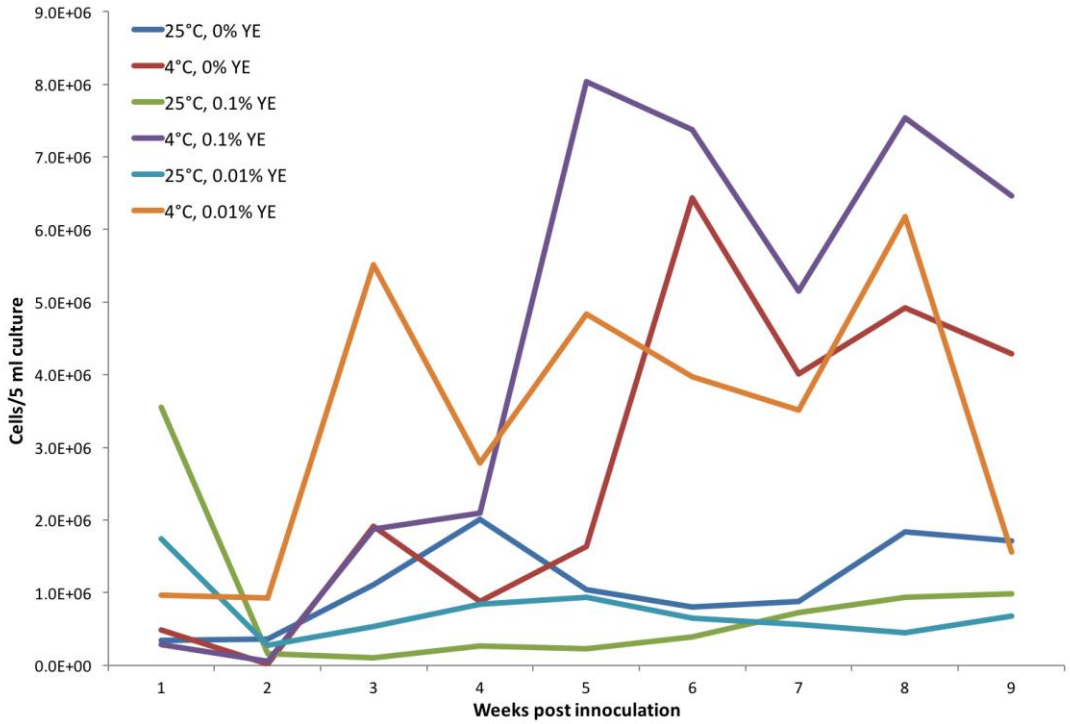


Figure 3.3.1 Cell counts of cultures under different growth conditions at log phase. *Cell counts are expressed as an average of two cultures. Cultures were either grown at room temperature or 4˚C, medium was supplemented with Yeast Extract (YE) at a final concentration of 0.1%, 0.01% or 0%.*

3.3.2 Resolution of contaminants from the *C. roenbergensis* transcriptome

The *C. roenbergensis* transcriptome was assembled into 40,858 unique transcripts from reads that were selected for by poly-A selection. Although poly-A selection should exclude the majority of bacterial transcripts, contaminants may still have been sequenced and assembled. To identify contaminating transcripts and remove them, an initial annotation of the transcripts was assigned using BLASTX against a non-redundant protein database. The 10,000 sequences with the highest score (based on E-value) were taken and their best BLAST hit interrogated to provide a positive benchmark with which to group sequences based on sequence identity. Any transcript with a top BLAST hit to a eukaryote with >40% sequence identity was classed as positive (confidently ascribed to *C. roenbergensis*), the negative group was defined as any transcript with a top BLAST hit >98% sequence identity to a bacterial sequence. Transcripts that did not fall into either of these categories were labelled unassigned (Figure 3.3.2).

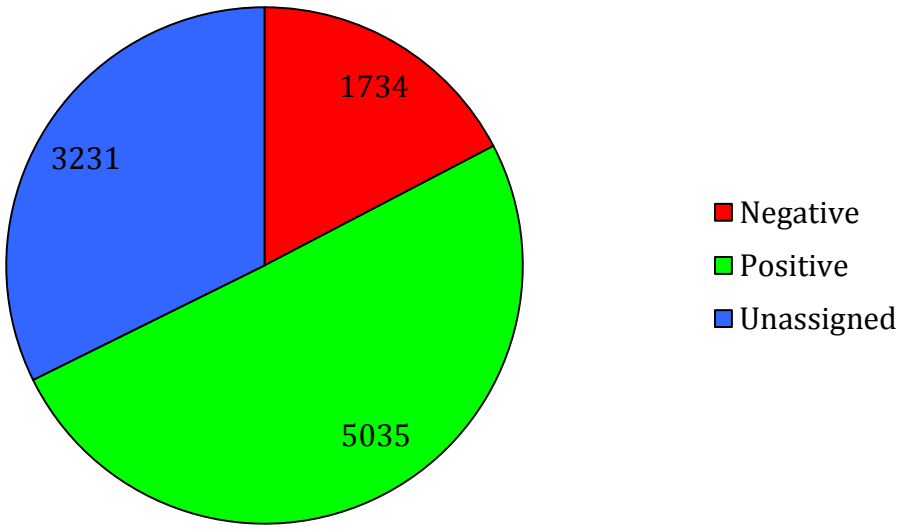


Figure 3.3.2 Pie chart showing proportions of the 10,000 transcripts with the highest E-value, which fall into each category. *Positives are those confidently ascribed to* C. roenbergensis, *negatives are most likely bacterial contamination and unassigned are those with low homology determined by sequence identity.*

The remaining 30,858 sequences that did not fall into the top 10,000 BLAST hits were also labelled unassigned and each transcript had read depth, base composition and codon usage (Nc) calculated. These values were plotted pairwise against each other (Figure 3.3.3) in order to identify if there was a clear distinction

between the positive and negative groups, and to identify which unassigned transcripts clustered with these groups. This would allow visual identification of the bacterial transcripts, which could then be excluded from the analysis. Base composition was calculated as GC percentage per transcript, read coverage as RPKM and codon usage as effective number of codons (Nc) per transcript.

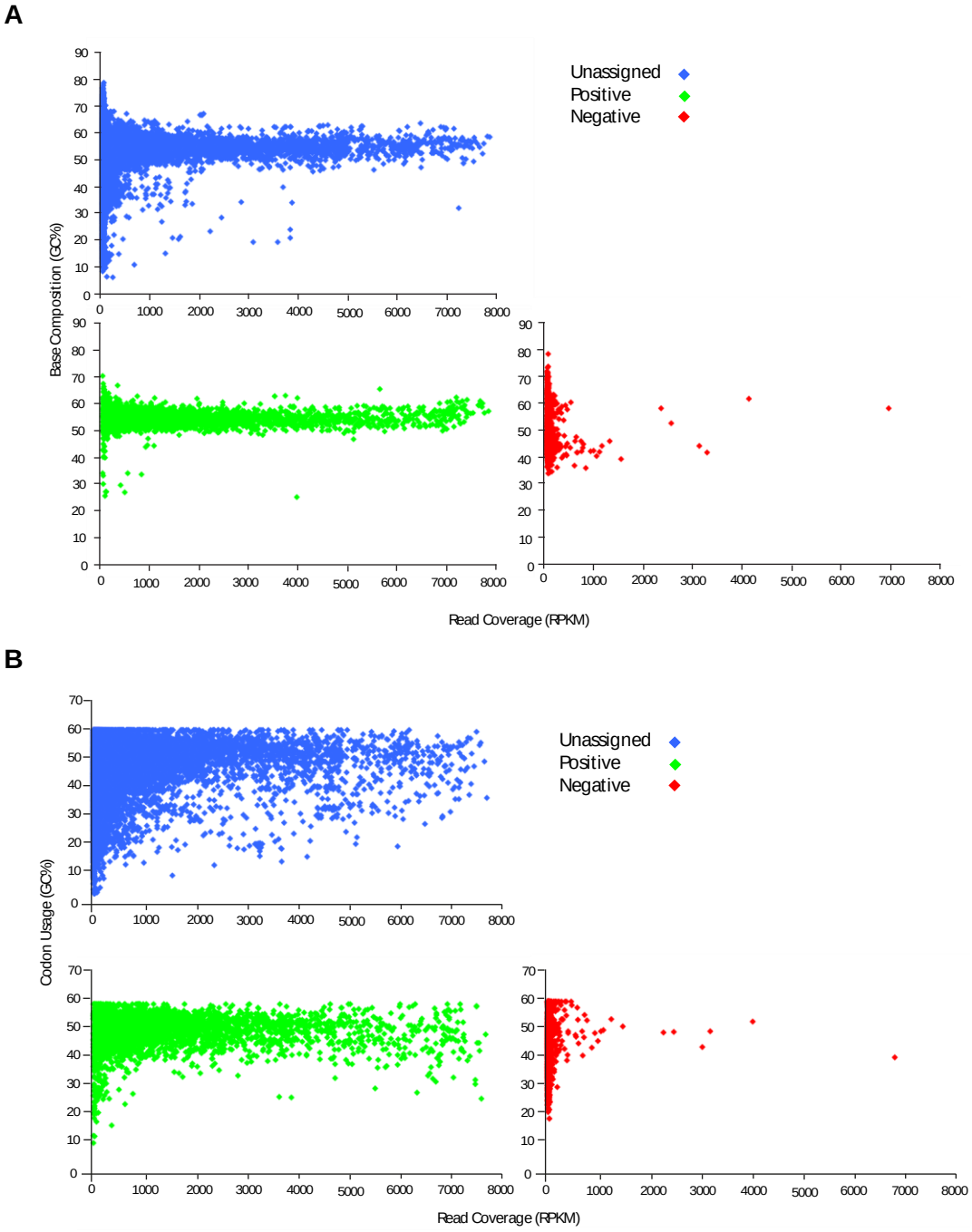


Figure 3.3.3 Pairwise comparison of codon usage, read coverage and base composition for 3 groups defined within the <u>C. roenbergensis</u> transcriptome. *Metrics were plotted pairwise against each other:* ***A.*** *read coverage vs. base composition,* ***B.*** *read coverage vs. codon usage.*

The results from this comparison show that it may be possible to partially separate these groups based on sequence composition, but there is no clear distinction. Ideally the two groups would cluster independently. Instead, I tried a third approach to separating bacterial and eukaryotic transcripts using kmer frequencies, calculated per transcript and applied to a Principle Components Analysis (PCA) in R. Each Kmer value used to calculate frequencies produced the same output, therefore only frequencies calculated with 3mers is shown (Figure 3.3.4).

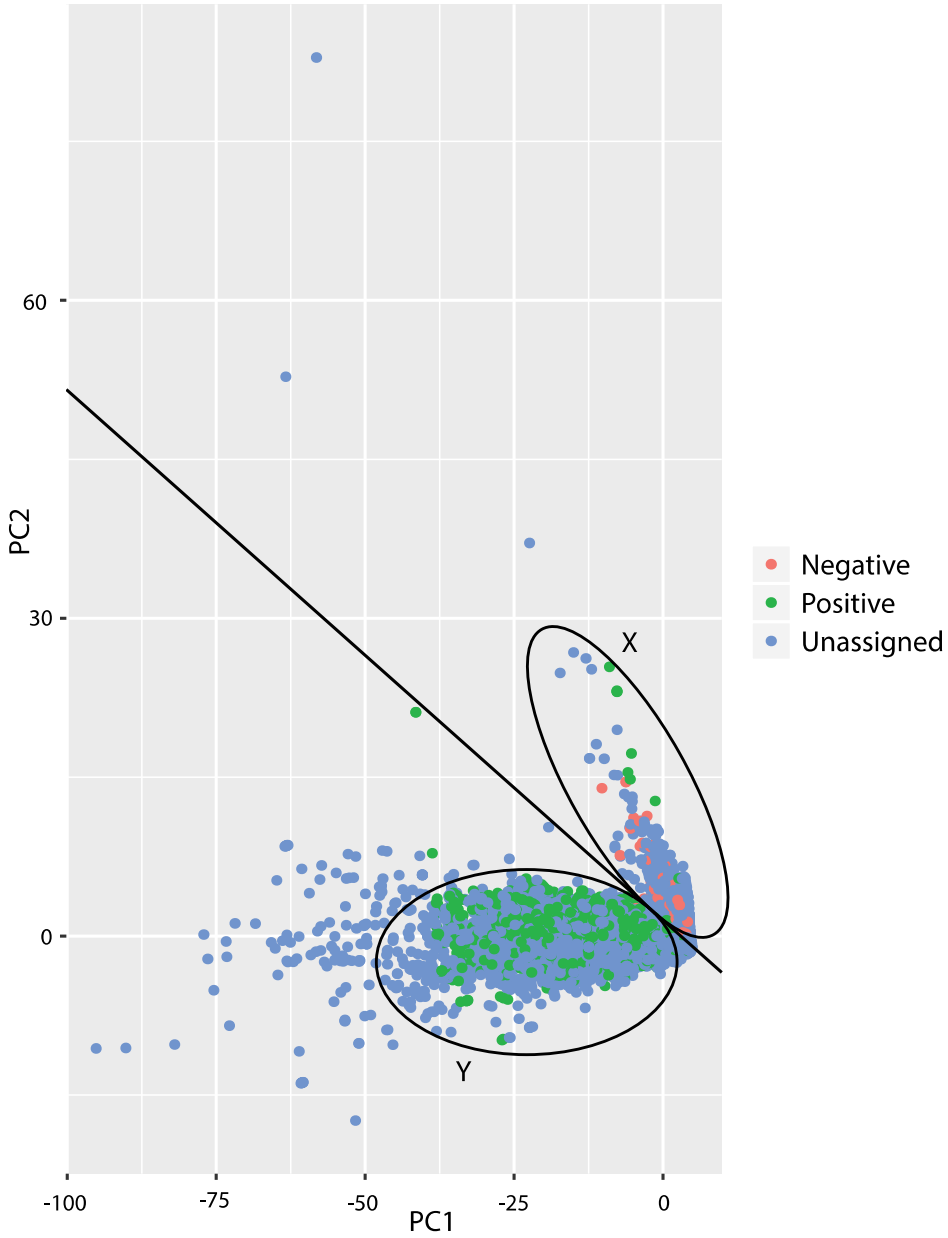


Figure 3.3.4 A principle components analysis of 3mer values. *PC1 and PC2 account for >95% of the variation. Sequences above the line were removed from the transcriptome; sequences below the line were included except for negative sequences below the line and positive sequences above the line.*

This approach proved to be very effective; it was possible to draw a single line between the positive and negative control groups, with unassigned transcripts aligning with one or other control quite clearly. Consequently, removing any transcript that was above the line, except for those already identified as positive, and retaining any transcript below the line removed 11,633 transcripts from the transcriptome. The positive transcripts that fell above the line were examined and found to code for mitochondrial proteins already described in *C. roenbergensis* including mitochondrial encoded ribosomal proteins (NP_051143.1), NADH dehydrogenase subunit 7 (NP_051138.1), an unidentified orf169 (NP_051135.1) and ATP synthase F1 subunit alpha (NP_051156.1). The removal of the contaminating bacterial sequences produced a transcriptome containing 28,952 transcripts. BUSCO was used to assess this transcriptome and determine the completeness. The transcripts scored 70.40% with BUSCO containing 203 complete BUSCOs, 61 duplicated BUSCOs, 38 fragmented and 127 absent BUSCOs.

In order to extract protein-coding sequences from the transcripts the longest open reading frame (ORF) from each transcript was extracted using TransDecoder. This protein dataset contained 18,439 proteins and scored 49.41% with BUSCO containing 137 complete BUSCOs, 32 duplicates, 75 fragmented BUSCOs and 217 absent. The protein dataset was annotated using InterProScan, which assigned 2,596 domain annotations to 6,400 transcripts, and was also the dataset utilised in the comparative analysis presented in Chapter 4.

3.3.3 KEGG mapping

As the protein dataset scored lower than the transcriptome when assessed with BUSCO, it was mapped to KEGG pathways to determine if there were any portions of the metabolism obviously absent. The protein dataset mapped 3,414 sequences to 349 pathways (Figure 3.3.5) which is comparable to the number of pathways mapped to by *P. lacertae* (4,490 protein mapped to 328 pathways) *Blastocystis* sp. STs (~3,750 proteins mapped to ~310 pathways), see Chapter 4, section 4.3.10. While fewer sequences were mapped they mapped to a broader range of pathways compared to both *P. lacertae* and *Blastocystis* sp. STs.

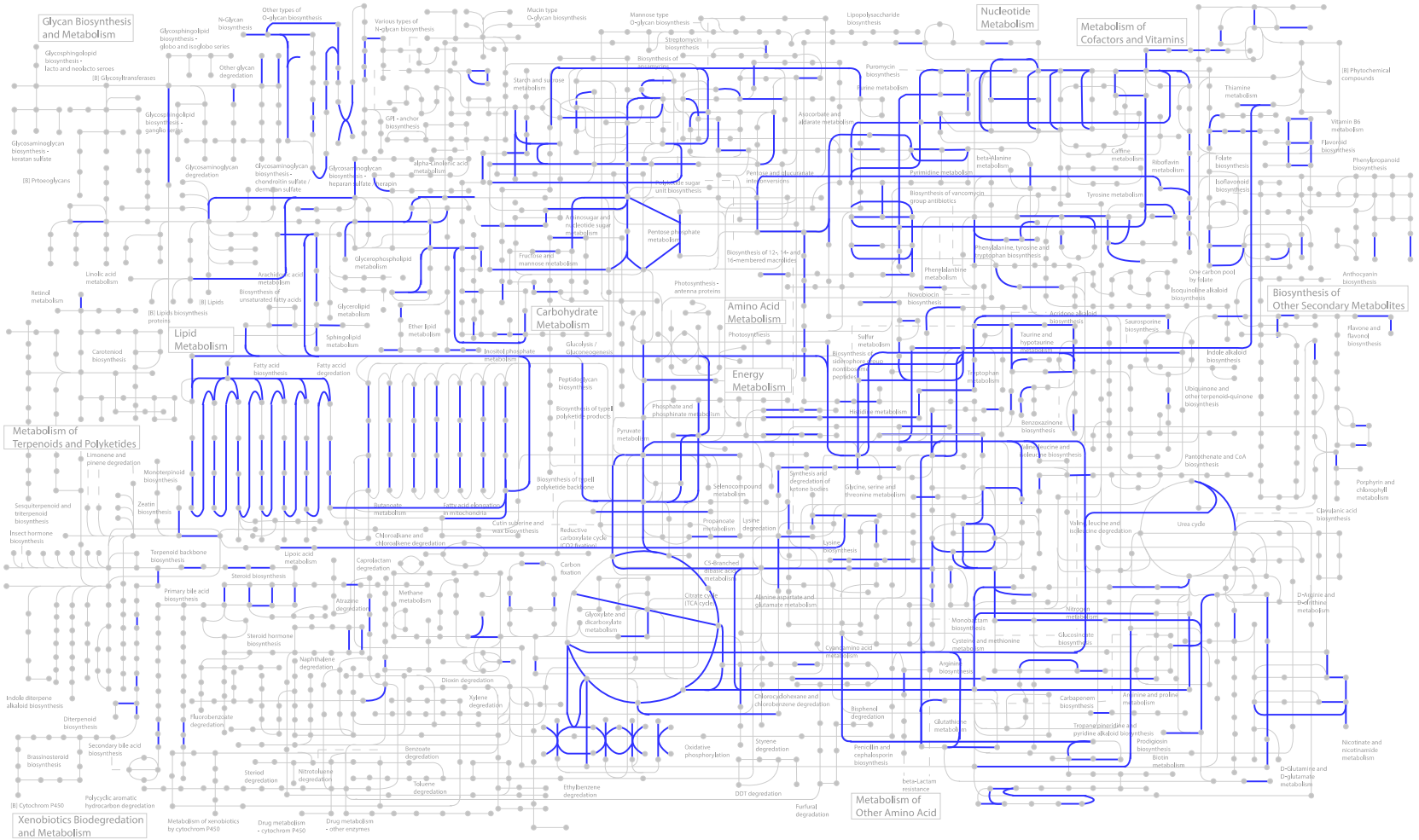


Figure 3.3.5 <u>C. roenbergensis</u> protein dataset mapped to KEGG pathways. *Blue lines indicate where proteins with KEGG annotations mapped.*

3.4 Discussion

In this chapter, I have produced a transcriptome for *C, roenbergensis* from xenic cultures and assessed the quality of this resource as a representative of this group of organisms and its usefulness as a free-living out group to the *Proteromonas – Blastocystis* sister taxon. Its usefulness was tested by obtaining independent measures of the assembly quality, transcriptome integrity and its completeness.

3.4.1 *C. roenbergensis* cultures

The cultures were optimised to produce the highest proportion of cells without allowing the bacterial food source to overgrow and choke the culture. Previous reports from attempts to culture *C. roenbergensis* have reported cell density of approximately 1 x 10^3 cells/ml [352], [370]. Here it has been shown that the population of heterotrophic flagellates can be altered by modification of maintenance protocols. The ability to grow highly populated cultures within a shorter time frame is useful when attempting to retrieve sequence data from any slow-growing and low-density culture. The approaches taken here do have drawbacks; while the increased richness in the medium triggered a longer log phase for *C. roenbergensis* it also increased the population of the bacterial food-source. This increases the mass of the contaminants, which must be filtered out during the process as the number of bacterial cells far outnumbers the protists. The changes in growth conditions is also likely to have a greater effect on some bacterial species than others, as could the change in temperature, and could lead to alterations in the make-up of the bacterial flora. This may have implications in the variety of bacteria available for feeding and as a result the expression profile of *C. roenbergensis*.

3.4.2 Resolution of the transcriptome

The RNA isolated from *C. roenbergensis* cultures was subjected to poly-A selection and therefore should contain relatively few bacterial contaminants. However, inefficiencies in the enrichment protocol meant that there may have been bacterial contamination carried into the sequencing [359]. There is also evidence that many bacteria, including *Escherichia coli* and *Caulobacter crescentus,* contain molecules that can be isolated using techniques developed for the capture of

eukaryotic mRNA which have a 3' poly adenylated tail [355]–[358]. In order to identify these post-sequencing, I used a BLAST search against known bacterial sequences. The threshold of >98% sequence identity was chosen in order to differentiate between bacterial sequences and bacterial-like sequences in the *C. roenbergensis* genome. Genuine bacterial sequences will be highly similar to existing environmental sequences in the database, whereas, bacterial-like sequences in *C. roenbergensis* will show higher divergence, as they will have diverged once transferred into the eukaryotic genome. We might expect there to be many bacterial-like sequences that genuinely belong to *C. roenbergensis* acquired via HGT through phagocytosis [371], [372]. HGTs play a major role in the make-up of many genomes including *Blastocystis* and therefore I attempted to preserve as many of these as possible [91], [283], [284]. As the resource is transcriptomic it is likely that there are HGT genes that are not constitutively expressed that will be absent from the dataset, our aim was to include as many as were available whilst preventing environmental contamination that would itself conflate the contribution of HGT.

While none were observed by light microscopy, it is conceivable that the culture contained not only a range of bacterial diversity but also eukaryotes. This would present a serious issue if true, as the approach taken here could not differentiate and subsequently remove these sequences. The approach taken here was to use sequence similarity to sequences in the database. A cut-off of >40% to a known eukaryote was used to identify 'positive' *Cafeteria* sequences. A large majority of these BLAST matches were to *Blastocystis* or other Stramenopiles (42%), and top matches to other eukaryotes had lower identities. Furthermore, annotation of the transcriptome has shown no large-scale duplication of genes, which would result from multiple eukaryotic genomes. Together, these observations mean that there is no good evidence for the presence of another eukaryote in the culture besides *C. roenbergensis*.

3.4.3 Completeness of the transcriptome

The final transcriptome size is comparable to some of the genome sizes published for other Stramenopiles [209], [305], [342]–[344], [373], [374]. However, as discussed in Chapter 2, the Stramenopiles are an extremely diverse group of organisms that exhibit vast differences in both genome size and gene content. BUSCO was also used to assess the completeness of the transcriptome.

BUSCO provided a score of 70.4% for the assembled transcriptome, which dropped to 49.41% after peptide prediction with TransDecoder. This drop in both number of sequences and completeness is concerning as potentially, up to 20% of the transcriptome has been lost through this method. TransDecoder attempts to extract protein-coding regions within transcripts by identifying all open reading frames (ORFs) over 300 bases long, beginning with an in-frame start codon and ending with a stop codon. The top 500 longest ORFs are used to build a 5th order Markov models, which are used to select the best ORF where more than one is available in a single transcript. If these conditions are not met then no protein-coding region is reported for the transcript.

The transcriptome mapped to 349 KEGG pathways, which is comparable to the genomes of *Proteromonas* and *Blastocystis*. This would indicate that the genome is not 50% incomplete, unless *C. roenbergensis* is usually far more metabolically competent than either intestinal counterpart. This analysis would suggest that the transcriptome represents a relatively complete gene set but that it may be missing copies of certain genes as fewer sequences were mapped than either *P. lacertae* or *Blastocystis*.

In spite of the disagreement between the completeness measures, this resource is a useful tool with which to investigate *Blastocystis* and *Proteromonas*. I conclude this because it was always the intention to generate a transcriptomic resource for *C. roenbergensis*, therefore this resource would always lack information about the gene content that may otherwise be provided by a genome. Use of this resource in a comparative analysis may be criticised if used to infer gains in the genomes of *P. lacertae* and *Blastocystis* because absence from *C. roenbergensis* cannot be attributed to gene loss. However, presence in *C. roenbergensis* can still be used to infer loss from *Blastocystis* and *Proteromonas*. In this way, the completeness of the transcriptome only affects the resolution with which losses can be inferred. It may also prove useful in the annotation process should a genome sequence become available.

3.4.4 Conclusions

In this chapter, I have presented the *C. roenbergensis* transcriptome. Bacterial contamination has been filtered using a bioinformatic approach and

suggests that this resource contains reliable information about this organism. The data may also be used as an annotation tool should a genome sequence become available.

Overall this transcriptome forms a useful outgroup to the *Proteromonas* – *Blastocystis* clade and will be used as a complementary resource in the comparative analysis discussed in Chapter 4. While it cannot be used to infer gain in either organism it may prove useful in identifying genes and gene families that have been subsequently lost from the genomes of other Stramenopiles.

Chapter 4. Comparative analysis of genome evolution in the Stramenopiles

In order to investigate the evolutionary history of *Blastocystis,* the newly generated resources from Chapters 2 and 3 were used in a comparative analysis with three *Blastocystis* sp. STs as well as other Stramenopile genomes. This analysis showed that the *Blastocystis* genomes are genuinely small within the Stramenopiles and that this is due to genome wide reduction in *Blastocystis* sp. STs, even in otherwise conserved gene families. *Blastocystis* sp. STs have also lost aspects of 'typical' Stramenopile morphology like the flagellar apparatus and other motility associated genes. Despite this, the *Blastocystis* genomes have also innovated and expanded a number of gene families including proteases and globular domain-containing proteins, which may hint at a process of genomic streamlining consistent with a change in life-strategy.

4.1 Introduction

In Chapters 2 and 3 I produced a genome sequence for *P. lacertae* and a transcriptome for *C. roenbergensis* respectively. Here I have used these new resources in a comparative analysis with the published genomes of *Blastocystis* sp. ST1, ST4 and ST7, along with the genomes of *Pythium ultimum*, *Phytophthora sojae*, *Saprolegnia diclina*, *Ectocarpus siliculosus* and *Thalassiosira pseudonana,* to investigate evolutionary changes that occurred in the ancestor of *Blastocystis,* and which might be associated with the evolution of symbiosis.

4.1.1 *Blastocystis* sp. STs genome sequences

Several genome assemblies for *Blastocystis* sp. subtypes have been published, though only three of these have annotations associated with them: ST1, ST4 and ST7. All three of these subtypes are found in humans but show different prevalence and preferences for zoonotic reservoirs (1.2.2) [71]. *Blastocystis* sp. ST1 is found in the stool of many mammals including primates, sheep, cows, dogs and pigs, *Blastocystis* sp. ST4 is found in rodents and *Blastocystis* sp. ST7 is associated with an avian reservoir [70].

Denoeud *et al.* 2011 provided the first assembly and annotation of a *Blastocystis* genome using subtype 7. The genome is 18.82 Mb with a GC content

of 45.2% and contains 6,020 genes [91]. The assembly is made up of 54 scaffolds representing 15 chromosomes. This represents a highly compact genome, both in terms of genome size and number of genes, when compared with previously published genomes for the Oomycetes and other Stramenopiles [91]. The genes account for 42% of the overall genome but showed an interesting structural feature whereby 19% of the genes (1,141) are arranged in around 400 duplicated blocks. The majority of these genes are present in two copies apart from histone and ferredoxin domain-containing proteins which are present in large multi-copy families.

Denoeud *et al.* 2011 also inferred the presence of 133 candidates for HGTs, some of which they suggest may be involved either with adaptation to the intestinal environment or parasitism itself. These included two copies of Major facilitator superfamily (MFS) transporters which may confer the ability to absorb nutrients in the intestinal environment or when attacking host tissues. They further noted 26 copies of immunoglobulin domain-containing proteins, including some that also had a cadherin domain and may be responsible for cell-cell adhesion, and five hydrolase domain-containing proteins that may be responsible for degradation of host tissues. Finally, Denoeud *et al.* 2011 also identified short-chain dehydrogenases and oxidoreductases that they suggested confer advantageous metabolic qualities especially in an anaerobic environment [91].

A large part of the original genome study was focused on the MROs present in *Blastocystis* sp. STs which appear to contain features consistent with both mitochondria and hydrogenosomes, having adapted to the anaerobic environment of the intestine and lack a complete oxidative phosphorylation chain and citric acid cycle [375]. Denoeud *et al.* 2011 also made an attempt to investigate potential virulence factors encoded in the *Blastocystis* genome. They were able to identify 75 proteins expressed; potentially at the cell surface and that may have a direct link with pathology such as proteases, hexoses, lectins, glycosyltransferases and protease inhibitors. The cysteine protease families were of particular interest as 96% of these genes are predicted to be surface expressed.

The genome sequence of *Blastocystis* sp. ST7 provided a platform and a standard against which other *Blastocystis* sp. STs genomes could be judged and compared. Wawrzyniak *et al.* 2015 [284] sequenced and assembled a 12.9 Mb genome for *Blastocystis* sp. ST4 with a GC content of 39.7%. The assembly contains 1,301 contigs and 5,707 genes. In comparison with the *Blastocystis* sp.

ST7 genome, this assembly contained fewer duplicated blocks and more than 30% of the genes annotated in *Blastocystis* sp. ST4 had no ortholog in *Blastocystis* sp. ST7, including seven secreted proteases. The differences between the two assemblies suggest that there may be high genetic diversity within the *Blastocystis* genus.

Shortly after the *Blastocystis* sp. ST4 genome became available, the Statens Serum Institut deposited the genome sequences of five *Blastocystis* sp. STs in the database, without publication (https://www.ncbi.nlm.nih.gov/genome/13540). No annotation was made available for these assemblies and so comparison between gene repertoire has not been done, however, the differences between the genome sizes and GC contents of each assembly does reflect the diversity already seen between *Blastocystis* sp. ST4 and ST7. *Blastocystis* sp. ST2 (a mammalian subtype (Figure 1.2.2)) was assembled into a 12.6 Mb genome with a GC content of 54%. *Blastocystis* sp. ST3 is the smallest assembled *Blastocystis* genome, 11.6 Mb, is most commonly found in human hosts and has a GC content of 52%. *Blastocystis* sp. ST6 is an avian subtype and has a genome size of 15.4 Mb and a GC content of 43.1%. *Blastocystis* sp. ST8 has a large host range including mammals and avians. Its genome is 12.2 Mb with a GC content of 39.7%. Finally, *Blastocystis* sp. ST9, which is only found in humans, has a genome size of 11.7 Mb and a GC content of 43%.

The most recent annotated genome to be sequenced was produced from *Blastocystis* sp. ST1 (ATCC 50177/Nand II) by Gentekaki *et al.* 2017 [376]. This genome assembly was 16.4 Mb, had a GC content of 53% and contained 1,092 contigs and 6,544 genes. This study was primarily concerned with defining the HGT gene content and inferring their roles in the adaptation of *Blastocystis* sp. STs to the intestinal environment. They showed that up to 2.5% (167 genes) of the *Blastocystis* sp. ST1 genome was acquired through HGT and that this proportion is variable among subtypes, both in gene content and in the divergence between orthologues.

Eme *et al.* 2017 [283] found that HGT has played an important role in facilitating the transition into the intestinal environment by providing genetic material to cope with oxygen stress, carbohydrate metabolism and acquisition, nitrogen metabolism and anaerobic metabolism has particularly benefitted from HGT acquisition. Additionally, they postulated that *Blastocystis* might have acquired genes for immune evasion and pathogenicity. The most striking example of this is

the presence of a cholesterol-alpha-glucosyltransferase. This is potentially involved in alpha-glycosylation of cholesterol which has been shown to promote escape from phagocytosis by *Helicobacter pylori* [377]. Interestingly, evidence for HGT is not restricted to donations from bacteria. They propose that a beta-1,3-galactosyltransferase was obtained from animals and is involved in molecular mimicry, potentially as camouflage from the host immune response. If true, this would be the first instance of a eukaryote utilising a mechanism such as this. Virulence factors associated with HGT included a type I polyketide synthase, potentially involved in host dysbiosis and inflammation, that was previously described by Denoeud *et al.* 2011 [91]. Eme *et al.* 2017 [283] confirmed its presence in all other *Blastocystis* subtypes and also described an O-methyltransferase and a malonyl transferase potentially involved in polyketide biosynthesis and also of bacterial origin. Finally, they described a bacterial-like tryptophanase, which most likely produces indole. Indole, previously thought to be produced by bacteria, is able to diffuse through the plasma membrane and influence a number of physiological processes including host-cell invasion, inflammation and virulence [378], [379].

From these genome sequences, it is clear that *Blastocystis* have the smallest Stramenopile genomes sequenced to date and that there is high diversity within the *Blastocystis* genus in terms of genome size, GC content and gene repertoire. The MROs of *Blastocystis* are interesting as they contain machinery common to both mitochondria and hydrogenosomes and so represent an opportunity to investigate the evolution of these organelles. The *Blastocystis* genomes may also be heavily influenced by HGT, which has facilitated the transition into the intestinal environment in terms of metabolism, persistence in the face of an immune response and, potentially, virulence.

4.1.2 *Blastocystis* and parasitism

There is much debate in the literature about whether *Blastocystis* sp. STs are genuinely parasitic or merely commensal. The greatest cause of controversy has been the prevalence of *Blastocystis* sp. in asymptomatic patients. There are roughly an equal number of publications that report either significantly higher prevalence of *Blastocystis* sp. STs in symptomatic patients or no difference at all [54]. *Blastocystis* has been linked to Irritable Bowel Syndrome (IBS) and patients often have up to double the infection rate than non-suffers [41], [103], [105], [107]. However, it may also be that *Blastocystis* are able to better colonise the gut of

people suffering from IBS [54]. This is compounded by the fact that IBS is often diagnosed by ruling out other common causes.

There has been very little data collected on the virulence of *Blastocystis*, due to the lack of an effective infection model. Lab mice have been deemed unsuitable because, while *Blastocystis* is able to cause symptoms during experimental infections, these are often different to those observed in humans and the infections themselves are self-limiting [102]. It has been shown, however, that *Blastocystis* cysteine proteases are able to modulate the immune response and induce host interleukin (IL)-8 [380] and that some proteases are able to cleave human immunoglobulin and disrupt the tight junctions of epithelial cells thereby increasing the permeability of the gut [381]–[383]. Secreted products were also shown to induce apoptosis in epithelial cells *in vitro* [384].

The genomes of the various *Blastocystis* subtypes that have been annotated revealed many gene families that have been associated with immune evasion and tissue degradation in other systems: hydrolases, proteases, adhesion-like proteins and protease inhibitors. Differences in copy number and expression between subtypes may explain differential infection rates and host specificity [91], [102], [283], [284], [385]. However, these gene families are also abundant in non-parasitic genome and thus their presence is not indicative of pathogenicity.

The lack of conclusive data describing *Blastocystis* as parasitic or not cannot be answered by comparative genomics alone, but neither does it compromise the comparative approach taken here. Whether or not *Blastocystis* spp. are capable of being the primary cause of a disease, they are still adapted to the intestinal environment and comparison with other organisms can help to understand how these adaptations evolved. Overall there is evidence to suggest that *Blastocystis* sp. STs have the ability to damage host tissues; but it remains to be conclusively shown whether *Blastocystis* sp. STs are a cause of disease, or an effect of more complex disorders with many causes and outcomes.

4.1.3 Other Stramenopile genome sequences

The Stramenopiles are an extremely diverse group of organisms (Figure 4.1.1). Ranging from photosynthetic lineages including single-celled diatoms and multicellular brown algae (kelp), to parasitic slime moulds and marine heterotrophs.

There is high variation in terms of morphology, genome structure and content, life-strategy and reproduction within this group [353].

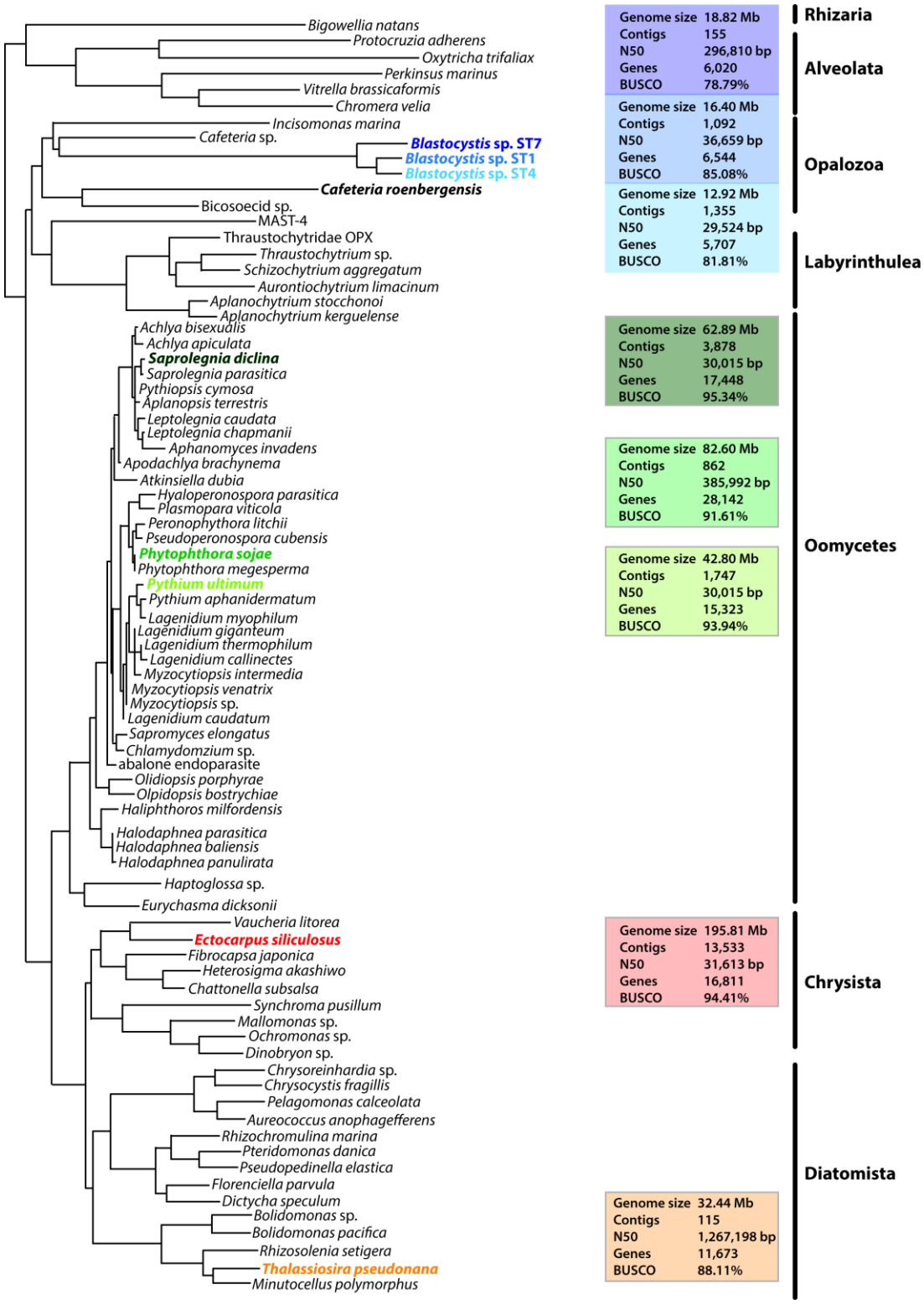


Figure 4.1.1 Stramenopile phylogeny showing the relationships between the genomes used in the comparative analysis. *This is a representation phylogeny redrawn from Derelle* et al. 2016 [353] *and Beakes* et al. *2012* [386]. *Taxa in bold were used in the comparative analysis showing the spread of genomes selected from across the Stramenopiles and coloured taxa show corresponding genome statistics for each assembly.*

The most highly studied group within the Stramenopiles are the Oomycetes, pathogens of plants and fish, which were investigated because of their significant economic impact on crop and fish farming. Over 80 species of *Phytophthora* have been named including *Phytophthora infestans, which* was responsible for the Irish potato famine in the mid- 19th century. *Phytophthora sojae* is a parasite of soybeans causing root and stem rot and costs the industry millions of (US) dollars each year [303]. *P. sojae* is hemibiotrophic and transitions through several stages during the course of an infection beginning with biotrophism but quickly becoming necrotrophic and obtaining nutrients from decaying plant tissue [303]. Tyler *et al.* 2006 [209] sequenced the 82 Mb genome of *P. sojae*. The genome was assembled into 82 scaffolds and 862 contigs and contains 26,489 genes, which scored 91% completeness with BUSCO. Initial analysis of the genome revealed an expansion of protein families associated with plant infections such as hydrolases, ABC transporters, toxins, protease inhibitors and, in particular, a superfamily of 700 proteins with similarity to known Oomycete virulence genes [209], [306].

Related to *Phytophthora* but also within the Oomycetes is *Pythium ultimum.* Members of this genus mostly inhabit soil although some are able to survive in aquatic environments. They are classed either as saprobes or opportunistic plant pathogens and cause a wide array of diseases for a wide variety of hosts, including animals [387], [388], though some preference for monocot or dicot hosts has been observed in plants. *P. ultimum* is a ubiquitous plant pathogen and one of the more pathogenic species for crops. Its genome was sequenced by Lévesque *et al.* 2010 [305] and assembled into 975 scaffolds with 1,747 contigs. The overall genome size is 42 Mb, almost half the size of *P. sojae*. However, the gene complement is 15,323, which accounts for 93% of the total genome according to BUSCO. The main difference between *Pythium* and *Phytophthora* is the process of zoospore production which differs because *Pythium* spp. produce them in vesicles whereas *Phytophthora* spp. produce them from sporangia [389].

The final Oomycete genome included in the analysis is *Saprolegnia diclina,* which is a more basal-branching member of the Class. This is an aquatic parasite of fish and amphibians which has received somewhat less attention than its land-based counterparts [390], [391]. Despite this, it is an economically and ecologically significant parasite targeting eggs and causing severe damage to farmed, as well as wild, species. Only recently has the cause of the infections been attributed to several different species of *Saprolegnia, S. diclina* among them [390]. The 62 Mb

genome was sequenced in 2013 by the Broad Institute and deposited in the NCBI database [392]. The assembly contained 390 scaffolds and 3,878 contigs. There are 17,448 genes annotated and these account for 95% of the genome as predicted by BUSCO.

A free-living out-group for the *Blastocystis* lineage is the phaeophyte alga *Ectocarpus* [353]. These are brown algae in the form of filamentous seaweed that are only very distantly related to green plants and are some of the only Stramenopiles to have evolved complex multicellularity. *E. siliculosus* is a model organism for brown algae and is closely related to the kelps. It is free-living and has a large 195 Mb genome made up of 1,561 scaffolds and 13,533 contigs. the genome contains 16,811 genes that includes a huge arsenal of signal transduction genes linked to its evolution of multicellularity [374]. The gene content has a BUSCO score of 94%.

Related to, but distinct from, the algae are the diatoms. These are unicellular, photosynthetic algae that have a defining morphological feature in the form of a silicified cell wall; a frustule, and unusual chloroplasts [342], [344], [393]. Like brown algae the photosynthetic characteristics of these organisms occurred independently of the acquisition of chloroplasts associated with green plants, though secondary endosymbiosis of a photosynthetic eukaryote. *Thalassiosira pseudonana* is diploid containing 24 chromosome pairs and contains all the metabolic capabilities to support a photoautotrophic organism. The genome is 32 Mb assembled into 64 scaffolds, 115 contigs and codes for 11,673 genes, which is 88% complete by BUSCO. Up to half of the predicted proteins have no orthologs in the database and may reflect novel functions undertaken in this highly unusual organism [342].

4.1.4 Aims of the chapter

In this chapter, I will utilise the *P. lacertae* genome (Chapter 2) and *C. roenbergensis* transcriptome (Chapter 3) in a comparative genomic analysis with *Blastocystis* in order to investigate the changes that occurred in the ancestor of *Blastocystis* that have facilitated the transition into the intestinal environment. There are four specific objectives:

1. Cluster homologous genes between representatives from across the Stramenopiles

2. Identify conserved gene families between *Blastocystis, P. lacertae* and *C. roenbergensis*
3. Identify gene gain and gene loss across the *Blastocystis* genomes using *P. lacertae* as an out-group
4. Identify adaptations in the ancestor of *Blastocystis* that are associated with adaptation to the intestinal environment

4.2 Materials and methods

4.2.1 Selection of appropriate diversity and quality of genome

At the time of writing the number of Stramenopile genomes in the NCBI database is 68. However, only 28 of these were associated with an annotation and were selected for further refinement. A mix of both parasitic and free-living Stramenopiles was required, to keep bias to a minimum, i.e. to prevent over-reliance on a particular subset of Stramenopiles.

Decisions on inclusion were also made based on the assembly statistics for each available assembly, as the quality of the annotation is dependent on the quality of the assembled contigs (Figure 4.1.1). Assemblies with comparatively low scaffold and contig counts and high N50 values were used preferentially. High sequence coverage and BUSCO scores for completeness were also preferred.

The clustering analysis was run using all three published and annotated genome sequences for *Blastocystis* sp. (ST1, ST4 and ST7), as well as the *P. lacertae* genome and the *C. roenbergensis* transcriptome, three parasitic oomycetes genome sequences (*Phytophthora sojae, Pythium ultimum* and *Saprolegnia diclina*), the free-living *Thalassiosira pseudonana* from the Bacillariophyta, and *Ectocarpus siliculosus* from PX clade. Together these sample all parts of Stramenopile diversity for which genome sequences are available.

4.2.2 The genomes used in the analysis

The genomes of *Blastocystis* sp. ST1 ATCC 50177/Nand II [376], *Blastocystis* sp. ST4 WR1 [284], *Blastocystis* sp. ST7 Singapore isolate B [91], *Pythium ultimum* DAOM BR144 [305], *Phytophthora sojae* strain: P6497 [209], *Saprolegnia diclina* VS20 [392], *Ectocarpus siliculosus* strain Ec32 [374] and *Thalassiosira pseudonana* CCMP1335 [342] were downloaded for the comparative analysis and *Bigelowiella natans* CCMP2755*, Bodo saltans* BSAL*, Chromera velia* CCMP2878*, Guillardia theta* CCMP2712*, Saccharomyces cerevisiae* S288c and *Tetrahymena thermophila* SB210 were downloaded for the phylodiversity analysis. All genomes were downloaded from NCBI.

4.2.3 Clustering of orthologous genes

OrthoMCL v2.0.9 [394] and OrthoFinder v0.2.8 [395] were used with default parameters to generate clusters of homologous genes either between species (orthologs) or within species (paralogs). OrthoMCL used an E-value threshold of 1e-5 for step 7: all-v-all BLAST. MySQL was used as the relational database. Sequences that OrthoMCL failed to cluster were assumed to be single-copy species-specific and were therefore grouped with each species unique repertoire.

A cluster was considered 'conserved' if they contained at least one sequence from *P. lacertae, C. roenbergensis* and one *Blastocystis* subtype; these groups may also contain sequences from another Stramenopile genome. A cluster was 'species-specific' if it only contained sequences from a single organism, (except *Blastocystis* where it must be present in at least one subtype but may contain representative sequences from multiple subtypes).

Clusters were considered to represent losses from the *Blastocystis* genomes if they were absent from all *Blastocystis* genomes, but present in both *P. lacertae* and *C. roenbergensis* and at least one other Stramenopile. Similarly, clusters that represent losses from *P. lacertae* and *C. roenbergensis* contain sequences from only *C. roenbergensis* or *P. lacertae* respectively, at least one *Blastocystis* and another Stramenopile.

Otherwise conserved clusters contained sequences from either *Blastocystis, P. lacertae* or *C. roenbergensis* and at least one other Stramenopile. Clusters containing sequences from only *Blastocystis* and *P. lacertae, Blastocystis* and *C. roenbergensis* or *P. lacertae* and *C. roenbergensis* were classed as being gained in the common ancestor of those organisms. All other clusters were those containing sequences from multiple Stramenopile genomes but none from *Blastocystis, P. lacertae* and *C. roenbergensis*. Such clusters are not relevant to this analysis.

4.2.4 Phylogenetic analyses

ClustalW [396] and MUSCLE [327] were used to align sequences with default parameters. Each alignment was also curated either manually or using Gblocks [328] which preserves only conserved blocks of residues. PhyML [245] generated maximum likelihood phylograms using automatic model selection and

providing bootstrap support. Ninja v1.2.2 [397] was used to build neighbour joining trees. Mr Bayes [398] was used to build Bayesian trees and to calculate posterior probability values. Default parameters were used unless otherwise stated.

4.2.5 Enrichment analyses

Domain annotations were assigned to each genome using InterProScan [317] and KEGG annotations by Ghost koala [324] was used to assign KO terms. Enrichment analyses were performed in R [366] using a hypergeometric test. This calculates the probability of a domain being in a particular group based on the occurrences of that domain in the whole genome. For example, it calculates the probability of drawing a while ball from a box containing white and black balls where the number of balls and their colours are known.

4.2.6 Phylodiversity

In order to determine if the *Blastocystis* sp. STs genomes were reduced in very highly conserved gene families, a reference panel of genomes from across the Eukaryota were selected and gene families conserved in all of them were extracted. These gene families were used to build baseline Neighbour-joining phylogenies and phylodiversity was calculated according to Faith 1992 [399] was calculated for each phylogeny (containing a conserved gene family) using the Picante package in R [400]. Phylodiversity was developed as an ecological measure of the species richness within a particular habitat. A higher score is awarded the less related a group of organisms is. I have used phylodiversity as a way of calculating the richness of gene families. This allows the value to be weighted to lower the contribution of genes that that are paralogous and increase the contribution of genes that represent different lineages. Orthologous sequences from *Blastocystis* sp. STs and *P. lacertae* were added to each phylogeny in turn to determine the contribution of these genomes to the increase in phylodiversity. The value for each *Blastocystis* sp. ST was subtracted from the *P. lacertae* value to get a comparison of which genome increased the phylodiversity of each gene family more.

4.3 Results

4.3.1 OrthoMCL clustering results

The clustering analysis using OrthoMCL assigned 113,129 genes from 10 genomes and transcriptomes into 21,700 orthologous groups, with an additional 39,631 sequences excluded from the analysis. The results are summarised in Figure 4.3.1. For the full table of results see S4.3.1.

It is assumed that the 'unclustered' sequences were excluded from the analysis because they have no orthologues or paralogues in any of the genomes included. Therefore, they have been labelled single copy species-specific sequences. OrthoMCL assigned 6,782 sequences from the *P. lacertae* genome, 9,273 from the *C. roenbergensis* transcriptome and 1,084, 894 and 1,039 from *Blastocystis* sp. ST1, ST4 and ST7 respectively to the unclustered group (Figure 4.3.1).

The 'conserved' sequences that feature in the non-overlapping sections of Figure 4.3.1 denote sequences from orthogroups that contain a representative sequence from one of the *P. lacertae, C. roenbergensis* or *Blastocystis* sp. genomes, and another Stramenopile. For example, orthogroup MCL1012 contains 187 sequences; 51 from *P. lacertae*, 66 from *P. sojae,* 42 from *P. ultimum*, 23 from *S. diclina,* 4 from *E. siliculosus* and 1 from *T. pseudonana* but none from any *Blastocystis* sp. STs or *C. roenbergensis*. Thus, these sequences are conserved throughout the Stramenopiles but do not overlap with the closest relatives of *P. lacertae* (Figure 4.3.1).

OrthoMCL also assigned a number of multi-copy species-specific groups containing 13,642 sequences for *P. lacertae*, 6,291 for *C. roenbergensis* and 1,899, 1,647 and 1,549 for *Blastocystis* sp. ST1, ST4 and ST7 respectively. These are species-specific and are labelled 'specific' in Figure 4.3.1 to differentiate them from 'conserved 'orthogroups (above). *P. lacertae* had the largest repertoire of species-specific genes of any genome included in the analysis. The largest of these contained over 800 proteins, though by definition they are specific to *P. lacertae* and therefore many share little homology to sequences in the database.

There is a core group of 892 conserved orthogroups that are found in at least one *Blastocystis* genome, the *C. roenbergensis* transcriptome and the *P. lacertae* genome. They may also be present in one or more of the other Stramenopile genomes. 527 of these contain representatives from all the genomes included in the analysis, the largest of which (MCL1021) contains 131 sequences overall, 8 of which are from *P. lacertae*, 3 are from *C. roenbergensis*, 8 from *Blastocystis* sp. ST1, 5 from *Blastocystis* sp. ST7 and 2 are from *Blastocystis* sp. ST4. 134 of these orthogroups represent single-copy orthogroups, containing only a single sequence from each genome.

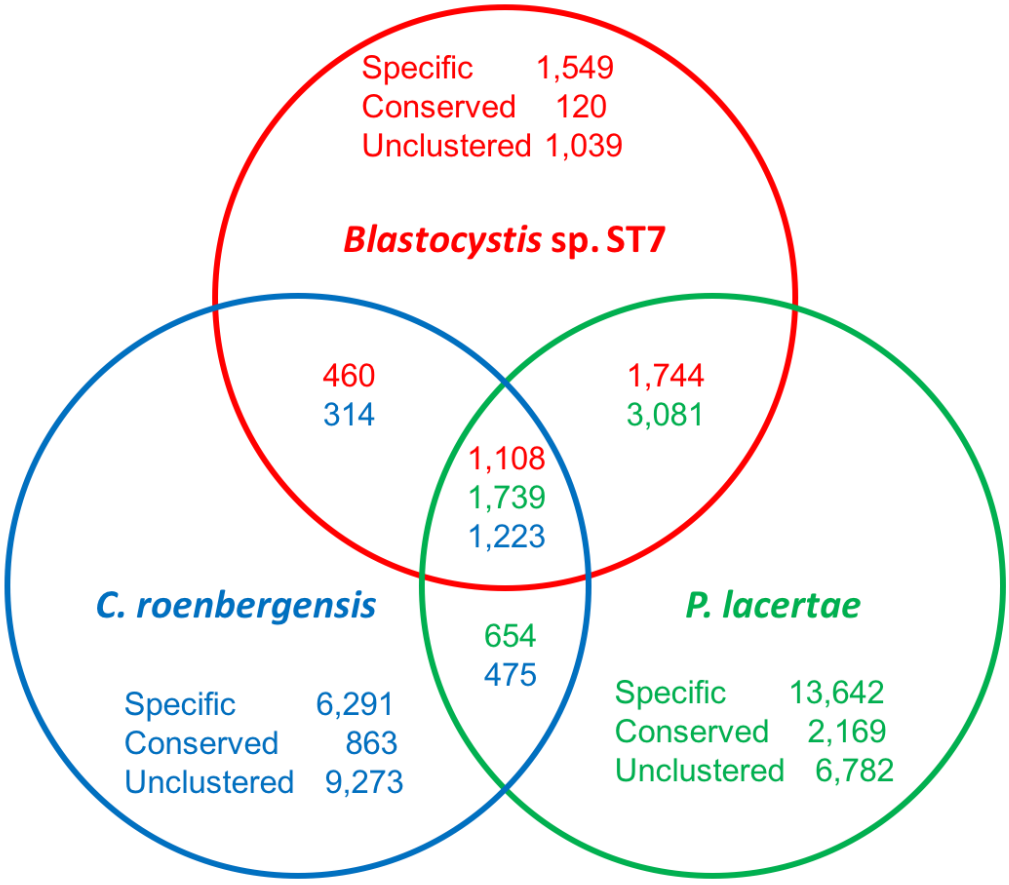


Figure 4.3.1 Venn diagram showing numbers of shared orthologs between Blastocystis sp. ST7, P. lacertae and C. roenbergensis. *Sequences clustered by OrthoMCL. 'Specific' refers to sequences from groups containing only one species, 'conserved' sequences are those with no orthologs in the other genomes shown here but which do have orthologs in other genomes included in the analysis and 'unclustered' sequences are assumed to represent single-copy species-specific proteins.*

Blastocystis sp. ST7 appears to share fewer genes with *C. roenbergensis* than *P. lacertae* does. This is an interesting result as the phylogeny (Figure 2.3.6) shows that *P. lacertae* and *Blastocystis* are equidistant from *C. roenbergensis* and it may therefore be expected they would share roughly equal number of genes. This also shows that, while the *Blastocystis* genomes are much smaller than the *P. lacertae* genome, the majority of this size difference is accounted for by the large species-specific repertoire of *P. lacertae*.

Interestingly, there were also substantial differences between the subtypes of *Blastocystis*, not just in terms of copy number within orthogroups but also in representation between orthogroups. The differences between the *Blastocystis* sp. STs species-specific repertoires is summarised in Figure 4.3.2. Only ~1000 *Blastocystis*-specific genes are shared between all subtypes and each genome appears to have its own repertoire of subtype-specific genes greater than the number of genes shared between all three.

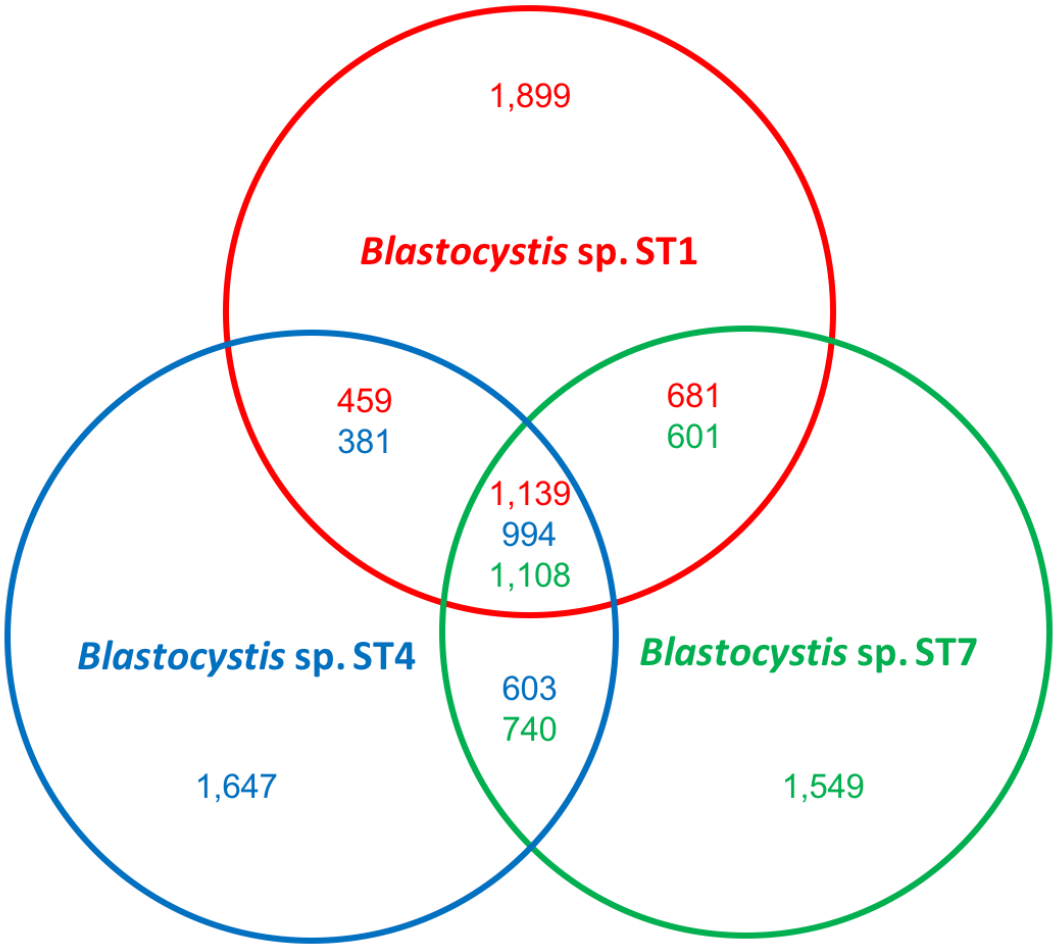


Figure 4.3.2 Venn diagram showing differences in orthologs between Blastocystis-specific sequences. *Sequences clustered by OrthoMCL and represent the number sequences from each subtype in Blastocystis-specific orthologous groups.*

As the *Blastocystis* sp. STs have smaller genomes than other Stramenopiles, the origin of those differences was investigated. The largest differences came from species-specific orthogroups in *P. lacertae*, however, examination of conserved orthogroups showed differences in the number of sequences contributed to conserved groups by *Blastocystis* sp. STs. To understand these differences, the orthogroups were further broken down into conserved groups with differing conservation profiles (Figure 4.3.3).

'Lineage specific' orthogroups were defined as any orthogroup that contained representatives from both *P. lacertae* and *Blastocystis* sp. STs only, such as MCL2961 which contains a single sequence from both *Blastocystis* sp. ST4 and ST7 as well as 8 sequences from *P. lacertae*. Only a single *Blastocystis* subtype

need be represented. In *C. roenbergensis* this was defined by including an orthogroup with a *C. roenbergensis* sequence and at least one other Stramenopile, but absent in *P. lacertae* and *Blastocystis*. 'Core Stramenopile' orthogroups must contain at least one sequence from each genome such as MCL1021 (above). 'Lost from Stramenopiles' are orthogroups which contain sequences from any two genomes of *P. lacertae, Blastocystis* sp. STs or *C. roenbergensis* but do not contain a sequence from another Stramenopile such as MCL5023 which contains a single sequence from *C. roenbergensis*, *P. lacertae* and *Blastocystis* sp. ST4 and two sequences each from *Blastocystis* sp. ST1 and ST7.

The final group, 'lost from sister species' are examples of orthogroups with representative sequences from *C. roenbergensis* or another Stramenopile and either *P. lacertae* or *Blastocystis* sp. but not both. While, *P. lacertae* contains more orthologues than all *Blastocystis* sp. STs in each of these conserved orthogroups, the difference is most striking in the 'lost from sister species' group (Figure 4.3.3). This group is indicative of cases where it is expected that there would be roughly equal numbers orthologues between *P. lacertae* and *Blastocystis* sp. but all *Blastocystis* subtypes included here appear to have fewer.

Conserved orthogroups were examined and found to routinely contain fewer representative sequences from *Blastocystis.* These are sequences that are otherwise conserved in the Stramenopiles but are present in lower copy number in all *Blastocystis* subtypes despite the differences between them.

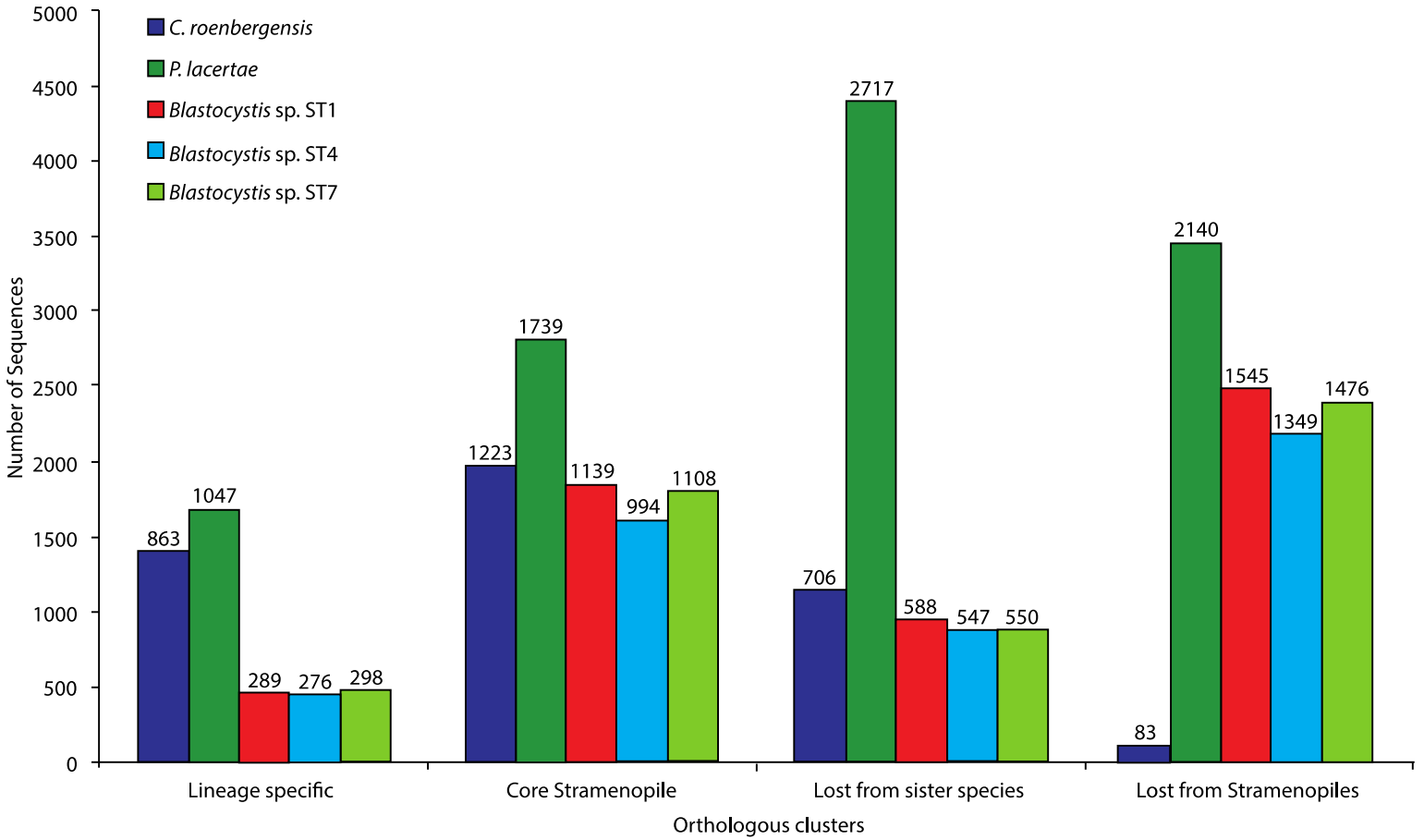


Figure 4.3.3 Number of sequences contained in different cluster categories assigned by OrthoMCL for multiple Stramenopile genomes. *Clusters have been defined as 'lineage specific' (only found in a given species), 'core Stramenopile' (found in all Stramenopile genomes), 'lost from sister species' (present in 2/3 species shown here), and 'lost from Stramenopile' (present in all three species shown here but absent in other Stramenopiles).*

From figure 4.3.3 it is clear that the number of conserved sequences in *P. lacertae* is consistently higher than any *Blastocystis* genome, while the number of conserved sequences lost from *Blastocystis* is substantially smaller than the corresponding number in *P. lacertae*. This suggests that the *Blastocystis* sp. STs genomes are missing sequences from highly conserved gene families, common across the large phylogenetic distances within the Stramenopiles. Comparison with *P. lacertae* demonstrates that the absence of these widely conserved genes is unique to *Blastocystis* and must have occurred after the separation of *Blastocystis* and *Proteromonas* from their common ancestor.

4.3.2 OrthoFinder clustering results

The clustering analysis using OrthoFinder assigned 152,760 genes from the 10 genomes and transcriptomes to 69,052 orthologous groups. With an additional 6,200 sequences excluded from the analysis. OrthoFinder failed to include 1,070 sequences from the *P. lacertae* genome, 0 from the *C. roenbergensis* transcriptome and 876, 745 and 696 from *Blastocystis* sp. ST1, ST4 and ST7 respectively. It also assigned a number of multi-copy species-specific groups containing 15,964 sequences for *P. lacertae*, 14,128 for *C. roenbergensis* and 1,703, 1,249 and 1,551 for *Blastocystis* sp. ST1, ST4 and ST7 respectively. *P. lacertae* had the largest repertoire of species-specific genes of any genome included in the analysis (Figure 4.3.4), though this number differed from that produced in the OrthoMCL analysis and is closely followed by the *C. roenbergensis*-specific sequences, which is a marked difference between OrthoMCL and OrthoFinder. For full table of results see S4.3.2.

Overall, OrthoFinder assigned more sequences from each genome to ‘conserved’ orthogroups (present in *P. lacertae, C. roenbergensis,* at least one *Blastocystis* genome and at least one other Stramenopile) and to specific orthogroups (present in only one genus), except for *Blastocystis* sp. ST1 and ST4.

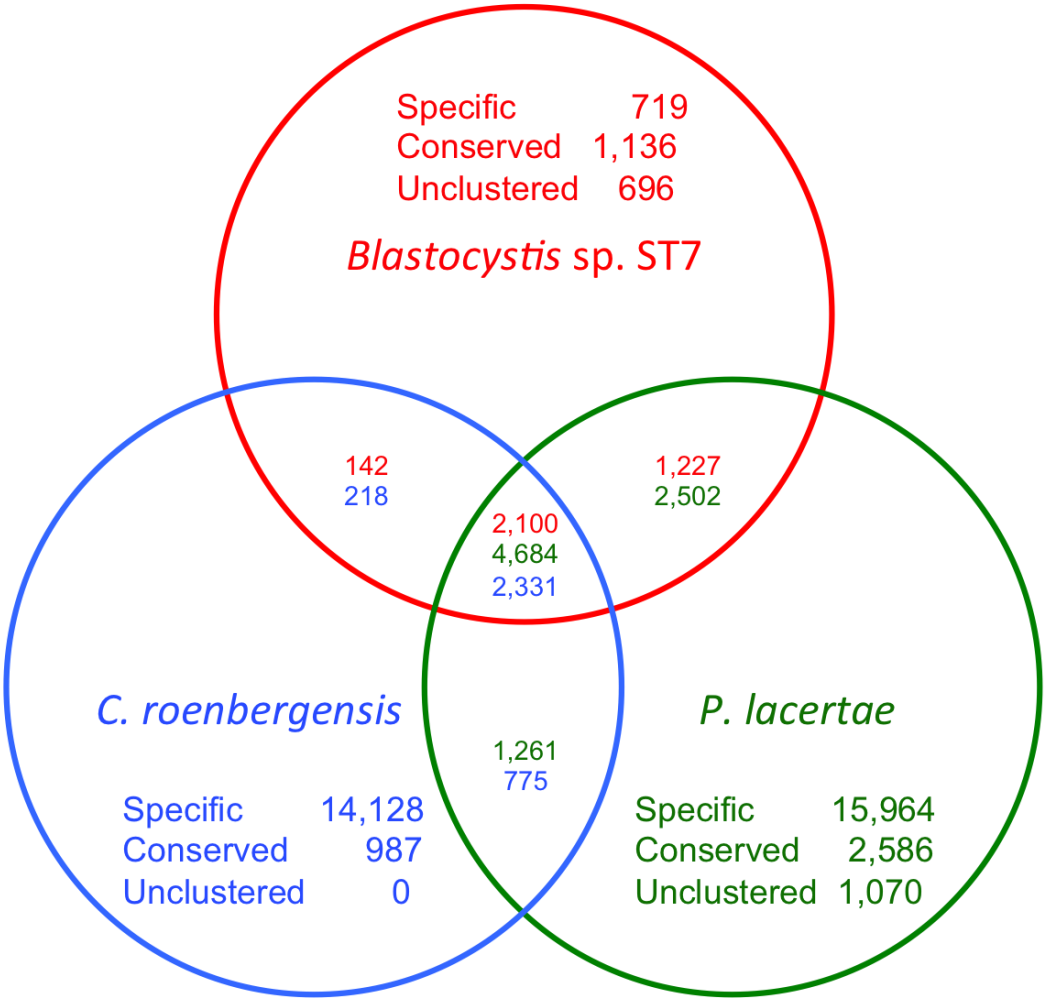


Figure 4.3.4 Venn diagram showing differences in orthologs between <u>Blastocystis</u>-specific sequences. *Sequences clustered by OrthoFinder. 'Specific' refers to sequences from groups containing only one species, 'conserved' sequences are those with no orthologs in the other genomes shown here but which do have orthologs in other genomes included in the analysis and 'unclustered' sequences are assumed to represent single-copy species-specific proteins.*

While the individual numbers of genes in the relationships between these three genomes are slightly different between OrthoMCL and OrthoFinder, the overall outcomes of the analyses are comparable: *Blastocystis* sp. ST7 still shares fewer genes with *C. roenbergensis* than *P. lacertae* and the *Blastocystis* genomes are still noticeably smaller when compared with the *P. lacertae* genome, although OrthoFinder has predicted a similarly sized species-specific repertoire in *C. roenbergensis* that's absent from the OrthoMCL output. It also appears as though *P. lacertae* is overrepresented in all shared groups, again consistent with the results from OrthoMCL.

The output from OrthoFinder also corroborates the findings that *Blastocystis* sp. STs contain fewer genes in conserved orthogroups (Figure 4.3.5). Although OrthoFinder has assigned fewer genes to the 'lineage specific' group, the greatest disparity between *Blastocystis* sp. and *P. lacertae* remains the 'lost from sister species' group.

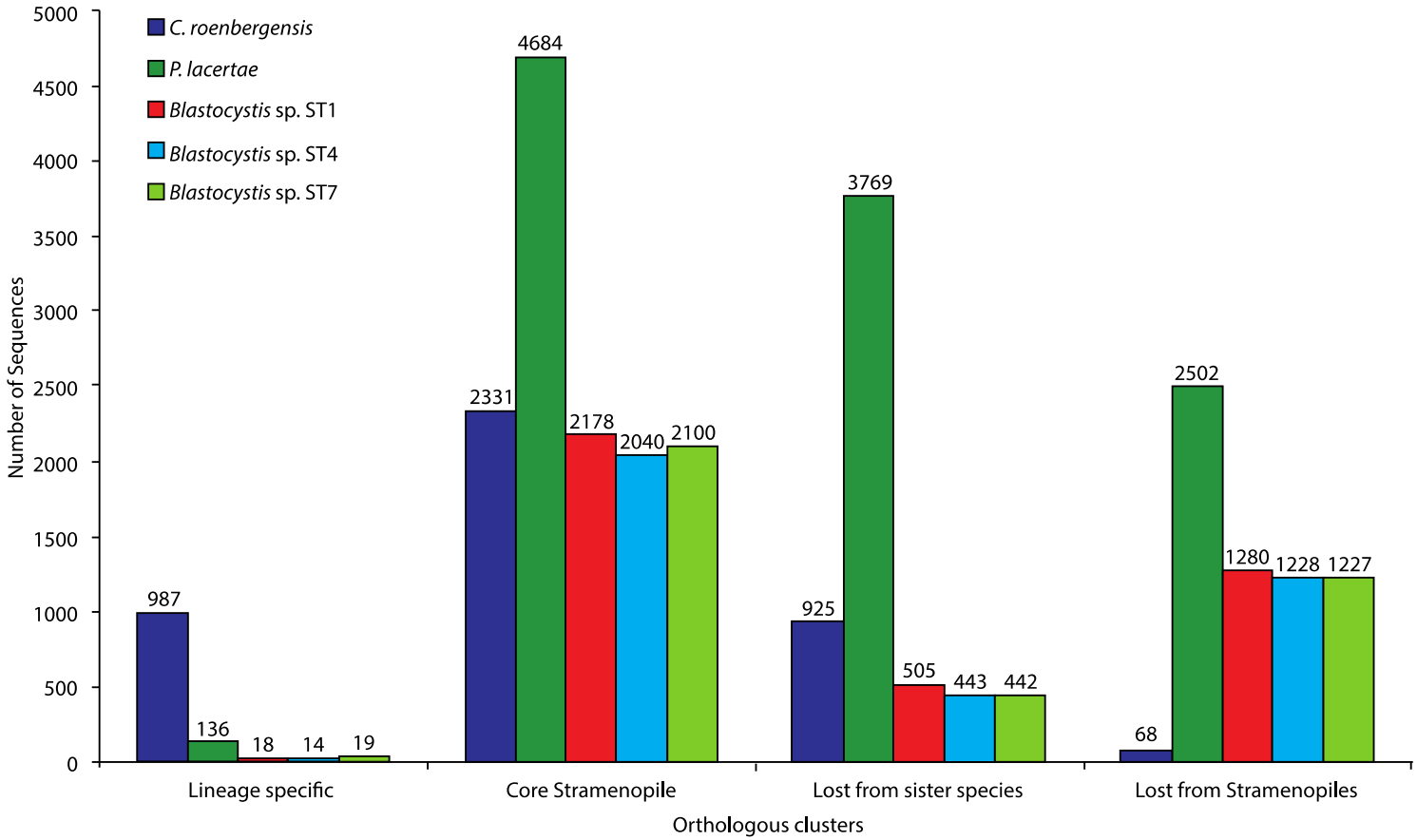


Figure 4.3.5 Number of sequences contained in different cluster categories assigned by OrthoFinder for multiple Stramenopile genomes. *Clusters have been defined as 'lineage specific' (only found in a given species), 'core Stramenopile' (found in all Stramenopile genomes), 'lost from sister species' (present in 2/3 species shown here), and 'lost from Stramenopile' (present in all three species shown here but absent in other Stramenopiles).*

While the outputs of OrthoMCL and OrthoFinder differ in the number of sequences assigned to orthogroups, the majority of the findings are consistent between the clustering programmes. *P. lacertae* contains a high proportion of species-specific genes, which may or may not be a similar proportion to *C. roenbergensis*. Despite this, all *Blastocystis* subtypes are routinely underrepresented in conserved gene clusters, especially with respect to sequences that appear to be absent with respect to *P. lacertae*.

4.3.3 *Blastocystis* gains are enriched for diverse domains

To examine the functional relevance of *Blastocystis*-specific genes identified in the clustering analysis, these sequences were subjected to a domain enrichment analysis (Table 4.3.1, S4.3.3, S4.3.4) in R. OrthoMCL and OrthoFinder gave different results in the clustering analysis with respect to the overall numbers of genes gained and lost. OrthoMCL found 156 domains to be significantly enriched in *Blastocystis* sp. STs, whereas OrthoFinder found only 35 significantly enriched domains (where $p < 0.001$).

While individual domains differ between the enrichment analyses for the OrthoMCL and OrthoFinder clusters, and therefore their associated protein sequence, there is some crossover between the enrichment results. For example, both enrichment analyses include Ankyrin or ANK repeat domains ($p = 1.30e^{-21}$ OrthoMCL, $p = 6.97e^{-3}$ OrthoFinder), globular domains like Putative Ig domain or Von Willebrand factor type A domain signature ($p = 4.52e^{-74}$ OrthoMCL, $p = 5.88e^{-3}$ OrthoFinder) as well as membrane associated domains: Major Facilitator Superfamily ($p = 1.02e^{-4}$ OrthoMCL) and src homology 3 domains ($p = 2.41e^{-4}$ OrthoFinder). These suggest that while a specific sequence may have been placed differently, based on differing thresholds for homology, some of the same types of protein may characterise *Blastocystis* gains according to both programmes.

The enrichment from both OrthoMCL and OrthoFinder show that *Blastocystis* sp. STs have acquired novel genes associated with a range of cellular functions but that a majority of these can be linked to factors required for survival in a host environment. Among these are peptidases from the C13 family (IPR001096) ($p = 1.75e^{-5}$ OrthoMCL, $p = 1.04e^{-3}$ OrthoFinder) (not included in the top 10 for OrthoMCL but still significant, S4.3.3). There are 17 sequences annotated with this domain in *Blastocystis* sp. ST1, 10 in *Blastocystis* sp. ST4 and 12 in *Blastocystis* sp.

ST7. Of the total 39 *Blastocystis* sequences, 17 (OrthoMCL) or 13 (OrthoFinder) are *Blastocystis*-specific. By comparison, the *P. lacertae* genome contains 21 sequences annotated with this domain and none in the *C. roenbergensis* transcriptome. Peptidases are known factors that enable persistence within a host and so expansion of these in *Blastocystis* may indicate biological significance.

Globular domains such as putative Ig domains (IPR008009) ($p = 4.52e^{-74}$ OrthoMCL) and Von Willebrand domain signatures (IPR002035) ($p = 5.88e^{-3}$ OrthoFinder) may also be important for cellular adhesion *in vivo*. *Blastocystis* sp. ST1, ST4 and ST7 contained 64, 158 and 59 putative Ig domain-containing proteins respectively in the OrthoMCL analysis. Of the 181 total putative Ig domain-containing proteins, 140 were specific to *Blastocystis* sp. STs. In contract to this, *P. lacertae* contains 136 sequences with this annotation compared to *C. roenbergensis*, which does not contain any. *Blastocystis* sp. ST1, ST4 and ST7 also contained two, five and two Von Willebrand domain-containing proteins respectively in the OrthoFinder analysis. Of the nine sequences four are specific to *Blastocystis* sp. STs however, *P. lacertae* does not contain any while *C. roenbergensis* contains 18 sequences with this annotation.

Table 4.3.1 Top 10 domains significantly enriched in *Blastocystis* spp. gains for OrthoMCL and OrthoFinder

OrthoMCL				OrthoFinder			
Domain Names	**InterPro domain**	**Gains**	**P-value**	**Domain Names**	**InterPro domain**	**Gains**	**P-value**
Putative Ig domain	IPR008009	140	4.52E-74	Ankyrin repeat-containing domain	IPR020683	16	6.97E-03
Hsp70 protein	IPR018181	73	5.32E-11	Peptidase C13 family	IPR001096	13	1.04E-03
Serine/threonine phosphatase family signature	IPR006186	57	1.67E-39	EF-hand domain	IPR002048	9	1.47E-03
Transforming protein P21 ras signature	PR00449	53	2.12E-22	Protein disulfide isomerase family	cd02961	8	5.85E-03
ankyrin repeats	IPR002110	40	1.30E-21	Src homology 3 domain	cd00174	5	2.41E-04
Major Facilitator Superfamily	IPR011701	31	1.02E-04	Glycosyltransferase Family 4	IPR028098	4	9.86E-03
zinc finger	IPR013085	29	6.06E-15	Aplha-amylase catalytic domain family	cd00551	3	<1.00E-100
Thioredoxin	IPR001853	29	1.11E-04	Von Willebrand factor type A domain signature	PR00453	3	5.88E-03
DnaJ domain signature	IPR001623	28	1.28E-07	Hemerythrin	cd12107	2	<1.00E-100
Tetratricopeptide repeats	IPR019734	27	1.53E-07	Leucine rich repeat	IPR001611	2	<1.00E-100

4.3.4 Putative Ig domains

Putative Ig domains were the most highly enriched result from OrthoMCL as well as being potentially linked to activity at the cell surface the repertoire of putative Ig domains was investigated (PF05345). The sequences were extracted from the InterProScan output, aligned with ClustalW and used to build a phylogeny with PhyML (Figure 4.3.6). As there are no representative species from *C. roenbergensis* it is difficult to infer lineages, however, the phylogeny does show expansion of *Blastocystis* sequences in each subtype, especially in *Blastocystis* sp. ST4. These may therefore be required for subtype-specific activities, and may play a role in determining host range

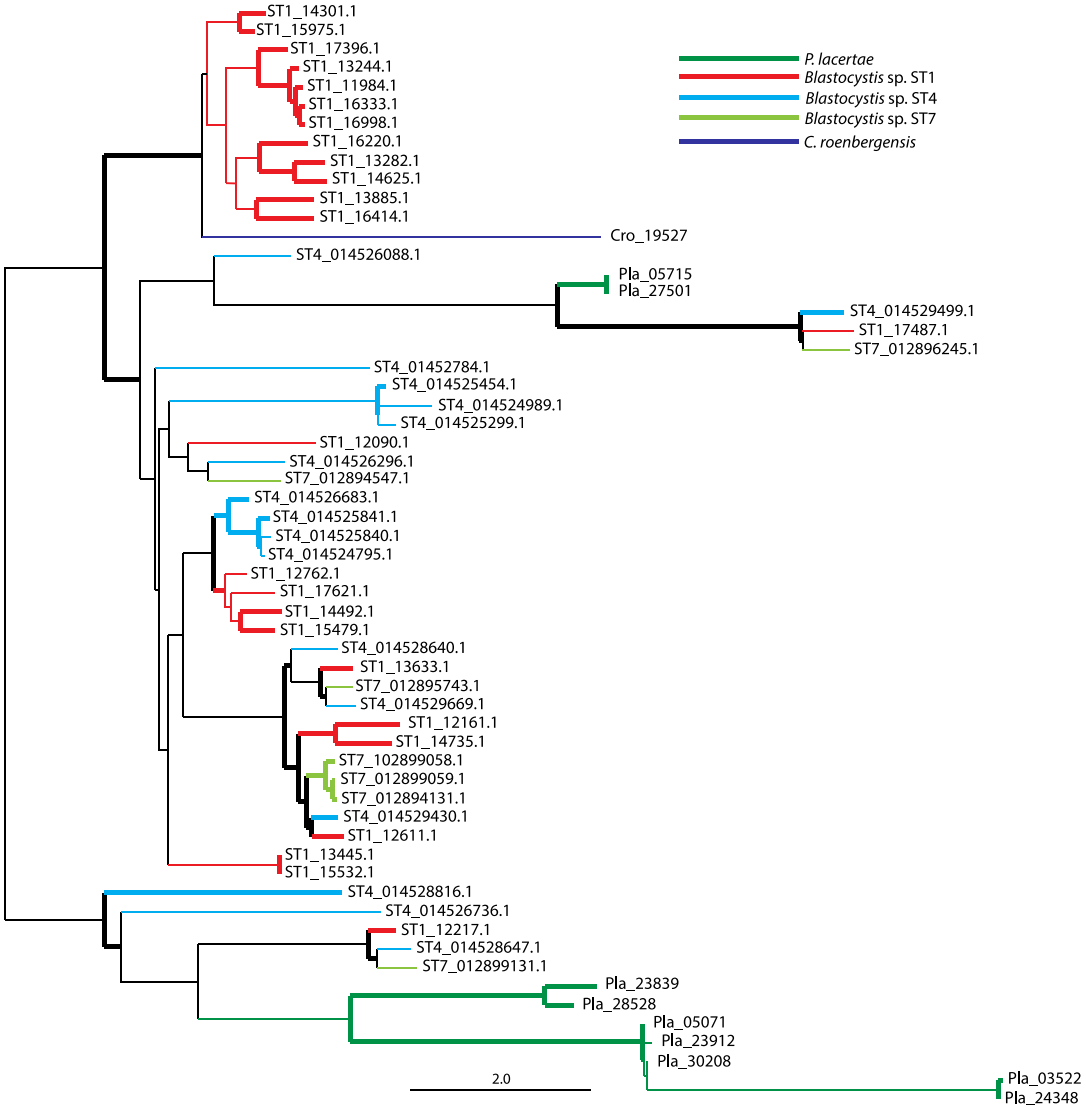


Figure 4.3.6 Maximum likelihood phylogeny of Ig domains-containing proteins built using WAG+G+F model and rooted at the midpoint*. Bold lines indicate bootstrap support >75. Dark green branches are* P. lacertae*, red branches are*

Blastocystis *sp. ST1, light green are* Blastocystis *sp. ST7, light blue are* Blastocystis *sp. ST4 and dark blue are* C. roenbergensis.

The tree appears to show *P. lacertae*-specific expansions as well as *Blastocystis*-specific ones. While this is difficult to determine in the absence of an outgroup, this tree may show the sensitivity of the enrichment analysis, in that it is able to detect 140 *Blastocystis*-specific putative Ig domain-containing proteins in the presence of *P. lacertae*-specific proteins.

4.3.5 Peptidase C13

To further investigate the results of the clustering analysis, all sequences from *Blastocystis* sp. STs*, P. lacertae* and *C. roenbergensis* containing a peptidase C13 family domain (PF01650) were extracted, aligned and used to build a phylogeny as above (Figure 4.3.7). This domain was selected because of its potential role in persistence within a host and virulence. Again, there were no sequences within the *C. roenbergensis* transcriptome annotated with this domain. The phylogeny shows again that there are species-specific expansions in both *Blastocystis* and in *P. lacertae*.

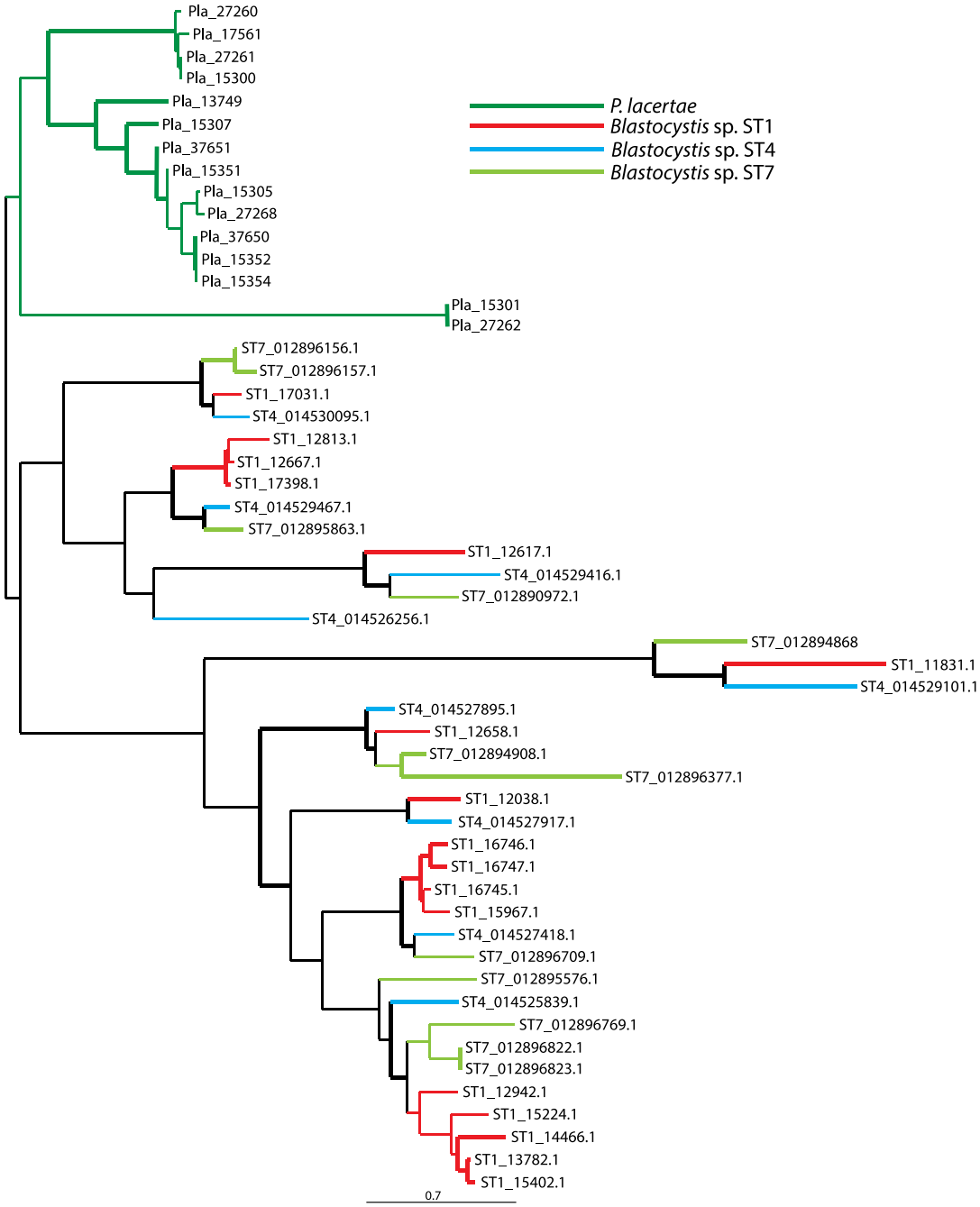


Figure 4.3.7 Maximum likelihood phylogeny of C13 family peptidase domains-containing proteins built using WAG+G+F model and rooted at the midpoint. *Bold lines indicate bootstrap support > 75.* P. lacertae *sequences are shown in dark green,* Blastocystis *sp. ST1 is shown in red,* Blastocystis *sp. ST4 show in light blue and* Blastocystis *sp. ST7 is shown in light green. There is clear separation of* P. lacertae *sequences from* Blastocystis *indicating independent expansion of this family in both* Proteromonas *and* Blastocystis.

The sequences included in this phylogeny show orthology between the *Blastocystis* subtypes and *P. lacertae*. Again, the lack of orthologues in *C. roenbergensis* makes the results of this difficult to interpret. In order to investigate

this, the orthogroups, produced by OrthoFinder, containing these sequences were examined. The 58 sequences correspond to nine orthogroups of which only six correspond to true *Blastocystis* gains. Orthogroup FDR02266 contains six sequences from *Blastocystis* sp. ST1, three from ST4 and one from ST7. FDR03217 contains five sequences form ST1, one from ST4 and two from ST7. FDR07097, FDR07101 and FDR07142 contain a single sequence from each *Blastocystis* subtype and FDR09058 contains a single sequence from ST1 and ST4 only. The clustering analysis, then, identified 29 *Blastocystis*-specific C13 peptidase domain-containing sequences of the 39 in all three genomes, however, the enrichment analysis identified only 13 of these as being significant.

These results have shown that while *Blastocystis* sp. STs contains expansions of these genes, they have also expanded independently in *P. lacertae* and are therefore not unique to *Blastocystis*.

4.3.6 *Blastocystis* losses are enriched for general functions

In order to interpret the functional significance of *Blastocystis* losses, conserved *P. lacertae* genes absent from *Blastocystis* were examined in a domain enrichment analysis. To qualify as a *Blastocystis* loss, a group must contain no representative sequences from any *Blastocystis* genome, at least one sequence from *P. lacertae* or *C. roenbergensis* and at least one sequence from another Stramenopile, to ensure that *Proteromonas*- and *Cafeteria*-specific sequences were not misinterpreted.

The analysis identified 260 domains that were significantly enriched among *Blastocystis* losses according to OrthoMCL, and 310 from the output of OrthoFinder. These domains were associated with diverse functions (Table 4.3.2, S4.3.5, S4.3.6). The most striking results were those for the EGF domains (IPR000742) ($p = 8.53e^{-12}$). However, closer examination of the orthogroups revealed that this result is most likely due to high copy number in the *P. lacertae* genome. HMMs were built and used to search the other Stramenopile genomes for homologues to the EGF domain-containing proteins. These scores were then used to build a network. Inclusion of all these scores resulted in a single cluster in which no detail could be discerned. By only taking the top 30% of HMM scores (138 - 9,650) meaningful clusters were defined that allowed for analysis of the data (Figure 4.3.8). This analysis showed not only that *Blastocystis* sp. STs contained homologues, but also

that *P. lacertae* contained a large species-specific expansion. This may be the cause for the appearance of this domain in the enrichment analysis. Significance (p-value) is based on the difference in copy number between the *P. lacertae* genes in the whole genome and the *P. lacertae* genes that form a cluster from which *Blastocystis* is absent. If the total number of genes annotated with a particular domain is high in the *P. lacertae* genome, in this case due to a species-specific expansion, then this may have the effect of skewing the p-value, resulting in inflated significance of the enrichment.

The only domains to feature in both analyses were the EF hand calcium-binding domain (IPR002048) ($p = 4.35e^{-5}$ OrthoMCL, $p = 2.20e^{-4}$ OrthoFinder), Trp/Asp WD repeats (IPR001680) ($p = 1.17e^{-4}$, OrthoMCL, $p = 5.22e^{-5}$ OrthoFinder) and Major Facilitator Superfamily MFS domain (IPR011701) ($p = 1.47e^{-7}$ OrthoMCL, $p = 1.10e^{-12}$ OrthoFinder).

Table 4.3.2 Top 10 domains significantly enriched in *Blastocystis* spp. losses for OrthoMCL and OrthoFinder

OrthoMCL				OrthoFinder			
Domain Names	**InterPro domain**	**Losses**	**P-value**	**Domain Names**	**InterPro domain**	**Losses**	**P-value**
EGF-like domain signature 1	IPR013032	43	8.53E-12	Ion transport protein	IPR005821	30	1.10E-26
Trp/Asp WD repeats profile	IPR001680	23	1.17E-04	EF-hand calcium-binding domain profile	IPR002048	30	2.20E-04
EF-hand calcium-binding domain profile	IPR002048	19	4.35E-05	Trp/Asp WD repeats circular profile	IPR001680	27	5.22E-05
Ion transport protein	IPR005821	11	1.99E-09	Fibronectin type-III domain profile	IPR003961	26	1.99E-03
Major Facilitator Superfamily	IPR011701	10	1.47E-07	Acyltransferase family	IPR004299	21	4.61E-31
Kinesin motor domain profile	IPR001752	8	7.52E-04	Major Facilitator Superfamily	IPR011701	20	1.10E-12
Caenorhabditis protein of unknown function DUF268	IPR004951	7	<1.00E-100	TTL domain profile	IPR004344	19	2.43E-24
GTP cyclohydrolase II	IPR032677	7	<1.00E-100	Tubulin/tyrosine ligase family	IPR004344	19	4.85E-23
Voltage-dependent L-type calcium channel IQ-associated	IPR031649	7	<1.00E-100	IQ motif profile	IPR000048	19	2.76E-03
TTL domain profile	IPR004344	7	1.12E-08	Zinc finger RING-type profile	IPR001841	16	5.03E-05

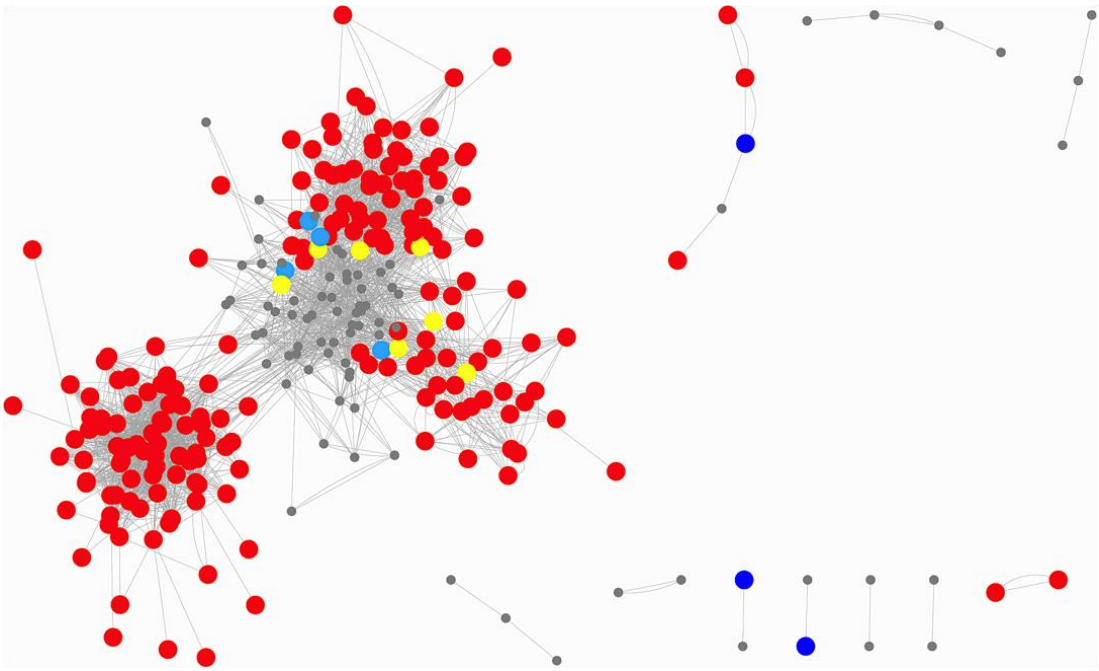


Figure 4.3.8 Network of EGF domains based on HMMER similarity scores*.* P. lacertae *sequences are shown in red,* Blastocystis *sp. STs sequences are shown in dark blue,* C. roenbergensis *sequences are shown in light blue and* Schizochytrium *sequences are shown in yellow. Grey points represent other Stramenopile sequences. This network was built using the top 30% of HMM scores allowing for definition between clusters. While* Blastocystis *sp. STs have three conserved homologues, there are at least two separate* P. lacertae*-specific expansions.*

Interestingly, MFS domain-containing proteins were also enriched among *Blastocystis* gains after OrthoMCL clustering (IPR011701). For *Blastocystis* gains, MFS domains were annotated on 99 sequences, of which 31 were specific to *Blastocystis* ($p = 1.02e^{-4}$). In contrast, in the *Blastocystis* losses for OrthoMCL, there are 72 MFS annotated sequences in the *P. lacertae* genome, of which 10 do not have an orthologue in *Blastocystis*. MFS transporters are a very diverse family of proteins and thus there may be gains and losses within different families labelled with the broad domain annotation of MFS. However, the OrthoFinder output shows enrichment of MFS domains of the same annotation as in the OrthoMCL *Blastocystis* gains (IPR011701) ($p = 1.10e^{-12}$). This shows 72 sequences with this annotation in the *P. lacertae* genome of which 20 lack an ortholog in *Blastocystis*. In order to test loss and gain within this family, the sequences were extracted from the InterProScan output, aligned with ClustalW and used to build a phylogeny (Figure 4.3.9).

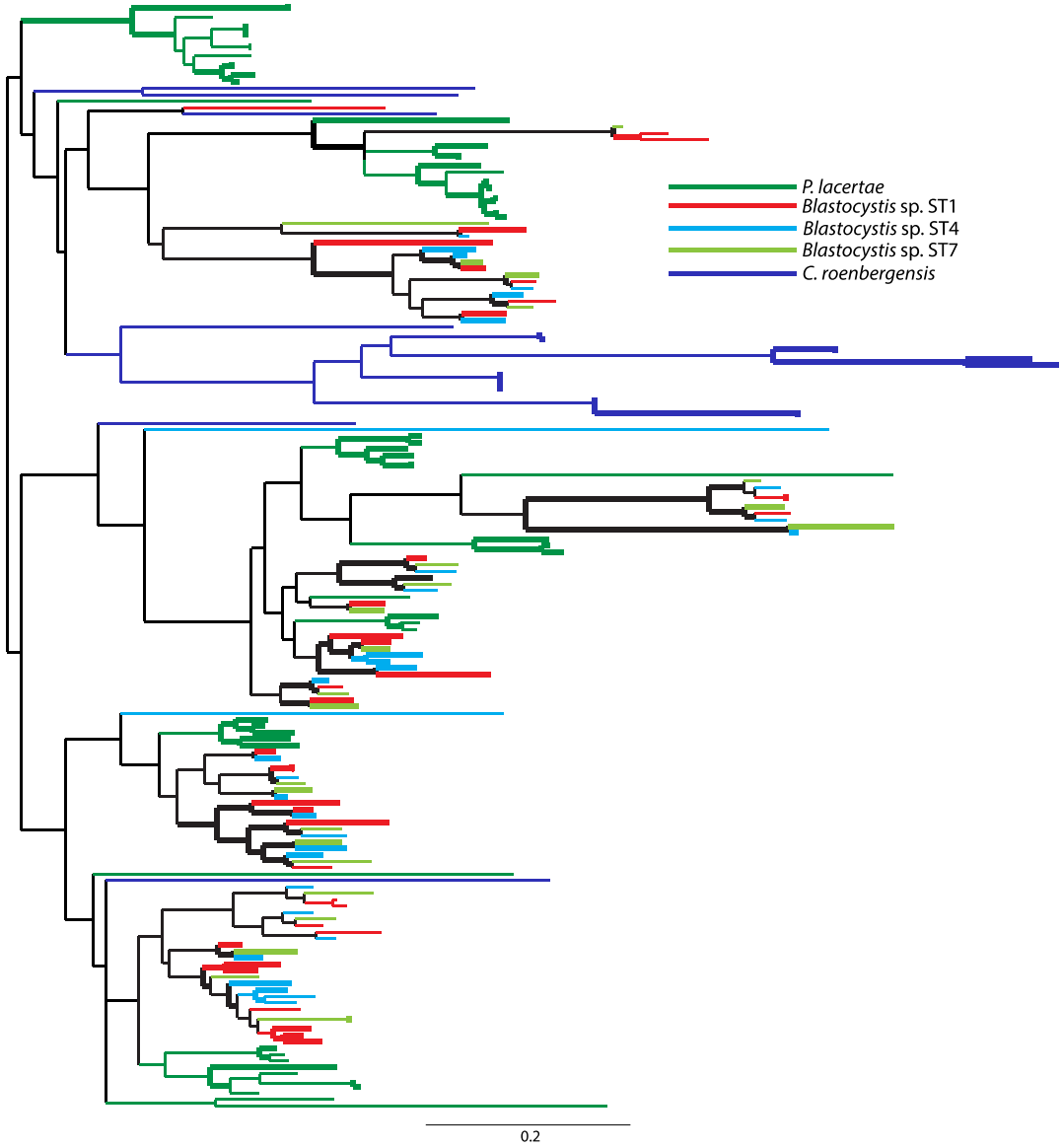


Figure 4.3.9 Maximum likelihood phylogeny of MFS domain-containing sequences built using VT+G+F model and rooted at the midpoint*.* P. lacertae *is shown in green,* Blastocystis sp. ST1 *is shown in red,* Blastocystis *sp. ST4 is shown in light blue,* Blastocystis *sp. ST7 is shown in light green and* C. roenbergensis *sequences are shown in dark blue.*

From the phylogeny, it is clear that there has been some species-specific expansion in both *P. lacertae* and *Blastocystis* sp. STs however; it is not obvious where the *Blastocystis* loss has occurred. The MFS enrichment in the *Blastocystis* losses may be the result of *P. lacertae*-specific expansions.

Overall, inferred *Blastocystis* sp. STs losses are highly diverse, but they are dependent on the genome content of *P. lacertae* and that *P. lacertae*-specific

expansions can suggest gene loss in *Blastocystis* where phylogenetic analysis demonstrates that it has not occurred.

4.3.7 *P. lacertae* gains and losses reflect wide scale genomic changes

In order to compare the *Blastocystis* gains and losses, enrichment analyses were also carried out on the *P. lacertae* genome. Enrichment of the OrthoMCL data contained 180 significantly enriched domains gained and 103 lost while the OrthoFinder data contained 263 significantly enriched domains gained and 222 lost. Losses were defined as any orthogroups with representatives from *C. roenbergensis* and at least one *Blastocystis* sp. STs as well as another Stramenopile. Gains were defined as any orthogroup containing only *P. lacertae* sequences.

The top 10 *P. lacertae* gains are shown in Table 4.3.3 (S4.3.7, S4.3.8, S4.3.9, S4.3.10) are broadly similar between OrthoMCL and OrthoFinder. Both show enrichment for similar domains such as BTB/POZ domains (IPR000210) ($p = 7.84e^{-4}$ OrthoMCL, $p = 1.34e^{-11}$ OrthoFinder), Ulp1 protease (IPR003653) ($p = 1.11e^{-6}$ OrthoMCL, $p = 8.19e^{-14}$ OrthoFinder) and Glycosyltransferase family 92 (IPR008166) ($p = 4.62e^{-4}$ OrthoMCL, $5.39e^{-5}$ OrthoFinder). These domains are associated with a range of functions including genome regulation, and cell surface-associated activity.

The top 10 *P. lacertae* losses are shown in Table 4.3.4 and again show consistency in the results between OrthoMCL and OrthoFinder. Both outputs show significant enrichment of ribosomal proteins (IPR001976) ($p = 9.67e^{-4}$ OrthoMCL, $p < 1.00e^{-100}$ OrthoFinder) and dynamin domains (IPR000375) ($p < 1.00e^{-100}$ OrthoMCL, $p < 1.00e^{-100}$ OrthoFinder) among a large array of other functions spread across cellular physiology.

Table 4.3.3 Top 10 domains significantly enriched in *P. lacertae* gains in OrthoMCL and OrthoFinder

OrthoMCL				OrthoFinder			
Domain Names	**InterPro domain**	**Gains**	**P-value**	**Domain Names**	**InterPro domain**	**Gains**	**P-value**
7 transmembrane sweet-taste receptor of 3 GCPR	IPR017978	109	4.20E-12	Periplasmic binding protein domain	PF13433	322	3.78E-03
BTB/POZ domain	IPR003131	87	7.84E-04	BTB/POZ domain	IPR003131	87	1.34E-11
Integrase catalytic domain profile	IPR001584	73	<1.00E-100	Integrase catalytic domain profile	IPR001584	73	<1.00E-100
Ulp1 protease family C-terminal catalytic domain	IPR003653	63	1.11E-06	Sperm-tail PG-rich repeat	IPR010736	66	4.28E-03
Glycosyltransferase family 92	IPR008166	33	4.62E-04	Ulp1 protease family C-terminal catalytic domain	IPR003653	63	8.19E-14
Rhodopsin-like GPCR transmembrane domain	IPR019336	19	2.89E-03	Glycosyltransferase family 92	IPR008166	30	5.39E-05
Mitochondrial 28S ribosomal protein S22	IPR019374	12	<1.00E-100	ATP P2X receptor	IPR001429	19	<1.00E-100
Opioid growth factor receptor OGFr conserved region	IPR006757	10	<1.00E-100	Zinc finger SWIM-type profile	IPR007527	17	3.39E-04
Domain of unknown function DUF303	IPR005181	9	<1.00E-100	Protein of unknown function DUF229	IPR004245	14	2.10E-04
Heat shock hsp20 proteins family profile	IPR002068	9	<1.00E-100	Domain in receptor targeting proteins Lin-2 and Lin-7	IPR004172	13	<1.00E-100

Table 4.3.4 Top 10 domains significantly enriched in *P. lacertae* losses for OrthoMCL and OrthoFinder

OrthoMCL				OrthoFinder			
Domain Names	**InterPro domain**	**Losses**	**P-value**	**Domain Names**	**InterPro domain**	**Losses**	**P-value**
Alpha-crystallin domain	cd06464	6	1.64E-03	Ribosomal protein S24e	IPR001976	11	<1.00E-100
Ribosomal protein L34e	IPR008195	4	9.67E-04	Homeobox KN domain	IPR008422	9	3.37E-16
La-type HTH domain profile	IPR006630	4	4.52E-03	Dynamin central region	IPR000375	8	<1.00E-100
Dynamin-like protein family	cd08771	3	<1.00E-100	Lipase class 3	IPR002921	8	5.74E-11
Phosphatases of the metallophosphatase superfamily	cd00144	3	4.16E-04	Calponin homology CH domain	IPR001715	8	9.48E-09
ADP ribosylation factor 1	cd04151	2	<1.00E-100	short chain dehydrogenase	IPR002347	8	6.09E-07
Cyclophilin-PPIL3-like	cd01928	2	<1.00E-100	Aminopeptidase I zinc metalloprotease M18 signature	IPR001948	7	1.11E-06
Diaphanous GTPase-binding Domain	IPR010473	2	<1.00E-100	DEAD/DEAH box helicase	IPR011545	6	1.41E-04
Aromatic amino acid lyase	IPR001106	2	<1.00E-100	Isoprenylcysteine carboxyl methyltransferase ICMT family	IPR007269	5	<1.00E-100
Dimerisation domain of Zinc Transporter	IPR027470	2	2.91E-03	Protein-S-isoprenylcysteine O-methyltransferase	IPR025770	5	<1.00E-100

These enrichment analyses show that the *P. lacertae* genome has gained and lost sequences from across a large array of functions compared to *Blastocystis* sp. STs. There is no obvious loss of specific function in *Blastocystis* sp. STs, rather the differences are varied so that the significant results belong to the most generic terms within the genome.

4.3.8 Metabolic comparison reveals loss of gene complexity in *Blastocystis* genome

The enrichment analysis of *Blastocystis* losses showed a number of metabolism-associated domains such as 3,4-dihydroxy-2-butanone 4-phosphate synthase (IPR000422, EC:4.1.99.12) ($p < 1.00e^{-100}$ OrthoMCL and OrthoFinder), semialdehyde dehydrogenase (IPR012280) ($p < 1.00e^{-100}$ OrthoMCL, $p = 1.78^{-3}$ OrthoFinder) and UTP-glucose-1-phosphate uridylyltransferase (IPR016267, EC:2.7.7.9) ($p = 7.12e^{-6}$ OrthoMCL, $p = 2.00e^{-5}$ OrthoFinder). In order to investigate the extent of the metabolic differences between *P. lacertae* and the *Blastocystis* sp. STs all four genomes and *C. roenbergensis* were annotated with KEGG terms and mapped to KEGG pathways and compared (Figure 4.3.10, S4.3.11).

The *Blastocystis* sp. ST1 genome mapped 4,143 genes to 335 pathways, *Blastocystis* sp. ST4 mapped 3,741 genes to 309 pathways, *and Blastocystis* sp. ST7 mapped 3,746 genes to 316 pathways. This is compared to *P. lacertae* (Chapter Two) and *C. roenbergensis* (Chapter Three), which mapped 4,490 genes to 328 pathways and 3,545 genes to 354 pathways respectively. The comparison showed that *Blastocystis* sp. STs have retained largely the same metabolic capabilities as *P. lacertae* as 291 pathways are shared between all *Blastocystis* sp. STs, *P. lacertae* and *C. roenbergensis*.

This suggests that *Blastocystis* sp. STs may have lost complexity from conserved metabolic pathways, which might indicate loss of redundancy. There is a difference of 347 genes between *Blastocystis* sp. ST1 and *P. lacertae,* which accounts for some of the metabolism-associated domains appearing in the enrichment analyses. There are seven pathways that are present in both *P. lacertae* and *C. roenbergensis* but absent from all *Blastocystis* sp. STs. 00591 Linoleic acid metabolism and 00550 Peptidoglycan biosynthesis both contain only one sequence from both *P. lacertae* and *C. roenbergensis* and therefore cannot be confidently attributed to *Blastocystis* losses. 00592 alpha-Linolenic acid metabolism and 05030 Cocaine addiction were mapped to by two sequences from both *P. lacertae* and *C.*

roenbergensis. It is worth noting that in the context of the Stramenopiles it is unlikely that the proteins that mapped to the 05030 Cocaine addiction pathway are involved in synaptic signalling, rather they contain ubiquitous effectors, such as protein kinase A (K04345, EC:2.7.11.11), that also comprise part of the signalling pathway involved in addiction in a human setting. 00130 Ubiquinone and other terpenoid-quinone biosynthesis was mapped to by a single sequence from *P. lacertae,* but by four sequences from *C. roenbergensis*. The two pathways that were most highly represented were 00261 Monobactam biosynthesis and 05032 Morphine addiction, which mapped three and four sequences and eight and three sequences from *P. lacertae* and *C. roenbergensis* respectively.

The sequences that mapped to the 00261 Monobactam biosynthesis pathway were: aspartate kinase (K00928, EC:2.7.2.4), aspartate-semialdehyde dehydrogenase (K00133, EC:1.2.1.11) and 4-hydroxy-tetrahydrodipicolinate synthase (K01714, EC:4.3.3.7) from *P. lacertae* and those plus 4-hydroxy-tetrahydrodipicolinate reductase (K00215, EC:1,17.1.8) from *C. roenbergensis*. The sequences that mapped to the 05032 Morphine addiction pathway were: guanine nucleotide-binding protein G(i) subunit alpha (K04630), beta-adrenergic-receptor-kinase (K00910, EC:2.7.11.15) and calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase (K13755, EC:3.1.4.17) from *C. roenbergensis* and those plus gamma-aminobutyric acid type B receptor (K04615) and protein kinase A (K04345, EC:2.7.11.11) from *P. lacertae*.

There are seven pathways found in one or more *Blastocystis* genome and absent in both *P. lacertae* and *C. roenbergensis.* Of these there is only one pathway that was mapped to by all three *Blastocystis* genomes, 04614 Renin-angiotensin system, which mapped only a single gene from each *Blastocystis* genome thimet oligopeptidase (K01392, EC:3.4.24.15). Only *Blastocystis* sp. ST1 and ST4 mapped a single sequence to 04950 Maturity onset diabetes of the young. *Blastocystis* sp. ST1 contains regulatory factor X 6 (K19521) and *Blastocystis* sp. ST4 contains MFS transporter, SP family, solute carrier family 2 (facilitated glucose transporter), member 2 (K07593).

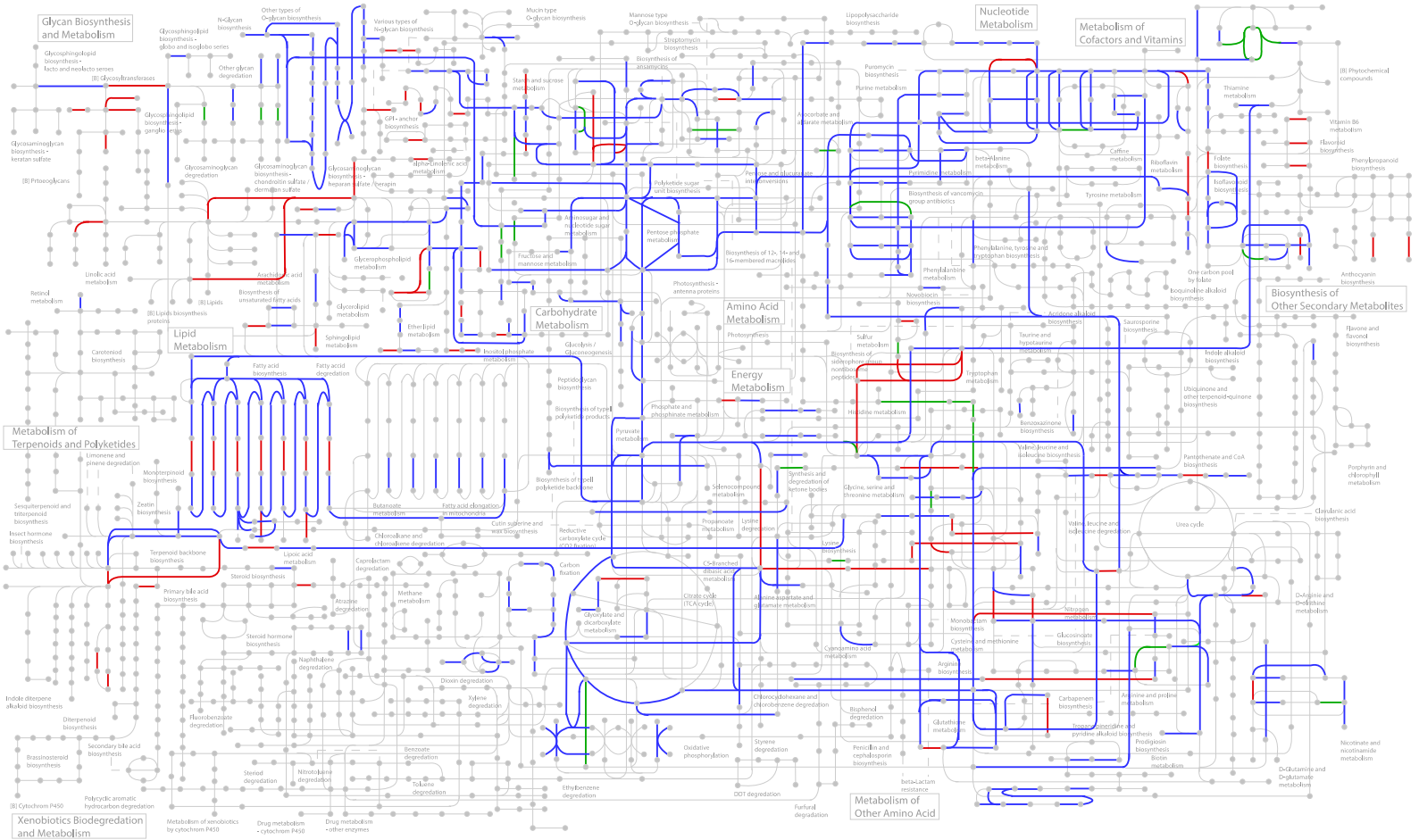


Figure 4.3.10 KEGG pathways mapped with <u>P. lacertae</u> and <u>Blastocystis</u> sp. ST7. *KEGG annotations mapped in blue are conserved between both genomes, in red are pathways mapped only in <u>P. lacertae</u> and green represents pathways mapped only in <u>Blastocystis</u> sp. ST7.*

The remaining *Blastocystis*-specific pathways were only mapped with one or two sequences but included, for *Blastocystis* sp. ST1; 00540 Lipopolysaccharide biosynthesis, 00524 Neomycin, kanamycin and gentamycin biosynthesis and 02026 Biofilm formation – *Escherichia coli* and for *Blastocystis* sp. ST7; 00253 tetracycline biosynthesis and 05150 *Staphylococcus aureus* infection.

While the majority of the pathways are blank (Figure 4.3.10 - grey) these pathways represent a fairly comprehensive account of all the metabolic pathways found in both eukaryotic and prokaryotic cells. The mapping of a single genome to a minority of these pathways is not unexpected though the similarity in the number of pathways mapped from each genome is. Despite the difference in genome sizes and the numbers of sequences mapped to KEGG between *Blastocystis* sp. STs and *P. lacertae,* they contain remarkably similar repertoires of pathways. If there is a fundamental difference between them, it is with respect to the number of genes involved in each pathway, the 'gene richness' of metabolism. *Blastocystis* apparently achieves a very comparable metabolism to *P. lacertae,* but with substantially fewer genes.

4.3.9 *Blastocystis* has lost the ability to form a flagellum

Though many different life cycles and morphological forms of *Blastocystis* have been described with varying degrees of confidence, there has been no report of any stage developing a flagellum. The flagellum is conserved across the Stramenopiles even in multicellular kelp where it appears only during zoosporosis [401]. The flagellum is also a useful tool during infection with some organisms requiring their flagellum to cause disease [402]. In order to confirm this observation about *Blastocystis* and to comment of the life stages, morphological forms and possible mode of infection, the genomes of *Blastocystis* sp. STs were screened for flagellar apparatus.

Screening of the *Blastocystis* genomes was done using reciprocal BLAST and HMMER to search for orthologs to 1,063 flagellar-associated proteins from Judelson *et al.* (2012) [403]. This list represents a range of functions and includes many sequences which have functions other than those associated with the flagellum in each organism. The purpose of this list was to ensure capture of all flagellar associated proteins. Reciprocal BLAST found 67 sequences in the ST4 genome and 113 in the ST7 genome, however, the identities of the hits revealed

that this list includes proteins with primary functions other than those associated with the flagellum and that all three *Blastocystis* genomes are lacking most kinesins and dyneins. The HMMER search revealed 1,732 sequences in the ST1 genome, 1,401 in the ST4 from and 1,495 in the ST7 genome with the same pattern of loss of components required to produce a functional flagellum.

To filter some of the sequences that may not be associated with a functional flagellum in all organisms, a subset of 17 sequences from Judelson *et al.* (2012) [403] were chosen based on presence in organisms with a flagellum and absence from organisms without a flagellum (Figure 4.3.11). Orthologs to these 17 'core' flagellar proteins were sought in the *Blastocystis* sp. STs genomes using BLAST and HMM search methods. None of these proteins had a reciprocal BLAST hit to any sequence in the *Blastocystis* sp. ST4 or ST7 genomes. However, a HMMER search found several proteins, some of which were annotated as dyneins, kinesins or actin binding in all three genomes. However, manual inspection of alignments for these genes revealed they were not true orthologs. In all examples, the reference sequence formed a clade including representatives from *P. lacertae, C. roenbergensis* or both with all *Blastocystis* sequences falling outside this group, showing that they were distant and partial homologues, rather than orthologous sequences.

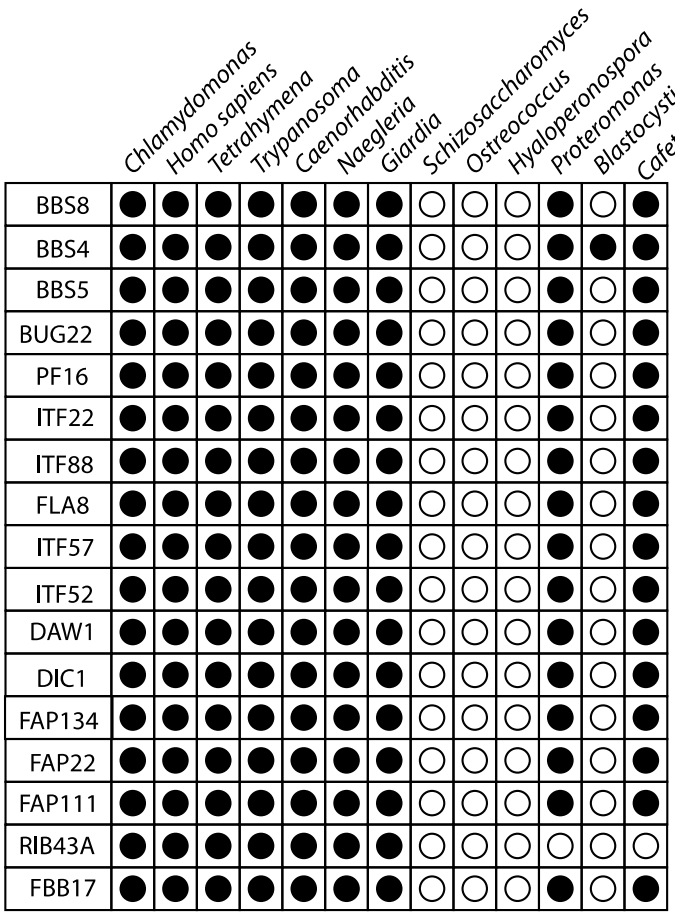


Figure 4.3.11 Table of 17 flagellar-associated proteins highly conserved in flagellates and absent from non-flagellates across the Eukaryota. Chlamydomonas, Homo, Tetrahymena, Trypanosoma, Caenorhabditis, Naegleria *and* Giardia *are genera that contain flagellates*. Schizosaccharomyces, Ostreococcus *and* Hyaloperonospora *are genera that do not contain flagellates. Full circles indicate presence of an ortholog; empty circles indicate absence of an ortholog. For full list of gene names see S4.3.12.*

From these results, it appears as though *Blastocystis* has completely lost the capacity to form a functioning flagellum at the genome level but that it may retain orthologues with other primary functions that no longer perform the flagellar-associated secondary functions.

4.3.10 Clustering of the putative secretome

To investigate changes at the cell surface, an *in silico* cell surface proteome was predicted using SignalP v4.1 to predict signal peptides, ModPred v1.0 to predict GPI anchor amidation sites and TMHMM v2.0c to predict transmembrane regions in the proteomes of the selected organisms. SignalP predicted 1,090 secreted proteins

for *P. lacertae,* 253 for *Blastocystis* sp. ST7, 257 for *Blastocystis* sp. ST4 and 1,386 for *C. roenbergensis*. ModPred predicted GPI anchor amidation sites with 'high' probability on 968 proteins from *P. lacertae,* 236 from *Blastocystis* sp. ST7, 277 from *Blastocystis* sp. ST4 and 993 from *C. roenbergensis.* Overall the two programmes predicted largely independent groups of proteins with only 36, 1, 8 and 82 proteins from *P. lacertae, Blastocystis* sp. ST7 and ST4 and *C. roenbergensis* appearing in both sets respectively. Overall, this approach defined a predicted cell surface proteome of 1,972 for *P. lacertae,* 480 and 515 for *Blastocystis* sp. ST7 and ST4 respectively and 2,188 for *C. roenbergensis.* Putative cell surface proteomes were estimated in the same way for *P. sojae, P. ultimum, S. diclina, T. pseudonana* and *E. siliculosus* and together these sequences were used in a separate clustering analysis with OrthoMCL.

The clustering analysis organised the 19,206 sequences into 3,142 orthologous groups while excluding 7,383 sequences in total. The results from the analysis are summarised in Figure 4.3.12 (S4.3.13) and show that the majority of sequences are specific to each species with very few shared between organisms. This may reflect large-scale changes at the cell surface, which is not unexpected, given the phylogenetic distances involved.

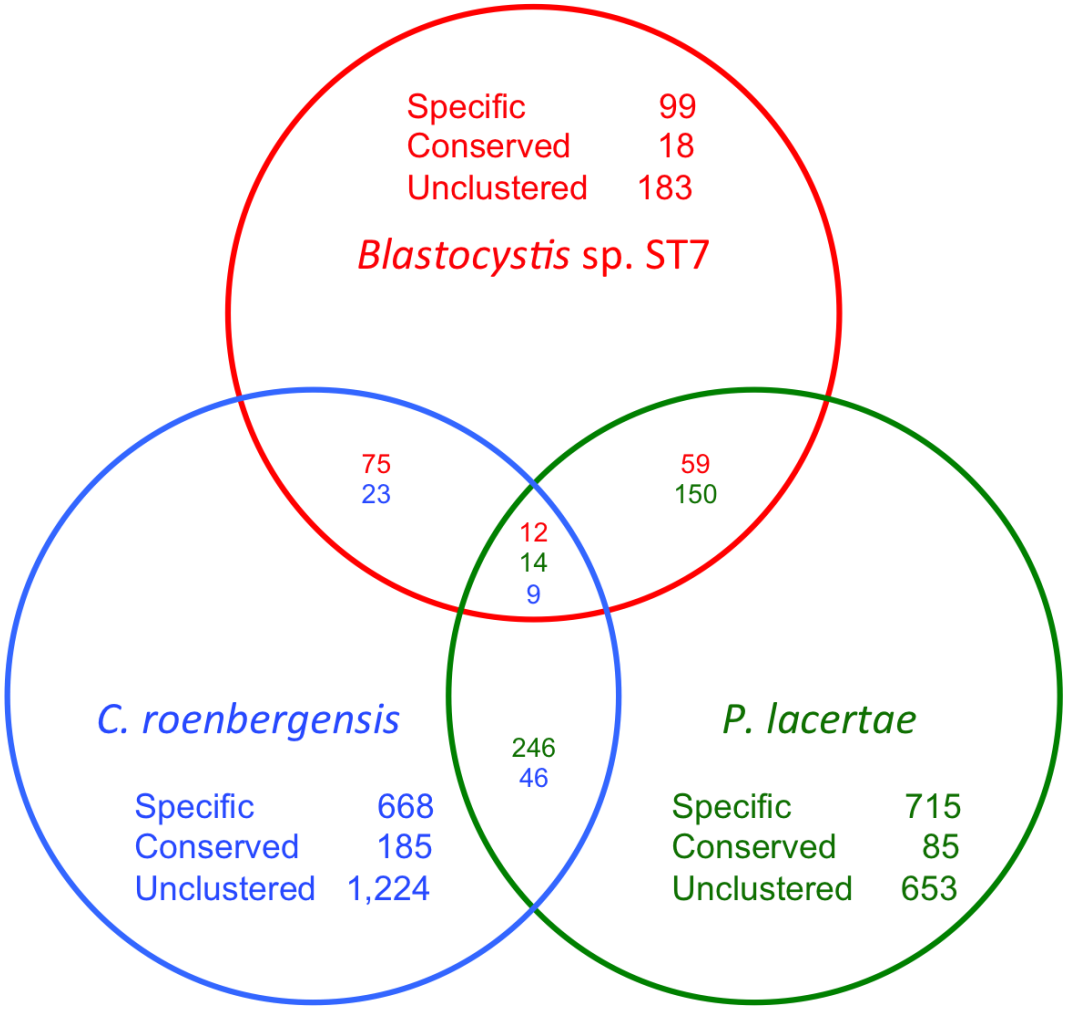


Figure 4.3.12 Venn diagram of clustering dome using a predicted cell surface proteome for each organism*.* *'Specific' refers to sequences from groups containing only one species, 'conserved' sequences are those with no orthologs in the other genomes shown here but which do have orthologs in other genomes included in the analysis and 'unclustered' sequences are assumed to represent single-copy species-specific proteins.*

Domain enrichment of the clustering results revealed that the top gains for *Blastocystis* sp. STs were similar to the results of the whole-genome enrichment analysis of *Blastocystis* gene gains (Table 4.3.5, S4.3.14, S4.3.15) including Ig-like (IPR032812, IPR008009) and glycosyl transferase for dystroglycan domains (PF13896). However, these results also identified other protease (IPR001096, IPR000668) and thioredoxin domains (IPR006338), which may have a number of roles at the cell surface.

Table 4.3.5 Top 10 domains significantly enriched in *Blastocystis* sp. ST4 and ST7 for putative cell surface proteome

Blastocystis sp. ST4				*Blastocystis* sp. ST7			
Domain Names	**InterPro domain**	**Gains**	**Prob.**	**Domain Names**	**InterPro domain**	**Gains**	**Prob.**
Putative Ig domain	IPR008009	16	1.39E-03	Peptidase C13 family	IPR001096	8	1.30E-10
Prokaryotic membrane lipoprotein lipid attachment site profile	PS51257	9	7.07E-07	Putative Ig domain	IPR008009	8	1.15E-03
Major Facilitator Superfamily	IPR011701	6	1.37E-03	Glycosyl-transferase for dystroglycan	PF13896	7	8.43E-06
Peptidase C13 family	IPR001096	6	5.78E-08	Thioredoxin	IPR006338	5	2.65E-03
Thioredoxin	IPR006338	6	2.23E-03	Papain family cysteine protease	IPR000668	5	3.32E-02
Cathepsin propeptide inhibitor domain I29	IPR013201	4	2.30E-04	Bacterial Ig_like domain	IPR032812	3	4.02E-06
Bacterial Ig-like domain	IPR032812	2	1.10E-04	Domain involved in innate immunity and lipid metabolism	IPR003172	3	<1.00E-100
Cyclophilin-type peptidyl-prolyl cis-trans isomerase signature	IPR020892	2	5.12E-03	Lung seven transmembrane receptor	IPR009637	3	1.26E-04
Endomembrane protein 70	IPR004240	2	1.02E-03	MD-2-related lipid-recognition domain	IPR003172	3	<1.00E-100
Homeobox domain profile	IPR001356	2	1.02E-03	Nop domain profile	IPR002687	3	4.02E-06

Clustering of the putative cell surface proteome of *Blastocystis* sp. STs and *P. lacertae* has shown that it is largely species-specific; however, proteases and globular domains are highly enriched suggesting that these families may be changing more rapidly than the background and therefore may be of biological importance

4.3.11 *Blastocystis* sp. STs have reduced phylodiversity across conserved gene groups

Up to this point, I have shown that patterns of gene loss and metabolic pathways offer circumstantial evidence of genome reduction in the ancestor of *Blastocystis* sp. STs. If the *Blastocystis* sp. STs genomes have indeed experienced consistent gene loss, this should be evident when phylogenetic diversity of conserved gene families in *Blastocystis* is directly compared to other eukaryotes, both parasitic and free-living. To this end, a panel of seven genomes were selected from across the eukaryotic spectrum. Proteins with domains conserved in each of these genomes and at least five paralogs in either *P. lacertae* or a *Blastocystis* sp. STs were extracted, aligned in MUSCLE and used to build neighbour joining trees in Ninja. 169 domains were extracted under these criteria and a phylogeny built for each. The relative contributions of *P. lacertae* and *Blastocystis* sequences to the phylodiversity of these conserved families, were calculated using Faith's phylodiversity [399].

Beginning with the phylodiversity value for the background tree composed of sequences from the reference panel of eukaryotes, the increase in phylodiversity when either *P. lacertae* or *Blastocystis* sp. STs were added was measured. The difference between increases due to *Blastocystis* and *P. lacertae* sequences respectively was then plotted against the total number of sequences present in the tree (Figure 4.3.13). In *Blastocystis* sp. ST1 53.8% of the gene families had lower phylodiversity than *P. lacertae*, *Blastocystis* sp. ST4 was less phylodiverse in 62.7% of gene families and *Blastocystis* sp. ST7 was less phylodiverse in 66.3%. While the total number of conserved gene families may not be less phylodiverse in *Blastocystis* sp. STs than in *P. lacertae*, the difference in phylodiversity between them is likely to be greater in cases where *Blastocystis* sp. STs are less phylodiverse. There are very few gene families with a phylodiversity value of less than -2.5, however, in each subtype there are examples of gene families with phylodiversity values greater than 2.5 for all subtypes.

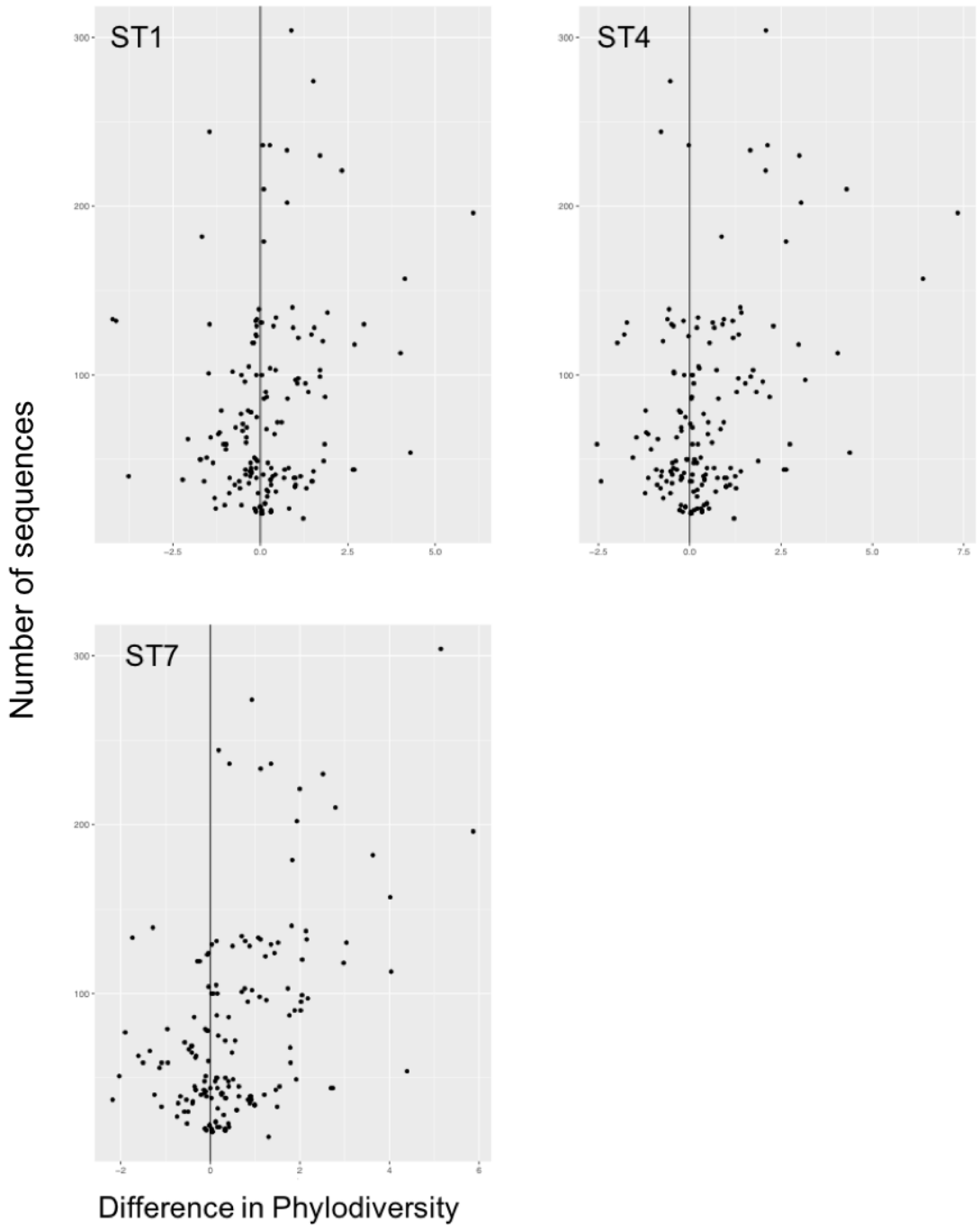


Figure 4.3.13 Phylodiversity of conserved gene families in* P. lacertae *versus* Blastocystis *sp. ST1, ST4 and ST7. *Positive values indicate* P. lacertae *has higher phylodiversity, negative values indicate* Blastocystis *has higher phylodiversity. Each point represents a phylodiversity score for a protein family conserved in multiple eukaryotic lineages. The proportion of gene families with lower phylodiversity in* Blastocystis *sp. STs is between 53 and 66%.*

In a slim majority of highly conserved gene families, *Blastocystis* sp. STs are less phylodiverse, but in cases where *Blastocystis* sp. STs are less phylodiverse, the scale of the reduction is larger in *Blastocystis* sp. STs than in *P. lacertae*.

4.4 Discussion

In this chapter, I have used the newly produced genome for *P. lacertae* (Chapter 2) and the transcriptome for *C. roenbergensis* (Chapter 3) in a comparative genomic analysis with *Blastocystis* sp. STs and other Stramenopiles. I have shown that several features of the *Blastocystis* sp. STs genomes indicate that they are reduced when compared to *P. lacertae* or other Stramenopiles, although various *P. lacertae*-specific expansions may complicate the interpretation of gene loss in *Blastocystis*. I have shown that the *Blastocystis* sp. STs genomes have reduced metabolism related gene richness as well as complete absence of genes required for a flagellum and other motility associated gene families.

4.4.1 Genome selection

Selection of Stramenopile genomes was based on capture of Stramenopile diversity and genome assembly and annotation quality. While the genomes utilised in the analysis are of high quality assembly and annotation, there is no representative genome for the Labyrinthulomycetes, the only Class of Stramenopiles not represented in this analysis (Figure 4.1.1). This clade is located between the Bicosoecida and the Oomycetes; hence a Labyrinthulomycete genome would add more resolution to this analysis, for example, when searching for orthologues to the C13 peptidase family (Figure 4.3.7).

Future work should aim to increase the diversity of Stramenopile genomes available, especially within the Slopalinida. *P. lacertae* is the deepest branching species currently known in this clade [291], but genome sequences from other members of this group would serve to aid identification of *P. lacertae*-specific gene families. These were a source of interference in the enrichment analyses and further knowledge of their orthology within the Slopalinida may serve to clear some of the ‘noise’ and allow more detailed insight into *Blastocystis* orthologues.

The *C. roenbergensis* transcriptome was used to define losses in both *P. lacertae* and *Blastocystis* sp. STs while this may have introduced bias dependent on transcription this was taken into account during the clustering. Only genes that have orthologs in *C. roenbergensis* were used to define losses, rather than absence in *C. roenbergensis* being used to infer gain. The lack of orthologues in *C. roenbergensis* for some of the phylogenetic analyses may have been due to the inherent

incompleteness of the transcriptome or genuine absence. However, without genomic information for *C. roenbergensis* or a related organism this question cannot be answered. A genome sequence could potentially be procured by successfully removing the bacterial component prior to sequencing, for example with cell sorting and single cell sequencing.

4.4.2 OrthoMCL Vs. OrthoFinder

While the OrthoFinder programme was developed as an improvement to the OrthoMCL package [395], the greatest difference between them in terms of the output from these analyses appears to be the number of sequences that are excluded and those assigned to single copy orthogroups. For OrthoMCL there are no single-copy groups (groups containing a single sequence from a single organism) and therefore it is assumed that these make up the cohort of sequences that do not form part of any group. For OrthoFinder, this becomes more of an issue as the programme does assign groups containing only a single sequence and so those that are not assigned to any group are more difficult to define. I have operated under the assumption that these sequences were excluded because of their uniqueness, though interrogation of the algorithm may show this not to be true. Regardless of the reason for the exclusion of a portion of sequences, the output of both programmes showed some overlap between the results produced from the enrichment analyses. This not only lends confidence to the results but also calls into question the advantage of one method over another.

Previous studies have shown a 25% difference in accuracy between these two programmes [395]. OrthoFinder claims to improve on the existing OrthoMCL package by transforming the BLAST bit scores to account for long sequences and long phylogenetic distances. It also uses a different algorithm for defining the similarity limits of an orthogroup. It is this final difference that most likely accounts for the differential exclusion of sequences. OrthoFinder uses the bit score from a reciprocal best hit (after normalisation for length) as the threshold for inclusion into the orthogroup. Presumably, hits that do not overcome this score in any orthogroup are either binned into their own orthogroup or excluded from the analysis. It is not clear what parameters define each of these behaviours. Refinement of the parameters used to run both OrthoFinder and OrthoMCL may reconcile the outputs from these packages and improve the resolution of individual genes, but the consensus between these methods means that there is confidence in the wider

implication of the results. In any case, trees and networks have been built using the results of BLAST and HMMER searches that manually corroborated the orthologous content.

4.4.3 Losses from *Blastocystis* sp. STs

The enrichment analyses were done either in the context of the *P. lacertae* genome, in the case of *Blastocystis* sp. STs or in the context of the *Blastocystis* sp. STs genomes in the case of *P. lacertae* for gains and losses. This skewed the enrichment analysis and produced artefacts in the enrichment analysis such as the EGF domains in the *Blastocystis* losses (Table 4.3.2, Figure 4.3.8). There is a large *P. lacertae*-specific expansion of EGF domain containing proteins, which is potentially responsible for the enrichment of this domain in this group. There are also other examples from across the enrichment analyses that distort the overall trend of loss from the *Blastocystis* genomes. When *Blastocystis* sp. ST gains were aligned to *P. lacertae* sequences with matching domain annotations, many of these sequences show orthology. This may suggest a previously unconsidered limitation of the clustering analysis whereby sequences that are too distantly related have been separated when more detailed analysis can show orthology. It may be possible to rectify this by utilising alternative enrichment tools [404] that are designed to account for genome bias such as GOrilla [405] though from our own data (not shown) GO term enrichment can only be applied to a fraction of each genome.

The losses from the *Blastocystis* sp. STs genomes were represented by a diverse array of gene families and functions from across cellular physiology. The most enriched domains were also small motifs that are often part of a protein with other primary functional domains. These are therefore some of the most abundant domains in eukaryotic cells. The fact that these domains feature heavily in the enrichment analyses may simply be due to their abundance in the genome, but more importantly, these suggest that the gene lost has occurred across the genome. Even in highly conserved gene families, while the number of conserved gene families was similar to *P. lacertae,* the scale of the reduction was much larger in *Blastocystis* sp. STs. This is consistent with the findings that *Blastocystis* sp. STs are reduced in very generic protein domains and in highly conserved gene families that have become redundant. If specific aspects of cell function were lost, these would feature in the analysis, as did motility and the flagellar apparatus (Sections

4.2.9, 4.3.11), however as genes have been lost from across the genome, the most generic domains in the genome come to the fore. This pattern is also consistent with other symbiont genomes and describes a streamlining process signifying adaptation to a restricted niche [206], [211], [218], [406].

4.4.4 *Blastocystis* sp. STs gains

The enrichment of Ig and peptidase domains in *Blastocystis* sp. STs gains in both the whole genome and the putative secretome suggests that these are biologically important to *Blastocystis* sp. STs. Peptidases in particular have a proven association with both the ability to persist in a host environment and with disease [238], [407]–[409]. The expansion of peptidases has occurred in each lineage of *Blastocystis* sp. rather than in the ancestor of all *Blastocystis* sp. STs (Figure 4.3.7) suggesting that these may play roles in determining different environmental tolerances such as host range and virulence. Genotyping of peptidase repertoire could therefore be used diagnostically to aid determination of infection outcome and even identification of zoonotic source of infections as has been done in *Streptococcal* infections [410].

Ig domains were also enriched in the context of both the whole genome and the putative secretome. These may represent a change in life-strategy for *Blastocystis* sp. STs. *P. lacertae* possesses a flagellum and is presumably able to move within the intestinal lumen, however, *Blastocystis* sp. STs have lost the ability to form a flagellum as shown in section 4.3.9. This change, accompanied by expansion of Ig-like domains may be linked to adhesion within the host gut. *Blastocystis* sp. STs have been observed adhering to food particles and to the epithelium itself [70]. This may even represent the cause of the suspected pathogenicity in *Blastocystis* sp. STs, adherence to the epithelium may cause an inflammatory response in a similar way to *Entamoeba* infections [157], [411], [412].

There is much conflict in the literature as to whether *Blastocystis* sp. STs are able to cause disease [93], [94], [96], [97]. While it is true there are patients suffering symptoms associated with the carriage of *Blastocystis* sp. STs, the vast majority of carriers are asymptomatic. There are also studies suggesting that *Blastocystis* plays an important role in maintaining a healthy gut microbiome in humans and that the absence of this organism is associated with reduced intestinal floral diversity [98]. I

have observed expansion of peptidases which may be associated with virulence, but may equally be utilised for immune evasion [381], [382].

Overall, the *Blastocystis* sp. STs genomes show evidence of genome-wide streamlining, consistent with adaptation to a restricted niche. The conclusions are suggestive of adaptation to a closer association with the host mucosa through expansion of cell surface-associated protein families and adhesion domain-containing proteins as well as loss of motility associated genes such as the flagellar apparatus. This proposed association with the host mucosa could potentially serve as the platform for influencing host factors and the association with disease.

4.4.5 The last common ancestor of *Proteromonas* and *Blastocystis*

From the analyses, it is possible to infer the genomic qualities of the last common ancestor of *Proteromonas* and *Blastocystis*. Despite the genomic differences observed, these organisms inhabit similar vertebrate intestinal environments, which may be indicative of an ancestor that inhabited an environment under similar conditions.

It is also possible to speculate that the last common ancestor possessed specific aspects of the cell apparatus such as the flagellum, and that these have been subsequently lost in *Blastocystis*. Motility is an important feature in intestinal protists with examples of a flagellum in *Trichomonas* and *Giardia* as well as other methods of motility such as pseudopodia locomotion in *Entamoeba* [413]. The absence of these in *Blastocystis* is suggestive of a highly derived phenotype. As both *Proteromonas* and *Blastocystis* contain independent expansions of Ig-like domain-containing families, it suggests that the ancestor of both had neither, and may have relied on motility alone or a subsequently lost subset of adhesins adapted for an entirely different environment. Our inferences of the last common ancestor are based on the contemporary genomes and environments. It is likely that the ancestor did not speciate under the conditions in which are observed the current descendants. The ancestor may have been adapted to an anaerobic environment outside of a host intestine all together. It is therefore also unlikely that the last ancestor was pathogenic.

The contemporary genome sequences of *P. lacertae* and *Blastocystis* allow a small window into their last common ancestor. Further sequencing of organisms in

this clade will allow more detailed insight into the evolution of these highly important organisms.

4.4.6 Conclusion

I have shown here that the *Blastocystis* sp. STs genomes are genuinely small in the context of the Stramenopiles and that this is due, in part, to genome wide reduction in the *Blastocystis* genomes. I have seen that reduction in the *Blastocystis* sp. STs genomes has influenced most, if not, all parts of the genome, but which has not resulted in the loss of metabolic function, rather a reduction in gene-richness'. Specific aspects of cellular function have also been lost, particularly with respect to the loss of the flagellum and other motility associated gene families. This may be indicative of a change in ecology whereby, in the absence of motility, *Blastocystis* sp. STs may have adapted to a more restrictive niche adhering to the host mucosa. The change in association with the mucosa may bestow an opportunistic quality, responsible for the symptoms seen in patients with compromised immune systems.

Chapter 5. Comparative genomics of *Entamoeba histolytica* virulence factors using free-living mastigamoebid relatives

In order to investigate the genome evolution of *Entamoeba*, a genomic resource for a free-living relative was required. In an effort to produce this resource, a transcriptome for *Mastigamoeba* sp. was produced. This resulted in two datasets, an inclusive dataset, which likely contained the majority of mastigamoebid sequences but in which there was also likely still bacterial contaminants from the cultures, and the exclusive dataset which was free from bacterial contamination but in which genuine *Mastigamoeba* sequences were absent. The comparison with *Entamoeba* revealed variation between out-groups that influenced the outcomes of analyses and evidence of gene expansion and reduction occurring within the same families. This is suggestive of high selection pressure acting on Cathepsins and Ras family proteins, highlighting their importance for colonisation and persistence within the host environment.

5.1 Introduction

Entamoeba histolytica is an intestinal symbiont responsible for invasive amoebiasis, a cause of global morbidity and mortality worldwide. Most infections are asymptomatic and virulent infection is rare, varying between strains of *E. histolytica*. The genome sequence of a virulent strain, HM-1:IMSS, was published in 2005 by Loftus *et al.* [414] and revised in 2010 by Lorenzi *et al.* [415]. Even after revision, many of the genes annotated in these approaches encode proteins of unknown function. Characterisation of these proteins by experimental procedure has been limited, as a model for knocking-out genes has yet to been achieved. Comparisons between virulent and avirulent species and strains of *Entamoeba* have identified a small number of virulence candidates though many of these are present in the genomes of avirulent strains, only differing in their expression levels in infection models [73].

While a comparative approach has been applied among *Entamoeba* species and strains, no comparisons have been made to characterise features that predate the evolution of the *Entamoeba* genus. In order to determine the derived features of the *E. histolytica* genome that might plausibly contribute to pathogenicity, I will use free-living mastigamoebid relatives to investigate changes that occurred in the ancestral *Entamoeba* genome that may have facilitated its transition to a symbiotic life strategy.

5.1.1 The *E. histolytica* genome

The initial assembly of the *E. histolytica* HM-1:IMSS genome by Loftus *et al*. [414] indicated a genome size of 23.8 Mb containing 9,938 genes. This assembly was made up of 888 scaffolds and almost a third of the genes annotated had no identifiable homologues in sequence databases. One of the more intriguing finds was that the metabolism has been influenced by secondary gene loss and acquisition of HGT genes. *E. histolytica* occupies an anaerobic environment and has lost a large proportion of mitochondrial machinery, including the mitochondrial genome, instead relying on laterally acquired bacterial-like fermentation genes to metabolise glucose from its environment [414]. These adaptations of the *E. histolytica* genome are similar to those found in other anaerobic parasites such as *Giardia lamblia* and *Trichomonas vaginalis* [416], [417]; such convergent evolution is thought to be driven by selection pressure applied within the intestinal environment. The first published assembly and annotation identified 96 HGT genes [414], which were identified as originating from bacterial lineages, most likely from the food sources of *E. histolytica.* The majority of these encode metabolic enzymes, which bestow increased capabilities especially with regards to carbohydrate and amino acid metabolism and may have facilitated the transition to an intestinal environment [414].

Reassembly and re-annotation of the *E. histolytica* genome slightly reduced the size of the assembly from 23 Mb to 20 Mb containing 8,201 genes [415]. This reduction was mostly due to the removal of artificial tandem duplications introduced in the initial assembly and 40% of the genes from the initial annotation required structural changes. This assembly and annotation was published and deposited in the database [418] allowing for outsourcing of work to add annotations and functional descriptors. At the time of writing the *E. histolytica* genome contains 8,306 proteins and has been used as the ‘gold standard’ for investigations into *E. histolytica* and disease.

5.1.2 *E. histolytica* genome and disease

Initial observations about the *E. histolytica* genome identified a number of potential candidates for virulence-associated genes. *E. histolytica* has been called a professional phagocyte [419] and has the ability to phagocytose a range of cell types including epithelial cells, erythrocytes, bacterial and immune cells. However,

in order to reach the intestinal epithelium to phagocytose host cells, it must first penetrate the mucosal layer in the intestinal lumen. It is thought that this is achieved by secretion of cysteine proteases, which break down the extracellular matrix and allow the amoeba to adhere directly to the epithelium [420]. These cysteine proteases may also play a role in immune evasion by 'clipping' bound antibodies attached to the cell surface [73]. Cysteine proteases are widely accepted as being useful tools in the arsenal of many symbionts [407], [408].

Ten new cysteine proteases with predicted cell-surface localisation were identified upon completion of the genome sequence [414]; however other studies have identified up to 20 cysteine proteases with predicted open reading frames [421]. At the time of writing 35 papain-family cysteine proteases, also known as cathepsins, have been identified in the *E. histolytica* genome [422]. These cysteine proteases, of which only a small subset are expressed *in vivo*, can be further separated into two distinct phylogenetic clades; EhCP-A and EhCP-B [421]. Both these clades show higher sequence similarity to cathepsin L-like enzymes than to cathepsin B-like enzymes and this is supported by the conservation of the ERFNIN motif in almost all sequences. Of the 35 cysteine proteases only four are expressed *in vitro* including EhCP-A5 which is strongly associated with virulence and the formation of liver abscesses [422]–[426].

The activity of the proteases may also be dependent on the proximity of the symbiont to the mucosal layer. Trophozoites have been observed binding to both the mucosal layer and to the epithelial cells underneath [151]. The trophozoite form of *E. histolytica* is the stage at which the symbiont is able to adhere to the gut epithelium and thereby cause disease. This attachment is thought to be mediated via an *N*-acetyl-D-galactosamine-binding lectin (Gal/GalNAc-lectin) complex [151]. There is evidence for a close relationship between cysteine proteases and adhesins and there are examples of these proteins forming dimeric complexes at the cell surface [427]. Some studies attribute the outcome of infection to the adhesion of this family of lectins to either the mucosal layer or epithelial cells [151].

The fact that human mucin is encoded by the MUC2 gene and is highly polymorphic has led to investigations into host factors determining the outcome of infection [151], [428]. Mucins are heavily modified in the Golgi by the addition of O-linked GalNAc residues. It is to these residues that the trophozoite adheres, which can prevent contact-dependent cytotoxicity [429]. In infection models *E. histolytica*

Gal/GalNAc lectins, as well as EhCP2 and EhCP5 are highly upregulated when exposed to the intestinal environment reflecting the importance of these factors during infection [140], [411].

Phagocytosis is a complex function which requires several genes with diverse functions. Contact-dependent toxicity, phagocytosis and motility are all functions that *E. histolytica* exhibits during an infection. These activities are mediated through the activity of Rho, Ras and Rab family proteins which are unusually abundant in the *E. histolytica* genome [414]. These families are upstream effectors of pathways associated with motility, effecting cytoskeleton rearrangement, and phagocytosis, mediating vesicular transport and fusion, required for invasion of host tissues [419]. They may also contribute to immune evasion whereby trophozoites are able to move bound antibodies to a pole of the cell, forming a 'uroid' which is subsequently shed into the environment as a vesicle [73].

5.1.3 Comparative genomics within *Entamoeba*

In 1993, *E. histolytica* was redefined as two species: the invasive *E. histolytica* and the avirulent *E. dispar* [129]. It was thought that this would explain the high proportion of asymptomatic carriers. However, even when the difference between these two morphologically identical species was taken into account, the majority of *E. histolytica* infections still did not result in invasive amoebiasis. Several studies have shown that within *E. histolytica,* strain variation can play an important role in determining the outcome of an infection and that distinct genotypes can be isolated from different regions of the same patient and show different rates of liver abscess formation [73], [424], [430]. The *Entamoeba* clade contains several distinct yet morphologically indistinguishable species most of which are capable of causing disease. Comparison with these non-pathogenic species have only identified a small number of factors linked to disease.

Other *Entamoeba* species that have been used in genome comparisons with *E. histolytica* mostly include the closest relative, *E. dispar* (Figure 5.1.1). It was thought to be a completely avirulent species, which nonetheless is still capable of inhabiting a human host, despite containing an almost identical gene repertoire to *E. histolytica* [129], [424], however recent studies have called for the avirulence of this species to be reassessed [431]. *E. moshkovskii* was thought to be a free-living representative of *Entamoeba* until it was isolated from human children [432], [433]

and its pathogenicity remains unresolved [434], [435]. *E. invadens* is more distantly related to *E. histolytica* and *E. dispar* [29] but is still capable of causing disease similar to *E. histolytica* in reptiles, and has been studied for its ability to encyst *in vitro* [436], something that has yet to be verified in human infective species of *Entamoeba*. All of these species of *Entamoeba* have published genomes available and have therefore been utilised in comparative analyses within the *Entamoeba* genus.

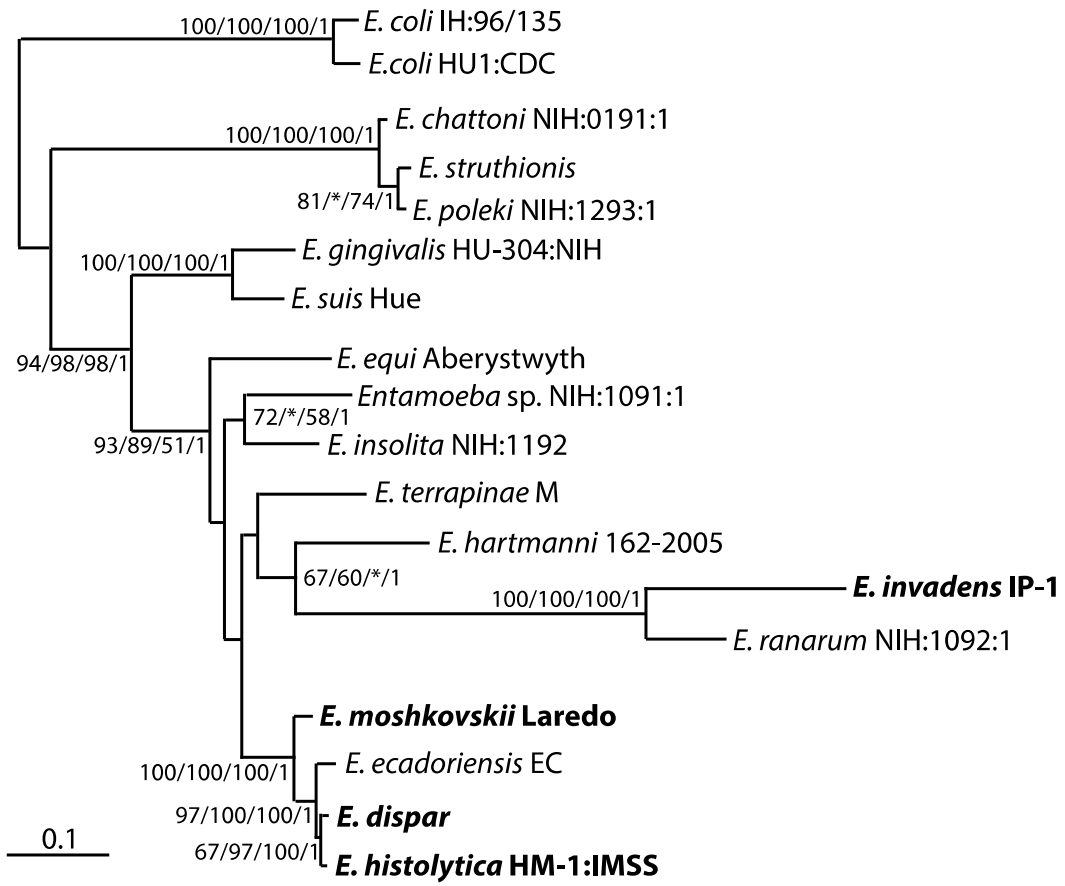


Figure 5.1.1 Bayesian phylogeny of* Entamoeba *showing the relationships between genomes used in this analysis. *Phylogeny redrawn based on SSU rRNA sequence data from Clark* et al. *2006* [132]. *Node support is shown as maximum-likelihood/parsimony/distance/Bayesian analysis. An asterisk (*) denotes where support is < 50%. Tree is rooted with* Entamoeba coli *and* Entamoeba *species used in this chapter are highlighted in bold.*

The identification of strains of *E. histolytica* with varying virulence profiles, such as *E. histolytica* Rahman, and the closely related, avirulent *E. dispar* has led to several comparative studies within the *Entamoeba* genus [29], [73], [420], [422], [430], [433], [437]. Most of these studies have focused on differences in sequence content and transcriptional levels between virulent and avirulent strains and species.

The findings of these studies have revealed that expression of genes associated with nitric oxide and reactive oxygen species (ROS) such as peroxiredoxin, superoxide dismutase and NADPH:flavin oxidoreductase are more highly expressed in virulent strains [73]. It is possible that the ancestor of *Entamoeba* was highly virulent but that this invasiveness has been tempered over time similar to infection caused by *Treponema pallidum* [438].

Comparative studies have also revealed a host of proteins that have weak support for influencing the virulence of a strain. This may suggest that the gene repertoire in the ancestral *Entamoeba* allowed colonisation of the host, but that colonisation in contemporary species is determined by epigenetic factors affecting expression of colonisation genes. Besides this debate however, these comparative studies within *Entamoeba* cannot provide insight into the evolutionary changes that coincide with the origin of the ancestral *Entamoeba*. All *Entamoeba* spp. live in close association with vertebrate guts, something we assume was also true of the common ancestor. To understand this common trait of *Entamoeba* spp. requires comparison with their nearest relatives, most of which are free-living. Here I will investigate this moment in evolutionary history using free-living mastigamoebid resources to illuminate genome evolution during the speciation of *Entamoeba*.

5.1.4 Free-living relatives of *Entamoeba*

In order to examine the ancestral *Entamoeba,* data was required for a free-living organism closely related to, but phylogenetically distinct from the Entamoebidae (Figure 1.3.1). This means that selection of the free-living out-group should be based on the phylogenetic proximity to this clade. Ideal candidates should be within the Pelobiontida clade and available for culture. These criteria allowed identification of a number of potential candidates.

Mastigella radicula AJC/RS/35 (ATCC 50342) was isolated from Stairfoot Quarry, Yorkshire, UK in 1990. To the best of my knowledge it has not appeared in the literature. *Mastigamoeba aflagellifera* AF065-Y (ATCC PRA-395) was isolated from the soil of a rice field in Aomori, Japan in 2006. A partial SSU rRNA sequence is available from NCBI under Archamoebae sp. AF065-Y [439]. This sequence was published only in the database and to the best of my knowledge has not appeared elsewhere in the literature. *Mastigamoeba* sp. WAC-6 (ATCC 50617) was isolated from freshwater Lake Waccamaw, North Carolina, USA in 1995 and was initially

deposited under the name *Conopodamoeba* sp. It also has not appeared in the literature. These organisms were selected because of their phylogenetic position relative to *E. histolytica* and their amenability in culture [133], [137].

As an important part of the comparative analysis an out-group was required from outside the Archamoebae in order to distinguish core amoeba genes that might be lost from *Entamoeba*, from mastigamoebid-specific genes that are uninformative about *Entamoeba*. Here I have used the genome of *Dictyostelium discoideum* AX4*.* This organism is a member of the Mycetozoa (Figure 1.3.1) and falls into a group of organisms commonly referred to as 'dictyostelid cellular slime moulds' [440]. This soil-dwelling amoeba has been the focus of many investigations because of its social nature, which enables it to exist as a free-living or a multicellular complex when starved. This trait has resulted in the study of this group of organisms, particularly *D. discoideum,* as it is a useful way of investigating motility, signalling and cellular interaction [441].

The genome sequence of *D. discoideum* was published by Eichinger *et al.* in 2005 [440] who estimated a genome size of 34 Mb. The assembly was organised into complete chromosomes and represents >95% of the chromosomal sequence and >99% of the genes encoded. The genome is quite gene dense, containing approximately 12,500 genes. The authors suggest that the high number of genes is related to the multicellular life stage exhibited by the organism and includes extensive secondary metabolic components useful in a social environment.

The genomic data available for this organism and its taxonomic position outside of the Archamoebae means this organism provides a good measure with which to identify core Archamoebae genes and gene families [137].

5.1.5 The *Mastigamoeba balamuthi* genome

Mastigamoeba balamuthi is a multi-nucleate, free-living member of the Archamoebae. Originally isolated from a water well in West Africa as *Phreatamoeba balamuthi*, it can be cultured with relative ease xenically or axenically in anaerobic conditions [442]. As *M. balamuthi* is related to *E. histolytica* as demonstrated by SSU rRNA sequence data [133], [137] it has been previously used to compare with the symbiont [443].

Several protist lineages contain modified forms of mitochondria such as mitosomes, mitochondria-like organelles and hydrogenosomes that have been specialised to suite anaerobic or microaerophilic environments. It is thought that the mitochondria in *M. balamuthi* represent a midpoint between mitochondria and the reduced mitosomes possessed by *E. histolytica* [286]. The findings from this study showed that *M. balamuthi* contains two copies of the NIF (nitrogen fixation) machinery, which were acquired by HGT in the ancestor of both *Mastigamoeba* and *Entamoeba,* and that *Entamoeba* has subsequently lost one of these copies [286]. Nyvltova *et al.* partially sequenced and produced a preliminary assembly of the *M. balamuthi* genome in order to confirm the presence of both copies of the NIF machinery components [286].

The genome assembly available for *M. balamuthi* was produced using both 454 and Illumina sequencing. These data were combined and assembled into a 49 Mb genome consisting of 2,462 contigs [286]. Nyvltova *et al.* did not systematically annotate this genome, though they did produce a transcriptome, which was mapped onto the genome assembly. By mapping this to the genome it was possible to validate both resources and obtain a rudimentary annotation of genes. As the genome was only partially sequenced, a robust assembly of the whole genome is not possible and is therefore of limited use. The data for *M. balamuthi* have been included here as a test of the integrity of the resource I have produced and as a second mastigamoebid data point for the investigation into *Entamoeba*.

5.1.6 Aims of the chapter

In this chapter, the aim is to produce transcriptomes for *Mastigella radicula*, *Mastigamoeba aflagellifera* and *Mastigamoeba* sp. in order to include a free-living mastigamoebid out-group to *Entamoeba*. Comparison of these transcriptomes with *E. histolytica*, *D. discoideum* and *M. balamuthi* to determine changes that occurred in the ancestral genome of *Entamoeba* will facilitate its transition from a free-living to a symbiotic life-strategy. There are six specific objectives:

1. Prepare high integrity RNA preps from mastigamoebid cultures
2. Sequence the mRNA with Illumina
3. Assemble the reads into a transcriptome
4. Remove contaminating transcripts from the assembly
5. Annotate the transcriptome and provide a measure of completeness

6. Conduct a comparative genomic analysis to identify virulence factors in *E. histolytica*

5.2 Materials and Methods

5.2.1 Cell culture maintenance

All cultures were maintained as per instructions provided for each organism by ATCC. *Mastigella radicula* (ATCC 50342) and *Mastigamoeba* sp. (ATCC 50617) were maintained in ATCC Medium 802: Sonneborn's Paramecium medium at 25.0 ˚C. The complete medium contained 2.5 g cerophyll in 1 l distilled water provided pre-prepared by Ward's Natural Science Establishment. Upon receipt of the medium it was boiled for 5 minutes after which 100 ml distilled water was added to compensate for evaporation. It was filtered through Whatman #1 paper and added 0.5 g Na_2HPO_4 (BDH).

The cultures contained bacterial food source present at the time of isolation and supplemented with *Klebsiella pneumoniae* subsp. *pneumoniae* (Schroeter) Trevistan (ATCC 13883), the strain designation is NCTC 9633 [NCDC 298-53, NCDC 410-68] hereby referred to as *K. pneumoniae*. Plastic 15 ml screw-capped flat-bottom tubes were inoculated with *K. pneumoniae* 24 hours prior to passage. Cultures containing *M. radicula* or *Mastigamoeba* sp. were rubbed with a sterile loop and shaken vigorously to dislodge cells adhering to the surface of the tube. 250 µl were then transferred to fresh tubes pre-inoculated with *K. pneumoniae* and incubated at 25 ˚C.

Mastigamoeba aflagellifera (ATCC PRA-395) was maintained in ATCC Medium 2832: Reduced YPD Medium at 25.0 ^{0}C. Complete medium contained 1.0 g KH_2PO_4, 1.0 g yeast extract, 10.0 g glucose (BDH), 0.5 g $MgSO_4$ x $7H_20$, 5.0 g peptone, 3.0 ml wheat germ oil and 0.3 g cysteine x HCL (Sigma-Aldrich) dissolved in distilled water and adjusted to pH 6.8-7.2. This medium was autoclaved and vacuum filtered to preserve axenic conditions. Final medium also contained 50 µg/ml gentamicin (Sigma-Aldrich).

Optimisation of culture growth was informed by cell counts for each species. Counts were estimated using an average cell number per field of an Axiovert 25 light microscope (x20) and were not used to calculate cell concentration, rather to monitor culture growth. Cultures were passaged and processed for RNA/DNA extraction at peak density.

5.2.2 DNA and RNA extraction and sequencing

Cultures at peak density were chilled, rubbed with a sterile loop and shaken vigorously to dislodge cells adhering to the tube. Cell were pelleted at 850 x g for 5 minutes then processed using either Qiagen DNeasy blood and tissue kit or Qiagen RNeasy kit according to manufacturer's protocol. RNA samples were treated on the column with RNase-free DNase according to the manufacturer's protocol. Both DNA and RNA were eluted in 30-40 µl of nuclease-free water. DNA and RNA quantification were measured using Tecan spectrophotometer and Magellan software according to the manufacturer's protocol.

Mastigamoeba sp. RNA samples were pooled and processed by the University of Liverpool's Centre for Genomic Research (CGR) using ribosomal RNA depletion (Ribozyme) according to the manufacturer's protocol. These were used to produce one Illumina RNASeq library using the strand-specific ScriptSeq kit. Paired-end sequencing (2x250 bp) was carried out using Illumina MiSeq platform generating in excess of 12M clusters. Post sequencing qc (fastqc) and trimming of adapters from raw reads resulted in reads produced 15,459,938-paired reads for *Mastigamoeba* sp.

5.2.3 Transcriptome assembly and annotation

Bowtie2 was used to map the trimmed reads against the *K. pneumoniae* genome, obtained from NCBI [439]. Reads that did not map (14,561,548-paired reads) were used to carry out a *de novo* assembly using Trinity v2.1.1 [312] with default parameters producing 75,506 unique transcripts.

Resolution and annotation of the transcriptome were attempted as described in Chapter 3. The negative control group was defined as any sequence with > 98% homology to a known bacterial sequence using BLAST. Failure to successfully resolve the bacterial contaminants from the transcriptome meant that a more stringent cut-off was required to filter bacterial contamination from the dataset. This was set at >70% sequence identity to a known bacterial sequence and the transcriptome was redefined as two datasets, the 'exclusive' set which contained only sequences of high confidence, >40% sequence identity to a known eukaryotic sequence, a top hit to an Archamoebae sequence or a reciprocal BLAST to the *M.*

balamuthi contigs. The 'inclusive' dataset included the exclusive dataset as well as any sequence that did not show a robust BLAST hit to a bacterium in the database.

5.2.4 Assessing completeness of transcriptome

The proteomes for *E. histolytica* HM-1:IMSS, *E. dispar* SAW760, *E. invadens* IP1 and *E. moshkovskii* Laredo were downloaded from AmoebaDB (http://amoebadb.org), part of the EuPathDB family of databases [418]. The *Dictyostelium discoideum* AX4 proteome was downloaded from NCBI [439]. A genome-guided assembly of the *M. balamuthi* transcriptome was used downloaded from http://mastigamoeba.img.cas.cz/

The clustering analysis was performed as described in Chapter 4. Other measures of completeness and phylogenomics were conducted as described in Chapter 3.

5.3 Results

5.3.1 Identification of bacterial sequences in *Mastigamoeba* sp. transcriptome

To remove sequences originating from the food source, *K. pneumoniae,* prior to assembly, the raw reads were mapped to the *K. pneumoniae* genome, using Bowtie. This accounted for 5.8% of the total number of RNASeq reads. These reads were removed and the unmapped reads were assembled into 75,506 transcripts using Trinity. This assembly was assessed by BUSCO, which returned a score of 34.73% complete (149 of 429 eukaryotic BUSCOs present). In order to remove the remaining contaminating bacterial sequences from the environmental flora, the transcripts were organised into groups based on their properties. Each transcript was assigned to one of three categories based on their homology to sequences in a non-redundant database and to the *M. balamuthi* contigs.

The positive group, those confidently ascribed as being *Mastigamoeba* sp. sequences, were defined as any transcript with a top BLAST hit to a eukaryotic sequence with >40% sequence identity, a BLAST hit to the *M. balamuthi* contigs or a top blast hit to an *Entamoeba* species. These criteria defined 3,394 transcripts. The negative group, those confidently ascribed to bacteria, were defined as any transcript with a top BLAST hit to any bacterial sequence with >98% sequence identity. This criterion defined 9,709 transcripts. The unassigned group contained any transcript, which did not satisfy the criteria of either the positive or negative groups. Transcripts in this group predominantly share little homology to any sequences in the database and therefore cannot be included or excluded based on homology alone, this group contained 62,403 transcripts (Figure 5.3.1).

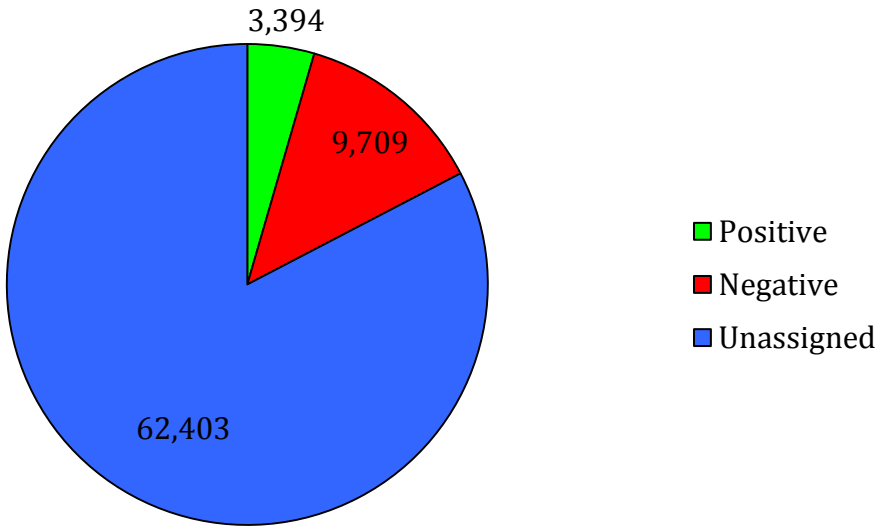
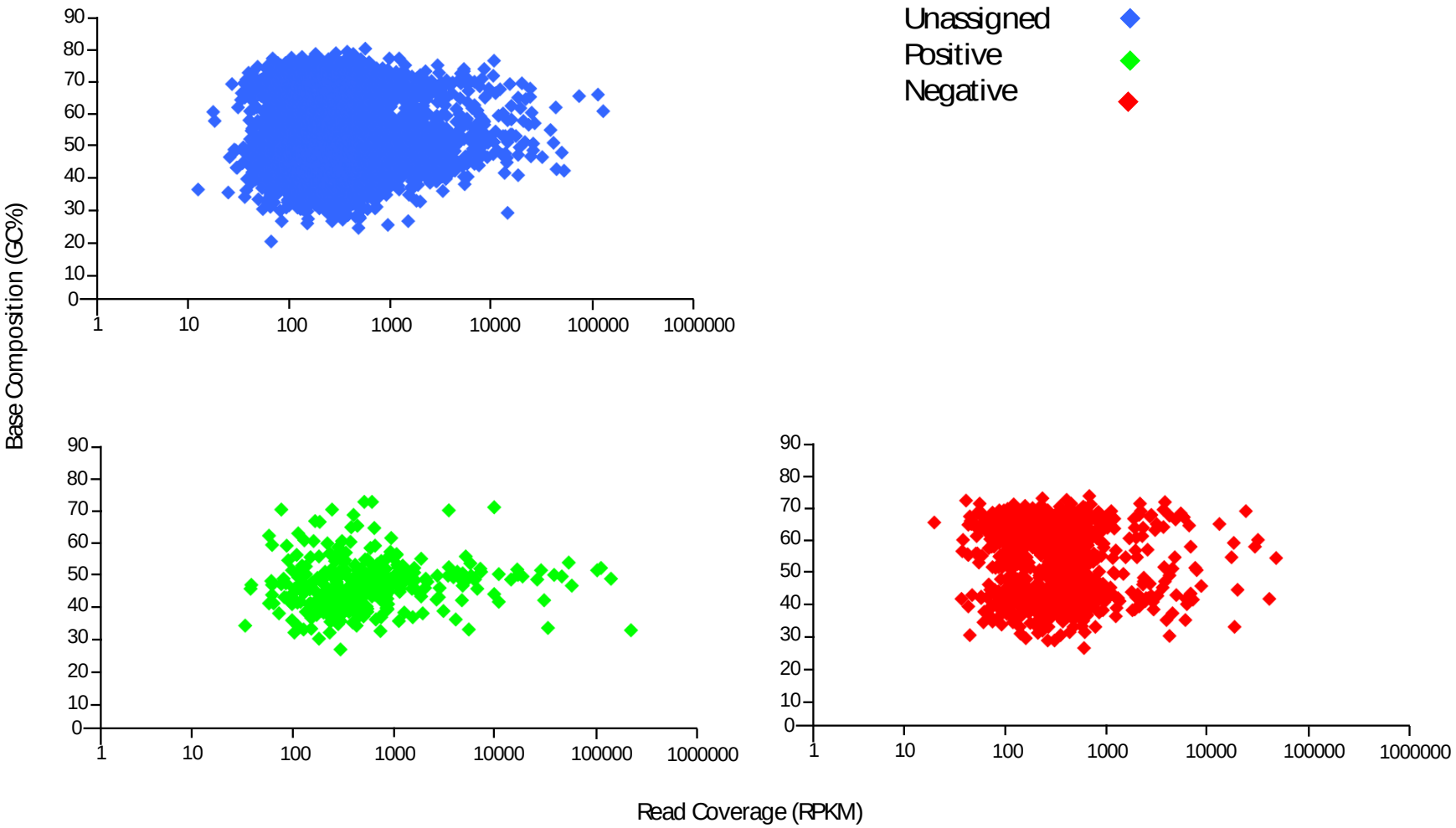


Figure 5.3.1 Pie chart showing proportions of the whole assembled transcriptome, which fall into each category. *Positives are those confidently ascribed to* Mastigamoeba, *negative are those confidently ascribed to bacteria and unassigned are those which cannot be placed in either positive or negative groups based on sequence homology.*

Negative transcripts may be immediately excluded from the analysis, however the unassigned group, which represents >80% of the assembled transcriptome cannot be automatically excluded, but also cannot be assumed to be genuinely mastigamoebid. Similarity of base composition, read depth and codon usage of any unassigned transcript to transcripts in either the positive or negative group may serve to inform on the eukaryotic or bacterial origin of each transcript. These metrics were calculated as an average per transcript and plotted pairwise in an attempt to visualise any distinction between positive and negative groups and similarity to these of transcripts in the unassigned group (Figure 5.3.2). Base composition is expressed as GC percentage per transcript, read coverage is expressed as number of reads mapped per transcript and codon usage is expressed as effective number of codons (Nc) per transcript.

A


Unassigned
Positive
Negative
Base Composition (GC%)
Read Coverage (RPKM)
0
10
20
30
40
50
60
70
80
90
1
10
100
1000
10000
100000
1000000

B

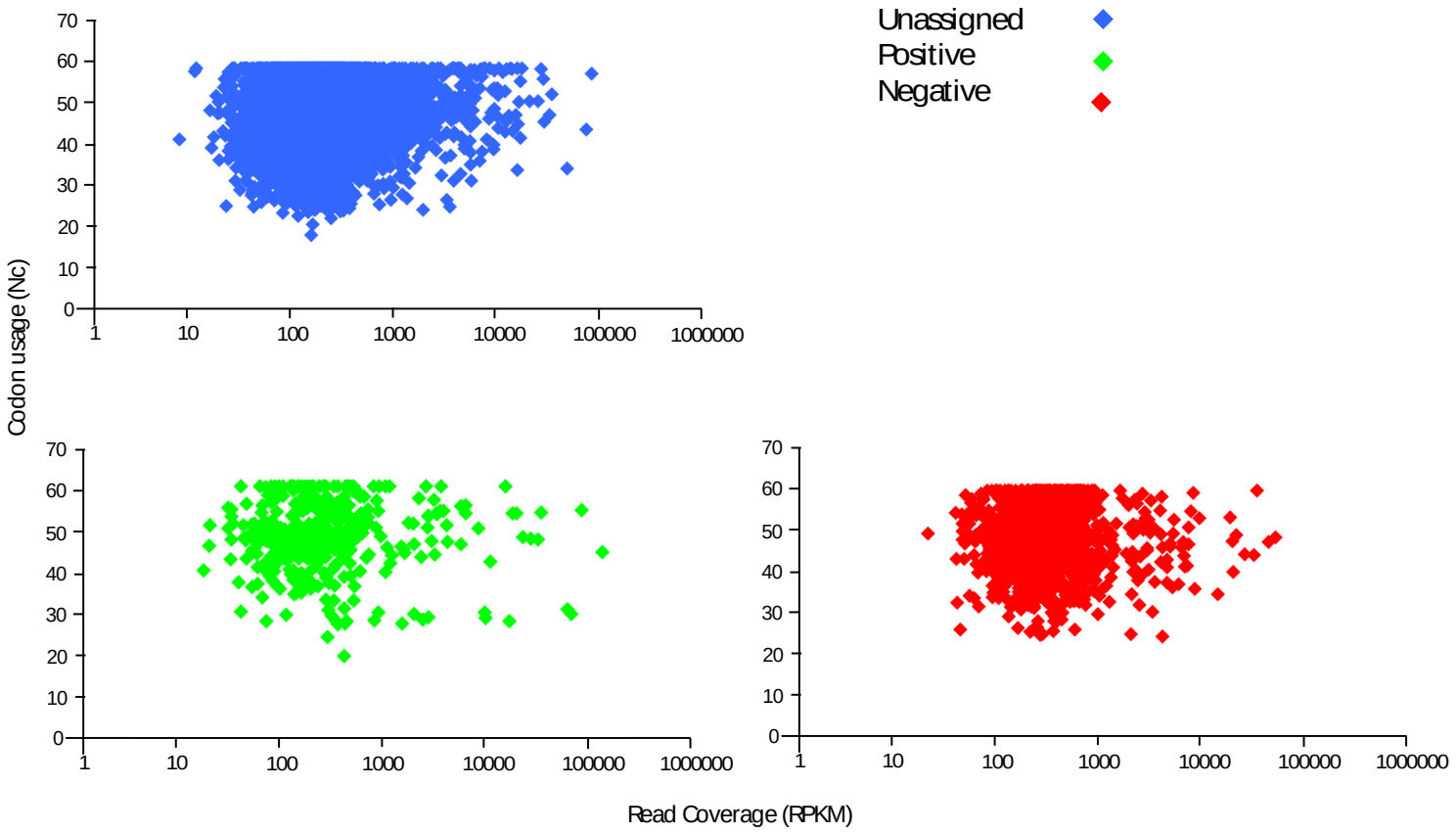


Figure 5.3.2 Pairwise comparison of codon usage, read coverage and base composition for 3 groups defined within the <u>Mastigamoeba</u> sp. transcriptome. *Metrics were plotted pairwise against each other:* ***A.*** *read coverage vs. base composition* ***B.*** *read coverage vs. codon usage.*

These results show that the codon usage, read coverage and base composition do not provide adequate separation of the transcripts. There is no distinction between the positive and negative groups and therefore I cannot infer similarity to either of these groups of transcripts in the unassigned group. Instead, I attempted to use kmer frequency to separate transcripts as in Chapter 3. The results of the kmer analysis are shown in Figure 5.3.3.

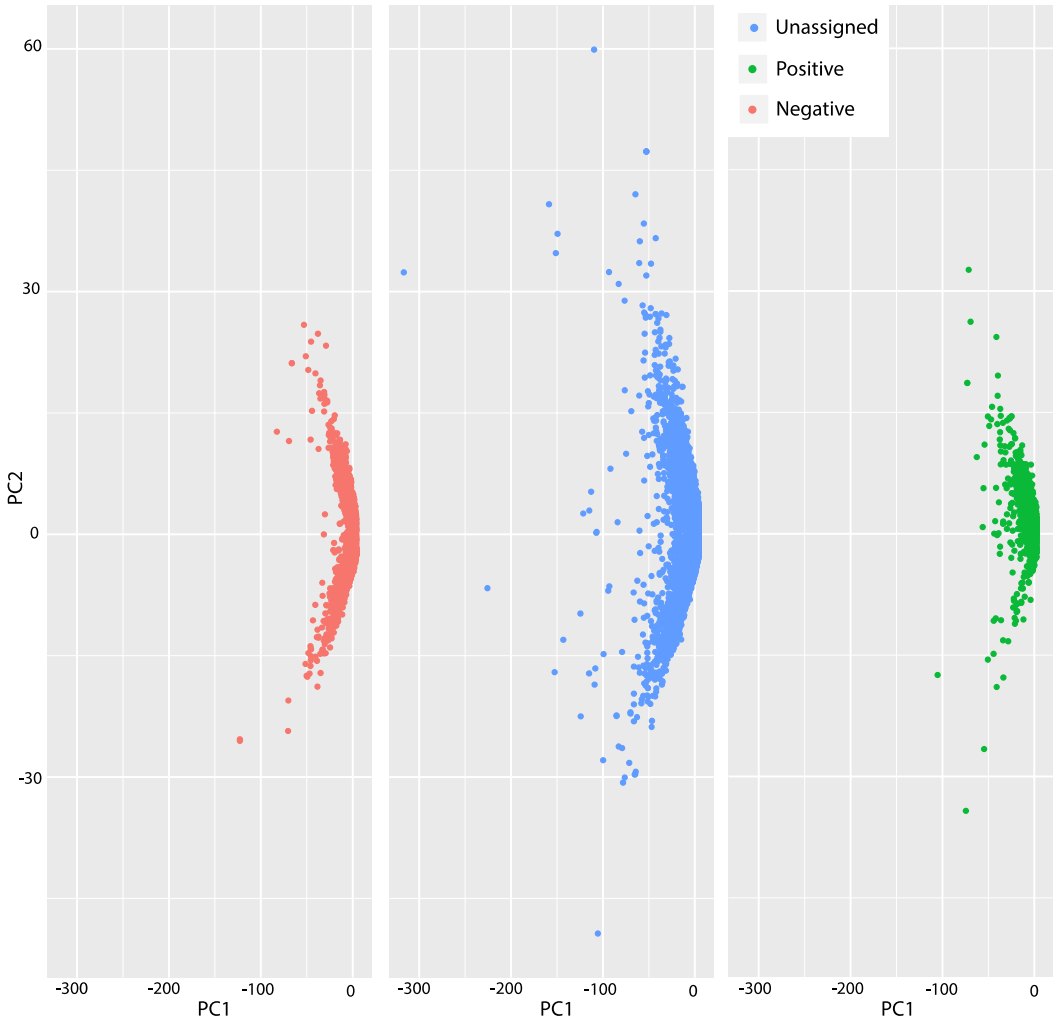


Figure 5.3.3 A principle components analysis of 3mer values. *PC1 and PC2 account for >95% of the variation. Negative transcripts are those with >98% sequence homology to a bacterial sequence. Positive sequences are those with >40% sequence homology to a eukaryote or a BLAST hit to the* M. balamuthi *contigs. Unassigned transcripts are those, which do not fall into either category.*

This approach also failed to produce adequate separation of the positive and negative groups. The unassigned group could therefore not be included in either the positive group as it likely contains contaminating bacterial sequences or the negative group as it may contain genuine *Mastigamoeba* sp. sequences. To resolve

this issue, the criterion for inclusion in the negative control group was relaxed from >98% sequence identity to a known bacterium to >70% sequence similarity [444]. This had the effect of reassigning 27,752 transcripts from the unassigned group to the negative group. Two datasets were then used in subsequent analyses: an inclusive dataset, which contains transcripts from both, the positive set and the modified unassigned set (37,685 transcripts) which scored 31.7% complete with BUSCO and an exclusive dataset, which contains only the positive set (3,394 transcripts) which scored 18.4% complete with BUSCO. These datasets were also passed through TransDecoder to extract protein sequences, which identified 16,135 protein sequences and a BUSCO score of 39.6% complete for the inclusive dataset and 1,953 protein sequences with a BUSCO score of 24.0% complete for the exclusive dataset.

Both of the transcriptome scores are lower than the protein scores. This may be due to the fact that BUSCO must first extract protein-coding sequences from transcripts in order to test for homology to the BUSCOs. TransDecoder may be more efficient at this process resulting in fewer BUSCOs being recovered from the transcripts.

In comparison, the *E. histolytica* genome scored 70.4%, the *D. discoideum* genome scored 95.3% and the *M. balamuthi* predicted proteome scored 68.9% with BUSCO. The *C. roenbergensis* transcriptome produced in Chapter 3 also scored 70.4%. It may therefore be possible to conclude that only a partial transcriptome for *Mastigamoeba* sp. was recovered in either dataset. In order to validate the transcriptome further, its content was examined in more detail.

5.3.2 *E. histolytica* has similar metabolic capabilities to *Mastigamoeba*

One method of validation for the *Mastigamoeba* sp. datasets was to investigate the metabolic capabilities of the transcriptome and compare them to those of *E. histolytica*. To this end the predicted proteome for *E. histolytica, M. balamuthi, D. discoideum* and the inclusive protein dataset for *Mastigamoeba* sp. were annotated using GhostKoala, which assigns KEGG annotations to protein sequences. These annotations were then mapped back to the KEGG pathways (Figure 5.3.4, S5.3.1).

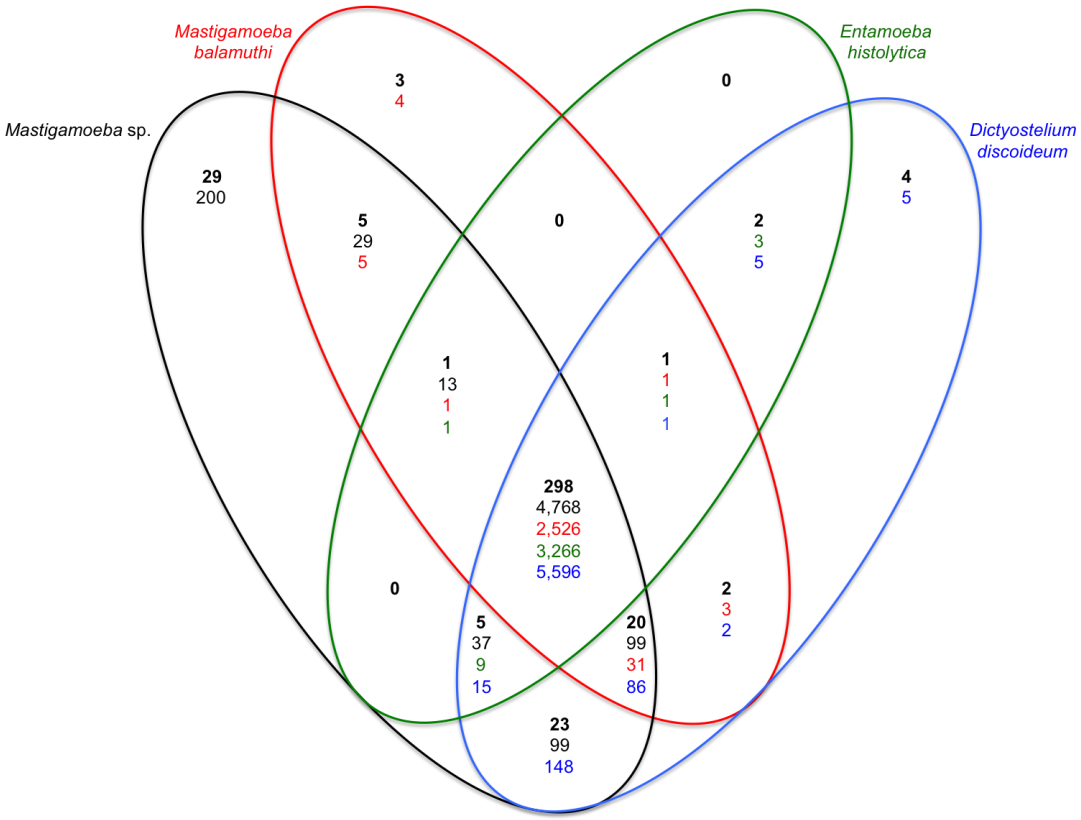


Figure 5.3.4 Venn diagram showing proteins mapped to KEGG pathways using the inclusive dataset for <u>Mastigamoeba</u> sp., <u>M. balamuthi</u>, <u>E. histolytica</u> and <u>D. discoideum</u>. *Values in bold represent metabolic pathways containing at least one protein, other values represent total proteins involved in all pathways.*

The KEGG comparison revealed that the majority of metabolic capabilities are shared between all four organisms. There are 20 pathways absent from *E. histolytica* but present in all other organisms but only 5 pathways absent from *M. balamuthi* and 1 pathway absent from both *Mastigamoeba* sp. and *D. discoideum* which are otherwise conserved. *Mastigamoeba* sp. has mapped to 29 unique pathways. *D. discoideum* has mapped to 4 specific pathways and *M. balamuthi* has mapped to 3 pathways. *E. histolytica* did not map to any species-specific pathways.

It might be expected that, if all the genomes were complete, the organisms that are more closely related will be more similar both in terms of the number of pathways mapped to and the number of individual proteins mapping to those pathways. However, it appears that the number of *Mastigamoeba* sp. proteins is overrepresented in all groups when compared with *M. balamuthi,* which may be due to incompleteness of one of more of the resources, for example *Mastigamoeba* sp. has 4,768 proteins that mapped to pathways conserved in all organisms whereas *M. balamuthi* has 2,526 proteins. These organisms are the most closely related and the

M. balamuthi genome was used to aid the identification of genuine *Mastigamoeba* transcripts and therefore such drastic disparity is potentially unexpected and may suggest that there is a high proportion of bacterial contaminants adding sequences to '*Mastigamoeba'* pathways, or that the *M. balamuthi* transcriptome captured only a portion of the protein-coding sequences.

To test whether the apparent overrepresentation of *Mastigamoeba* sp. was due to the inclusion of bacterial transcripts, the analysis was repeated using the exclusive dataset. This provides an estimation of how much the unassigned group is contributing to the predicted *Mastigamoeba* sp. metabolism (Figure 5.3.5, S5.3.2).

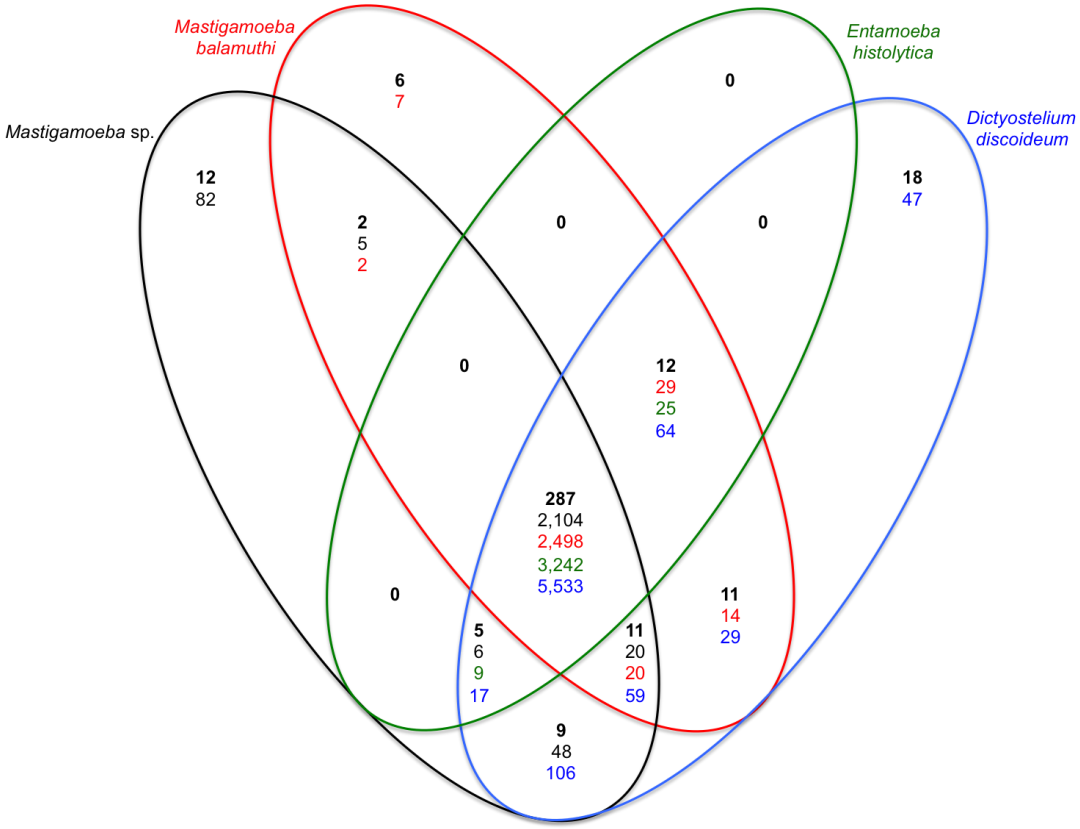


***Figure 5.3.5 Venn diagram showing number of proteins mapped to KEGG pathways using the exclusive dataset for* <u>Mastigamoeba</u> *sp.,* <u>M. balamuthi, E. histolytica</u> *and* <u>D. discoideum</u>.** *Values in bold represent metabolic pathways containing at least one protein, other values represent total proteins involved in all pathways.*

The results from this second analysis seem to support the findings of the previous one with the majority of pathways being shared between all four organisms, though the number of *Mastigamoeba* sp. proteins mapping to these pathways has dropped from 4,768 to 2,104 sequences. There are still 12 pathways

that are specific to *Mastigamoeba* sp. but this is now comparable to *D. discoideum,* which mapped to 18 specific pathways. *M. balamuthi* contains 6 species-specific pathways and *E. histolytica* did not map to a specific pathway as seen in the previous analysis.

Mastigamoeba sp. is no longer overrepresented in conserved pathways when compared with *M. balamuthi*. This supports the idea that the inclusive dataset contains contaminating sequences that reduce the validity of the *Mastigamoeba* transcriptome. The results of these comparisons show that the majority of metabolic capabilities are shared between all four organisms. However, there is disparity in the number of genes participating in shared pathways with *D. discoideum* containing the most and *Mastigamoeba* sp. containing the fewest. It also appears that the *Mastigamoeba* sp. predicted proteome maps to specific pathways not present in the transcriptome of its closest relative *M. balamuthi* suggesting that the *M. balamuthi* genome may be missing components of otherwise conserved metabolism.

5.3.3 Clustering analysis of orthologous proteins

In order to assess how many core amoeba and mastigamoebid genes have been included in the *Mastigamoeba* sp. transcriptome, and to determine to what extent the *Mastigamoeba* sp. and *M. balamuthi* resources were similar, they were used in a clustering analysis provided by OrthoMCL and OrthoFinder along with the proteomes for *E. histolytica*, *M. balamuthi* and *D. discoideum* (Figure 5.3.6, S5.3.3, S5.3.4). For this analysis, the inclusive protein dataset for *Mastigamoeba* sp. was used as contaminating bacterial sequences will most likely cluster together into bacterial specific groups or be excluded from the analysis as single copy, species-specific genes.

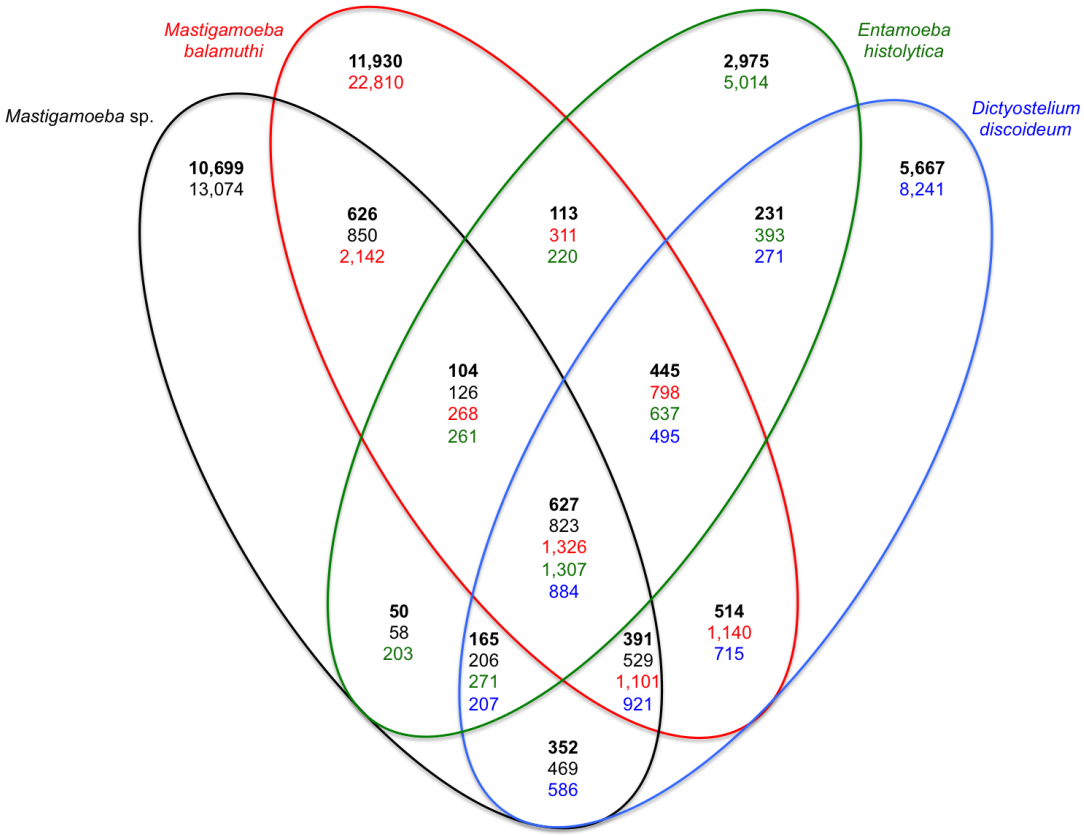


***Figure 5.3.6 Venn diagram showing numbers of species-specific and orthologous proteins for* Mastigamoeba *sp.*, M. balamuthi, E. histolytica *and* D. discoideum**. *Numbers were produced by OrthoMCL and represent the number proteins from each species in orthologous groups. Numbers in bold represent the number of orthologous groups. Species-specific groups include unclustered proteins are assumed to represent single-copy species-specific proteins.*

OrthoMCL and OrthoFinder differed slightly in their estimation of organising clusters of proteins into orthologous groups. Most notably OrthoMCL failed to cluster 10,699 proteins from *Mastigamoeba* sp., 11,930 from *M. balamuthi,* 2,975 from *E. histolytica* and 5,667 from *D. discoideum.* It is assumed that these sequences represent species-specific single-copy genes that were excluded because they cannot be placed into an orthologous or a paralogous cluster. OrthoFinder did not fail to cluster proteins, however, unlike OrthoMCL it does assign single-copy species-specific proteins to individual clusters. OrthoFinder assigned 12,184 *Mastigamoeba* sp. proteins, 21,471 *M. balamuthi* proteins, 4,688 *E. histolytica* proteins and 7,630 *D. discoideum* proteins to single-copy species-specific clusters. The output from OrthoMCL was preferred, as it managed to cluster the greatest proportion of each genome or transcriptome and this allows us to comment on a higher proportion of genes.

Overall OrthoMCL identified 627 clusters, which contains orthologs from all species. These clusters contain a total of 4,340 genes, which represent ancestral conserved protein lineages present in the common ancestor of all four species. The number of proteins in the species-specific clusters for each organism is correlated with the total genome size. *Mastigamoeba* sp. contains 1,239 groups with 3,613 multi-copy species-specific proteins and 9,461 single copy species-specific proteins the highest proportion of single-copy species-specific proteins of any organism included here.

While the number of proteins in species-specific clusters is correlated with overall genome size, this pattern is not observed in conserved groups. *E. histolytica*, when compared to *Mastigamoeba* sp. and *D. discoideum* is consistently overrepresented. This is most obvious when looking at proteins in clusters conserved across all four species where *E. histolytica* is overrepresented by approximately 400 proteins (1,307 compared to 823 and 884) but is true across other conserved clusters as well. Conversely, in conserved clusters *Mastigamoeba* sp. is underrepresented when compared to *D. discoideum.*

The phylogenetic proximity of two organisms is sometimes correlated with the number of homologous clusters, though this is an unreliable measure of phylogenetic relationship. It is interesting to note that *Mastigamoeba* sp. and *M. balamuthi* are the most closely related organisms in this analysis and share 626 clusters only between themselves, which is the highest proportion of pairwise conserved orthogroups. While this result conforms to the inference of phylogeny, other relationships do not. The clusters generated by OrthoMCL appear to suggest that *E. histolytica* is more related to *D. discoideum* than to either of the *Mastigamoeba* species (231 groups compared to 50 and 113). It also suggests that both *Mastigamoeba* species are more related to *D. discoideum* than to *E. histolytica* (352 compared to 50 and 514 compared to 113 for *Mastigamoeba* sp. and *M. balamuthi* respectively). This result contradicts the established phylogeny in the literature and is mirrored in the output produced by OrthoFinder as well. This may reflect the incompleteness of both the mastigamoebid resources.

Overall, the *Mastigamoeba* sp. transcriptome contains a higher than expected proportion of single-copy species-specific proteins, though the total number of species-specific proteins (including those in paralogous clusters) conforms to the correlation observed with genome size. It can also be seen that

Mastigamoeba sp. is underrepresented and *E. histolytica* is overrepresented in conserved orthogroups.

5.3.4 Acquisition of HGTs in *Entamoeba* occurred in the common ancestor of *Mastigamoeba* and *Entamoeba*

As a final measure of how complete the transcriptome was and in order to investigate horizontal gene content in the *Mastigamoeba* genomes, 96 previously identified HGTs in *Entamoeba* [285], [414] were used to search the inclusive *Mastigamoeba* sp. transcriptome, the *M. balamuthi* predicted proteome and the *M. balamuthi* contigs using reciprocal BLAST. This approach identified homologues for 67 of these in *Mastigamoeba* sp. and 50 in *M. balamuthi*. Those missing from the M*. balamuthi* transcriptome were searched for in the *M. balamuthi* contigs. This revealed 90 homologues within the genome.

From this it may be concluded that a majority of the previously identified HGTs were present in the ancestor of both *Entamoeba* and *Mastigamoeba* but that the mastigamoebid transcriptomes contain only a partial record of these. Interestingly, the *Mastigamoeba* sp. transcriptome appears to contain more HGTs than the *M. balamuthi* transcriptome, despite the greater number of *M. balamuthi* transcripts. This is yet another indication that the *Mastigamoeba* sp. transcriptome is a useful resource to identify variation between mastigamoebids.

5.3.5 Species-specific *E. histolytica* proteins are enriched for Ras superfamily domain-containing proteins

The clustering analysis described in the preceding section identified 2,757 genes in 718 paralogous clusters that are specific to *E. histolytica*. In order to determine if these proteins correspond to specific functions, a domain enrichment analysis was performed despite many lacking any annotation or similarity to sequences in the database. InterProScan was used to assign domain annotations the proteins and these were tested for enrichment in the context of the whole *E. histolytica* genome. The enrichment analysis was a hypergeometric distribution analysis performed in R. This analysis showed 183 domains enriched with a probability of < 0.001 (Table 5.3.1, S5.3.5, S5.3.6). These domains represent a diverse range of functions including domains linked to metabolism, signalling, trafficking and transport, structural domains, DNA binding and ribosomal subunits.

However, a particularly prominent result shows that the *E. histolytica* genome contains an expanded repertoire of Ras superfamily domain-containing proteins (PR00449) ($p = 7.65e^{-32}$). A result that has also been noted in the literature [414], [436], [445]. This may reflect an expansion of these effectors, which are associated with a broad range of function including motility, phagocytosis and virulence.

Table 5.3.1 Top 10 domains significantly enriched in *E. histolytics* gains

Domain name	InterPro domain	Gains	P-value
Transforming protein P21 ras signature	PR00449	386	7.65E-32
EF-hand calcium-binding domain profile	IPR002048	63	2.95E-05
TLDc domain	IPR006571	39	2.22E-16
Dbl homology (DH) domain profile	IPR000219	38	2.14E-05
Hsp70 protein	IPR018181	38	8.27E-05
SH3 domain signature	IPR001452	37	6.02E-05
Pumilio RNA_binding repeat profile	IPR001313	37	6.42E-04
Pleckstrin homology domain profile	IPR001849	26	7.60E-03
DENN AEX-3 domain	IPR001194	22	9.06E-12
NHL repeat profile	IPR013017	19	<1.00E-100

Inclusion of the Hsp70 protein domain (IPR018181) ($p = 8.27e^{-5}$) may reflect adaptation to the intestinal environment and the temperature stresses that occur during the transmission cycle of *E. histolytica*. Other domains such as SH3 domain (IPR001452) ($p = 6.02e^{-5}$) are membrane associated and may hint at changes in the cell surface proteome that are associated with the host environment. Overall, the *E. histolytica* genome seems to be enriched for a wide variety of functions, but especially for Ras and Ras-associated gene families.

5.3.6 *Entamoeba* contains genus-specific expansion of Ras-family proteins

In order to investigate the expansion of Ras superfamily domain-containing proteins identified as being significantly enriched in *E. histolytica*-specific groups assigned by OrthoMCL, a single *Entamoeba* Ras domain-containing protein (EHI151610) was compared, using BLAST, to the transcriptome of *Mastigamoeba* sp. and the proteomes of *E. histolytica, M. balamuthi* and *D. discoideum* as well as *E. dispar, E. moshkovskii* and *E. invadens* to extract orthologous sequences (E-value <0.001 and >35% sequence identity). This approach identified 18 orthologs from *Mastigamoeba* sp., 11 from *M. balamuthi,* 30 from *D. discoideum,* 48 from *E. histolytica,* 47 from *E. dispar,* 56 from *E. moshkovskii* and 42 from *E. invadens.*

Simply from the numbers of recovered sequences it is possible to show that *Entamoeba* appears to contain more copies of this gene. These sequences were aligned using ClustalW, manually curated and used to build a maximum likelihood phylogeny (Figure 5.3.7).

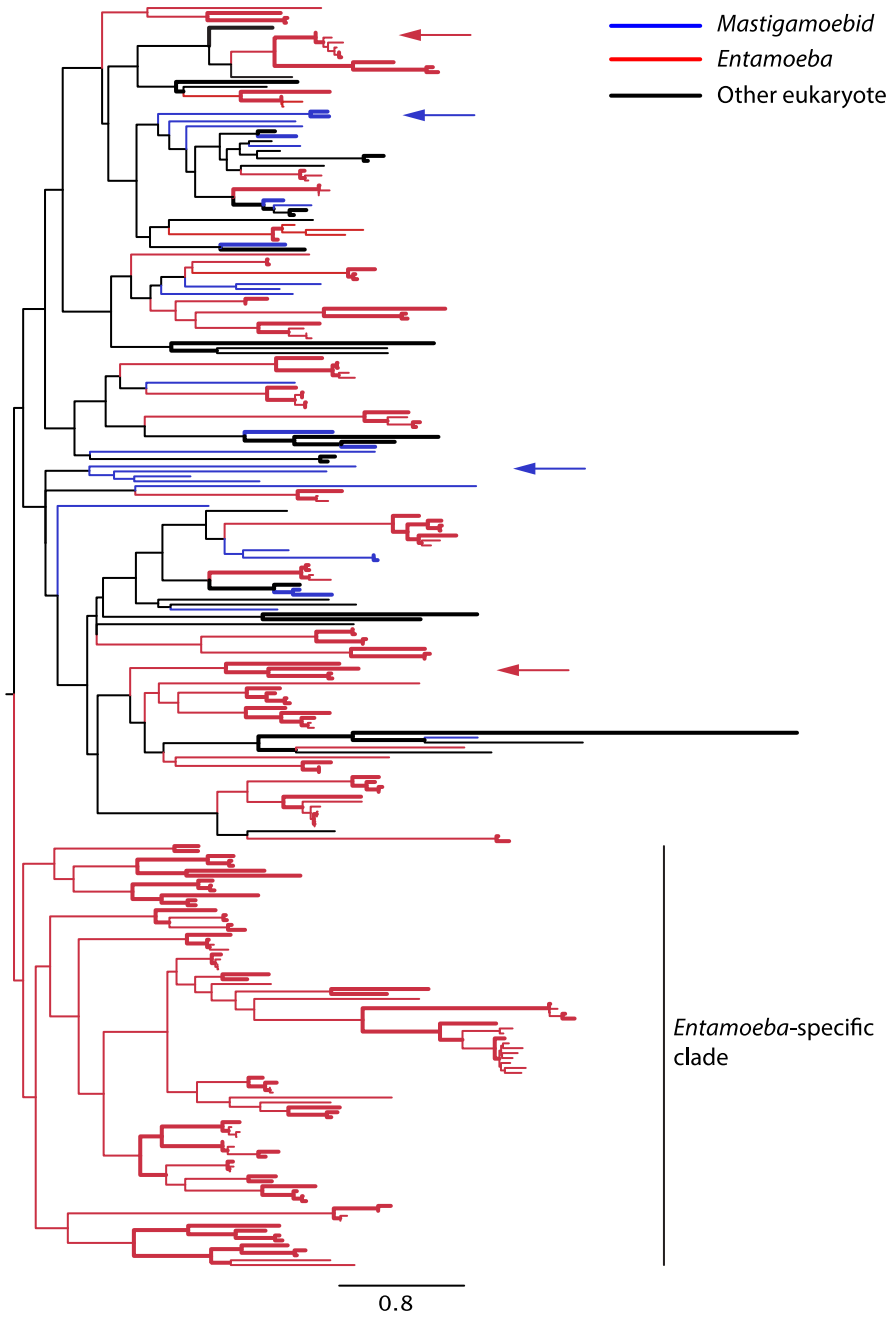


Figure 5.3.7 Maximum likelihood phylogeny of Ras-family amino acid sequences built using LG+G+F model and rooted with the Entamoeba-specific clade. *Node robustness was assessed with non-parametric bootstraps 100 replicates, bold lines indicate bootstrap support >75. Red lines represent Entamoeba sequences and blue lines represent Mastigamoeba sequences and black lines represent D. discoideum sequences. Red arrows denote Entamoeba expansion, blue arrows denote Entamoeba absence.*

The phylogenetic analysis revealed that *Entamoeba* has retained almost all of the ancestral lineages of the Ras superfamily, and in many instances, has expanded these lineages (marked by red arrows Figure 5.3.7). There are also examples where no ortholog was identified in either of the mastigamoebid transcriptomes; however, this may not be due to gene loss, rather due to the limited information available from the transcriptomes. The striking feature of this tree is the *Entamoeba*-specific clade, which is most likely the reason for the identification of this family in the enrichment analysis. The absence of any orthologous sequences from *Mastigamoeba* or *D. discoideum* suggests a large expansion in the ancestor of *Entamoeba*.

5.3.7 *E. histolytica* losses are enriched for diverse domain functions

In addition to identifying genes specific to *E. histolytica,* the clustering analysis also identified 391 clusters that indicate loss from *E. histolytica*. They contained 529 proteins from *Mastigamoeba* sp., 1,101 from *M. balamuthi* and 921 from *D. discoideum* only. InterProScan provided domain annotations for the 583 *Mastigamoeba* sp. proteins, which were used in an enrichment analysis in the context of the *Mastigamoeba* sp. transcriptome. This enrichment analysis showed 144 domains enriched with a probability of < 0.001 (Table 5.3.2, S5.3.7, S5.3.8) that represent a vast array of functions including metabolism, vesicular regulation and DNA binding proteins.

Table 5.3.2 Top 10 domains significantly enriched in *E. histolytica* losses

Domain name	InterPro domain	Losses	P-value
Receptor for egg jelly domain	IPR002859	5	<1.00E-100
Peptidase C1A CathepsinB	cd02620	4	<1.00E-100
2Fe-2S ferredoxin-type iron-sulfur binding region signature	IPR006058	3	<1.00E-100
Biotin carboxylase C-terminal domain	IPR005482	3	<1.00E-100
Vacuolar sorting protein 9 VPS9 domain	IPR003123	3	<1.00E-100
3'5'-cyclic nucleotide phosphodiesterase	IPR002073	2	<1.00E-100
AP endonucleases family 2 profile	IPR001719	2	<1.00E-100
CO dehydrogenase flavoprotein C-terminal domain	IPR005107	2	<1.00E-100
Cobalamin adenosyltransferase	IPR016030	2	<1.00E-100
dDENN domain profile	IPR005112	2	<1.00E-100

This list of enriched domains is based on a small number of proteins but may still provide clues about the losses from the *E. histolytica* genome. Interestingly,

although vesicle associated domains are enriched among gene losses (IPR003123) ($p < 1.00e^{-100}$); they were also enriched in the *E. histolytica* gains (IPR003123) ($p = 7.73e^{-5}$). This may be due to expansion and contraction within the same gene family, or between families with the same domain annotation. The most abundant loss domain (IPR002859) ($p < 1.00e^{-100}$) is for a receptor for egg jelly domain that has no known function, though it does contain polycystic kidney disease domain (PKD) regions, which may be involved in protein-protein or protein-carbohydrate interactions [446].

The loss of the peptidase domain (cd02620) ($p < 1.00e^{-100}$) may suggest a biologically important change, as cathepsins are known for their ability to aid persistence within a host environment and for causing disease. In order to investigate these losses, the sequences with this domain annotation were extracted and used to build a phylogeny (Figure 5.3.8).

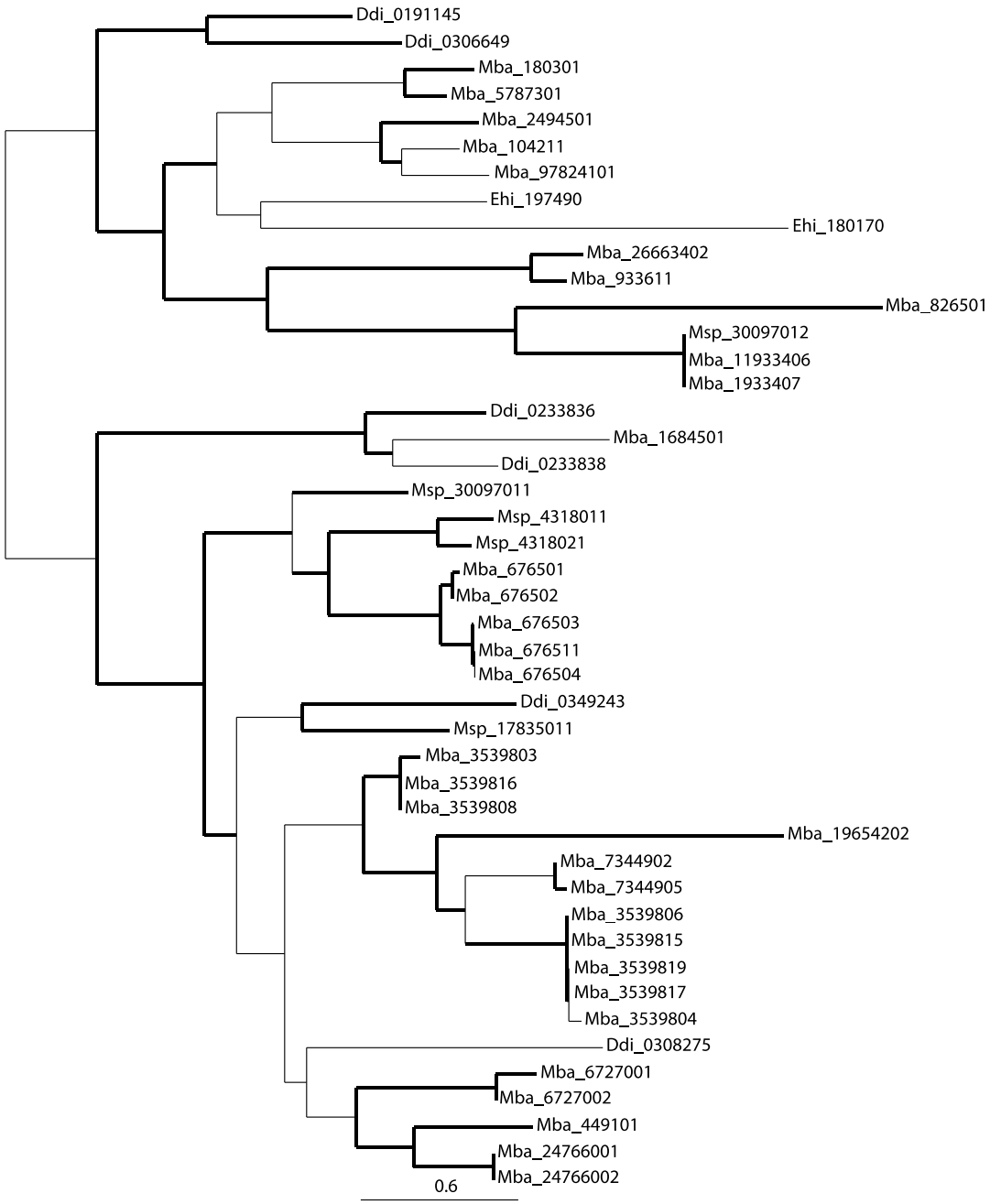


Figure 5.3.8 Maximum likelihood phylogeny of Peptidase C1A Cathepsin B domain-containing proteins built using BLOSUM62+G+F model and rooted at the midpoint*. Node robustness assess by non-parametric bootstraps 100 replicates, nodes in bold denote support >75.*

From the phylogeny, it appears as though the majority of the sequences are *M. balamuthi*-specific acquisitions. However, there appear to be four lineages represented in the tree, of which *E. histolytica* is present in only one in two copies. This confirms the enrichment analysis results in that *E. histolytica* possesses fewer

lineages of C1A peptidases than *Mastigamoeba* sp., *M. balamuthi* and *D. discoideum*.

5.3.8 *Entamoeba* has an expanded arsenal of cathepsins, which show reduced phylodiversity when compared to free-living relatives

In order to identify orthologs of cathepsins in the other species each proteome was searched for its cathepsin repertoire. The cluster produced by OrthoMCL assigned only 15 sequences to the same one as EHI181230 and all were paralogs. Therefore, cathepsin sequences were extracted from all four genomes by comparing a single cathepsin sequence (EHI181230/EhCP112) from the literature [414], [421], [427], [447]–[449] using BLAST. This approach identified 24 cathepsin sequences from the *E. histolytica* genome, 14 from *D. discoideum* and 8 and 7 from *Mastigamoeba* sp. and *M. balamuthi* respectively. This method was also applied to other members of the *Entamoeba* genus: *E. dispar, E. invadens* and *E. moshkovshii* yielding 16, 30 and 22 cathepsin sequences respectively as well as 13 sequences from *Naegleria fowleri* which served as a more distant outgroup. As with the Ras family proteins (above), the number of orthologs recovered from each species suggests that members of the *Entamoeba* genus may contain more copies of these proteins. These sequences were aligned using ClustalW, manually curated and used to build a maximum likelihood phylogeny (Figure 5.3.9).

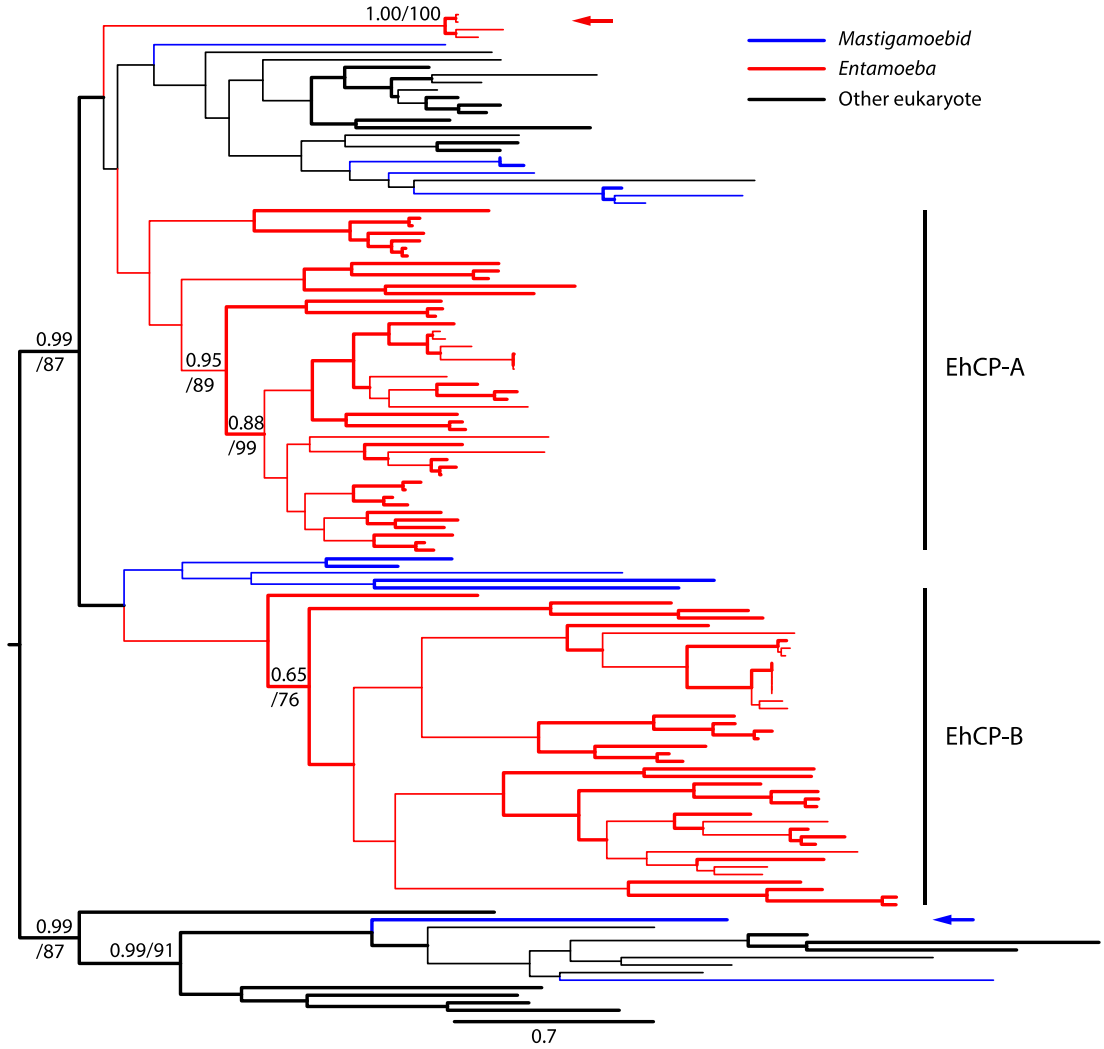


Figure 5.3.9 Maximum likelihood phylogeny of cathepsin amino acid sequences sampled from Archamoebae built using WAG+G model. *Node robustness was assessed with non-parametric bootstraps; branches with bootstrap support >75 are shown with bold lines. Posterior probability values were estimated using Bayesian inference. The two clades of* Entamoeba *sequences are marked except for the single cluster, which does not fall into these clades marked with a red arrow. Red lines indicate* Entamoeba *sequences, blue lines indicate* Mastigamoeba *sequences and black lines are other eukaryotes. The two clades of* Entamoeba *sequences are marked except for the single cluster, which does not fall into these clades marked with a red arrow*, Mastigamoeba *sequences with no homologues in* Entamoeba *marked with a blue arrow.*

This phylogeny shows that all *Entamoeba* cathepsins appear to cluster into two distinct clades, EhCP-A and EhCP-B. There is however, a single *E. histolytica* sequence and its orthologs which appears to branch away from the other clades; EHI_062480 marked by a red arrow (Figure 5.3.9). However, there are no robust nodes to support the observation that these clades are separate and that the outlier clade (containing EHI_062480) does not cluster with either of these clades and

therefore its placement is not certain. It can also be seen that there is a clade in which no *Entamoeba* sequences are found but that contains orthologs from *Mastigamoeba* and *Naegleria*, which may represent ancestral lineages lost from *Entamoeba* (Figure 5.3.9, blue arrow).

To add confidence to our observations, a panel of genomes representing an array of organisms from across the eukaryota were screened for their cathepsin repertoire using BLAST. These sequences were added to the existing alignmnent and the phylogeny was reestimated (Figure 5.3.10).

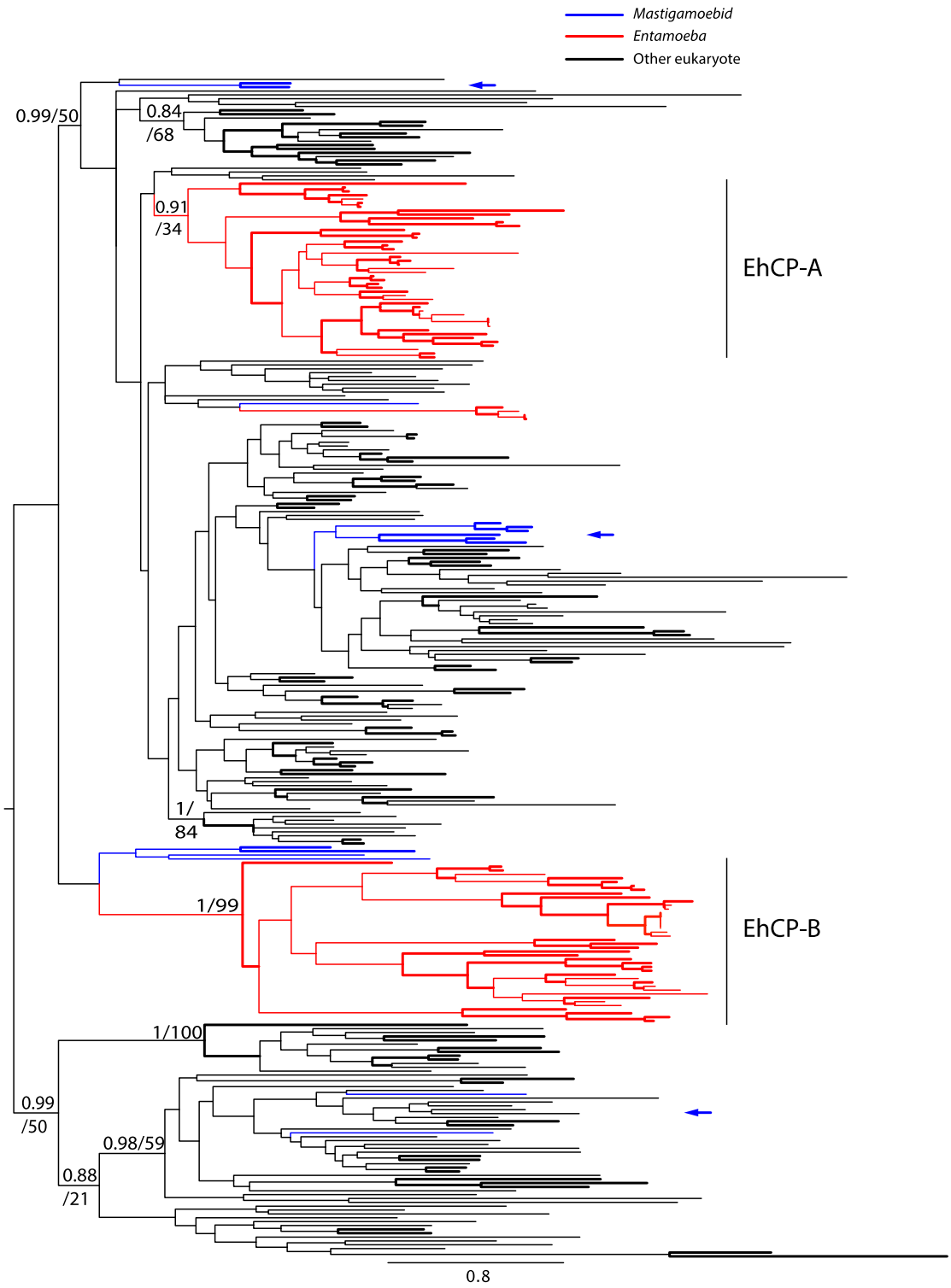


Figure 5.3.10 Maximum likelihood phylogeny of cathepsin amino acid sequences, sampled from across the Eukaryota built using WAG+G model. *Node robustness was assessed with non-parametric bootstraps; branches with bootstrap support >75 are shown with bold lines Posterior probability values were estimated using Bayesian inference. Red lines indicate* Entamoeba *sequences, blue lines indicate* Mastigamoeba *sequences and black lines are other eukaryotes. The two clades of* Entamoeba *sequences are marked; lineages absent from* Entamoeba *are marked with blue arrows.*

This phylogeny further shows that the *Entamoeba*-specific clades cluster more closely to sequences from other eukaryotes than to each other. This confirms that almost all *Entamoeba* cathepsins can be separated into two clades with distinct ancestral lineages. These clades have also been expanded in the ancestor of *Entamoeba* as the majority of the nodes within the clade are organised into orthologous groups of the *Entamoeba* species. Coupled with the lack of *Mastigamoeba* sequences within these clades, this shows that this expansion occurred after the split from *Mastigamoeba*.

EHI062480 and its orthologs still clusters outside of EhCP-A and EhCP-B together with representative sequences from *M. balamuthi* and other eukaryotes (marked by red arrow). To test this, the tree was constrained to include this clade within EhCP-A and tested with log likelihood ratios in RaxML. The difference in log-likelihood between the unconstrained tree (-lnL = 133176.3) and a constrained tree (-lnL = 133194.8) was 18.5 (df = 1; $p < 0.01$) on the X^2 distribution. Therefore constraining EHI062480 to fall within the established cathepsin-L clade produces a significantly reduced likelihood.

Interestingly, *M. balamuthi* cathepsin sequences represent three distinct lineages, whereas *Mastigamoeba* sp. sequences represent six lineages of which two are shared between both. Given the state of completeness of these data, this variation between the mastigamoebid resources is most likely due to missing data but might be due to be species variation. This observation shows that both of these resources may be incomplete, but that they are not incomplete in the same way. Using both transcriptomes in combination provides a better overview of mastigamoebid genomics than either in isolation.

In all other clades *Mastigamoeba* and *D. disciodeum* all have a far more diverse set of sequences which cluster more closely with distant eukaryotes than with sequences from *Entamoeba.* This suggests that the loss of cathepsin repertoire in *Entamoeba* occurred after the split from *Mastigamoeba.* These suggest that changes which happened in the ancestor of *Entamoeba* after its split from *Mastigamoeba* have contributed to both the expansion and lack of diversity observed.

5.4 Discussion

In this chapter I have sequenced and assembled a transcriptome for *Mastigameoba* sp. and used this in a comparative analysis with *E. histolytica*. I have attempted to validate the transcriptome and shown that it can be used to infer character states in the last common ancestor with *Entaoeba.* I have shown that species variation can also impact the outcomes of analyses. I have also reported on the Ras and cathepsin repertoires of *Entamoeba* spp. and infered reduction and expansion in the ancestor of *Entamoeba*.

5.4.1 Validity of the *Mastigamoeba* sp. transcriptome

The greatest issue in drawing any conclusion from the data presented here is that the *Mastigamoeba* sp. transcriptome used for these comparisons either contains sequences which are of bacterial origin, in the inclusive dataset, or which only represents a small protion of the genuine proteome, in the exclusive dataset. I have used multiple methods to validate the content of the transcriptome including an analysis of metabolism and identification of orthologous sequences between *Mastigamoeba* sp., *M. balamuthi, E. histolytica* and *D. discoideum*.

The metabolic validation of the transcriptome showed that there were few differences between the inclusive and exclusive datasets considering that the number of *Mastogamoeba* sp. genes annotated by KEGG was more than double (2,266 up to 5,245) (Figure 5.3.4, Figure 5.3.5). The most striking difference is between the species-specific pathways in *Mastigamoeba* sp. which increased from 82 genes in 12 pathways to 200 genes in 29 pathways. The pathways that were most differentially mapped to by the exclusive dataset were 02040 Flagellar assembly, containing 17 sequences and 01503 Cationic antimicrobial peptide resistance, containing 11 sequences. The remaning pathways mapped three or fewer sequences. These sequences may well represent contaminating bacterial sequences, however, it is also possible that these are genuine *Mastigamoeba* sequences. *Mastigamoeba* sp. and *M. balamuthi* both posess the ability to form flagellar apparatus [134]. It is possible that the *M. balamuthi* resource utilised does not contain these sequences. To test if they were present in the genome assembly, the *Mastigamoeba* sp. sequences were compared to the *M. balamuthi* contigs using BLAST and this showed that the *M. balamuthi* genome does not contain the sequences from the 02040 Flagellar assembly pathway. However, the antimicrobial

genes found homology to two *Mastigamoeba* sp. sequences annotated as UDP-N-acetylglucosamine acyltransferase (K00677, EC:2.3.1.129) and peptidyl-prolyl cis-trans isomerase A (cyclophilin A) (K03767, EC:5.2.1.8). This highlights both a limitation of using a transcriptomic resource and a benefit of using the *Mastigamoeba* sp. transcriptome. While there is considerable overlap between these transcriptomes, there are also aspects that may only be captured by using both. This method shows that the inclusive dataset allows for the identification of conserved pathways without much fear of including a large proportion of bacterial contamination. It has also identified sequences that may otherwise have been missed from the *M. balamuthi* genome.

The second validation method was a clustering analysis which showed that the majority of the *Mastigamoeba* sp. transcriptome is species-specific. This is most likely due to the inclusion of the unassigned group which, by definition does not have orthologues in the other genomes. Also, the majority of the *M. balamuthi*, *E. histolytica* and *D. discoideum* proteins are species-specific too. In this respect, the *Mastigamoeba* sp. transcriptome appears to conform to the patterns of its closest relatives. We must consider the possibility that these sequences represent contaminating bacterial sequences. However, the presence of a larger group of species-specific sequences in the *M. balamuthi* genome suggests that a high proportion of expanded sequences may be a feature of the Mastigamobids. In the conserved clusters *Mastigamoeba* sp. appeared to be underepresented in comparison to all other species (Figure 5.3.6). This is most likely a result of the low mastigamoebid content of the genome rather than a true reflection of *Mastigamoeba* sp. conserved genes families.

The conclusion from both of these validations is that while many genuine *Mastigamoeba* sp. sequences are missing from the inclusive dataset, it is unlikely that a high proportion of bacterail transcripts has distroted the results. The *Mastigamoeba* sp. transcriptome can uncover shortcomings in the existing *M. balamuthi* transcriptome and is therefore a useful resource for investigation into *Entamoeba* evolution.

5.4.2 Ras family proteins are expanded specifically in *Entamoeba*

The enrichment analysis identified numerous *E. histolytica*-specific gene families, the most compelling of which concerned an expansion Ras , Rho and Rab

domain-containing proteins. Further analysis of this particular family (Figure 5.3.7, Table 5.3.1), revealed that *E. histolytica* posesses a large assemblage of species-specific Ras family proteins. while this may plausibly be attributed to incomplete transcriptomes on the aprts of the mastigamoebids, there is also no ortholog from *D. discoideum*.

This expansion of this family of proteins is potentially associated with the phagocytic and trogocytic qualities of *E. histolytica*. Phagocytosis is an important aspect of *E. histolytica* virulence [419] enabling the symbiont to feed on surrounding bacteria [450] as well as a number of host cells including epithelial cells [419]. It has even been shown that phagocytic-deficient variants are less virulent *in vivo* [451]. Amebic trogocytosis is another mechanism by which *E. histolytica* has been shown to disrupt epithelial cells upon contact. It involves taking 'pieces' of the target cell in a similar manner to phagocytosis, except that the whole cell is not ingested [155], [156]. The expansion of these gene families may be what allows *E. histolytica* to engulf such a variety of different cell types and in a number of different ways.

Expansion of this family may also be associated with cytoskeleton rearrangament for the purpose of motility. *E. histolytica* is able to migrate via the blood stream to other organs where it is able to form abcesses [157], [431]. It is also able to move via bleb and pseudopodia formation [413] which may be polarised by members of the Ras superfamily [452]. This is different from the motility exhibited by the mastigamoebids which locomote using flagellar apparatus [134]. If the expansion of Ras family proteins is associated with the acquisition of a type of locomotion that enables *E. histolytica* to disseminate outside of the intestine and cause ALAs then presence or expression profiles of this family may aid diagnostic or treatment targets specific to *Entamoeba*.

Here I have shown that proteins in the Ras superfamily are expanded in the genome of *Entamoeba* spp. Further, I have shown that this expansion occurred in the ancestor of *Enatmoeba* and not in the last common ancestor with *Mastigamoeba* suggesting that this may be an adaptation with biological significance for persistence within a host environment or even with respect to virulence *in vivo* by expanding the range of cell types that *Entamoeba* spp. are able to engulf, or by bestowing migratory capabilities linked to the formation of ALAs.

5.4.3 Cathepsin gain in the ancestor of *Entamoeba*

There are estimated to be approximately 35 cysteine proteases in the *E. histolytica* genome [422]. Our method for extracting orthologous sequences only captured 24 of these (Table 5.4.1). There are several reasons why these may have been excluded from the analysis, the first of which being that the BLAST search may not have been sensitive enough to include all the papain-family proteins. Secondly, some of these sequences may have been removed from the alignment because they did not align. In either case there may be sequences missing from the phylogeny from all species. There may also be untranscribed sequences absent from both mastigamoebid transcriptomes (Figures 5.3.9 and 5.3.10).

Table 5.4.1 Gene names and ID's for *E. histolytica* cathepsins.
Names in **bold** are not in the analysis. CP denotes cysteine protease.

AmoebaDB	Old Name	Current Name	Annotation
EHI074180	EhCP1	EhCP-A1	CP1, putative
EHI062480	EhCP17	EhCP-A10	CP, putative
EHI197490	EhCP19	EhCP-A11	CP, putative
EHI180170	-	EhCP-A12	CP, putative
EHI033710	EhCP2	EhCP-A2	CP2
EHI159610	EhCP3	EhCP-A3	CP, putative
EHI050570	EhCP4	EhCP-A4	CP, putative
EHI168240	EhCP5	EhCP-A5	CP, putative
EHI151440	EhCP6	EhCP-A6	CP, putative
EHI039610	EhCP8	EhCP-A7	CP, putative
EHI151400	EhCP9	EhCP-A8	CP, putative
EHI096740	EhCP10	EhCP-A9	CP, putative
EHI117650	EhCP7	EhCP-B1	CP, putative
EHI180650	-	EhCP-B10	CP, putative
EHI179600	EhCP11	EhCP-B2	CP, putative
EHI140220	EhCP12	EhCP-B3	CP, putative
EHI030720	EhCP13	EhCP-B4	CP, putative
EHI200690	EhCP14	EhCP-B5	CP, putative
EHI126170	EhCP15	EhCP-B6	CP, putative
EHI091450	EhCP16	EhCP-B7	CP, putative
EHI097900	EhCP18	EhCP-B8	CP, putative
EHI181230	EhCP112	EhCP-B9	CP, putative
EHI006920	EhCP-C1	-	Papain family CP domain containing protein
EHI010850	-	-	CP, putative
EHI019390	-	-	CP, pseudogene
EHI084060	-	-	CP, putative
EHI108240	-	-	CP, putative
EHI121160	-	-	CP, putative
EHI123950	-	-	CP, putative
EHI127470	-	-	CP, pseudogene
EHI132640	-	-	CP, pseudogene
EHI138460	EhCP-C3	-	Papain family CP domain containing protein
EHI144040	-	-	CP, pseudogene
EHI160330	-	-	CP, putative
EHI182260	-	-	CP, putative

Our study shows that the two groups of *Entamoeba* cathepsin, EhCP-A and EhCP-B are more closely related to distant eukaryotes than to each other. This has been hinted at in the literature [421] which used neighbour joining phylogenies to show both *Entamoeba* cathepsin groups EhCP-A and EhCP-B have outgroups provided by sequences from across the eukaryota. Here, I have confirmed this with a more robust phylogenetic analysis but also raised questions about the classification of EHI062480 as EhCP-A. I have also shown that this pattern is not conserved in the free-living outgroups, suggesting this is an adaptation which

evolved in the ancestral *Entamoeba*, and therefore may be associated with adaptation to a host environment and virulence.

Cathepsins have previously been associated closely with the severity, outcome and the mode of *E. histolytica* infection [152], [422], [423], [427], [430], [449]. Cathepsins are not only associated with degradation of the mucosa, but also of immune evasion by clipping of bound antibodies. The greatly expanded repertoire of cathepsins in *E. histolytica* form an arsenal which enable the symbiont to invade and persist within the host. The loss of other lineages otherwise present in *Mastigamoeba* suggest that these have been redundant after the shift to a parasitic or commensal lifestyle. Characterisation of the function of the mastigamoebid proteins that cluster with these expanded repertoires may suggest what function these proteases served in the ancestor and subsequently how they have evolved to fulfil the functions they serve in the current situation.

In summary, all *Entamoeba* cathepsins are derived from three ancestral lineages with orthologs in mastigamoebids. However, mastigamoebids possess an additional four lineages that are present in other eukaryotes but absent from all *Entamoeba* spp., and therefore represent losses from the ancestral symbiont. Thus, cathepsin phylodiversity is reduced in *Entamoeba* in comparison with *Mastigamoeba* but two of the lineages that *Entamoeba* spp. have retained are expanded. All cathepsins are derrived from three ancestral lineages, two of which were expanded in the ancestor of *Entamoeba*. Functional categorisation of the 'ancestral', mastigamoebid proteins may elucidate the role these proteins played in the ancestor of *Entamoeba* and may give clues about how the current *Entamoeba* sequences have evolved to alter their function.

5.4.4 HGT events occurred in the ancestor of *Entamoeba*

I have found evidence to suggest that *Entamoeba* and *Mastigamoeba* have inherited some of the same HGTs from a common ancestor. The difference in content may be due to the limitation of using a transcriptome to identify genes with low or transient expression or that may have been removed during the resolution of the bacterial sequences. Despite there being two transcriptomes it is concievable that some orthologous genes are not expressed *in vitro* and therefore are not present in either transcriptome. However, when the genome assembly was investigated, there were still some HGTs missing. While, possible that these are due

to sequencing gaps, it is possible that some of the HGT events ocurred after *Entamoeba* split form *Mastigamoeba*. It may be these HGTs that confer qualities to *Entamoeba* that enable it to lead a different lifestyle from *Mastigamoeba*. This observation supports the findings of previous studies which have found that while some lineages of HGT are indeed not conserved between *Entamoeba* and *Mastigamoeba*, there are HGTs which are present in all Archamoebae and potentially even shared with organisms as distantly realted as *Dictyosteluim* [437], [443].

Identification of HGTs is highly dependent on sampling and the older a gene acquisition by HGT is, the less idenity the sequence shares with contemporary bacterial sequences. Our, cutoff of >70% sequence identity should have preserved genuine HGTs shared with *Entamoeba* or even older [444]. HGT in and between protists requires additional sampling and comprehensive studies in order to extract the identity, timing and donor of genes in this continuing and dynamic process and transcriptomic resources cannot be used to infer loss [453].

5.4.5 Conclusions

In this chapter I have presented an assembled transcriptome for *Mastigamoeba* sp. and attempted to remove contaminating bacterial sequences. I have validated the transcriptome with multiple approaches and shown this resource exposes variation between mastigamoebid transcriptomes that can affect the outcomes of analyses and is therefore a useful complementary resource to the existing *M. balamuthi* genome. These resources have shown the *E. histolytica* genome has an expansion of Ras superfamily domain-containing proteins that may be involved in the pathogenicity of *Entamoeba* spp. Finally I have confirmed the reduction in cathepsin diversity in the ancestral Entamoeba, followed by masisve expansion of the remaining cathepsin-L using appropriate outgroups.

Overall, the *Mastigamoeba* transcriptome has helped to identify changes that occurred in the ancestor of *Entamoeba* that may be associated with the ability of *Entamoeba* spp. to persist in a different environment to that of the last common ancestor with M*astigamoeba* and may be important factors in diagnosis and treatment of *Entamoeba* infections.

Chapter 6. General discussion and concluding remarks

Following recent advances in sequencing technology, comparative genomics now represents an important tool in the study of parasites and other host-associated symbionts [212]. The majority of previous studies have focused on comparisons within parasitic clades involving closely related organisms and are therefore unable to provide information on the emergence of symbiosis [206], [250], and studies that have utilised a free-living out-group have often focused on vector-borne diseases [206], [261], [454]. This thesis has applied the concept of comparative genomics using a free-living comparator to investigate the evolution of symbiosis within two abundant intestinal symbionts of humans.

6.1 Genome production and future work

In Chapter 2, a genome sequence of *P. lacertae*, a free-living relative of *Blastocystis,* was produced. I concluded that the 52.2 Mb draft genome of *P. lacertae,* consisting of 1,449 contiguous contigs with an N50 of 92,586 bp is comparable, in terms of its contiguity and coverage, to other published genomes sequenced using the same technology. I also assessed the completeness of the gene annotation and showed that the 35,706 genes account for >85% of the total BUSCOs. This genome represents a much-needed data point, not only for the sampling of this under-studied clade, but also for protist biology in general.

The *P. lacertae* genome and its annotation may be used as a reference against which future genomes from this clade may be compared. Future work should aim to improve the sampling within this group of organisms and provide genome sequences for species of the *Karotomorpa* and *Protoopalina* genera [291], which may help to explore the repertoire of 13,642 currently labelled '*P. lacertae* specific' genes. Many of these will be uniquely derived in *Proteromonas,* but some will inevitably represent authentic losses from *Blastocystis,* and so will be relevant to understanding the latter's evolution. Better understanding of the ecology of *P. lacertae* would also serve to place features of the genome into biological context, such as why the flagellar apparatus has been retained in *P. lacertae,* but lost in *Blastocystis*, and what the large species-specific gene families are contributing to.

6.2 Transcriptome production and future work

In addition to the *P. lacertae* genome, I have produced transcriptomes for *C. roenbergensis* and *Mastigamoeba* sp. in Chapters 3 and 5 respectively. Both these transcriptomes were generated from xenic cultures, which contained an unknown number of bacterial food sources. As a result, the data collected was akin to metatranscriptomic data, and was treated accordingly. In the case of *C. roenbergensis* the resolution of the bacterial contaminants was achieved using sequence homology and kmer frequencies producing a transcriptome consisting of 28,952 transcripts, and scoring >70% with BUSCO. Conversely, the *Mastigamoeba* sp. transcriptome was resolved using sequence homology alone and contained only >30% of the BUSCO content. This means that there is insufficient confidence in *Mastigamoeba* sp. specific genes in the transcriptome because they may be the result of contamination. However, this resource was still able to provide information about losses in *Entamoeba,* for instance of cathepsin.

A genome sequence for *C. roenbergensis* may help to answer some of the remaining questions about the *Blastocystis* transition from a free-living environment in terms of genome structure and non-coding elements. However, in order to achieve this, the cultures must either be grown on a controlled subset of bacteria as a food source, or successfully separated from the bacterial component prior to sequencing. After sequencing, the transcriptome presented here will be a useful tool in the annotation process that is still the rate-limiting step for genome production.

In addition to the previously published *Mastigamoeba balamuthi* genome [286], here I presented a transcriptome of *Mastigamoeba* sp. for use as a free-living out group to *Entamoeba* and as a second *Mastigamoeba* representative which allowed for identification of high variability within the *Mastigamoeba* genus. I present a transcriptome consisting of 10,507 transcripts, which are confidently ascribed to *Mastigamoeba* sp. and 40,930 transcripts of unconfirmed origin based on sequence homology. While these organisms can be difficult to culture, I have shown that the typical ‘lab strains’ such as *M. balamuthi* may not be entirely representative of the other species in the genus or other genera in the clade. While the sequencing of important pathogens will always take priority, care should be taken to avoid assumptions about representative genomes and reduction of diverse clades to a single or a few token free-living out-groups.

A concerted attempt to annotate the *M. balamuthi* genome would undoubtedly identify aspects of the genome that may have been absent from this analysis but mapping of the newly sequenced transcriptome to the *M. balamuthi* contigs may uncover similarities and differences between the *Mastigamoeba* species and may even reveal previously unidentified *M. balamuthi* genes. There are also other genera that may prove useful in rounding out the sampling bias seen in this clade of organisms and add more data to investigate Pelobiontida specific gene families such as *Pelomyxa* and *Rhizomastix* [137].

6.3 Comparative genomic analyses

The addition of *Mastigamoeba* sp. to the genomic resources available for the study of the Archamoebae facilitated a comparative analysis between *M. balamuthi, Mastigamoeba* sp., *E. histolytica* and *D. discoideum*. I found very little evidence of reduction of metabolic genes in *E. histolytica* when compared to the free-living out-groups, but saw expansion of Ras superfamily, potentially indicative of mechanisms that allow *Entamoeba* to lyse host cells, invade host epithelium and migrate to other host organs.

I confirmed that the cathepsin repertoire in *E. histolytica* is separated into two distinct clades but I showed that this adaptation is specific to the ancestor of *Entamoeba* and also called into question the established phylogeny of *E. histolytica* cathepsins and show that while two cathepsin lineages have been expanded, other lineages have been lost when compared to other eukaryotic cathepsin diversity. The dichotomy of expansion and contraction within the same gene family is indicative of extreme selective pressures acting on these genes.

The production of the *P. lacertae* genome and the *C. roenbergensis* transcriptome allowed a comparative analysis with the previously published genomes of three *Blastocystis* subtypes and five other Stramenopile genomes. Simple size comparison between *Blastocystis* and the other genomes showed that the *Blastocystis* genomes are reduced and using domain annotations it was possible to infer this reduction was genome wide. However, when looking in detail at the metabolic component, there was very little reduction in the capabilities of *Blastocystis* compared with *P. lacertae*, what was observed was a loss of gene-richness in metabolic pathways, perhaps indicative of a loss of functional redundancy. I also observed loss of complete cellular components such as the loss

of the flagellum. Once one of the definitive morphological features of the Stramenopiles, the flagella apparatus and other motility-associated gene families are absent from the *Blastocystis* genomes. I assessed the phylodiversity of gene families conserved across eukaryotes and found that even highly conserved gene families, *Blastocystis* sp. STs were more likely to exhibit more extensive reduction in cases where the phylodiversity was greater in *P. lacertae*. I have shown that the *Blastocystis* genomes contain a record of genomic streamlining that appears to have occurred in the ancestor of *Blastocystis*. I have proposed a possible ecology for *Blastocystis* based on its lack of flagellum and other motility associated gene families and the expansion of adhesin-like molecules at the cell surface.

Future work should include examination of the non-coding region of the genomes. This work has focused on proteome data available from the genomic and transcriptomic resources, non-coding aspects of the genome may improve our understanding of regulation influencing changes in protein-coding regions. I have worked to improve the understanding of two under-studied groups of organisms but, as I have shown, variation between these is capable of altering interpretations. Further sampling and sequencing of the genomes of these organisms will not only increase the knowledge of these organisms, but of the whole clade as species and genus-specific genes and gene families are identified. Functional categorisation of *Blastocystis* specific gene families may also deepen our understanding of the role of *Blastocystis* both in healthy and diseased individuals. As more *Blastocystis* genomes are sequenced and become available it may also be possible to determine, with the aid of the out-groups produced and presented here, the genotypic differences that drive host specificity, preference and disease outcome between *Blastocystis* subtypes and isolates.

The genome and transcriptomes presented here are an example of the effort being made to collect genetic data on a huge range of organisms. Traditionally, model organisms have been selected based on traits that have more to do with their manageability, rather than their genetic features or phylogenetic position. Sequencing technology is becoming increasingly faster, more efficient and able to sequence longer pieces of DNA. Eventually, sequencing, assembly and annotation will occur in the same step as genomes are read end-to-end. This will dispense with the need for assembly programmes or metagenomic approaches to separating the genomes of organisms, as individual genomes will be sequenced in a single piece. At this point, judging the quality of an assembly will be moot. Therefore, the

interpretation will be influenced by the diversity of genomes available and the knowledge of how representative they are about the wider clade.

I have conducted these analyses *in silico* and provided predictions and direction for future studies that should aim to localise and categorise the function of potentially important proteins identified in this study. The addition of a viable animal infection model for both *Blastocystis* and *E. histolytica* would also help to generate much needed understanding of how these organisms behave *in vivo*.

6.4 Genomic reduction

In this thesis, I have made frequent reference to genome reduction, specifically to the loss of genes and gene families, rather than reduction of the physical size or capabilities of the genome. Genomic reduction is widespread across many organisms [455], but is especially obvious in single-celled organisms that inhabit a very stable environment, such as symbionts of both bacterial and eukaryotic origins [456]. These represent some of the most extreme examples of genomic reduction and there have been several attempts to explain why there might be so much pressure for symbiont genomes to become reduced [457]. Explanations for this widespread phenomenon include the need to conserve resources during replication or to keep replication time to a minimum [458]–[460]. While this selective pressure is not so obvious in multicellular organisms, which do not compete for resources in the same way, there is clear evidence for selective pressure acting, particularly on protist and bacterial genomes, to clear non-functional genes from the genome.

Genes accumulate mutations and become non-functional when they are non-essential to the survival or reproduction of the cell. In a free-living environment, cells need to be able to cope with large-scale environmental changes, including among others; salinity, temperature, predation and nutrients. These conditions may change suddenly and for extended periods of time. The genomes of free-living organisms must therefore retain their versatility and redundancy in order to persist in a constantly changing environment. Symbionts, however, rely on their host to accommodate the majority of environmental changes and thus, redundancies in symbiont genomes are truly superfluous and are subsequently lost. In *Buchnera*, an endosymbiont of aphids, this is exemplified by large reduction in genome size and

metabolism, except for certain amino acid biosynthesis pathways, as the host provides a consistent flow of the necessary nutrients required by the cell, bar the essential amino acids that the endosymbiont provides for the host [461], [462]. In tapeworms, the parasite has lost both its gut and light detection organs. The gut is unnecessary as it is able to take up nutrients from the host and even light detection has been lost because detection of this stimuli cannot influence the survival of the tapeworm [463]. The closer the association to the host, the less need there is for versatility and consequently, the greater the genomic reduction (e.g. mitochondria).

In parasitic lineages, this produces an interesting pattern. While genomic reduction is not ubiquitous across all parasitic lineages, many parasites do have reduced genomes. The Oomycetes [209], [210], [464] and nematodes [219], [220], [465] are examples of parasites that have comparatively large (unreduced) genomes. If the link between reduction and parasitism were true then why do these parasites not exhibit reduction in their genomes? This is linked not only to the level of association with the host, but also to the duration of that association. Both the Oomycetes and nematodes are free-living for a portion of their life-cycles [466], [467]. For the Oomycetes, this occurs during the zoospore stage, after sexual reproduction, when cells are disseminated into the aquatic environment in order to search out new hosts. At this point they must persist in the fluid conditions of a free-living environment. Nematodes also have a free-living life-stage after the eggs develop prior to infection of another host [468]. As a result, these parasitic lineages cannot dispense of their versatility and thus retain large genomes containing redundancies.

In symbionts that are parasitic throughout their life cycles, such as Apicomplexa and Trypanosomatids that parasitise a vector, and *Entamoeba* and *Blastocystis* that encyst in order to transmit, there is no need for such versatility and redundancy. These symbiont lineages are prime examples of genomic reduction [216], [223], [250]. In vector borne parasites, despite the need for versatility between hosts, there is not the same range of potential environmental conditions to cope with. Each host provides homeostasis of certain conditions and the parasite transmits directly between them. In intestinal symbionts, transmission is often faecal-oral and requires only the formation of a cyst [52], [73]. Cysts are resistant to environmental changes, without the need for versatility, they are metabolically dormant, content to wait until they are ingested by another host. This, minimal requirement, may help to explain why intestinal symbionts have such highly reduced

genomes even when compared to parasites with alternative modes of transmission. If true, then it is logical to enquire why the *P. lacertae* genome is so large compared to the *Blastocystis* genomes. If they inhabit similar environments, why has the *P. lacertae* genome not undergone a similar phenotype of reduction? The answer to this, is to admit that little is known about the natural ecology of *P. lacertae* and while it has been isolated from the faeces of lizards, this may only form part of its natural life cycle and it must be conceded that it may well inhabit environments outside the host for extended periods of time. The need of some symbionts to maintain their versatility for habitation of a free-living environment, even for a portion of their life cycle, may be the root of their respective lack of genome reduction.

6.5 Genomic expansion

This thesis has also identified a number of species-specific gene family expansions. Symbionts require a number of innovations in order to overcome the two major obstacles to inhabiting a host: the host immune system and transmission between hosts. Parasites also have the added need to invade and thus require the genes to facilitate this. It is precisely these parasite-specific tools that are used against them in the immune response and vaccine development. Cell surface expressed proteins are the most sought-after candidates for vaccine development; provided they are conserved among strains or even species of the parasite and are expressed constitutively. Proteins that are expressed in alternative life-stages or that are different between species may also be useful as diagnostic markers that allow differentiation between species or strains that may affect the outcome of the disease.

6.6 Free-living out-groups

The use of free-living organisms in this thesis means that it is possible to investigate the origin of symbiosis within a particular clade. This approach has been taken before to investigate the trypanosomes [206], Strongyloides [469] and Apicomplexa [250]. These studies reached similar conclusions about how symbiosis has arisen in each of these clades: adaptations that have arisen in the ancestor of the symbionts have provided a platform that enabled the transition from free-living to symbiotic. In the Kinetoplastida, ancestral bodonin, used for adhesion to prey, formed the basis for pathogenic adhesion to host cells as well as repurposing of the flagellum, which occurred in the ancestor of the Apicomplexa and facilitates their

invasive abilities [206].

This thesis has added two clades to this list, the intestinal Stramenopiles and the Archamoebae. Its findings also reflect the findings of the previous studies, which suggest that adaptations in the ancestors of symbiotic clades allowed the transition into the host environment. In the case of both the intestinal Stramenopiles and the Archamoebae, the other members of the clade are marine or freshwater dwelling. It is therefore reasonable to assume that the common respective ancestors of the intestinal Stramenopiles and Archamoebae also fell into this category. However, *Blastocystis* and *Entamoeba* are both anaerobic and so we must consider an aquatic, yet anaerobic environment that suited the transition into the intestine of an ancestral host. This was potentially sediment surrounding a freshwater drinking source. It may be that *P. lacertae* still spends some of its time in an environment similar to this.

6.7 The origins of parasitism

It may be interesting to speculate on the emergence of symbiosis in this clade, given the predicted habitat of these organisms prior to the evolution of symbiosis. Mechanisms for immune evasion may have arisen from methods initially useful for escape from binding agents released by bacteria or fungal competitors and contemporary virulence factors may have initially suited the degradation of bacterial biofilms or other protein-rich sources of nutrients. The encystation for transmission was likely already an evolved response to unfavourable condition such as oxygen stress or desiccation and this was simply used to bypass the hostile conditions in the stomach until the more favourable condition of the intestine were reached. After being passed in the stool of the host, encystation again allows the organism to wait until ingestion by another host. Successive passage through this system may have produced variants that became better suited to this environment until a variant emerged that was only able to persist within a host.

6.8 Final thoughts

Genomic reduction is seen in some, but not all symbionts depending on their ancestral state and the degree of obligate biotrophy throughout their life cycles. But all symbiont genomes are adapted for survival within the host and transmission between hosts. This thesis has provided a first insight into how these processes

have shaped the *Blastocystis* genome (loss of motility, genomic streamlining, expansion of adhesive proteins), and the *Entamoeba* genome (HGT, expansion of proteases and of Ras family proteins).

These are processes and adaptations that have been seen in other symbiotic clades, suggesting that there are some aspects of becoming a symbiont that are reproducible in broadly different organisms. This most likely represents the common challenge of interacting with hosts, principally immunity, and the similar physical constraints of surviving and prospering in a living environment.

References

[1] R. E. Ley, C. A. Lozupone, M. Hamady, R. Knight, and J. I. Gordon, "**Worlds within worlds: evolution of the vertebrate gut microbiota,**" *Nature Reviews Microbiology*, vol. 6, no. 10, pp. 776–788, 2008.

[2] H. Liu, X. Guo, R. Gooneratne, R. Lai, C. Zeng, F. Zhan, and W. Wang, "**The gut microbiome and degradation enzyme activity of wild freshwater fishes influenced by their trophic levels.,**" *Scientific reports*, vol. 6, no. September 2015, p. 24340, Apr. 2016.

[3] F. Backhed, "**Host-Bacterial Mutualism in the Human Intestine,**" *Science*, vol. 307, no. 5717, pp. 1915–1920, 2005.

[4] A. Alberdi, O. Aizpurua, K. Bohmann, M. L. Zepeda-Mendoza, and M. T. P. Gilbert, "**Do Vertebrate Gut Metagenomes Confer Rapid Ecological Adaptation?,**" *Trends in Ecology and Evolution*, vol. 31, no. 9. Secretariat of the Convention on Biological Diversity, pp. 689–699, Sep-2016.

[5] L. C. Robosky, D. F. Wells, L. A. Egnash, M. L. Manning, M. D. Reily, and D. G. Robertson, "**Metabonomic identification of two distinct phenotypes in Sprague-Dawley (Crl:CD(SD)) rats,**" *Toxicological Sciences*, vol. 87, no. 1, pp. 277–284, 2005.

[6] C. M. Rohde, D. F. Wells, L. C. Robosky, M. L. Manning, C. B. Clifford, M. D. Reily, and D. G. Robertson, "**Metabonomic evaluation of schaedler altered microflora rats,**" *Chemical Research in Toxicology*, vol. 20, no. 10, pp. 1388–1392, Oct. 2007.

[7] J. L. Sonnenburg, "**Glycan Foraging in Vivo by an Intestine-Adapted Bacterial Symbiont,**" *Science*, vol. 307, no. 5717, pp. 1955–1959, 2005.

[8] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon, "**Human gut microbiome viewed across age and geography,**" *Nature*, vol. 486, no. 7402, pp. 222–227, May 2012.

[9] P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon, "**An obesity-associated gut microbiome with increased capacity for energy harvest,**" *Nature*, vol. 444, no. 7122, pp. 1027–1131, Dec. 2006.

[10] T. Olszak, D. An, S. Zeissig, M. P. Vera, J. Richter, A. Franke, J. N. Glickman, R. Siebert, R. M. Baron, D. L. Kasper, and R. S. Blumberg, "**Microbial Exposure During Early Life Has Persistent Effects on Natural Killer T Cell Function,**" *Science*, vol. 336, no. 6080, pp. 489–493, Apr. 2012.

[11] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon, "**Microbial ecology: Human gut microbes associated with obesity,**" *Nature*, vol. 444, no. 7122, pp. 1022–1023, Dec. 2006.

[12] A. L. Kau, P. P. Ahern, N. W. Griffin, A. L. Goodman, and J. I. Gordon, "**Human nutrition, the gut microbiome and the immune system.,**" *Nature*, vol. 474, no. 7351, pp. 327–336, 2011.

[13] J. Dicksved, J. Halfvarson, M. Rosenquist, G. Järnerot, C. Tysk, J. Apajalahti, L. Engstrand, and J. K. Jansson, "**Molecular analysis of the gut microbiota of identical twins with Crohn's disease,**" *The ISME Journal*, vol. 2, no. 7, pp. 716–727, Jul. 2008.

[14] D. N. Frank, A. L. St. Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace, "**Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases,**" *Proceedings of the National Academy of Sciences*, vol. 104, no. 34, pp. 13780–13785, Aug. 2007.

[15] A. Gonzalez, J. Stombaugh, C. Lozupone, P. J. Turnbaugh, J. I. Gordon, and R. Knight, "**The mind-body-microbial continuum,**" *Dialogues in Clinical Neuroscience*, vol. 13, no. 1, pp. 55–62, 2011.

[16] J. R. Lupton, "**Microbial degradation products influence colon cancer risk: the butyrate controversy.,**" *The Journal of nutrition*, vol. 134, no. 2, pp. 479–482, Feb. 2004.

[17] J. Qin, R. Li, J. Raes, M. Arumugam, S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, and H. Yang, "**A human gut microbial gene catalog established by metagenomic sequencing,**" *Nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[18] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "**The Human Microbiome Project,**" *Nature*, vol. 449, no. October, pp. 804–810, 2007.

[19] R. E. Ley, M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight, and J. I. Gordon, "**Evolution of Mammals and Their Gut Microbes,**" *Science*, vol. 320, no. 5883, pp. 1647–1651, 2008.

[20] P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, M. Sargent, S. R. Gill, K. E. Nelson, and D. A. Relman, "**Diversity of the Human Intestinal Microbial Flora,**" vol. 308, no. 5728, pp. 1635–1638, 2006.

[21] H.-P. Horz, "**Archaeal Lineages within the Human Microbiome: Absent, Rare or Elusive?,**" *Life (Basel, Switzerland)*, vol. 5, no. 2, pp. 1333–45, May 2015.

[22] C. A. Lozupone, J. I. Stombaugh, J. I. Gordon, J. K. Jansson, and R. Knight, "**Diversity, stability and resilience of the human gut microbiota,**" *Nature*, vol. 489, no. 7415, pp. 220–230, 2012.

[23] L. W. Parfrey, W. A. Walters, and R. Knight, "**Microbial eukaryotes in the human microbiome: Ecology, evolution, and future directions,**" *Frontiers in Microbiology*, vol. 2, no. JULY, pp. 1–6, 2011.

[24] A. Graham, **Ecological rules governing helminth-microparasite co-infection**, vol. 105. 2008.

[25] K. D. Lafferty, "**Interacting Parasites,**" *Science*, vol. 330, no. 6001, pp. 187–188, 2010.

[26] R. Radek, "**Flagellates, bacteria, and fungi assoociated with termites: diversity and function in nutrition- a review,**" pp. 183–196, 1999.

[27] H. Wildschutte, D. M. Wolfe, A. Tamewitz, and J. G. Lawrence, "**Protozoan predation, diversifying selection, and the evolution of antigenic diversity in Salmonella,**" *Proceedings of the National Academy of Sciences*, vol. 101, no. 29, pp. 10644–10649, 2004.

[28] J. R. Marchesi, "**Prokaryotic and eukaryotic diversity of the human gut,**" *Advances in Applied Microbiology*, vol. 72, no. C, pp. 43–62, 2010.

[29] C. R. Stensvold, M. Lebbad, E. L. Victory, J. J. Verweij, E. Tannich, M. Alfellani, P. Legarraga, and C. G. Clark, "**Increased sampling reveals novel lineages of entamoeba: Consequences of genetic diversity and host specificity for taxonomy and molecular detection,**" *Protist*, vol. 162, no. 3, pp. 525–541, 2011.
[30] B. S. Pritt and C. G. Clark, "**Amebiasis,**" *Mayo Clinic Proceedings*, vol. 83, no. 10, pp. 1154–1160, Oct. 2008.
[31] Ü. Z. Ok, M. Cirit, A. Üner, E. Ok, F. Akçiçek, A. Başçi, and M. A. Özcel, "**Cryptosporidiosis and blastocystosis in renal transplant recipients,**" *Nephron*, vol. 75, no. 2, pp. 171–174, 1997.
[32] J. Ackers, S. Gillepsie, and R. Pearson, **Trichomonads**. Chichester: John Wiley & Sons, 2001.
[33] A. Alexeieff, "**Sur la nature des formations dites 'kystes de Trichomonas intestinalis,'**" *Comptes Rendus Hebdomadaires des Séances et Mémoires de la Société de Biologie*, vol. 2, no. 71, pp. 296–298, 1911.
[34] E. Brumpt, "**Côlite à Tetramitus Mesnili (Wenyon 1910) et Côlite a Trichomonas intestinalis Leuckart 1879. Blastocystis hominis n. sp. et formes voisines,**" *Bulletin de la Société de Pathologie Exotique*, vol. 5, pp. 725–730, 1912.
[35] C. R. Stensvold, G. K. Suresh, K. S. W. Tan, R. C. A. Thompson, R. J. Traub, E. Viscogliosi, H. Yoshikawa, and C. G. Clark, "**Terminology for Blastocystis subtypes - a consensus,**" *Trends in Parasitology*, vol. 23, no. 3, pp. 93–96, Mar. 2007.
[36] C. H. Zierdt, W. S. Rude, and B. S. Bull, "**Protozoan characteristics of Blastocystis hominis.,**" *American Journal of Clinical Pathology*, vol. 48, no. 5, pp. 495–501, 1967.
[37] T. Jelinek, G. Peyerl, T. L6scher, F. Von Sonnenburg, and H. D. Nothdurft, "**The Role of Blastocystis hominis as a Possible Intestinal Pathogen in Travellers,**" *Journal of Infection*, vol. 35, pp. 63–66, 1997.
[38] K. S. W. Tan, "**Blastocystis in humans and animals: New insights using modern methodologies,**" *Veterinary Parasitology*, vol. 126, no. 1–2 SPEC.ISS. Elsevier, pp. 121–144, Dec-2004.
[39] L. O. Andersen and C. R. Stensvold, "**Blastocystis in Health and Disease: Are We Moving from a Clinical to a Public Health Perspective?,**" *Journal of Clinical Microbiology*, vol. 54, no. 3. American Society for Microbiology, pp. 524–528, Mar-2016.
[40] E. T. Baldo, V. Y. Belizario, W. U. De Leon, H.-H. Kong, and D.-I. Chung, "**Infection status of intestinal parasites in children living in residential institutions in Metro Manila, the Philippines,**" *The Korean Journal of Parasitology*, vol. 42, no. 2, p. 67, 2004.
[41] O. Cirioni, A. Giacometti, D. Drenaggi, F. Ancarani, and G. Scalise, "**Prevalence and clinical relevance of Blastocystis hominis in diverse patient cohorts.,**" *European journal of epidemiology*, vol. 15, no. 4, pp. 389–393, 1999.
[42] A. C. Florez, D. A. Garcia, L. Moncada, and M. Beltran, "**Prevalence of microsporidia and other intestinal parasites in patients with HIV infection, Bogota, 2001,**" *Biomedica*, vol. 23, no. 3, pp. 274–282, Sep. 2003.
[43] B. L. Herwaldt, K. R. De Arroyave, S. P. Wahlquist, A. M. De Merida, A. S.

Lopez, and D. D. Juranek, "**Multiyear prospective study of intestinal parasitism in a cohort of Peace Corps volunteers in Guatemala,**" *Journal of Clinical Microbiology*, vol. 39, no. 1, pp. 34–42, Jan. 2001.
[44] K. Pegelow, R. Gross, K. Pietrzik, W. Lukito, A. L. Richards, and D. Fryauff, **Parasitological and nutritional situation of school children in the Sukaraja district, West Java, Indonesia**, vol. 28. 1997.
[45] W. Saksirisampant, V. Wiwanitkit, A. Ampavasiri, S. Nuchprayoon, and S. Yenthakam, "**Intestinal parasitic infestations among children in an orphanage in Pathum Thani province,**" *Journal of the Medical Association of Thailand*, vol. 86, no. SUPPL. 2, 2003.
[46] P. Taamasri, M. Mungthin, R. Rangsin, B. Tongupprakarn, W. Areekul, and S. Leelayoova, "**Transmission of intestinal blastocystosis related to the quality of drinking water,**" *Southeast Asian Journal of Tropical Medicine and Public Health*, vol. 31, no. 1, pp. 112–117, 2000.
[47] L. C. Wang, "**Changing patterns in intestinal parasitic infections among Southeast Asian laborers in Taiwan,**" *Parasitology Research*, vol. 92, no. 1, pp. 18–21, Jan. 2004.
[48] J. J. Windsor, L. Macfarlane, G. Hughes-Thapa, S. K. A. Jones, and T. M. Whiteside, "**Incidence of Blastocystis hominis in faecal samples submitted for routine microbiological analysis,**" *British Journal of Biomedical Science*, vol. 59, no. 3, pp. 154–157, Jan. 2002.
[49] O. M. Amin, "**Seasonal prevalence of intestinal parasites in the United States during 2000.,**" *The American journal of tropical medicine and hygiene*, vol. 66, no. 6, pp. 799–803, 2002.
[50] K. S. W. Tan, "**New insights on classification, identification, and clinical relevance of Blastocystis spp.,**" *Clinical Microbiology Reviews*, vol. 21, no. 4, pp. 639–665, 2008.
[51] C. H. Zierdt, "**Blastocystis hominis--past and future.,**" *Clinical microbiology reviews*, vol. 4, no. 1, pp. 61–79, Jan. 1991.
[52] P. F. L. L. Boreham and D. J. Stenzel, "**Blastocystis in Humans and Animals: Morphology, Biology, and Epizootiology,**" *Advances in Parasitology*, vol. 32, no. C, pp. 1–70, Jan. 1993.
[53] D. J. Stenzel and P. F. L. L. Boreham, "**Blastocystis hominis revisited,**" *Clinical Microbiology Reviews*, vol. 9, no. 4, pp. 563–584, Oct. 1996.
[54] C. G. Clark, M. van der Giezen, M. A. Alfellani, and C. R. Stensvold, "**Recent developments in Blastocystis research.,**" *Advances in parasitology*, vol. 82, pp. 1–32, Jan. 2013.
[55] A. A. Vdovenko, "**Blastocystis hominis: Origin and significance of vacuolar and granular forms,**" *Parasitology research*, vol. 86, pp. 8–10, Feb. 2000.
[56] H. Yoshikawa, K. Yoshida, A. Nakajima, K. Yamanari, S. Iwatani, and I. Kimata, "**Fecal-oral transmission of the cyst form of Blastocystis hominis in rats,**" *Parasitology Research*, vol. 94, no. 6, pp. 391–396, 2004.
[57] K. T. Moe, M. Singh, J. Howe, L. C. Ho, S. W. Tan, G. C. Ng, X. Q. Chen, and E. H. Yap, "**Observations on the ultrastructure and viability of the cystic stage of Blastocystis hominis from human feces,**" *Parasitology Research*, vol. 82, no. 5, pp. 439–444, May 1996.
[58] H. K. Tan and C. H. Zierdt, "**Ultrastructure of Blastocystis hominis,**" *Zeitschrift für Parasitenkunde (Berlin, Germany)*, vol. 42, pp. 315–324, Dec.

1973.
[59] L. A. Dunn, P. F. L. Boreham, and D. J. Stenzel, "**Ultrastructural variation of Blastocystis hominis stocks in culture,**" *International Journal for Parasitology*, vol. 19, no. 1, pp. 43–56, 1989.
[60] C. H. Zierdt and H. K. Tan, "**Ultrastructure and light microscope appearance of Blastocystis hominis in a patient with enteric disease,**" *Zeitschrift für Parasitenkunde (Berlin, Germany)*, vol. 50, pp. 277–283, Nov. 1976.
[61] D. J. Stenzel and P. F. L. Boreham, "**A cyst-like stage of Blastocystis hominis,**" *International Journal for Parasitology*, vol. 21, no. 5, pp. 613–615, 1991.
[62] A. M. Johnson, A. Thanou, P. F. L. Boreham, and P. R. Baverstock, "**Blastocystis hominis: phylogenetic affinities determined by rRNA sequence comparison.,**" *Experimental parasitology*, vol. 68, no. 3, pp. 283–288, 1989.
[63] J. D. Silberman, M. L. Sogin, D. D. Leipe, and C. G. Clark, "**Human parasite finds taxonomic home.**" p. 398, 1996.
[64] N. Arisue, T. Hashimoto, H. Yoshikawa, Y. Nakamura, G. Nakamura, F. Nakamura, T.-A. Yano, and M. Hasegawa, "**Phylogenetic Position of Blastocystis hominis and of Stramenopiles Inferred from Multiple Molecular Sequence Data,**" vol. 49, no. 1, pp. 42–53, 2002.
[65] J. D. Hoevers and K. F. Snowden, "**Analysis of the ITS region and partial ssu and lsu rRNA genes of Blastocystis and Proteromonas lacertae.,**" *Parasitology*, vol. 131, no. Pt 2, pp. 187–196, 2005.
[66] D. J. Patterson, "**Stramenopiles: chromophytes from a protistan perspective,**" in *The Chromophyte Algae: Problems and Perspectives*, 1989, pp. 357–379.
[67] D. D. Leipe, P. O. Wainright, J. H. Gunderson, D. Porter, D. J. Patterson, F. Valois, S. Himmerich, and M. L. Sogin, "**The stramenopiles from a molecular prospective: 16S-like rRNA sequences from Labyrinthuloidesminuta and Cafeteria roenbergensis,**" *Phycologia*, vol. 33, no. 5, pp. 369–377, Sep. 1994.
[68] M. Moriya, T. Nakayama, and I. Inouye, "**A New Class of the Stramenopiles, Placididea Classis nova: Description of Placidia cafeteriopsis gen. et sp. nov.,**" *Protist*, vol. 153, no. 2, pp. 143–156, 2002.
[69] M. A. Alfellani, C. R. Stensvold, A. Vidal-Lapiedra, E. S. U. Onuoha, A. F. Fagbenro-Beyioku, and C. G. Clark, "**Variable geographic distribution of Blastocystis subtypes and its potential implications,**" *Acta Tropica*, vol. 126, no. 1, pp. 11–18, 2013.
[70] I. Wawrzyniak, P. Poirier, E. Viscogliosi, M. Dionigia, C. Texier, F. Delbac, and H. El Alaoui, "**Blastocystis, an unrecognized parasite: an overview of pathogenesis and diagnosis.,**" *Therapeutic advances in infectious disease*, vol. 1, no. 5, pp. 167–178, Oct. 2013.
[71] M. A. Alfellani, A. S. Jacob, N. O. Perea, R. C. Krecek, D. Taner-Mulla, J. J. Verweij, B. Levecke, E. Tannich, C. G. Clark, and C. R. Stensvold, "**Diversity and distribution of Blastocystis sp. subtypes in non-human primates.,**" *Parasitology*, vol. 140, no. 8, pp. 966–971, Jul. 2013.
[72] R. D. Adam, "**Biology of Giardia lamblia,**" *Clinical Microbiology Reviews*, vol. 10, no. 1. pp. 19–34, 1997.

[73] I. W. Wilson, G. D. Weedall, and N. Hall, "**Host-Parasite interactions in Entamoeba histolytica and Entamoeba dispar: What have we learned from their genomes?,**" *Parasite Immunology*, vol. 34, no. 2–3, pp. 90–99, 2012.
[74] A. J. Mazurie, J. M. Alves, L. S. Ozaki, S. Zhou, D. C. Schwartz, and G. A. Buck, "**Comparative genomics of cryptosporidium,**" *International Journal of Genomics*, vol. 2013, p. 832756, 2013.
[75] K. Kain, M. Noble, H. J. Freeman, and R. L. Barteluk, "**Epidemiology and clinical features associated with Blastocystis hominis infection,**" *Diagnostic microbiology and infectious disease*, vol. 8, pp. 235–244, Jan. 1988.
[76] M. Lambert, J. Gigi, and C. Bughin, "**Persistent diarrhoea and Blastocystis hominis.,**" *Acta clinica Belgica*, vol. 47, no. 2, pp. 129–130, 1992.
[77] R. A. Miller and B. H. Minshew, "**Blastocystis hominis: An Organism in Search of a Disease,**" *Reviews of Infectious Diseases*, vol. 10, no. 5, pp. 930–938, Sep. 1988.
[78] L. Nimri and R. Batchoun, "**Intestinal colonization of symptomatic and asymptomatic schoolchildren with Blastocystis hominis,**" *Journal of Clinical Microbiology*, vol. 32, no. 11. pp. 2865–2866, 1994.
[79] M. Singh, K. Suresh, L. C. Ho, G. C. Ng, and E. H. Yap, "**Elucidation of the life cycle of the intestinal protozoan Blastocystis hominis,**" *Parasitology Research*, vol. 81, no. 5, pp. 446–450, 1995.
[80] C. H. Zierdt, "**Studies of Blastocystis hominis,**" *The Journal of Protozoology*, vol. 20, no. 1, pp. 114–121, Feb. 1973.
[81] X. Zhang, J. Y. Qiao, X. J. Zhou, F. R. Yao, and Z. C. Wei, "**Morphology and reproductive mode of Blastocystis hominis in diarrhea and in vitro,**" *Parasitology Research*, vol. 101, no. 1, pp. 43–51, Jun. 2007.
[82] S. K. Govind, A. A. Khairul, H. V Smith, J. J. Windsor, D. J. Stenzel, L. Macfarlane, S. K. Govind, A. A. Khairul, and H. V Smith, "**Multiple reproductive processes in Blastocystis hominis,**" *Trends in parasitology*, vol. 18, no. 12, pp. 289–290, Dec. 2002.
[83] X. Q. Chen, M. Singh, J. Howe, L. C. Ho, S. W. Tan, and E. H. Yap, "**In vitro encystation and excystation of Blastocystis ratti,**" *Parasitology*, vol. 118, no. 2, pp. 151–160, 1999.
[84] K. T. Moe, M. Singh, J. Howe, L. C. Ho, S. W. Tan, X. Q. Chen, and E. H. Yap, "**Development of Blastocystis hominis cysts into vacuolar forms in vitro,**" *Parasitology Research*, vol. 85, no. 2, pp. 103–108, Jan. 1999.
[85] S. W. Tan, M. Singh, E. H. Yap, L. C. Ho, K. T. Moe, J. Howe, and G. C. Ng, "**Colony formation of Blastocystis hominis in soft agar,**" *Parasitology Research*, vol. 82, no. 4. Springer-Verlag, pp. 375–377, Apr-1996.
[86] S. W. Tan, M. Singh, K. T. Thong, L. C. Ho, K. T. Moe, X. Q. Chen, G. C. Ng, and E. H. Yap, "**Clonal growth of Blastocytosis hominis in soft agar with sodium thioglycollate,**" *Parasitology Research*, vol. 82, no. 8. Springer-Verlag, pp. 737–739, Sep-1996.
[87] R. Fayer, T. Elsasser, R. Gould, G. Solano, J. Urban, and M. Santin, "**Blastocystis tropism in the pig intestine,**" *Parasitology Research*, vol. 113, no. 4, pp. 1465–1472, Apr. 2014.
[88] S. S. R. Ajjampur, C. W. Png, W. N. Chia, Y. Zhang, and K. S. W. Tan, "**Ex Vivo**

and In Vivo Mice Models to Study Blastocystis spp. Adhesion, Colonization and Pathology: Closer to Proving Koch's Postulates," *PLOS ONE*, vol. 11, no. 8, p. e0160458, Aug. 2016.

[89] J. A. Upcroft, L. A. Dunn, L. S. Dommett, A. Healey, P. Upcroft, and P. F. L. Boreham, "**Chromosomes of Blastocystis hominis,**" *International Journal for Parasitology*, vol. 19, no. 8, pp. 879–883, Dec. 1989.

[90] J. A. Carbajal, L. Del Castillo, M. D. Lanuza, J. Villar, and R. Borrás, "**Karyotypic diversity among Blastocystis hominis isolates,**" *International Journal for Parasitology*, vol. 27, no. 8, pp. 941–945, Aug. 1997.

[91] F. Denoeud, M. Roussel, B. Noel, I. Wawrzyniak, C. Da Silva, M. Diogon, E. Viscogliosi, C. Brochier-Armanet, A. Couloux, J. Poulain, B. Segurens, V. Anthouard, C. Texier, N. Blot, P. Poirier, G. C. Ng, K. S. W. Tan, F. Artiguenave, O. Jaillon, J.-M. Aury, F. Delbac, P. Wincker, C. P. Vivarès, and H. El Alaoui, "**Genome sequence of the stramenopile Blastocystis, a human anaerobic parasite.,**" *Genome biology*, vol. 12, no. 3, p. R29, Jan. 2011.

[92] N. Andiran, Z. C. Acikgoz, S. Turkay, and F. Andiran, "**Blastocystis hominis-an emerging and imitating cause of acute abdomen in children,**" *Journal of Pediatric Surgery*, vol. 41, no. 8, pp. 1489–1491, Aug. 2006.

[93] S. Leelayoova, R. Rangsin, P. Taamasri, T. Naaglor, U. Thathaisong, and M. Mungthin, "**Evidence of waterborne transmission of Blastocystis hominis,**" *American Journal of Tropical Medicine and Hygiene*, vol. 70, no. 6, pp. 658–662, 2004.

[94] M. Carrascosa, J. Martinez, and J. L. Perez-Castrilld, "**Hemorriiagic Proctosigmoiditis and Blastocystis hominis Infection,**" *Annals of Internal Medicine*, vol. 124, no. 2, pp. 278–279, Jan. 1996.

[95] Y. Levy, J. George, and Y. Shoenfeld, "**Severe Blastocystis hominis in an elderly man,**" *Journal of Infection*, vol. 33, no. 1, pp. 57–59, Jul. 1996.

[96] K. Leder, M. E. Hellard, M. I. Sinclair, C. K. Fairley, and R. Wolfe, "**No correlation between clinical symptoms and Blastocystis hominis in immunocompetent individuals,**" *Journal of Gastroenterology and Hepatology (Australia)*, vol. 20, no. 9, pp. 1390–1394, Sep. 2005.

[97] A. Tungtrongchitr, S. Manatsathit, C. Kositchaiwat, J. Ongrotchanakun, N. Munkong, P. Chinabutr, S. Leelakusolvong, and W. Chaicumpa, "**Blastocystis hominis infection in irritable bowel syndrome patients,**" *Southeast Asian J Trop Med Public Health*, vol. 35, no. 3, pp. 705–710, Sep. 2004.

[98] L. W. Parfrey, W. A. Walters, C. L. Lauber, J. C. Clemente, D. Berg-Lyons, C. Teiling, C. Kodira, M. Mohiuddin, J. Brunelle, M. Driscoll, N. Fierer, J. A. Gilbert, and R. Knight, "**Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity,**" *Frontiers in Microbiology*, vol. 5, no. JUN, p. 298, 2014.

[99] P. D. Scanlan and J. R. Marchesi, "**Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces,**" *The ISME Journal*, vol. 2, no. 12, pp. 1183–1193, 2008.

[100] A. M. Petersen, C. R. Stensvold, H. Mirsepasi, J. Engberg, A. Friis-Møller, L.

J. Porsbo, A. M. Hammerum, I. Nordgaard-Lassen, H. V. Nielsen, and K. A. Krogfelt, "**Active ulcerative colitis associated with low prevalence of Blastocystis and Dientamoeba fragilis infection.,**" *Scandinavian journal of gastroenterology*, vol. 48, no. 5, pp. 638–639, May 2013.

[101] K. F. Boorom, H. Smith, L. Nimri, E. Viscogliosi, G. Spanakos, U. Parkar, L.-H. Li, X.-N. Zhou, U. Z. Ok, S. Leelayoova, and M. S. Jones, "**Oh my aching gut: irritable bowel syndrome, Blastocystis, and asymptomatic infection.,**" *Parasites & vectors*, vol. 1, no. 1, p. 40, 2008.

[102] K. S. W. Tan, H. Mirza, J. D. W. Teo, B. Wu, and P. a Macary, "**Current Views on the Clinical Relevance of Blastocystis spp.,**" *Current infectious disease reports*, vol. 12, no. 1, pp. 28–35, Jan. 2010.

[103] P. Poirier, I. Wawrzyniak, C. P. Vivarès, F. Delbac, and H. El Alaoui, "**New insights into Blastocystis spp.: A potential link with irritable bowel syndrome,**" *PLoS Pathogens*, vol. 8, no. 3, p. e1002545, Mar. 2012.

[104] P. D. Scanlan, "**Blastocystis: Past pitfalls and future perspectives,**" *Trends in Parasitology*, vol. 28, no. 8. pp. 327–334, Aug-2012.

[105] J. Yakoob, W. Jafri, M. A. Beg, Z. Abbas, S. Naz, M. Islam, and R. Khan, "**Blastocystis hominis and Dientamoeba fragilis in patients fulfilling irritable bowel syndrome criteria,**" *Parasitology Research*, vol. 107, no. 3, pp. 679–684, Aug. 2010.

[106] S. A. Fouad, M. M. A. Basyoni, R. A. Fahmy, and M. H. Kobaisi, "**The pathogenic role of different Blastocystis hominis genotypes isolated from patients with irritable bowel syndrome,**" *Arab Journal of Gastroenterology*, vol. 12, no. 4, pp. 194–200, Dec. 2011.

[107] D. E. Jimenez-Gonzalez, W. A. Martinez-Flores, J. Reyes-Gordillo, M. E. Ramirez-Miranda, Sara Arroyo-Escalante, M. Romero-Valdovinos, D. Stark, V. Souza-Saldivar, F. Martinez-Hernandez, A. Flisser, A. Olivo-Diaz, and P. Maravilla, "**Blastocystis infection is associated with irritable bowel syndrome in a Mexican patient population,**" *Parasitology Research*, vol. 110, no. 3, pp. 1269–1275, Mar. 2012.

[108] R. Nagel, L. Cuttell, C. R. Stensvold, P. C. Mills, H. Bielefeldt-Ohmann, and R. J. Traub, "**Blastocystis subtypes in symptomatic and asymptomatic family members and pets and response to therapy.,**" *Internal medicine journal*, vol. 42, no. 11, pp. 1187–1195, Nov. 2012.

[109] C. Vogelberg, C. R. Stensvold, S. Monecke, A. Ditzen, K. Stopsack, U. Heinrich-Gr??fe, and C. P??hlmann, "**Blastocystis sp. subtype 2 detection during recurrence of gastrointestinal and urticarial symptoms,**" *Parasitology International*, vol. 59, no. 3, pp. 469–471, Sep. 2010.

[110] D. M. Abdel-Hameed and O. M. Hassanin, "**Proteaese activity of Blastocystis hominis subtype3 in symptomatic and asymptomatic patients.,**" *Parasitology research*, vol. 109, no. 2, pp. 321–327, Aug. 2011.

[111] N. M. Zuel-Fakkar, D. M. Abdel Hameed, and O. M. Hassanin, "**Study of Blastocystis hominis isolates in urticaria: A case-control study,**" *Clinical and Experimental Dermatology*, vol. 36, no. 8, pp. 908–910, Dec. 2011.

[112] R. Verma and K. Delfanian, "**Blastocystis hominis associated acute urticaria.,**" *The American journal of the medical sciences*, vol. 346, no. 1, pp. 80–81, Jul. 2013.

[113] E. M. Hussein, A. M. Hussein, M. M. Eida, and M. M. Atwa, "**Pathophysiological variability of different genotypes of human Blastocystis hominis Egyptian isolates in experimentally infected rats,**" *Parasitology Research*, vol. 102, no. 5, pp. 853–860, Apr. 2008.
[114] Y. Yan, S. Su, R. Lai, H. Liao, J. Ye, X. Li, X. Luo, and G. Chen, "**Genetic variability of Blastocystis hominis isolates in China,**" *Parasitology Research*, vol. 99, no. 5, pp. 597–601, Oct. 2006.
[115] M. Özyurt, Ö. Kurt, K. Mølbak, H. V. Nielsen, T. Haznedaroglu, and C. R. Stensvold, "**Molecular epidemiology of Blastocystis infections in Turkey,**" *Parasitology International*, vol. 57, no. 3, pp. 300–306, Sep. 2008.
[116] M. V. Domínguez-Márquez, R. Guna, C. Muñoz, M. T. Gómez-Muñoz, and R. Borrás, "**High prevalence of subtype 4 among isolates of Blastocystis hominis from symptomatic patients of a health district of Valencia (Spain),**" *Parasitology Research*, vol. 105, no. 4, pp. 949–955, Oct. 2009.
[117] C. R. Stensvold, M. A. Alfellani, S. Nørskov-Lauritsen, K. Prip, E. L. Victory, C. Maddox, H. V Nielsen, and C. G. Clark, "**Subtype distribution of Blastocystis isolates from synanthropic and zoo animals and identification of a new subtype,**" *International Journal for Parasitology*, vol. 39, no. 4, pp. 473–479, 2009.
[118] J. Forsell, M. Granlund, C. R. Stensvold, G. C. Clark, and B. Evengard, "**Subtype analysis of Blastocystis isolates in Swedish patients,**" *European Journal of Clinical Microbiology and Infectious Diseases*, vol. 31, no. 7, pp. 1689–1696, Jul. 2012.
[119] A. Katsarou-Katsari, C. M. C. M. Vassalos, K. Tzanetou, G. Spanakos, C. Papadopoulou, and N. Vakalis, "**Acute urticaria associated with amoeboid forms of Blastocystis sp. subtype 3,**" *Acta Dermato-Venereologica*, vol. 88, no. 1. pp. 80–81, 2008.
[120] R. Valsecchi, P. Leghissa, and V. Greco, "**Cutaneous lesions in Blastocystis hominis infection.,**" *Acta dermato-venereologica*, vol. 84, no. 4, pp. 322–323, 2004.
[121] H. Yoshikawa, Z. Wu, I. Kimata, M. Iseki, I. K. M. D. Ali, M. B. Hossain, V. Zaman, R. Haque, and Y. Takahashi, "**Polymerase chain reaction-based genotype classification among human Blastocystis hominis populations isolated from different countries,**" *Parasitology Research*, vol. 92, no. 1, pp. 22–29, Jan. 2004.
[122] F. Dogruman-Al, H. Dagci, H. Yoshikawa, Ö. Kurt, and M. Demirel, "**A possible link between subtype 2 and asymptomatic infections of Blastocystis hominis,**" *Parasitology Research*, vol. 103, no. 3, pp. 685–689, Aug. 2008.
[123] L. Souppart, G. Sanciu, A. Cian, I. Wawrzyniak, F. Delbac, M. Capron, E. Dei-Cas, K. Boorom, L. Delhaes, and E. Viscogliosi, "**Molecular epidemiology of human Blastocystis isolates in France,**" *Parasitology Research*, vol. 105, no. 2, pp. 413–421, Jul. 2009.
[124] S. Jantermtor, P. Pinlaor, K. Sawadpanich, S. Pinlaor, A. Sangka, C. Wilailuckana, W. Wongsena, and H. Yoshikawa, "**Subtype identification of Blastocystis spp. isolated from patients in a major hospital in northeastern Thailand,**" *Parasitology Research*, vol. 112, no. 4, pp. 1781–1786, Apr. 2013.

[125] T. Roberts, D. Stark, J. Harkness, and J. Ellis, "**Subtype distribution of Blastocystis isolates identified in a Sydney population and pathogenic potential of Blastocystis,**" *European Journal of Clinical Microbiology and Infectious Diseases*, vol. 32, no. 3, pp. 335–343, Mar. 2013.
[126] B. Böhm-Gloning, J. Knobloch, and B. Walderich, "**Five subgroups of Blastocystis hominis from symptomatic and asymptomatic patients revealed by restriction site analysis of PCR-amplified 16S-like rDNA.,**" *Tropical medicine & international health : TM & IH*, vol. 2, no. 8, pp. 771–778, Aug. 1997.
[127] Y. Kaneda, N. Horiki, X. J. Cheng, Y. Fujita, M. Maruyama, and H. Tachibana, "**Ribodemes of Blastocystis Hominis isolated in Japan,**" *American Journal of Tropical Medicine and Hygiene*, vol. 65, no. 4, pp. 393–396, Oct. 2001.
[128] A. E. Pinilla, M. C. López, and D. F. Viasus, "**Historia del protozoo Entamoeba histolytica,**" *Revista Medica de Chile*, vol. 136, no. 1. Sociedad Médica de Santiago, pp. 118–124, Jan-2008.
[129] L. S. Diamond and C. G. Clark, "**A Redescription of Entamoeba Histolytica Schaudinn, 1903 (Emended Walker, 1911) Separating It From Entamoeba Dispar Brumpt, 1925,**" *Journal of Eukaryotic Microbiology*, vol. 40, no. 3, pp. 340–344, May 1993.
[130] T. Cavalier-Smith, "**A 6-kingdom classification and a unified phylogeny.,**" *In: HEA Schenk, Schwemmler W (eds) Endocytobiology II. De Gruyter, Berlin*, pp. 1027–1034, 1983.
[131] T. Pánek, E. Zadrobílková, G. Walker, M. W. Brown, E. Gentekaki, M. Hroudová, S. Kang, A. J. Roger, A. K. Tice, Č. Vlček, and I. Čepička, "**First multigene analysis of Archamoebae (Amoebozoa: Conosa) robustly reveals its phylogeny and shows that Entamoebidae represents a deep lineage of the group,**" *Molecular Phylogenetics and Evolution*, vol. 98, no. February, pp. 41–51, 2016.
[132] C. G. Clark, F. Kaffashian, B. Tawari, J. J. Windsor, A. Twigg-Flesner, M. C. G. Davies-Morel, J. Blessmann, F. Ebert, B. Peschel, A. Le Van, C. J. Jackson, L. Macfarlane, and E. Tannich, "**New insights into the phylogeny of Entamoeba species provided by analysis of four new small-subunit rRNA genes,**" *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, no. 9, pp. 2235–2239, Sep. 2006.
[133] C. R. Stensvold, M. Lebbad, and C. G. Clark, "**Last of the human protists: The phylogeny and genetic diversity of Iodamoeba,**" *Molecular Biology and Evolution*, vol. 29, no. 1, pp. 39–42, Jan. 2012.
[134] G. Walker, A. G. B. Simpson, V. Edgcomb, M. L. Sogin, D. J. Patterson, G. B. Simpson, V. Edgcomb, M. L. Sogin, and D. J. Patterson, "**Ultrastructural identities of Mastigamoeba punctachora, Mastigamoeba simplex and Mastigella commutans and assessment of hypotheses of relatedness of the pelobionts (Protista),**" *European Journal of Protistology*, vol. 37, no. 1, pp. 25–49, Jan. 2001.
[135] E. Ska Zadrobílková, G. Walker, I. Cep Cka, E. Zadrobílková, G. Walker, and I. Čepička, "**Morphological and Molecular Evidence Support a Close Relationship Between the Free-living Archamoebae Mastigella and Pelomyxa,**" *Protist*, vol. 166, no. 1, pp. 14–41, Feb. 2015.

[136] T. Cavalier-Smith, E. E. Y. Chao, and B. Oates, "**Molecular phylogeny of Amoebozoa and the evolutionary significance of the unikont Phalansterium,**" *European Journal of Protistology*, vol. 40, no. 1, pp. 21–48, May 2004.
[137] E. Š. Ptáčková, A. Y. Kostygov, L. V Chistyakova, L. Falteisek, A. O. Frolov, D. J. Patterson, G. Walker, and I. Cepicka, "**Evolution of Archamoebae: Morphological and Molecular Evidence for Pelobionts Including Rhizomastix, Entamoeba, Iodamoeba, and Endolimax,**" *Protist*, vol. 164, no. 3, pp. 380–410, May 2013.
[138] A. Benjak, C. Sala, and R. C. Hartkoorn, "**Whole-transcriptome sequencing for high-resolution transcriptomic analysis in Mycobacterium tuberculosis,**" in *Mycobacteria Protocols: Third Edition*, Humana Press, New York, NY, 2015, pp. 17–30.
[139] I. K. M. Ali, C. G. Clark, and W. A. Petri, "**Molecular epidemiology of amebiasis,**" *Infection, Genetics and Evolution*, vol. 8, no. 5, pp. 698–707, Sep. 2008.
[140] C. Ximnez, P. Moran, L. Rojas, A. Valadez, and A. Gomez, "**Reassessment of the epidemiology of amebiasis: State of the art,**" *Infection, Genetics and Evolution*, vol. 9, no. 6, pp. 1023–1032, Dec. 2009.
[141] K. Watanabe and W. A. Petri Jr., "**Molecular biology research to benefit patients with Entamoeba histolytica infection,**" *Mol Microbiol*, vol. 98, no. 2, pp. 208–217, Oct. 2015.
[142] D. C. Savage, "**Microbial Ecology of the Gastrointestinal Tract,**" *Annual Review of Microbiology*, vol. 31, no. 1, pp. 107–133, Oct. 1977.
[143] H. Voigt, J. C. Olivo, P. Sansonetti, and N. Guillén, "**Myosin IB from Entamoeba histolytica is involved in phagocytosis of human erythrocytes.,**" *Journal of cell science*, vol. 112, pp. 1191–1201, 1999.
[144] J. A. Walsh, "**Problems in Recognition and Diagnosis of Amebiasis: Estimation of the Global Magnitude of Morbidity and Mortality,**" *Inpharma Weekly*, vol. NA;, no. 1214, p. 17, 1999.
[145] R. Lozano, M. Naghavi, K. Foreman, S. Lim, K. Shibuya, V. Aboyans, J. Abraham, T. Adair, R. Aggarwal, S. Y. Ahn, M. Alvarado, H. R. Anderson, L. M. Anderson, K. G. Andrews, C. Atkinson, L. M. Baddour, S. Barker-Collo, D. H. Bartels, M. L. Bell, E. J. Benjamin, D. Bennett, K. Bhalla, B. Bikbov, A. Bin Abdulhak, G. Birbeck, F. Blyth, I. Bolliger, S. Boufous, C. Bucello, M. Burch, P. Burney, J. Carapetis, H. Chen, D. Chou, S. S. Chugh, L. E. Coffeng, S. D. Colan, S. Colquhoun, K. E. Colson, J. Condon, M. D. Connor, L. T. Cooper, M. Corriere, M. Cortinovis, K. C. De Vaccaro, W. Couser, B. C. Cowie, M. H. Criqui, M. Cross, K. C. Dabhadkar, N. Dahodwala, D. De Leo, L. Degenhardt, A. Delossantos, J. Denenberg, D. C. Des Jarlais, S. D. Dharmaratne, E. R. Dorsey, T. Driscoll, H. Duber, B. Ebel, P. J. Erwin, P. Espindola, M. Ezzati, V. Feigin, A. D. Flaxman, M. H. Forouzanfar, F. G. R. Fowkes, R. Franklin, M. Fransen, M. K. Freeman, S. E. Gabriel, E. Gakidou, F. Gaspari, R. F. Gillum, D. Gonzalez-Medina, Y. A. Halasa, D. Haring, J. E. Harrison, R. Havmoeller, R. J. Hay, B. Hoen, P. J. Hotez, D. Hoy, K. H. Jacobsen, S. L. James, R. Jasrasaria, S. Jayaraman, N. Johns, G. Karthikeyan, N. Kassebaum, A. Keren, J. P. Khoo, L. M. Knowlton, O. Kobusingye, A. Koranteng, R. Krishnamurthi, M. Lipnick, S. E. Lipshultz, S. L. Ohno, J. Mabweijano, M. F. MacIntyre, L. Mallinger, L. March, G. B. Marks, R. Marks, A. Matsumori, R. Matzopoulos,

B. M. Mayosi, J. H. McAnulty, M. M. McDermott, J. McGrath, G. A. Mensah, T. R. Merriman, C. Michaud, M. Miller, T. R. Miller, C. Mock, A. O. Mocumbi, A. A. Mokdad, A. Moran, K. Mulholland, M. N. Nair, L. Naldi, K. M. V. Narayan, K. Nasseri, P. Norman, M. O'Donnell, S. B. Omer, K. Ortblad, R. Osborne, D. Ozgediz, B. Pahari, J. D. Pandian, A. P. Rivero, R. P. Padilla, F. Perez-Ruiz, N. Perico, D. Phillips, K. Pierce, C. A. Pope, E. Porrini, F. Pourmalek, M. Raju, D. Ranganathan, J. T. Rehm, D. B. Rein, G. Remuzzi, F. P. Rivara, T. Roberts, F. R. De León, L. C. Rosenfeld, L. Rushton, R. L. Sacco, J. A. Salomon, U. Sampson, E. Sanman, D. C. Schwebel, M. Segui-Gomez, D. S. Shepard, D. Singh, J. Singleton, K. Sliwa, E. Smith, A. Steer, J. A. Taylor, B. Thomas, I. M. Tleyjeh, J. A. Towbin, T. Truelsen, E. A. Undurraga, N. Venketasubramanian, L. Vijayakumar, T. Vos, G. R. Wagner, M. Wang, W. Wang, K. Watt, M. A. Weinstock, R. Weintraub, J. D. Wilkinson, A. D. Woolf, S. Wulf, P. H. Yeh, P. Yip, A. Zabetian, Z. J. Zheng, A. D. Lopez, and C. J. L. Murray, "**Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: A systematic analysis for the Global Burden of Disease Study 2010,**" *The Lancet*, vol. 380, no. 9859, pp. 2095–2128, Dec. 2012.

[146] "**WHO/PAHO/UNESCO report. A consultation with experts on amoebiasis. Mexico City, Mexico 28-29 January, 1997.,**" *Epidemiological bulletin*, vol. 18, no. 1, pp. 13–14, Mar. 1997.

[147] R. Haque, I. M. Ali, R. B. Sack, B. M. Farr, G. Ramakrishnan, and W. A. Petri, "**Amebiasis and mucosal IgA antibody against the Entamoeba histolytica adherence lectin in Bangladeshi children.,**" *The Journal of infectious diseases*, vol. 183, no. 12, pp. 1787–1793, 2001.

[148] D. Stark, S. J. Van Hal, G. Matthews, J. Harkness, and D. Marriott, "**Invasive amebiasis in men who have sex with men, Australia,**" *Emerging Infectious Diseases*, vol. 14, no. 7, pp. 1141–1143, Jul. 2008.

[149] C.-C. Hung, S.-Y. Chang, and D.-D. Ji, "**Entamoeba histolytica infection in men who have sex with men,**" *The Lancet Infectious Diseases*, vol. 12, no. 9, pp. 729–736, Sep. 2012.

[150] K. Watanabe, H. Gatanaga, A. Aleyla Escueta-de Cadiz, J. Tanuma, T. Nozaki, and S. Oka, "**Amebiasis in HIV-1-infected Japanese men: Clinical features and response to therapy,**" *PLoS Neglected Tropical Diseases*, vol. 5, no. 9, 2011.

[151] W. a Petri, R. Haque, and B. J. Mann, "**The bittersweet interface of parasite and host: lectin-carbohydrate interactions during human invasion by the parasite Entamoeba histolytica.,**" *Annual review of microbiology*, vol. 56, pp. 39–64, 2002.

[152] M. E. Lidell, D. M. Moncada, K. Chadee, and G. C. Hansson, "**Entamoeba histolytica cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel.,**" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 24, pp. 9298–9303, 2006.

[153] M. Lejeune, J. M. Rybicka, and K. Chadee, "**Recent discoveries in the pathogenesis and immune response toward Entamoeba histolytica.,**" *Future microbiology*, vol. 4, no. 1, pp. 105–118, Feb. 2009.

[154] S. M. Becker, K.-N. Cho, X. Guo, K. Fendig, M. N. Oosman, R. Whitehead, S. M. Cohn, and E. R. Houpt, "**Epithelial cell apoptosis facilitates**

Entamoeba histolytica infection in the gut.," *The American journal of pathology*, vol. 176, no. 3, pp. 1316–1322, 2010.
[155] K. S. Ralston, "**Taking a bite: Amoebic trogocytosis in Entamoeba histolytica and beyond,**" *Current Opinion in Microbiology*, vol. 28, pp. 26–35, 2015.
[156] K. S. Ralston, "**Chew on this: Amoebic trogocytosis and host cell killing by Entamoeba histolytica,**" *Trends in Parasitology*, vol. 31, no. 9. NIH Public Access, pp. 442–452, Sep-2015.
[157] S. L. Stanley, "**Amoebiasis,**" in *Lancet*, 2003, vol. 361, no. 9362, pp. 1025–1034.
[158] J. Santi-Rocca, M. C. Rigothier, and N. Guillén, "**Host-microbe interactions and defense mechanisms in the development of amoebic liver abscesses,**" *Clinical Microbiology Reviews*, vol. 22, no. 1. pp. 65–75, 2009.
[159] Z. Yang and J. Huang, "**De novo origin of new genes with introns in Plasmodium vivax,**" *FEBS Letters*, vol. 585, no. 4, pp. 641–644, Feb. 2011.
[160] D. Guerzoni and A. McLysaght, "**De novo origins of human genes,**" *PLoS Genetics*, vol. 7, no. 11. 2011.
[161] S. Ohno, **Evolution by Gene Duplication**. Springer-Verlag, 1970.
[162] M. Hurles, "**Gene duplication: The genomic trade in spare parts,**" *PLoS Biology*, vol. 2, no. 7. 2004.
[163] H. Kaessmann, "**Origins, evolution, and phenotypic impact of new genes,**" *Genome Research*, vol. 20, no. 10. Cold Spring Harbor Laboratory Press, pp. 1313–1326, Oct-2010.
[164] M. Lynch, "**The Evolutionary Fate and Consequences of Duplicate Genes,**" *Science*, vol. 290, no. 5494, pp. 1151–1155, 2000.
[165] D. A. Kleinjan, R. M. Bancewicz, P. Gautier, R. Dahm, H. B. Schonthaler, G. Damante, A. Seawright, A. M. Hever, P. L. Yeyati, V. van Heyningen, and P. Coutinho, "**Subfunctionalization of Duplicated Zebrafish pax6 Genes by cis-Regulatory Divergence,**" *PLoS Genetics*, 2008.
[166] C. Deng, C.-H. C. Cheng, H. Ye, X. He, and L. Chen, "**Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict,**" *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 50, pp. 21593–21598, 2010.
[167] J. S. Taylor and J. Raes, "**Duplication and Divergence: The Evolution of New Genes and Old Ideas,**" *Annual Review of Genetics*, vol. 38, no. 1, pp. 615–643, 2004.
[168] J. Zhang, "**Evolution by gene duplication: An update,**" *Trends in Ecology and Evolution*, vol. 18, no. 6. Elsevier, pp. 292–298, Jun-2003.
[169] J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler, "**Segmental duplications: organization and impact within the current human genome project assembly.,**" *Genome research*, vol. 11, no. 6, pp. 1005–1017, 2001.
[170] F. A. Kondrashov, "**Gene duplication as a mechanism of genomic adaptation to a changing environment,**" *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1749, pp. 5048–5057, 2012.
[171] J. A. Bailey and E. E. Eichler, "**Primate segmental duplications: crucibles of evolution, diversity and disease,**" *Nature Reviews Genetics*, vol. 7, no. 7, pp. 552–564, Jul. 2006.

[172] D. D. Wu, D. M. Irwin, and Y. P. Zhang, "**De novo origin of human protein-coding genes,**" *PLoS Genetics*, vol. 7, no. 11, 2011.
[173] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson, "**The B73 Maize Genome: Complexity, Diversity, and Dynamics,**" *Science*, vol. 326, no. 5956, pp. 1112–1115, Nov. 2009.
[174] Y. Van de Peer, "**Tetraodon genome confirms Takifugu findings: most fish are ancient polyploids.,**" *Genome biology*, vol. 5, no. 12, p. 250, 2004.
[175] C. Song, S. J. Liu, J. Xiao, W. G. He, Y. Zhou, Q. B. Qin, C. Zhang, and Y. Liu, "**Polyploid organisms,**" *Science China Life Sciences*, vol. 55, no. 4. pp. 301–311, 2012.
[176] P. R. Marri, W. Hao, and G. B. Golding, "**Gene gain and gene loss in Streptococcus: Is it driven by habitat?,**" *Molecular Biology and Evolution*, vol. 23, no. 12, pp. 2379–2391, 2006.
[177] T. Kaplan, "**The Role of Horizontal Gene Transfer in Antibiotic Resistance,**" *Eukaryon*, vol. 10, no. March, pp. 80–81, 2014.
[178] C. M. Diez, K. Roessler, and B. S. Gaut, "**Epigenetics and plant genome evolution,**" *Current Opinion in Plant Biology*, vol. 18, no. 1. pp. 1–8, 2014.
[179] E. Jablonka and M. J. Lamb, "**The changing concept of epigenetics.,**" *Annals of the New York Academy of Sciences*, vol. 981, pp. 82–96, Dec. 2002.
[180] J. A. Law and S. E. Jacobsen, "**Establishing, maintaining and modifying DNA methylation patterns in plants and animals,**" *Nature Reviews Genetics*, vol. 11, no. 3, pp. 204–220, Mar. 2010.
[181] R. H. Dowen, M. Pelizzola, R. J. Schmitz, R. Lister, J. M. Dowen, J. R. Nery, J. E. Dixon, and J. R. Ecker, "**Widespread dynamic DNA methylation in**

response to biotic stress," *Proceedings of the National Academy of Sciences*, vol. 109, no. 32, pp. E2183--E2191, Aug. 2012.
[182] H. Stroud, M. V. C. Greenberg, S. Feng, Y. V Bernatavichute, and S. E. Jacobsen, "**Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome,**" *Cell*, vol. 152, no. 1–2, pp. 352–364, 2013.
[183] V. Klimeš, E. Gentekaki, A. J. Roger, and M. Eliáš, "**A large number of nuclear genes in the human parasite blastocystis require mRNA polyadenylation to create functional termination codons,**" *Genome Biology and Evolution*, vol. 6, no. 8, pp. 1956–1961, Jul. 2014.
[184] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden, and I. G. Young, "**Sequence and organization of the human mitochondrial genome,**" *Nature*, vol. 290, no. 5806, pp. 457–465, Apr. 1981.
[185] J. H. Chang and L. Tong, "**Mitochondrial poly(A) polymerase and polyadenylation,**" *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, vol. 1819, no. 9–10. pp. 992–997, Sep-2012.
[186] W. M. Fitch, "**Distinguishing Homologous from Analogous Proteins,**" *Systematic Zoology*, vol. 19, no. 2, p. 99, Jun. 1970.
[187] A. M. Altenhoff, R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz, "**Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs,**" *PLoS Computational Biology*, vol. 8, no. 5, p. 8, 2012.
[188] E. C. Dimmer, R. P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M. J. Martin, B. Bely, P. Browne, W. M. Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M. C. Blatter, E. Boutet, S. Braconi-Quintaje, L. Breuza, A. Bridge, E. Coudert, A. Estreicher, L. Famiglietti, S. Ferro-Rojas, M. Feuermann, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, J. James, S. Jimenez, F. Jungo, G. Keller, P. Lemercier, D. Lieberherr, P. Masson, M. Moinat, I. Pedruzzi, S. Poux, C. Rivoire, B. Roechert, M. Schneider, A. Stutz, S. Sundaram, M. Tognolli, L. Bougueleret, G. Argoud-Puy, I. Cusin, P. Duek-Roggli, I. Xenarios, and R. Apweiler, "**The UniProt-GO Annotation database in 2011,**" *Nucleic Acids Research*, vol. 40, no. D1, pp. D565--70, Jan. 2012.
[189] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, "**Large-scale gene function analysis with the PANTHER classification system,**" *Nature Protocols*, vol. 8, no. 8, pp. 1551–1566, Jul. 2013.
[190] L. Arvestad, A.-C. C. Berglund, J. Lagergren, and B. Sennblad, "**Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution,**" *Proceedings of the eighth annual international conference on Computational molecular biology - RECOMB '04*, pp. 326–335, 2004.
[191] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney, "**EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates,**" *Genome Research*, vol. 19, no. 2, pp. 327–335, 2009.
[192] O. Akerborg, B. Sennblad, L. Arvestad, and J. Lagergren, "**Simultaneous**

Bayesian gene tree reconstruction and reconciliation analysis.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5714–5719, 2009.
[193] M. D. Rasmussen and M. Kellis, "**Unified modeling of gene duplication, loss, and coalescence using a locus tree,**" *Genome Research*, vol. 22, no. 4, pp. 755–765, 2012.
[194] B. Boussau, G. J. Szollosi, L. Duret, M. Gouy, E. Tannier, and V. Daubin, "**Genome-scale coestimation of species and gene trees,**" *Genome Research*, vol. 23, no. 2, pp. 323–330, 2013.
[195] T. J. C. Anderson, J. Patel, and M. T. Ferdig, "**Gene copy number and malaria biology,**" *Trends in Parasitology*, vol. 25, no. 7. pp. 336–343, 2009.
[196] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "**Origins and functional impact of copy number variation in the human genome,**" *Nature*, vol. 464, no. 7289, pp. 704–712, 2010.
[197] J. H. Xu, J. L. Bennetzen, and J. Messing, "**Dynamic gene copy number variation in collinear regions of grass genomes,**" *Molecular Biology and Evolution*, vol. 29, no. 2, pp. 861–871, 2012.
[198] J. Sun, H. Jiang, R. Flores, and J. Wen, "**Gene duplication in the genome of parasitic Giardia lamblia,**" *BMC Evolutionary Biology*, vol. 10, p. 49, 2010.
[199] R. D. Emes and Z. Yang, "**Duplicated paralogous genes subject to positive selection in the genome of Trypanosoma brucei,**" *PLoS ONE*, vol. 3, no. 5, 2008.
[200] S. Nair, D. Nash, D. Sudimack, A. Jaidee, M. Barends, A. C. Uhlemann, S. Krishna, F. Nosten, and T. J. C. Anderson, "**Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites,**" *Molecular Biology and Evolution*, vol. 24, no. 2, pp. 562–573, 2007.
[201] J. Wasmuth, J. Daub, J. M. Peregrín-Alvarez, C. A. M. Finney, and J. Parkinson, "**The origins of apicomplexan sequence innovation,**" *Genome Research*, vol. 19, no. 7, pp. 1202–1213, 2009.
[202] M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda, "**Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences,**" *Systematic Biology*, vol. 28, no. 2, pp. 132–163, Jun. 1979.
[203] R. D. M. Page, "**Maps Between Trees and Cladistic Analysis of Historical Associations among Genes,Organisms, and Areas,**" *Systematic Biology*, vol. 43, no. 1, pp. 58–77, Mar. 1994.
[204] D. Durand, B. V Halldórsson, and B. Vernot, "**A hybrid micro-macroevolutionary approach to gene tree reconstruction.,**" *Journal of Computational Biology*, vol. 13, no. 2, pp. 320–335, Mar. 2006.
[205] M. D. Rasmussen and M. Kellis, "**A Bayesian approach for fast and accurate gene tree reconstruction,**" *Molecular Biology and Evolution*,

vol. 28, no. 1, pp. 273–290, Jan. 2011.
[206] A. P. Jackson, T. D. Otto, M. Aslett, S. D. Armstrong, F. Bringaud, A. Schlacht, C. Hartley, M. Sanders, J. M. Wastling, J. B. Dacks, A. Acosta-Serrano, M. C. Field, M. L. Ginger, and M. Berriman, "**Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism,**" *Current Biology*, vol. 26, no. 2, pp. 161–172, 2016.
[207] Y. Yang and D. Luo, "**The origin of parasitism gene in nematodes: Evolutionary analysis through the construction of domain trees,**" *Evolutionary Bioinformatics*, vol. 2013, no. 9, pp. 453–466, 2013.
[208] M. Blaxter and G. Koutsovoulos, "**The evolution of parasitism in Nematoda,**" *Parasitology*, vol. 142, no. S1, pp. S26--S39, 2015.
[209] P. Brumer, H. Rabitz, S. Y. Lee, R. A. Mathies, R. W. Schoenlein, L. A. Peteanu, R. A. Mathies, C. V Shank, F. Gai, P. A. Anfinrud, T. Saito, H. Ohtani, K. Heyne, R. Diller, E. Tajkhorshid, K. Schulten, V. S. Batista, K. Ohara, M. Abe, K. Nakagami, Y. Fujimura, G. Krampert, P. Niklaus, P. Nuernberger, G. Gerber, P. Brumer, W. Wohlleben, R. J. Cogdell, D. Zeidler, M. Motzkus, D. Oesterhelt, K. C. Hasson, J. C. Mcdonald, P. A. Anfinrud, A. M. Nagy, R. J. D. Miller, V. V Savransky, N. V Tkachenko, V. I. Chukharev, R. Diller, M. Stockburger, E. S. Imasheva, R. Govindjee, T. G. Ebrey, P. C. Becker, C. V Shank, R. A. Harris, R. A. Mathies, N. Friedman, M. Ottolenghi, M. Sheves, S. Ruhman, M. Yamashita, A. Suguro, H. Shigekawa, B. M. Tyler, S. Tripathy, X. Zhang, P. Dehal, R. H. Y. Jiang, A. Aerts, F. D. Arredondo, L. Baxter, D. Bensasson, J. L. Beynon, J. Chapman, C. M. B. Damasceno, A. E. Dorrance, D. Dou, A. W. Dickerman, I. L. Dubchak, M. Garbelotto, M. Gijzen, S. G. Gordon, F. Govers, N. J. Grunwald, W. Huang, K. L. Ivors, R. W. Jones, S. Kamoun, K. Krampis, K. H. Lamour, M.-K. Lee, W. H. McDonald, M. Medina, H. J. G. Meijer, E. K. Nordberg, D. J. Maclean, M. D. Ospina-Giraldo, P. F. Morris, V. Phuntumart, N. H. Putnam, S. Rash, J. K. C. Rose, Y. Sakihama, A. a Salamov, A. Savidor, C. F. Scheuring, B. M. Smith, B. W. S. Sobral, A. Terry, T. a Torto-Alalibo, J. Win, Z. Xu, H. Zhang, I. V Grigoriev, D. S. Rokhsar, and J. L. Boore, "**Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis.,**" *Science (New York, N.Y.)*, vol. 313, no. 5791, pp. 1261–1266, Sep. 2006.
[210] B. J. Haas, S. Kamoun, M. C. Zody, R. H. Y. Jiang, R. E. Handsaker, L. M. Cano, M. Grabherr, C. D. Kodira, S. Raffaele, T. Torto-Alalibo, T. O. Bozkurt, A. M. V Ah-Fong, L. Alvarado, V. L. Anderson, M. R. Armstrong, A. Avrova, L. Baxter, J. Beynon, P. C. Boevink, S. R. Bollmann, J. I. B. Bos, V. Bulone, G. Cai, C. Cakir, J. C. Carrington, M. Chawner, L. Conti, S. Costanzo, R. Ewan, N. Fahlgren, M. a Fischbach, J. Fugelstad, E. M. Gilroy, S. Gnerre, P. J. Green, L. J. Grenville-Briggs, J. Griffith, N. J. Grünwald, K. Horn, N. R. Horner, C.-H. Hu, E. Huitema, D.-H. Jeong, A. M. E. Jones, J. D. G. Jones, R. W. Jones, E. K. Karlsson, S. G. Kunjeti, K. Lamour, Z. Liu, L. Ma, D. Maclean, M. C. Chibucos, H. McDonald, J. McWalters, H. J. G. Meijer, W. Morgan, P. F. Morris, C. a Munro, K. O'Neill, M. Ospina-Giraldo, A. Pinzón, L. Pritchard, B. Ramsahoye, Q. Ren, S. Restrepo, S. Roy, A. Sadanandom, A. Savidor, S. Schornack, D. C. Schwartz, U. D. Schumann, B. Schwessinger, L. Seyer, T. Sharpe, C. Silvar, J. Song, D. J. Studholme, S. Sykes, M. Thines, P. J. I. van de Vondervoort, V. Phuntumart, S. Wawra, R. Weide, J. Win, C. Young, S. Zhou, W. Fry, B. C. Meyers, P. van West, J. Ristaino, F. Govers, P. R. J. Birch, S. C.

Whisson, H. S. Judelson, and C. Nusbaum, "**Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans.,**" *Nature*, vol. 461, no. 7262, pp. 393–398, Sep. 2009.
[211] J. D. Debarry and J. C. Kissinger, "**Jumbled genomes: Missing apicomplexan synteny,**" *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2855–2871, Oct. 2011.
[212] A. P. Jackson, "**The evolution of parasite genomes and the origins of parasitism,**" *Parasitology*, vol. 142, no. S1, pp. S1--S5, 2015.
[213] C. H. Zierdt, "**Cytochrome Free Mitochondria of an Anaerobic Protozoan Blastocystis hominis,**" *The Journal of Protozoology*, vol. 33, no. 1, pp. 67–69, Feb. 1986.
[214] S. Nakjang, T. A. Williams, E. Heinz, A. K. Watson, P. G. Foster, K. M. Sendra, S. E. Heaps, R. P. Hirt, and T. Martin Embley, "**Reduction and expansion in microsporidian genome evolution: new insights from comparative genomics.,**" *Genome biology and evolution*, vol. 5, no. 12, pp. 2285–2303, 2013.
[215] P. J. Keeling and C. H. Slamovits, "**Causes and effects of nuclear genome reduction,**" *Current Opinion in Genetics and Development*, vol. 15, no. 6. pp. 601–608, Dec-2005.
[216] A. P. Jackson, "**Genome evolution in trypanosomatid parasites.,**" *Parasitology*, vol. 142, no. S1, pp. 1–17, Feb. 2014.
[217] J. Jerlström-Hultqvist, O. Franzén, J. Ankarklev, F. Xu, E. Nohýnková, J. O. Andersson, S. G. Svärd, and B. Andersson, "**Genome analysis and comparative genomics of a Giardia intestinalis assemblage E isolate.,**" *BMC genomics*, vol. 11, p. 543, Oct. 2010.
[218] J. C. Kissinger and J. DeBarry, "**Genome cartography: Charting the apicomplexan genome,**" *Trends in Parasitology*, vol. 27, no. 8. pp. 345–354, 2011.
[219] E. Ghedin, S. Wang, D. Spiro, E. Caler, Q. Zhao, J. Crabtree, J. E. Allen, A. L. Delcher, D. B. Guiliano, D. Miranda-Saavedra, S. V Angiuoli, T. Creasy, P. Amedeo, B. Haas, N. M. El-Sayed, J. R. Wortman, T. Feldblyum, L. Tallon, M. Schatz, M. Shumway, H. Koo, S. L. Salzberg, S. Schobel, M. Pertea, M. Pop, O. White, G. J. Barton, C. K. S. Carlow, M. J. Crawford, J. Daub, M. W. Dimmic, C. F. Estes, J. M. Foster, M. Ganatra, W. F. Gregory, N. M. Johnson, J. Jin, R. Komuniecki, I. Korf, S. Kumar, S. Laney, B.-W. Li, W. Li, T. H. Lindblom, S. Lustigman, D. Ma, C. V Maina, D. M. a Martin, J. P. McCarter, L. McReynolds, M. Mitreva, T. B. Nutman, J. Parkinson, J. M. Peregrín-Alvarez, C. Poole, Q. Ren, L. Saunders, A. E. Sluder, K. Smith, M. Stanke, T. R. Unnasch, J. Ware, A. D. Wei, G. Weil, D. J. Williams, Y. Zhang, S. a Williams, C. Fraser-Liggett, B. Slatko, M. L. Blaxter, and A. L. Scott, "**Draft genome of the filarial nematode parasite Brugia malayi.,**" *Science (New York, N.Y.)*, vol. 317, no. 5845, pp. 1756–1760, 2007.
[220] M. Mitreva, D. P. Jasmer, D. S. Zarlenga, Z. Wang, S. Abubucker, J. Martin, C. M. Taylor, Y. Yin, L. Fulton, P. Minx, S.-P. Yang, W. C. Warren, R. S. Fulton, V. Bhonagiri, X. Zhang, K. Hallsworth-Pepin, S. W. Clifton, J. P. McCarter, J. Appleton, E. R. Mardis, and R. K. Wilson, "**The draft genome of the parasitic nematode Trichinella spiralis,**" *Nature Genetics*, vol. 43, no. 3, pp. 228–235, 2011.
[221] M. Berriman, B. J. Haas, P. T. LoVerde, R. A. Wilson, G. P. Dillon, G. C.

Cerqueira, S. T. Mashiyama, B. Al-Lazikani, L. F. Andrade, P. D. Ashton, M. A. Aslett, D. C. Bartholomeu, G. Blandin, C. R. Caffrey, A. Coghlan, R. Coulson, T. A. Day, A. Delcher, R. DeMarco, A. Djikeng, T. Eyre, J. A. Gamble, E. Ghedin, Y. Gu, C. Hertz-Fowler, H. Hirai, Y. Hirai, R. Houston, A. Ivens, D. A. Johnston, D. Lacerda, C. D. Macedo, P. McVeigh, Z. Ning, G. Oliveira, J. P. Overington, J. Parkhill, M. Pertea, R. J. Pierce, A. V Protasio, M. A. Quail, M.-A. Rajandream, J. Rogers, M. Sajid, S. L. Salzberg, M. Stanke, A. R. Tivey, O. White, D. L. Williams, J. Wortman, W. Wu, M. Zamanian, A. Zerlotini, C. M. Fraser-Liggett, B. G. Barrell, N. M. El-Sayed, J. Rogers, and M.-A. Rajandream, "**The genome of the blood fluke Schistosoma mansoni,**" *Nature*, vol. 460, no. 7253, pp. 352–358, Jul. 2009.

[222] M. J. Gardner, 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, "**Genome sequence of the human malaria parasite Plasmodium falciparum,**" *Nature*, vol. 419, no. 6906, pp. 498–511, Oct. 2002.

[223] P. J. Keeling, "**Reduction and compaction in the genome of the apicomplexan parasite Cryptosporidium parvum,**" *Developmental Cell*, vol. 6, no. 5, pp. 614–616, May 2004.

[224] A. H. Fairlamb, "**Novel biochemical pathways in parasitic protozoa.,**" *Parasitology*, vol. 99 Suppl, no. Supplement 1, pp. S93--112, 1989.

[225] J. J. Marr, R. L. Berens, and D. J. Nelson, "**Purine metabolism in Leishmania donovani and Leishmania braziliensis.,**" *Biochimica et biophysica acta*, vol. 544, no. 2, pp. 360–371, Dec. 1978.

[226] W. E. Gutteridge and M. Gaborak, "**A re-examination of purine and pyrimidine synthesis in the three main forms of Trypanosoma cruzi,**" *International Journal of Biochemistry*, vol. 10, no. 5, pp. 415–422, 1979.

[227] H. F. Hassan and G. H. Coombs, "**Purine and pyrimidine metabolism in parasitic protozoa,**" *FEMS Microbiology Letters*, vol. 54, no. 1, pp. 47–83, Feb. 1988.

[228] H. P. De Koning, D. J. Bridges, and R. J. S. Burchmore, "**Purine and pyrimidine transport in pathogenic protozoa: From biology to therapy,**" *FEMS Microbiology Reviews*, vol. 29, no. 5. pp. 987–1020, Nov-2005.

[229] N. P. Robinson, N. Burman, S. E. Melville, and J. D. Barry, "**Predominance of duplicative VSG gene conversion in antigenic variation in African trypanosomes.,**" *Molecular and cellular biology*, vol. 19, no. 9, pp. 5839–5846, Sep. 1999.

[230] N. M. El-Sayed, P. J. Myler, G. Blandin, M. Berriman, J. Crabtree, G. Aggarwal, E. Caler, H. Renauld, E. A. Worthey, C. Hertz-Fowler, E. Ghedin, C. Peacock, D. C. Bartholomeu, B. J. Haas, A.-N. Tran, J. R. Wortman, U. C. M. Alsmark, S. Angiuoli, A. Anupama, J. Badger, F. Bringaud, E. Cadag, J. M. Carlton, G. C. Cerqueira, T. Creasy, A. L. Delcher, A. Djikeng, T. M. Embley,

C. Hauser, A. C. Ivens, S. K. Kummerfeld, J. B. Pereira-Leal, D. Nilsson, J. Peterson, S. L. Salzberg, J. Shallom, J. C. Silva, J. Sundaram, S. Westenberger, O. White, S. E. Melville, J. E. Donelson, B. Andersson, K. D. Stuart, and N. Hall, "**Comparative genomics of trypanosomatid parasitic protozoa.,**" *Science (New York, N.Y.)*, vol. 309, no. 5733, pp. 404–409, Jul. 2005.

[231] A. P. Jackson, A. Berry, M. Aslett, H. C. Allinson, P. Burton, J. Vavrova-Anderson, R. Brown, H. Browne, N. Corton, H. Hauser, J. Gamble, R. Gilderthorp, L. Marcello, J. McQuillan, T. D. Otto, M. A. Quail, M. J. Sanders, A. van Tonder, M. Ginger, M. C. Field, J. D. Barry, C. Hertz-Fowler, and M. Berriman, "**Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species,**" *Pnas*, vol. 109, no. 9, pp. 3416–3421, Feb. 2012.

[232] J. M. Carlton, J. H. Adams, J. C. Silva, S. L. Bidwell, H. Lorenzi, E. Caler, J. Crabtree, S. V Angiuoli, E. F. Merino, P. Amedeo, Q. Cheng, R. M. R. Coulson, B. S. Crabb, H. A. Del Portillo, K. Essien, T. V Feldblyum, C. Fernandez-becerra, P. R. Gilson, A. H. Gueye, X. Guo, S. Kang'a, T. W. A. Kooij, M. Korsinczky, E. V.-S. V.-S. Meyer, V. Nene, I. Paulsen, O. White, S. A. Ralph, Q. Ren, T. J. Sargeant, S. L. Salzberg, C. J. Stoeckert, S. A. Sullivan, M. M. Yamamoto, S. L. Hoffman, J. R. Wortman, M. J. Gardner, M. R. Galinski, J. W. Barnwell, C. M. Fraser-Liggett, A. Hernando, K. Essien, T. V Feldblyum, C. Fernandez-becerra, E. V.-S. V.-S. Meyer, V. Nene, I. Paulsen, O. White, S. A. Sullivan, M. M. Yamamoto, S. L. Hoffman, S. Kang'a, T. W. A. Kooij, M. Korsinczky, E. V.-S. V.-S. Meyer, V. Nene, I. Paulsen, O. White, S. A. Ralph, Q. Ren, T. J. Sargeant, S. L. Salzberg, C. J. Stoeckert, S. A. Sullivan, M. M. Yamamoto, S. L. Hoffman, J. R. Wortman, M. J. Gardner, M. R. Galinski, J. W. Barnwell, and C. M. Fraser-Liggett, "**Comparative genomics of the neglected human malaria parasite Plasmodium vivax,**" *Nature*, vol. 455, no. 7214, pp. 757–763, Oct. 2008.

[233] B. N. Adhikari, J. P. Hamilton, M. M. Zerillo, N. Tisserat, C. A. Lévesque, and C. R. Buell, "**Comparative Genomics Reveals Insight into Virulence Strategies of Plant Pathogenic Oomycetes,**" *PLoS ONE*, vol. 8, no. 10, p. e75072, Oct. 2013.

[234] A. Cortés, J. Sotillo, C. Muñoz-Antolí, J. Molina-Durán, J. G. Esteban, R. Toledo, A. Corté S, J. Sotillo, C. Muñoz-Antolí, J. Molina-Durán, J. G. Esteban, and R. Toledo, "**Antibody trapping: A novel mechanism of parasite immune evasion by the trematode Echinostoma caproni,**" *PLoS Neglected Tropical Diseases*, vol. 11, no. 7, p. e0005773, Jul. 2017.

[235] C. Verwaerde, C. Auriault, J. L. Neyrink, and A. Capron, "**Properties of Serine Proteases of Schistosoma mansoni Schistosomula Involved in the Regulation of IgE Synthesis,**" *Scandinavian Journal of Immunology*, vol. 27, no. 1, pp. 17–24, Jan. 1988.

[236] R. Campos-Rodriguezp and A. Jarillo-Luna, "**The pathogenicity of Entamoeba histolytica is related to the capacity of evading innate immunity,**" *Parasite Immunology*, vol. 27, no. 1–2. pp. 1–8, Jan-2005.

[237] M. J. Blackman, "**Proteases in host cell invasion by the malaria parasite,**" *Cellular Microbiology*, vol. 6, no. 10. pp. 893–903, Oct-2004.

[238] L. D. Sibley, "**The roles of intramembrane proteases in protozoan parasites,**" *Biochimica et Biophysica Acta - Biomembranes*, vol. 1828, no.

12. pp. 2908–2915, 2013.
[239] R. C. Hardison, "**Comparative genomics,**" *PLoS Biology*, vol. 1, no. 2. Public Library of Science, p. E58, Nov-2003.
[240] J. Felsenstein, "**Phylogenies from Molecular Sequences: Inference and Reliability,**" *Annual Review of Genetics*, vol. 22, no. 1, pp. 521–565, Dec. 1988.
[241] J. Kim, "**Improving the accuracy of phylogenetic estimation by combining different methods,**" *Systematic Biology*, vol. 42, no. 3, pp. 331–340, Sep. 1993.
[242] J. J. Wiens, M. R. Servedio, S. Biology, and N. Jun, "**Accuracy of Phylogenetic Analysis Including and Excluding Polymorphic Characters,**" *Systematic Biology*, vol. 46, no. 2, pp. 332–345, Jun. 1997.
[243] S. Poe, "**Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods.,**" *Systematic biology*, vol. 52, no. 3, pp. 423–428, Jun. 2003.
[244] B. Rannala, J. P. Huelsenbeck, Z. Yang, and R. Nielsen, "**Taxon Sampling and the Accuracy of Large Phylogenies,**" *Syst. Biol*, vol. 47, no. 4, pp. 702–710, Dec. 1998.
[245] S. Guindon, O. Gascuel, and B. Rannala, "**A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood,**" *Systematic Biology*, vol. 52, no. 5, pp. 696–704, Oct. 2003.
[246] R. Desper and O. Gascuel, "**Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle,**" *Journal of Computational Biology*, vol. 9, no. 5, pp. 687–705, Oct. 2002.
[247] J. H. Degnan and N. A. Rosenberg, "**Gene tree discordance, phylogenetic inference and the multispecies coalescent,**" *Trends in Ecology and Evolution*, vol. 24, no. 6. pp. 332–340, Jun-2009.
[248] A. R. Nabhan and I. N. Sarkar, "**The impact of taxon sampling on phylogenetic inference: A review of two decades of controversy,**" *Briefings in Bioinformatics*, vol. 13, no. 1. Oxford University Press, pp. 122–134, Jan-2012.
[249] J. E. de la Torre-Barcena, S. O. Kolokotronis, E. K. Lee, D. W. Stevenson, E. D. Brenner, M. S. Katari, G. M. Coruzzi, and R. DeSalle, "**The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data,**" *PLoS ONE*, vol. 4, no. 6, 2009.
[250] Y. H. Woo, H. Ansari, T. D. Otto, C. M. Klinger, M. Kolisko, J. Michálek, A. Saxena, D. Shanmugam, A. Tayyrov, A. Veluchamy, S. Ali, A. Bernal, J. del Campo, J. Cihlář, P. Flegontov, S. G. Gornik, E. Hajdušková, A. Horák, J. Janouškovec, N. J. Katris, F. D. Mast, D. Miranda-Saavedra, T. Mourier, R. Naeem, M. Nair, A. K. Panigrahi, N. D. Rawlings, E. Padron-Regalado, A. Ramaprasad, N. Samad, A. Tomčala, J. Wilkes, D. E. Neafsey, C. Doerig, C. Bowler, P. J. Keeling, D. S. Roos, J. B. Dacks, T. J. Templeton, R. F. Waller, J. Lukeš, M. Oborník, and A. Pain, "**Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites,**" *eLife*, vol. 4, no. JUNE, 2015.
[251] J. Ankarklev, O. Franzén, D. Peirasmaki, J. Jerlström-Hultqvist, M. Lebbad, J. Andersson, B. Andersson, and S. G. Svärd, "**Comparative genomic analyses of freshly isolated Giardia intestinalis assemblage A**

isolates," *BMC Genomics*, vol. 16, no. 1, p. 697, Sep. 2015.
[252] H. Cai, Z. Zhou, J. Gu, Y. Wang, H. Cai, Z. Zhou, J. Gu, and Y. Wang, "**Comparative Genomics and Systems Biology of Malaria Parasites Plasmodium,**" *Current bioinformatics*, vol. 7, no. 4, pp. 1–22, Dec. 2013.
[253] M. Sistrom, B. Evans, R. Bjornson, W. Gibson, O. Balmer, and P. Ma, "**ComparativeGenomicsRevealsMultipleGeneticBackgrounds of Human Pathogenicity in the Trypanosoma brucei Complex,**" *Genome biology and evolution*, vol. 6, no. 10, pp. 2811–2819, Oct. 2014.
[254] D. A. Tschoeke, G. L. Nunes, R. Jardim, J. Lima, A. S. R. Dumaresq, M. R. Gomes, L. de M. Pereira, D. R. Loureiro, P. H. Stoco, H. L. de Matos Guedes, A. B. de Miranda, J. Ruiz, A. Pitaluga, F. P. Silva, C. M. Probst, N. J. Dickens, J. C. Mottram, E. C. Grisard, and A. M. R. Dávila, "**The comparative genomics and phylogenomics of Leishmania amazonensis parasite,**" *Evolutionary Bioinformatics*, vol. 10, pp. 131–153, 2014.
[255] S. M. Teixeira, R. M. C. de Paiva, M. M. Kangussu-Marcolino, and W. D. DaRocha, "**Trypanosomatid comparative genomics: Contributions to the study of parasite biology and different parasitic diseases,**" *Genetics and Molecular Biology*, vol. 35, no. 1. pp. 1–17, 2012.
[256] A. J. Reid, S. J. Vermont, J. A. Cotton, D. Harris, G. A. Hill-Cawthorne, S. Könen-Waisman, S. M. Latham, T. Mourier, R. Norton, M. A. Quail, M. Sanders, D. Shanmugam, A. Sohal, J. D. Wasmuth, B. Brunk, M. E. Grigg, J. C. Howard, J. Parkinson, D. S. Roos, A. J. Trees, M. Berriman, A. Pain, and J. M. Wastling, "**Comparative genomics of the apicomplexan parasites Toxoplasma gondii and neospora caninum: Coccidia differing in host range and transmission strategy,**" *PLoS Pathogens*, vol. 8, no. 3, 2012.
[257] G. Pan, J. Xu, T. Li, Q. Xia, S.-L. Liu, G. Zhang, S. Li, C. Li, H. Liu, L. Yang, T. Liu, X. Zhang, Z. Wu, W. Fan, X. Dang, H. Xiang, M. Tao, Y. Li, J. Hu, Z. Li, L. Lin, J. Luo, L. Geng, L. Wang, M. Long, Y. Wan, N. He, Z. Zhang, C. Lu, P. J. Keeling, J. Wang, Z. Xiang, and Z. Zhou, "**Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation,**" *BMC Genomics*, vol. 14, no. 1, p. 186, 2013.
[258] A. J. Walker, M. Ressurreição, and R. Rothermel, "**Exploring the function of protein kinases in schistosomes: Perspectives from the laboratory and from comparative genomics,**" *Frontiers in Genetics*, vol. 5, no. JUL, p. 229, 2014.
[259] J. M. Feng, J. Sun, D. D. Xin, and J. F. Wen, "**Comparative analysis of the 5S rRNA and its associated proteins reveals unique primitive rather than parasitic features in Giardia lamblia,**" *PLoS ONE*, vol. 7, no. 6, 2012.
[260] O. H. Cissé, M. Pagni, and P. M. Hauser, "**Comparative genomics suggests that the human pathogenic fungus pneumocystis jirovecii acquired obligate biotrophy through gene loss,**" *Genome Biology and Evolution*, vol. 6, no. 8, pp. 1938–1948, Jul. 2014.
[261] R. S. Coyne, L. Hannick, D. Shanmugam, J. B. Hostetler, D. Brami, V. S. Joardar, J. Johnson, D. Radune, I. Singh, J. H. Badger, U. Kumar, M. Saier, Y. Wang, H. Cai, J. Gu, M. W. Mather, A. B. Vaidya, D. E. Wilkes, V. Rajagopalan, D. J. Asai, C. G. Pearson, R. C. Findly, H. W. Dickerson, M. Wu, C. Martens, Y. Van de Peer, D. S. Roos, D. M. Cassidy-Hanley, and T. G. Clark,

"**Comparative genomics of the pathogenic ciliate Ichthyophthirius multifiliis, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control,**" *Genome Biology*, vol. 12, no. 10, p. R100, Oct. 2011.

[262] G. E. Moore, "**Cramming more components onto integrated circuits,**" *Electronics*, vol. 38, no. 8, 1965.

[263] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law, "**Comparison of next-generation sequencing systems,**" *Journal of Biomedicine and Biotechnology*, vol. 2012, no. August 2014, 2012.

[264] F. S. Collins, M. Morgan, and A. Patrinos, "**The Human Genome Project: lessons from large-scale biology.,**" *Science (New York, N.Y.)*, vol. 300, no. 5617, pp. 286–290, Apr. 2003.

[265] R. Kumar, Y. Ichihashi, S. Kimura, D. H. Chitwood, L. R. Headland, J. Peng, J. N. Maloof, and N. R. Sinha, "**A High-Throughput Method for Illumina RNA-Seq Library Preparation,**" *Frontiers in Plant Science*, vol. 3, no. August, pp. 1–10, 2012.

[266] E. R. Mardis, "**The impact of next-generation sequencing technology on genetics,**" *Trends in Genetics*, vol. 24, no. 3. pp. 133–141, 2008.

[267] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E. Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S.

VandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "**Accurate whole human genome sequencing using reversible terminator chemistry,**" *Nature*, vol. 456, no. 7218, pp. 53–59, Nov. 2008.

[268] M. L. Metzker, "**Sequencing technologies — the next generation,**" *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[269] N. J. Loman, R. V Misra, T. J. Dallman, C. Constantinidou, S. E. Gharbia, J. Wain, and M. J. Pallen, "**Performance comparison of benchtop high-throughput sequencing platforms,**" *Nature Biotechnology*, vol. 30, no. 5, pp. 434–439, 2012.

[270] H. Y. K. Lam, M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder, "**Performance comparison of whole-genome sequencing platforms,**" *Nature Biotechnology*, vol. 30, no. 1, pp. 78–82, Dec. 2011.

[271] Y. C. Chen, T. Liu, C. H. Yu, T. Y. Chiang, and C. C. Hwang, "**Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly,**" *PLoS ONE*, vol. 8, no. 4, 2013.

[272] E. E. Schadt, S. Turner, and A. Kasarskis, "**A window into third-generation sequencing,**" *Human Molecular Genetics*, vol. 19, no. R2, 2010.

[273] K. J. Travers, C. S. Chin, D. R. Rank, J. S. Eid, and S. W. Turner, "**A flexible and efficient template format for circular consensus sequencing and SNP detection,**" *Nucleic acids research*, vol. 38, no. 15, p. e159, 2010.

[274] "**Media Resources - Pacific Biosciences.**" .

[275] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, "**Real-Time DNA Sequencing from Single Polymerase Molecules,**" *Science*, vol. 323, no. 5910, pp. 133–138, Jan. 2009.

[276] A. Rhoads and K. F. Au, "**PacBio Sequencing and Its Applications,**" 2015.

[277] "**AllSeq. Pacific Biosciences.**" 2017.

[278] Illumina, "**HiSeq 2500 Specifications Key performance parameters.**" .

[279] J. Korlach, "**Understanding Accuracy in SMRT® Sequencing,**" *Kpoasqb.Pacb.Com*, no. iii, pp. 1–9, 2013.

[280] R. P. Schaudies, **Biological Idenification**. WOODHEAD, 2017.

[281] S. D. Brown, S. M. Utturkar, T. S. Magnuson, A. E. Ray, F. L. Poole, W. A. Lancaster, M. P. Thorgersen, M. W. W. Adams, and D. A. Elias, "**Complete Genome Sequence of Pelosinus sp. Strain UFO1 Assembled Using Single-Molecule Real-Time DNA Sequencing Technology.,**" *Genome announcements*, vol. 2, no. 5, pp. 2–3, Sep. 2014.

[282] Y. Terabayashi, A. Juan, H. Tamotsu, N. Ashimine, K. Nakano, M. Shimoji, A. Shiroma, K. Teruya, K. Satou, and T. Hirano, "**First Complete Genome**

Sequence of Salmonella enterica subsp. enterica Serovar Typhimurium Strain ATCC 13311 (NCTC 74), a Reference Strain of Multidrug Resistance, as Achieved by Use of PacBio Single-Molecule Real-Time Technology," *Genome announcements*, vol. 2, no. 5, pp. e00986--14, Oct. 2014.

[283] L. Eme, E. Gentekaki, B. Curtis, J. M. Archibald, and A. J. Roger, "**Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut,**" *Current Biology*, vol. 14, no. 0, pp. 1113–1126, 2017.

[284] I. Wawrzyniak, D. Courtine, M. Osman, C. Hubans-Pierlot, A. Cian, C. Nourrisson, M. Chabe, P. Poirier, A. Bart, V. Polonais, P. Delgado-Viscogliosi, H. El Alaoui, A. Belkorchia, T. van Gool, K. S. W. Tan, S. Ferreira, E. Viscogliosi, and F. Delbac, "**Draft genome sequence of the intestinal parasite Blastocystis subtype 4-isolate WR1,**" *Genomics Data*, vol. 4, pp. 22–23, 2015.

[285] C. G. Clark, U. C. M. Alsmark, M. Tazreiter, Y. Saito-Nakano, V. Ali, S. Marion, C. Weber, C. Mukherjee, I. Bruchhaus, E. Tannich, M. Leippe, T. Sicheritz-Ponten, P. G. Foster, J. Samuelson, C. J. Noël, R. P. Hirt, T. M. Embley, C. A. Gilchrist, B. J. Mann, U. Singh, J. P. Ackers, S. Bhattacharya, A. Bhattacharya, A. Lohia, N. Guillén, M. Duchêne, T. Nozaki, and N. Hall, "**Structure and Content of the Entamoeba histolytica Genome,**" *Advances in Parasitology*, vol. 65, no. 7, pp. 51–190, 2007.

[286] E. Nyvltova, R. Sutak, K. Harant, M. Sedinova, I. Hrdy, J. Paces, C. Vlcek, and J. Tachezy, "**NIF-type iron-sulfur cluster assembly system is duplicated and distributed in the mitochondria and cytosol of Mastigamoeba balamuthi,**" *Proceedings of the National Academy of Sciences*, vol. 110, no. 18, pp. 7371–7376, 2013.

[287] L. Eichinger, J. A. Pachebat, G. Glöckner, M.-A. Rajandream, R. Sucgang, M. Berriman, J. Song, R. Olsen, K. Szafranski, Q. Xu, B. Tunggal, S. Kummerfeld, M. Madera, B. A. Konfortov, F. Rivero, A. T. Bankier, R. Lehmann, N. Hamlin, R. Davies, P. Gaudet, P. Fey, K. Pilcher, G. Chen, D. Saunders, E. Sodergren, P. Davis, A. Kerhornou, X. Nie, N. Hall, C. Anjard, L. Hemphill, N. Bason, P. Farbrother, B. Desany, E. Just, T. Morio, R. Rost, C. Churcher, J. Cooper, S. Haydock, N. van Driessche, A. Cronin, I. Goodhead, D. Muzny, T. Mourier, A. Pain, M. Lu, D. Harper, R. Lindsay, H. Hauser, K. James, M. Quiles, M. Madan Babu, T. Saito, C. Buchrieser, A. Wardroper, M. Felder, M. Thangavelu, D. Johnson, A. Knights, H. Loulseged, K. Mungall, K. Oliver, C. Price, M. A. Quail, H. Urushihara, J. Hernandez, E. Rabbinowitsch, D. Steffen, M. Sanders, J. Ma, Y. Kohara, S. Sharp, M. Simmonds, S. Spiegler, A. Tivey, S. Sugano, B. White, D. Walker, J. Woodward, T. Winckler, Y. Tanaka, G. Shaulsky, M. Schleicher, G. Weinstock, A. Rosenthal, E. C. Cox, R. L. Chisholm, R. Gibbs, W. F. Loomis, M. Platzer, R. R. Kay, J. Williams, P. H. Dear, A. A. Noegel, B. Barrell, and A. Kuspa, "**The genome of the social amoeba Dictyostelium discoideum,**" *Nature*, vol. 435, no. 7038, pp. 43–57, May 2005.

[288] S. H. H. edited by John J. Lee Eugene C. Bovee, **An illustrated guide to the protozoa**. Lawrence, Kansas : Society of Protozoologists : Allen Press Inc., [1985] ©1985.

[289] N. ÜÇÜNCÜ, B. GÖÇMEN, and S. ÜÇÜNCÜ, "**PROTEROMONAS FOTOS -**

Dikenli Keler Laudakia stellio stellio (Linnaeus, 1758) (Reptilia Lacertilia) Rektumunda Yaşayan Protozoonlar ve Yapıları.pdf," *Türkiye Parazitoloji Dergisi 25*, vol. 25, no. 1, pp. 79–83, 2001.
[290] J. Kulda, "**Axenic cultivation of Proteromonas lacertae-viridis (Grassi 1879).,**" *Journal of Protozoology*, vol. 20, no. 3, pp. 536–537, 1973.
[291] M. Kostka, I. Cepicka, V. Hampl, and J. Flegr, "**Phylogenetic position of Karotomorpha and paraphyly of Proteromonadidae.,**" *Molecular phylogenetics and evolution*, vol. 43, no. 3, pp. 1167–1170, Jun. 2007.
[292] T. Cavalier-Smith and E. E.-Y. Chao, "**Phylogeny and Megasystematics of Phagotrophic Heterokonts (Kingdom Chromista),**" *Journal of Molecular Evolution*, vol. 62, no. 4, pp. 388–420, 2006.
[293] M. Kostka, V. Hampl, I. Cepicka, and J. Flegr, "**Phylogenetic position of Protoopalina intestinalis based on SSU rRNA gene sequence,**" *Molecular Phylogenetics and Evolution*, vol. 33, no. 1, pp. 220–224, 2004.
[294] V. Pérez-Brocal, R. Shahar-Golan, and C. G. Clark, "**A linear molecule with two large inverted repeats: the mitochondrial genome of the stramenopile Proteromonas lacertae.,**" *Genome biology and evolution*, vol. 2, pp. 257–266, Jan. 2010.
[295] G. Brugerolle and C. F. Bardele, **Cortical cytoskeleton of the flagellateProteromonas lacertae: Interrelation between microtubules, membrane and somatonemes**, vol. 142. 1988.
[296] G. Brugerolle and J. P. Mignot, **Handbook of Protoctista**. Boston: Jones and Barlett Publishers, 1990.
[297] G. Brugerolle and L. Joyon, "**Étude cytologique ultrastructurale des genres Proteromonas et Karotomorpha (Zoomoastigophorea Protero- monadida Grassé 1952).,**" *Protistologica*, vol. 11, pp. 531–546, 1975.
[298] J. Yin, J. Howe, and K. S. W. Tan, "**Staurosporine-induced programmed cell death in Blastocystis occurs independently of caspases and cathepsins and is augmented by calpain inhibition,**" *Microbiology*, vol. 156, no. 5, pp. 1284–1293, May 2010.
[299] D. D. Leipe, S. M. Tong, C. L. Goggin, S. B. Slemenda, N. J. Pieniazek, and M. L. Sogin, "**16S-like rDNA sequences from Developayella elegans, Labyrinthuloides haliotidis, and Proteromonas lacertae confirm that the stramenopiles are a primarily heterotrophic group,**" *European Journal of Protistology*, vol. 32, no. 4, pp. 449–458, 1996.
[300] N. Arisue, T. Hashimoto, and H. Yoshikawa, "**Sequence heterogeneity of the small subunit ribosomal RNA genes among blastocystis isolates.,**" *Parasitology*, vol. 126, no. Pt 1, pp. 1–9, 2003.
[301] S. Kamoun, "**A catalogue of the effector secretome of plant pathogenic oomycetes.,**" *Annual review of phytopathology*, vol. 44, pp. 41–60, Jan. 2006.
[302] S. Raffaele, J. Win, L. M. Cano, and S. Kamoun, "**Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of Phytophthora infestans.,**" *BMC genomics*, vol. 11, no. 1, p. 637, Jan. 2010.
[303] B. B. Horowitz and M. D. Ospina-Giraldo, "**The pectin methylesterase gene complement of Phytophthora sojae: Structural and functional analyses, and the evolutionary relationships with its oomycete**

homologs," *PLoS ONE*, vol. 10, no. 11, pp. 1–16, 2015.
[304] R. Stam, J. Jupe, A. J. M. Howden, J. a Morris, P. C. Boevink, P. E. Hedley, and E. Huitema, "**Identification and Characterisation CRN Effectors in Phytophthora capsici Shows Modularity and Functional Diversity.,**" *PloS one*, vol. 8, no. 3, p. e59517, Jan. 2013.
[305] C. A. Lévesque, H. Brouwer, L. Cano, J. P. Hamilton, C. Holt, E. Huitema, S. Raffaele, G. P. Robideau, M. Thines, J. Win, M. M. Zerillo, G. W. Beakes, J. L. Boore, D. Busam, B. Dumas, S. Ferriera, S. I. Fuerstenberg, C. M. M. Gachon, E. Gaulin, F. Govers, L. Grenville-Briggs, N. Horner, J. Hostetler, R. H. Y. Jiang, J. Johnson, T. Krajaejun, H. Lin, H. J. G. Meijer, B. Moore, P. Morris, V. Phuntmart, D. Puiu, J. Shetty, J. E. Stajich, S. Tripathy, S. Wawra, P. van West, B. R. Whitty, P. M. Coutinho, B. Henrissat, F. Martin, P. D. Thomas, B. M. Tyler, R. P. De Vries, S. Kamoun, M. Yandell, N. Tisserat, and C. R. Buell, "**Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire.,**" *Genome biology*, vol. 11, no. 7, p. R73, 2010.
[306] A. Savidor, R. S. Donahoo, O. Hurtado-Gonzales, M. L. Land, M. B. Shah, K. H. Lamour, and W. H. McDonald, "**Cross-species global proteomics reveals conserved and unique processes in Phytophthora sojae and Phytophthora ramorum.,**" *Molecular & cellular proteomics : MCP*, vol. 7, no. 8, pp. 1501–1516, Aug. 2008.
[307] C. G. Clark and L. S. Diamond, "**Methods for cultivation of luminal parasitic protists of clinical importance,**" *Clinical Microbiology Reviews*, vol. 15, no. 3, pp. 329–341, 2002.
[308] G. Marçais and C. Kingsford, "**A fast, lock-free approach for efficient parallel counting of occurrences of k-mers,**" *Bioinformatics*, vol. 27, no. 6, pp. 764–770, 2011.
[309] B. Liu, Y. Shi, J. Yuan, X. Hu, H. Zhang, N. Li, Z. Li, Y. Chen, D. Mu, and W. Fan, "**Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects,**" *arXiv*, p. 1308.2012, 2013.
[310] C. L. Weiß, M. Pais, L. M. Cano, S. Kamoun, and H. A. Burbano, "**nQuire: A Statistical Framework For Ploidy Estimation Using Next Generation Sequencing,**" *bioRxiv*, p. 143537, 2017.
[311] C. Trapnell, L. Pachter, and S. L. Salzberg, "**TopHat: Discovering splice junctions with RNA-Seq,**" *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.
[312] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, and A. Regev, "**De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity,**" *Nature protocols*, vol. 8, no. 8, p. 10.1038/nprot.2013.084, Aug. 2013.
[313] M. Stanke, R. Steinkamp, S. Waack, and B. Morgenstern, "**AUGUSTUS: A web server for gene finding in eukaryotes,**" *Nucleic Acids Research*, vol. 32, no. WEB SERVER ISS., pp. 309–312, 2004.
[314] I. Korf, "**Gene finding in novel genomes.,**" *BMC bioinformatics*, vol. 5, p. 59, 2004.
[315] T. Carver, S. R. Harris, M. Berriman, J. Parkhill, and J. a. McQuillan,

"**Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data,**" *Bioinformatics*, vol. 28, no. 4, pp. 464–469, 2012.
[316] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "**Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research,**" *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.
[317] E. M. Zdobnov and R. Apweiler, "**InterProScan--an integration platform for the signature-recognition methods in InterPro.,**" *Bioinformatics (Oxford, England)*, vol. 17, no. 9, pp. 847–848, 2001.
[318] a Krogh and B. Larsson, "**Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes,**" *Journal of molecular …*, vol. 305, no. 3, pp. 567–580, 2001.
[319] J. D. Bendtsen, H. Nielsen, G. Von Heijne, and S. Brunak, "**Improved prediction of signal peptides: SignalP 3.0,**" *Journal of Molecular Biology*, vol. 340, no. 4, pp. 783–795, 2004.
[320] V. Pejaver, W. L. Hsu, F. Xin, A. K. Dunker, V. N. Uversky, and P. Radivojac, "**The structural and functional signatures of proteins that undergo multiple events of post-translational modification,**" *Protein Science*, vol. 23, no. 8, pp. 1077–1093, 2014.
[321] S. Griffiths-Jones, "**Annotating non-coding RNAs with Rfam.,**" *Current Protocols in Bioinformatics*, vol. Chapter 12, p. Unit 12.5, 2005.
[322] A. F. A. Smit and R. Hubley, "**RepeatModeler Open-1.0.**" .
[323] Z. Xu and H. Wang, "**LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons,**" *Nucleic Acids Research*, vol. 35, no. SUPPL.2, 2007.
[324] M. Kanehisa, Y. Sato, and K. Morishima, "**BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences,**" *Journal of Molecular Biology*, vol. 428, no. 4, pp. 726–731, 2016.
[325] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "**KEGG for integration and interpretation of large-scale molecular data sets,**" *Nucleic Acids Research*, vol. 40, no. D1, 2012.
[326] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V Kriventseva, and E. M. Zdobnov, "**BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs,**" *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, 2015.
[327] R. C. Edgar, "**MUSCLE: Multiple sequence alignment with high accuracy and high throughput,**" *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.
[328] J. Castresana, "**Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.,**" *Molecular biology and evolution*, vol. 17, no. 4, pp. 540–552, Apr. 2000.
[329] V. Lefort, J.-E. Longueville, and O. Gascuel, "**SMS: Smart Model Selection in PhyML,**" *Molecular Biology and Evolution*, 2017.
[330] H. Akaike, "**Information Theory and an Extension of the Maximum Likelihood Principle BT - Selected Papers of Hirotugu Akaike,**" E. Parzen, K. Tanabe, and G. Kitagawa, Eds. New York, NY: Springer New York, 1998, pp. 199–213.

[331] J. M. Archibald and A. G. B. Simpson, **Handbook of the Protists Second Edition**. .
[332] V. Zaman, J. Howe, and M. Ng, "**Observations on the surface coat of Blastocystis hominis,**" *Parasitology Research*, vol. 83, pp. 731–733, 1997.
[333] D. J. Sienzel, P. F. L. Boreham, and R. McDougall, "**Ultrastructure of Blastocystis hominis in human stool samples,**" *International Journal for Parasitology*, vol. 21, no. 7, pp. 807–812, 1991.
[334] V. Zaman, J. Howe, M. Ng, and T. K. Goh, "**Scanning electron microscopy of the surface coat of Blastocystis hominis.,**" *Parasitology research*, vol. 85, no. 12, pp. 974–976, 1999.
[335] M. W. Lightowlers and M. D. Rickard, "**Excretory–secretory products of helminth parasites: effects on host immune responses,**" *Parasitology*, vol. 96, no. S1, pp. S123--S166, 1988.
[336] L. A. Baxt, R. P. Baker, U. Singh, and S. Urban, "**An Entamoeba histolytica rhomboid protease with atypical specificity cleaves a surface lectin involved in phagocytosis and immune evasion,**" *Genes and Development*, vol. 22, no. 12, pp. 1636–1646, Jun. 2008.
[337] J. Calderon, M. de Lourdes Munoz, and H. M. Acosta, "**Surface redistribution and release of antibody-induced caps in entamoebae,**" *J Exp Med*, vol. 151, no. 1, pp. 184–193, Jan. 1980.
[338] K. Judge, M. Hunt, S. Reuter, A. Tracey, M. A. Quail, J. Parkhill, and S. J. Peacock, "**Comparison of bacterial genome assembly software for MinION data and their applicability to medical microbiology,**" *Microbial Genomics*, vol. 2, no. 9, 2016.
[339] Y.-C. Liao, S.-H. Lin, and H.-H. Lin, "**Completing bacterial genome assemblies: strategy and performance comparisons,**" *Scientific Reports*, vol. 5, no. 1, p. 8747, 2015.
[340] N. Ye, X. Zhang, M. Miao, X. Fan, Y. Zheng, D. Xu, J. Wang, L. Zhou, D. Wang, Y. Gao, Y. Wang, W. Shi, P. Ji, D. Li, Z. Guan, C. Shao, Z. Zhuang, Z. Gao, J. Qi, and F. Zhao, "**Saccharina genomes provide novel insight into kelp biology,**" *Nature Communications*, vol. 6, p. 6986, Apr. 2015.
[341] X. Wang, Z. Shao, W. Fu, J. Yao, Q. Hu, and D. Duan, "**Chloroplast genome of one brown seaweed, Saccharina japonica (Laminariales, Phaeophyta): Its structural features and phylogenetic analyses with other photosynthetic plastids,**" *Marine Genomics*, vol. 10, pp. 1–9, Jun. 2013.
[342] E. V. Armbrust, J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. Zhou, A. E. Allen, K. E. Apt, M. Bechner, M. A. Brzezinski, B. K. Chaal, A. Chiovitti, A. K. Davis, M. S. Demarest, J. C. Detter, T. Glavina, D. Goodstein, M. Z. Hadi, U. Hellsten, M. Hildebrand, B. D. Jenkins, W. W. Y. Lau, T. W. Lane, F. W. Larimer, J. C. Lippmeier, S. Lucas, A. Montsant, M. Obornik, M. S. Parker, B. Palenik, G. J. Pazour, P. M. Richardson, T. A. Rynearson, M. A. Saito, D. C. Schwartz, K. Thamatrakoln, K. Valentin, A. Vardi, F. P. Wilkerson, and D. S. Rokhsar, "**The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism,**" *Architecture*, vol. 306, no. October, pp. 79–86, 2004.
[343] R. Radakovits, R. E. Jinkerson, S. I. Fuerstenberg, H. Tae, R. E. Settlage, J. L. Boore, and M. C. Posewitz, "**Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropis gaditana,**"

Nature Communications, vol. 3, p. 686, Feb. 2012.
[344] M. Lommer, M. Specht, A.-S. Roy, L. Kraemer, R. Andreson, M. A. Gutowska, J. Wolf, S. V Bergner, M. B. Schilhabel, U. C. Klostermeier, R. G. Beiko, P. Rosenstiel, M. Hippler, and J. LaRoche, "**Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation,**" *Genome Biology*, vol. 13, no. 7, p. R66, 2012.
[345] T. Fenchel and D. J. Patterson, "**Cafeteria roenbergensis nov. gen., nov. sp., a heterotrophic microflagellate from marine plankton,**" *Marine Microbial Food Webs*, vol. 3, no. 1, pp. 9–19, 1988.
[346] C. J. O 'kelly1 and D. J. Patterson, "**The flagellar apparatus of Cafeteria roenbergensis Fenchel & Patterson, 1988 (Bicosoecales = Bicosoecida),**" *European Journal of Protistology*, vol. 32, pp. 216–226, 1996.
[347] N. Vørs, "**Marine Heterotrophic Amoebae, Flagellates and Heliozoa From Belize (Central America) and Tenerife (Canary Islands), With Descriptions of New Species, Luffisphaera Bulbochaete N. Sp., L. Longihastis N. Sp., L. Turriformis N. Sp. and Paulinella Intermedia N,**" *Journal of Eukaryotic Microbiology*, vol. 40, no. 3, pp. 272–287, 1993.
[348] H. R. Preisig, N. Vors, and G. Hallfors, **Diversity of heterotrophic heterokont flagellates.** Oxford: Clarendon Press, 1991.
[349] J. Larsen and D. J. Patterson, "**Some flagellates (Protista) from tropical marine sediments,**" *Journal of Natural History*, vol. 24, no. 4, pp. 801–937, Aug. 1990.
[350] N. Vørs, "**Heterotrophic amoebae, flagellates and heliozoa from the Tvärminne area, Gulf of Finland, in 1988–1990,**" *Ophelia*, vol. 36, no. 1, pp. 1–109, Sep. 1992.
[351] N. Vørs, "**Heterotrophic amoebae, flagellates and heliozoa from Arctic marine waters (North West Territories, Canada and West Greenland),**" *Polar Biology*, vol. 13, no. 2, pp. 113–126, 1993.
[352] D. J. Patterson, K. Nygaard, G. Steinberg, and C. M. Turley, "**HETEROTROPHIC FLAGELLATES AND OTHER PROTISTS ASSOCIATED WITH OCEANIC DETRITUS THROUGHOUT THE WATER COLUMN IN THE MID NORTH ATLANTIC,**" *mar. biol. Ass. U.K*, vol. 73, pp. 67–95, 2017.
[353] R. Derelle, P. López-García, H. Timpano, and D. Moreira, "**A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (Heterokonts),**" *Molecular Biology and Evolution*, vol. 33, no. 11, pp. 2890–2898, 2016.
[354] A. Stechmann, K. Hamblin, V. Pérez-Brocal, D. Gaston, G. S. Richmond, M. van der Giezen, C. G. Clark, and A. J. Roger, "**Organelles in Blastocystis that Blur the Distinction between Mitochondria and Hydrogenosomes,**" *Current Biology*, vol. 18, no. 8, pp. 580–585, 2008.
[355] H. Nakazato, S. Venkatesan, and M. Edmonds, "**Polyadenylic acid sequences in E. coli messenger RNA,**" *Nature*, vol. 256, no. 5513, pp. 144–146, Jul. 1975.
[356] P. R. Srinivasan, M. Ramanarayanan, and E. Rabbani, "**Presence of polyriboadenylate sequences in pulse-labeled RNA of Escherichia coli,**" *Proc Natl Acad Sci U S A*, vol. 72, no. 8, pp. 2910–2914, Aug. 1975.
[357] N. Ohta, M. Sanders, and A. Newton, "**Poly(adenylic acid) sequences in**

the RNA of Caulobacter crescenus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72, no. 6, pp. 2343–2346, Jun. 1975.
[358] N. Sarkar, "**POLYADENYLATION OF mRNA IN PROKARYOTES,**" *Microbiology*, vol. 142, no. 11, pp. 173–197, Nov. 1997.
[359] M. J. Strong, G. Xu, L. Morici, S. Splinter Bon-Durant, M. Baddoo, Z. Lin, C. Fewell, C. M. Taylor, and E. K. Flemington, "**Microbial Contamination in Next Generation Sequencing: Implications for Sequence-Based Analysis of Clinical Samples,**" *PLoS Pathogens*, vol. 10, no. 11, 2014.
[360] G. J. Dick, A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, A. P. Yelton, and J. F. Banfield, "**Community-wide analysis of microbial genome sequence signatures,**" *Genome Biology*, vol. 10, no. 8, p. R85, 2009.
[361] C. C. Laczny, N. Pinel, N. Vlassis, and P. Wilmes, "**Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction,**" *Scientific Reports*, vol. 4, no. 1, p. 4516, 2015.
[362] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson, "**GroopM: an automated tool for the recovery of population genomes from related metagenomes,**" *PeerJ*, vol. 2, p. e603, 2014.
[363] I. Sharon, M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman, and J. F. Banfield, "**Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization,**" *Genome Research*, vol. 23, no. 1, pp. 111–120, 2013.
[364] P. Puigbò, E. Guzmán, A. Romeu, and S. Garcia-Vallvé, "**OPTIMIZER: A web server for optimizing the codon usage of DNA sequences,**" *Nucleic Acids Research*, vol. 35, no. SUPPL.2, 2007.
[365] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz, "**A Bioinformatician's Guide to Metagenomics,**" *Microbiology and Molecular Biology Reviews*, vol. 72, no. 4, pp. 557–578, 2008.
[366] R. RDC Team, "**R: A Language and Environment for Statistical Computing,**" *R Foundation for Statistical Computing*, vol. 1, no. 2.11.1. Vienna, p. 409, 2012.
[367] RStudio Team., "**Integrated Development for R. RStudio, Inc.,**" *R. RStudio, Inc., Boston, MA.* Boston, 2015.
[368] P. Rice, I. Longden, and A. Bleasby, "**EMBOSS: The European Molecular Biology Open Software Suite,**" *Trends in Genetics*, vol. 16, no. 1, pp. 276–277, 2000.
[369] B. Langmead, "**Aligning short sequencing reads with Bowtie,**" *Current Protocols in Bioinformatics*, no. SUPP.32, 2010.
[370] R. Massana, J. Del Campo, C. Dinter, and R. Sommaruga, "**Crash of a population of the marine heterotrophic flagellate Cafeteria roenbergensis by viral infection,**" *Environmental Microbiology*, vol. 9, no. 11, pp. 2660–2669, 2007.
[371] W. F. Doolittle, Y. Boucher, C. L. NesbO, C. J. Douady, J. O. Andersson, and A. J. Roger, "**How big is the iceberg of which organellar genes in nuclear genomes are but the tip?,**" *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 358, no. 1429, pp. 39–58, 2003.
[372] W. Ford Doolittle, "**You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes,**" *Trends in*

Genetics, vol. 14, no. 8. pp. 307–311, Aug-1998.
[373] B. Paquin, M. J. Laforest, L. Forget, I. Roewer, Z. Wang, J. Longcore, and B. F. Lang, "**The fungal mitochondrial genome project: evolution of fungal mitochondrial genomes and their gene expression.,**" *Current genetics*, vol. 31, no. 5, pp. 380–395, May 1997.
[374] J. M. Cock, L. Sterck, P. Rouzé, D. Scornet, A. E. Allen, G. Amoutzias, V. Anthouard, F. Artiguenave, J.-M. Aury, J. H. Badger, B. Beszteri, K. Billiau, E. Bonnet, J. H. Bothwell, C. Bowler, C. Boyen, C. Brownlee, C. J. Carrano, B. Charrier, G. Y. Cho, S. M. Coelho, J. Collén, E. Corre, C. Da Silva, L. Delage, N. Delaroque, S. M. Dittami, S. Doulbeau, M. Elias, G. Farnham, C. M. M. Gachon, B. Gschloessl, S. Heesch, K. Jabbari, C. Jubin, H. Kawai, K. Kimura, B. Kloareg, F. C. Küpper, D. Lang, A. Le Bail, C. Leblanc, P. Lerouge, M. Lohr, P. J. Lopez, C. Martens, F. Maumus, G. Michel, D. Miranda-Saavedra, J. Morales, H. Moreau, T. Motomura, C. Nagasato, C. a Napoli, D. R. Nelson, P. Nyvall-Collén, A. F. Peters, C. Pommier, P. Potin, J. Poulain, H. Quesneville, B. Read, S. a Rensing, A. Ritter, S. Rousvoal, M. Samanta, G. Samson, D. C. Schroeder, B. Ségurens, M. Strittmatter, T. Tonon, J. W. Tregear, K. Valentin, P. von Dassow, T. Yamagishi, Y. Van de Peer, and P. Wincker, "**The Ectocarpus genome and the independent evolution of multicellularity in brown algae,**" *Nature*, vol. 465, no. 7298, pp. 617–621, 2010.
[375] I. Wawrzyniak, M. Roussel, M. Diogon, A. Couloux, C. Texier, K. S. W. Tan, C. P. Vivarès, F. Delbac, P. Wincker, and H. El Alaoui, "**Complete circular DNA in the mitochondria-like organelles of Blastocystis hominis,**" *International Journal for Parasitology*, vol. 38, no. 12, pp. 1377–1382, 2008.
[376] E. Gentekaki, B. A. Curtis, C. W. Stairs, D. E. Salas-leiva, E. K. Herman, L. Eme, M. Elia, M. C. Arias, M. J. Klute, H. Suga, B. Malik, A. W. Pightling, M. Kolisko, R. A. Rachubinski, A. Schlacht, D. M. Soanes, A. D. Tsaousis, M. Archibald, S. G. Ball, J. B. Dacks, C. G. Clark, M. Van Der, and A. J. Roger, **Extreme genome diversity in the hyper- prevalent parasitic eukaryote Blastocystis ´ r Klimes ´ s**. 2017.
[377] C. Wunder, Y. Churin, F. Winau, D. Warnecke, M. Vieth, B. Lindner, U. Zähringer, H.-J. Mollenkopf, E. Heinz, and T. F. Meyer, "**Cholesterol glucosylation promotes immune evasion by Helicobacter pylori,**" *Nature Medicine*, vol. 12, no. 9, pp. 1030–1038, Sep. 2006.
[378] G. Li and K. D. Young, "**Indole production by the tryptophanase TnaA in escherichia coli is determined by the amount of exogenous tryptophan,**" *Microbiology (United Kingdom)*, vol. 159, no. 2, pp. 402–410, Feb. 2013.
[379] J. K. Nicholson, E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia, and S. Pettersson, "**Host-gut microbiota metabolic interactions.,**" *Science (New York, N.Y.)*, vol. 336, no. 6086, pp. 1262–1267, Jun. 2012.
[380] H. Y. Long, A. Handschack, W. König, and A. Ambrosch, "**Blastocystis hominis modulates immune responses and cytokine release in colonic epithelial cells.,**" *Parasitology research*, vol. 87, pp. 1029–1030, 2001.
[381] M. K. Puthia, A. Vaithilingam, J. Lu, and K. S. W. Tan, "**Degradation of human secretory immunoglobulin A by Blastocystis.,**" *Parasitology*

research, vol. 97, no. 5, pp. 386–389, Nov. 2005.
[382] S. W. S. Sio, M. K. Puthia, A. S. Y. Lee, J. Lu, and K. S. W. Tan, "**Protease activity of Blastocystis hominis.,**" *Parasitology research*, vol. 99, no. 2, pp. 126–130, Jul. 2006.
[383] H. Dagci, S. Ustun, M. S. Taner, G. Ersoz, F. Karacasu, and S. Budak, "**Protozoon infections and intestinal permeability,**" *Acta Tropica*, vol. 81, no. 1, pp. 1–5, Jan. 2002.
[384] M. K. Puthia, S. W. S. Sio, J. Lu, and K. S. W. Tan, "**Blastocystis ratti induces contact-independent apoptosis, F-actin rearrangement, and barrier function disruption in IEC-6 cells.,**" *Infection and immunity*, vol. 74, no. 7, pp. 4114–4123, Jul. 2006.
[385] H. Mirza and K. S. W. Tan, "**Blastocystis exhibits inter- and intra-subtype variation in cysteine protease activity,**" *Parasitology Research*, vol. 104, no. 2, pp. 355–361, Jan. 2009.
[386] G. W. Beakes, S. L. Glockling, and S. Sekimoto, "**The evolutionary phylogeny of the oomycete 'fungi,'**" *Protoplasma*, vol. 249, no. 1. pp. 3–19, 2012.
[387] F. N. Martin and J. E. Loper, "**Critical Reviews in Plant Sciences Soilborne Plant Diseases Caused by Pythium spp.: Ecology, Epidemiology, and Prospects for Biological Control Soilborne Plant Diseases Caused by Pythium spp.: Ecology, Epidemiology, and Prospects for Biological Control,**" *Critical Reviews in Plant Sciences*, vol. 18, no. 182, pp. 111–181, Mar. 1999.
[388] G. A. Saunders, J. O. Washburn, D. E. Egerter, and J. R. C. N.-2178 Anderson, "**Pathogenicity of fungi isolated from field-collected larvae of the western treehole mosquito, Aedes sierrensis (Diptera: Culicidae),**" *J. Invert. Pathol.*, vol. 52, no. 2, pp. 360–363, Sep. 1988.
[389] A. J. van der Plaats-Niterink, **Monograph of the Genus Pythium**. Centraalbureau voor Schimmelcultures, 1981.
[390] M. J. Fernández-Benéitez, M. E. Ortiz-Santaliestra, M. Lizana, and J. Diéguez-Uribeondo, "**Saprolegnia diclina: Another species responsible for the emergent disease 'Saprolegnia infections' in amphibians,**" *FEMS Microbiology Letters*, vol. 279, no. 1, pp. 23–29, 2008.
[391] E. Thoen, T. Vrålstad, E. Rolén, R. Kristensen, Evensen, and I. Skaar, "**Saprolegnia species in Norwegian salmon hatcheries: Field survey identifies S. diclina sub-clade IIIB as the dominating taxon,**" *Diseases of Aquatic Organisms*, vol. 114, no. 3, pp. 189–198, 2015.
[392] R. H. Y. Jiang, I. de Bruijn, B. J. Haas, R. Belmonte, L. Löbach, J. Christie, G. van den Ackerveken, A. Bottin, V. Bulone, S. M. Díaz-Moreno, B. Dumas, L. Fan, E. Gaulin, F. Govers, L. J. Grenville-Briggs, N. R. Horner, J. Z. Levin, M. Mammella, H. J. G. Meijer, P. Morris, C. Nusbaum, S. Oome, A. J. Phillips, D. van Rooyen, E. Rzeszutek, M. Saraiva, C. J. Secombes, M. F. Seidl, B. Snel, J. H. M. Stassen, S. Sykes, S. Tripathy, H. van den Berg, J. C. Vega-Arreguin, S. Wawra, S. K. Young, Q. Zeng, J. Dieguez-Uribeondo, C. Russ, B. M. Tyler, and P. van West, "**Distinctive Expansion of Potential Virulence Genes in the Genome of the Oomycete Fish Pathogen Saprolegnia parasitica,**" *PLoS Genetics*, vol. 9, no. 6, p. e1003272, Jun. 2013.
[393] M. Tachibana, A. E. Allen, S. Kikutani, Y. Endo, C. Bowler, and Y. Matsuda, "**Localization of putative carbonic anhydrases in two marine**

diatoms, Phaeodactylum tricornutum and Thalassiosira pseudonana," *Photosynthesis Research*, vol. 109, no. 1, pp. 205–221, 2011.
[394] S. Fischer, B. P. Brunk, F. Chen, X. Gao, O. S. Harb, J. B. Iodice, D. Shanmugam, D. S. Roos, and C. J. Stoeckert, "**Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups,**" *Current Protocols in Bioinformatics*, no. SUPPL.35, 2011.
[395] D. M. Emms and S. Kelly, "**OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy,**" *Genome Biology*, vol. 16, no. 1, p. 157, 2015.
[396] K. Chaichoompu, S. Kittitornkun, and S. Tongsima, "**MT-ClustalW: Multithreading multiple sequence alignment,**" in *20th International Parallel and Distributed Processing Symposium, IPDPS 2006*, 2006, vol. 2006.
[397] T. J. Wheeler, "**Large-scale neighbor-joining with NINJA,**" in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5724 LNBI, pp. 375–389.
[398] F. Ronquist and J. P. Huelsenbeck, "**MrBayes 3: Bayesian phylogenetic inference under mixed models,**" *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.
[399] D. P. Faith, "**Conservation evaluation and phylogentic diversity,**" *Biological Conservation*, vol. 61, pp. 1–10, 1992.
[400] S. W. Kembel, P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. D. Ackerly, S. P. Blomberg, and C. O. Webb, "**Picante: R tools for integrating phylogenies and ecology,**" *Bioinformatics*, vol. 26, no. 11, pp. 1463–1464, Jun. 2010.
[401] G. B. Bouck, "**Extracellular microtubules. The origin, structure, and attachment of flagellar hairs in Fucus and Ascophyllum antherozoids.,**" *Journal of Cell Biology*, vol. 40, no. 2, pp. 446–460, 1969.
[402] M. Tomich, C. A. Herfst, J. W. Golden, and C. D. Mohr, "**Role of flagella in host cell invasion by Burkholderia cepacia,**" *Infection and Immunity*, vol. 70, no. 4, pp. 1799–1806, Apr. 2002.
[403] H. S. Judelson, J. Shrivastava, and J. Manson, "**Decay of Genes Encoding the Oomycete Flagellar Proteome in the Downy Mildew Hyaloperonospora arabidopsidis,**" *PLoS ONE*, vol. 7, no. 10, 2012.
[404] L. Shi Jing, F. Fathiah Muzaffar Shah, M. Saberi Mohamad, K. Moorthy, S. Deris, Z. Zakaria, and S. Napis, "**A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data,**" *Current Proteomics*, vol. 12, no. 1, pp. 14–27, 2015.
[405] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "**GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists,**" *BMC Bioinformatics*, vol. 10, no. 1, p. 48, Feb. 2009.
[406] J. D. DeBarry and J. C. Kissinger, "**A survey of innovation through duplication in the reduced genomes of twelve parasites,**" *PLoS ONE*, vol. 9, no. 6, 2014.
[407] M. Sajid and J. H. McKerrow, "**Erratum to 'Cysteine proteases of**

parasitic organisms,'" *Molecular and Biochemical Parasitology*, vol. 121, no. 1, p. 159, 2002.
[408] J. H. McKerrow, "**Parasite proteases,**" *Experimental Parasitology*, vol. 68, no. 1, pp. 111–115, 1989.
[409] T. C. O'Brien, Z. B. Mackey, R. D. Fetter, Y. Choe, A. J. O'Donoghue, M. Zhou, C. S. Craik, C. R. Caffrey, and J. H. McKerrow, "**A parasite cysteine protease is key to host protein degradation and iron acquisition,**" *Journal of Biological Chemistry*, vol. 283, no. 43, pp. 28934–28943, Oct. 2008.
[410] A. Shet, E. L. Kaplan, D. R. Johnson, and P. P. Cleary, "**Immune Response to Group A Streptococcal C5a Peptidase in Children : Implications for Vaccine Development,**" *The Journal of infectious diseases*, vol. 55455, pp. 809–817, 2003.
[411] V. Sanchez, J. Serrano-Luna, E. Ram??rez-Moreno, V. Tsutsumi, and M. Shibayama, "**Entamoeba histolytica: Overexpression of the gal/galnac lectin, ehcp2 and ehcp5 genes in an in vivo model of amebiasis,**" *Parasitology International*, vol. 65, no. 6, pp. 665–667, 2016.
[412] K. Nakada-Tsukui and T. Nozaki, "**Immune response of amebiasis and immune evasion by Entamoeba histolytica,**" *Frontiers in Immunology*, vol. 7, no. MAY. 2016.
[413] A. Aguilar-Rojas, J. C. Olivo-Marin, and N. Guillen, "**The motility of Entamoeba histolytica: finding ways to understand intestinal amoebiasis,**" *Current Opinion in Microbiology*, vol. 34, pp. 24–30, 2016.
[414] B. Loftus, I. Anderson, R. Davies, U. C. M. Alsmark, J. Samuelson, P. Amedeo, P. Roncaglia, M. Berriman, R. P. Hirt, B. J. Mann, T. Nozaki, B. Suh, M. Pop, M. Duchene, J. Ackers, E. Tannich, M. Leippe, M. Hofer, I. Bruchhaus, U. Willhoeft, A. Bhattacharya, T. Chillingworth, C. Churcher, Z. Hance, B. Harris, D. Harris, K. Jagels, S. Moule, K. Mungall, D. Ormond, R. Squares, S. Whitehead, M. A. Quail, E. Rabbinowitsch, H. Norbertczak, C. Price, Z. Wang, N. Guillén, C. Gilchrist, S. E. Stroup, S. Bhattacharya, A. Lohia, P. G. Foster, T. Sicheritz-Ponten, C. Weber, U. Singh, C. Mukherjee, N. M. El-Sayed, W. A. Petri, C. G. Clark, T. M. Embley, B. Barrell, C. M. Fraser, and N. Hall, "**The genome of the protist parasite Entamoeba histolytica,**" *Nature*, vol. 433, no. 7028, pp. 865–868, Feb. 2005.
[415] H. A. Lorenzi, D. Puiu, J. R. Miller, L. M. Brinkac, P. Amedeo, N. Hall, and E. V Caler, "**New assembly, reannotation and analysis of the Entamoeba histolytica genome reveal new genomic features and protein content information,**" *PLoS Neglected Tropical Diseases*, vol. 4, no. 6, 2010.
[416] K. F. Baum, R. L. Berens, J. J. Marr, J. A. Harrington, and T. Spector, "**Purine deoxynucleoside salvage in Giardia lamblia,**" *Journal of Biological Chemistry*, vol. 264, no. 35, pp. 21087–21090, 1989.
[417] M. S. Abrahamsen, "**Complete Genome Sequence of the Apicomplexan, Cryptosporidium parvum,**" *Science*, vol. 304, no. 5669, pp. 441–445, Apr. 2004.
[418] C. Aurrecoechea, A. Barreto, J. Brestelli, B. P. Brunk, E. V Caler, S. Fischer, B. Gajria, X. Gao, A. Gingle, G. Grant, O. S. Harb, M. Heiges, J. Iodice, J. C. Kissinger, E. T. Kraemer, W. Li, V. Nayak, C. Pennington, D. F. Pinney, B. Pitts, D. S. Roos, G. Srinivasamoorthy, C. J. Stoeckert, C. Treatman, and H. Wang, "**AmoebaDB and MicrosporidiaDB: functional genomic**

resources for Amoebozoa and Microsporidia species," *Nucleic Acids Research*, vol. 39, no. Database issue, pp. D612--D619, Jan. 2011.
[419] H. Voigt and N. Guillén, "**New insights into the role of the cytoskeleton in phagocytosis of Entamoeba histolytica.,**" *Cellular microbiology*, vol. 1, no. 3, pp. 195–203, 1999.
[420] E. Tannichz, H. Scholzes, R. Nickel, and R. D. Horstmann, "**Homologous Cysteine Proteinases of Pathogenic and Nonpathogenic Entamoeba histolytica,**" vol. 266, no. 8, pp. 4798–4803, 2000.
[421] I. Bruchhaus, B. J. Loftus, N. Hall, and E. Tannich, "**The intestinal protozoan parasite Entamoeba histolytica contains 20 cysteine protease genes, of which only a small subset is expressed during in vitro cultivation,**" *Eukaryotic Cell*, vol. 2, no. 3, pp. 501–509, Jun. 2003.
[422] J. Matthiesen, A. K. Bär, A. K. Bartels, D. Marien, S. Ofori, L. Biller, E. Tannich, H. Lotter, and I. Bruchhaus, "**Overexpression of specific cysteine peptidases confers pathogenicity to a nonpathogenic Entamoeba histolytica clone,**" *mBio*, vol. 4, no. 2, pp. 1–11, 2013.
[423] M. Tillack, N. Nowak, H. Lotter, R. Bracha, D. Mirelman, E. Tannich, and I. Bruchhaus, "**Increased expression of the major cysteine proteinases by stable episomal transfection underlines the important role of EhCP5 for the pathogenicity of Entamoeba histolytica,**" *Molecular and Biochemical Parasitology*, vol. 149, no. 1, pp. 58–64, 2006.
[424] D. Bansal, P. Ave, S. Kerneis, P. Frileux, O. Boché, A. C. Baglin, G. Dubost, A. S. Leguern, M. C. Prevost, R. Bracha, D. Mirelman, N. Guillén, and E. Labruyère, "**An ex-vivo human intestinal model to study Entamoeba histolytica pathogenesis,**" *PLoS Neglected Tropical Diseases*, vol. 3, no. 11, 2009.
[425] R. Thibeaux, A. Dufour, P. Roux, M. Bernier, A. C. Baglin, P. Frileux, J. C. Olivo-Marin, N. Guillén, and E. Labruyère, "**Newly visualized fibrillar collagen scaffolds dictate Entamoeba histolytica invasion route in the human colon,**" *Cellular Microbiology*, vol. 14, no. 5, pp. 609–621, 2012.
[426] S. Ankri, T. Stolarsky, R. Bracha, D. Mirelman, S. Ankri, T. Stolarsky, R. Bracha, F. Padilla-vaca, and D. Mirelman, "**Antisense Inhibition of Expression of Cysteine Proteinases Affects Entamoeba histolytica - Induced Formation of Liver Abscess in Hamsters Antisense Inhibition of Expression of Cysteine Proteinases Affects Entamoeba histolytica -Induced Formation of Liver A,**" vol. 67, no. 1, pp. 421–423, 1999.
[427] G. Garcia-Rivera, M. A. Rodriguez, R. Ocadiz, M. C. Martinez-Lopez, R. Arroyo, A. Gonzalez-Robles, and E. Orozco, "**Entamoeba histolytica: A novel cysteine protease and an adhesin form the 112 kDa surface protein,**" *Molecular Microbiology*, vol. 33, no. 3, pp. 556–568, 1999.
[428] L. E. Vinall, A. S. Hill, P. Pigny, W. S. Pratt, N. Toribara, J. R. Gum, Y. S. Kim, N. Porchet, J. P. Aubert, and D. M. Swallow, "**Variable number tandem repeat polymorphism of the mucin genes located in the complex on 11p15.5,**" *Human Genetics*, vol. 102, no. 3, pp. 357–366, 1998.
[429] K. Chadee, W. A. Petri, D. J. Innes, and J. I. Ravdin, "**Rat and human colonic mucins bind to and inhibit adherence lectin of Entamoeba histolytica,**" *Journal of Clinical Investigation*, vol. 80, no. 5, pp. 1245–

1254, 1987.
[430] P. H. Davis, J. Schulze, and S. L. Stanley, "**Transcriptomic comparison of two Entamoeba histolytica strains with defined virulence phenotypes identifies new virulence factor candidates and key differences in the expression patterns of cysteine proteases, lectin light chains, and calmodulin,**" *Molecular and Biochemical Parasitology*, vol. 151, no. 1, pp. 118–128, 2007.
[431] C. Ximenez, R. Cerritos, L. Rojas, S. Dolabella, P. Moran, M. Shibayama, E. Gonzalez, A. Valadez, E. Hernandez, O. Valenzuela, A. Limon, O. Partida, and E. F. Silva, "**Human amebiasis: Breaking the paradigm?,**" *International Journal of Environmental Research and Public Health*, vol. 7, no. 3, pp. 1105–1120, 2010.
[432] M. M. De Luca, B. C. Pezzani, M. A. Cordoba, and J. A. Basualdo, "**Characterization and quantitation of parasite species in the effluents of the Berisso main sewage channel, Buenos Aires, Argentina,**" *Zentralbl Hyg Umweltmed*, vol. 200, no. 4, pp. 349–357, Oct. 1997.
[433] C. G. Clark and L. S. Diamond, "**Intraspecific variation and phylogenetic relationships in the genus Entamoeba as revealed by riboprinting,**" *Journal of Eukaryotic Microbiology*, vol. 44, no. 2, pp. 142–154, 1997.
[434] I. K. M. Ali, M. B. Hossain, S. Roy, P. F. Ayeh-Kumi, W. A. Petri, R. Haque, and C. G. Clark, "**Entamoeba moshkovskii infections in children in Bangladesh,**" *Emerging Infectious Diseases*, vol. 9, no. 5, pp. 580–584, 2003.
[435] C. Shimokawa, M. Kabir, M. Taniuchi, D. Mondal, S. Kobayashi, I. K. M. Ali, S. U. Sobuz, M. Senba, E. Houpt, R. Haque, W. A. Petri, and S. Hamano, "**Entamoeba moshkovskii is associated with diarrhea in infants and causes diarrhea and colitis in mice,**" *Journal of Infectious Diseases*, vol. 206, no. 5, pp. 744–751, 2012.
[436] K. Nakada-Tsukui, Y. Saito-Nakano, A. Husain, and T. Nozaki, "**Conservation and function of Rab small GTPases in Entamoeba: Annotation of E. invadens Rab and its use for the understanding of Entamoeba biology,**" *Experimental Parasitology*, vol. 126, no. 3, pp. 337–347, 2010.
[437] J. O. Andersson, R. P. Hirt, P. G. Foster, and A. J. Roger, "**Evolution of four gene families with patchy phylogenetic distributions: influx of genes into protist genomes.,**" *BMC evolutionary biology*, vol. 6, no. 1, p. 27, 2006.
[438] E. Tognotti, "**The rise and fall of syphilis in renaissance Europe,**" *Journal of Medical Humanities*, vol. 30, no. 2, pp. 99–113, 2009.
[439] W. Li, A. Cowley, M. Uludag, T. Gur, H. McWilliam, S. Squizzato, Y. M. Park, N. Buso, and R. Lopez, "**The EMBL-EBI bioinformatics web and programmatic tools framework.,**" *Nucleic acids research*, vol. 43, no. Web Server issue, pp. W580--W580--4, 2015.
[440] S. P. David, N. V. Murthy, E. A. Rabiner, M. R. Munafó, E. C. Johnstone, R. Jacob, R. T. Walton, and P. M. Grasby, "**The genome of the social amoeba Dictyostelium discoideum,**" *Cancer Research*, vol. 25, no. 10, pp. 2586–2590, 2007.
[441] J. E. Strassmann and D. C. Queller, "**How social evolution theory impacts**

our understanding of development in the social amoeba Dictyostelium," *Development Growth and Differentiation*, vol. 53, no. 4, pp. 597–607, 2011.
[442] L. A. Chavez, W. Balamuth, and T. Gong, "**A Light and Electron Microscopical Study of a New, Polymorphic Free-Living Amoeba, Phreatamoeba balamuthi n. g., n. sp.,**" *The Journal of Protozoology*, vol. 33, no. 3, pp. 397–404, 1986.
[443] E. Nyvltova, C. W. Stairs, I. Hrdy, J. Ridl, J. Mach, J. Paces, A. J. Roger, and J. Tachezy, "**Lateral gene transfer and gene duplication played a key role in the evolution of mastigamoeba balamuthi hydrogenosomes,**" *Molecular Biology and Evolution*, vol. 32, no. 4, pp. 1039–1055, 2015.
[444] C. Ku and W. F. Martin, "**A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule,**" *BMC Biology*, vol. 14, no. 1, p. 89, Dec. 2016.
[445] Y. Saito-Nakano, B. J. Loftus, N. Hall, and T. Nozaki, "**The diversity of Rab GTPases in Entamoeba histolytica,**" in *Experimental Parasitology*, 2005, vol. 110, no. 3 SPEC. ISS., pp. 244–252.
[446] S. M. Nauli, F. J. Alenghat, Y. Luo, E. Williams, P. Vassilev, X. Li, A. E. H. Elia, W. Lu, E. M. Brown, S. J. Quinn, D. E. Ingber, and J. Zhou, "**Polycystins 1 and 2 mediate mechanosensation in the primary cilium of kidney cells,**" *Nature Genetics*, vol. 33, no. 2, pp. 129–137, 2003.
[447] X. Madriz, M. B. Martínez, M. A. Rodríguez, G. Sierra, C. Martínez-López, A. M. Riverón, L. Flores, and E. Orozco, "**Expression in fibroblasts and in live animals of Entamoeba histolytica polypeptides EhCP112 and EhADH112,**" *Microbiology*, vol. 150, no. 5, pp. 1251–1260, 2004.
[448] L. I. Quintas-Granados, E. Orozco, L. G. Brieba, R. Arroyo, and J. Ortega-López, "**Purification, refolding and autoactivation of the recombinant cysteine proteinase EhCP112 from Entamoeba histolytica,**" *Protein Expression and Purification*, vol. 63, no. 1, pp. 26–32, 2009.
[449] R. Ocadiz, E. Orozco, E. Carrillo, L. I. Quintas, J. Ortega-Lopez, R. M. Garcia-Perez, T. Sanchez, B. A. Castillo-Juarez, G. Garcia-Rivera, and M. A. Rodriguez, "**EhCP112 is an Entamoeba histolytica secreted cysteine protease that may be involved in the parasite-virulence,**" *Cellular Microbiology*, vol. 7, no. 2, pp. 221–232, 2005.
[450] R. Bracha, D. Kobiler, and D. Mirelman, "**Attachment and ingestion of bacteria by trophozoites of Entamoeba histolytica,**" *Infection and Immunity*, vol. 36, no. 1, pp. 396–406, Apr. 1982.
[451] M. A. Rodriguez and E. Orozco, "**Isolation and characterization of phagocytosis- and virulence-deficient mutants of Entamoeba histolytica,**" *J Infect Dis*, vol. 154, no. 1, pp. 27–32, Jul. 1986.
[452] A. Hernandez-Flores, M. de J. Almaraz-Barrera, D. Lozano-Amado, J. Correa-Basurto, A. Rojo-Dominguez, E. Luna-Rivera, M. Schnoor, N. Guillen, R. Hernandez-Rivas, and M. Vargas, "**A new nucleocytoplasmic RhoGAP protein contributes to control the pathogenicity of Entamoeba histolytica by regulating EhRacC and EhRacD activity,**" *Cellular Microbiology*, vol. 18, no. 11, pp. 1653–1672, Nov. 2016.
[453] J. Huang, "**Horizontal gene transfer in eukaryotes: The weak-link model,**" *BioEssays*, vol. 35, no. 10, pp. 868–875, 2013.
[454] B. Nerima, D. Nilsson, and P. Mäser, "**Comparative genomics of**

metabolic networks of free-living and parasitic eukaryotes.," *BMC Genomics*, vol. 11, p. 217, Mar. 2010.

[455] Y. I. Wolf and E. V Koonin, "**Genome reduction as the dominant mode of evolution,**" *BioEssays*, vol. 35, no. 9, pp. 829–837, Sep. 2013.

[456] H. G. Morrison, 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, "**Genomic Minimalism in the Early Diverging Intestinal Parasite Giardia lamblia,**" *Science*, vol. 317, no. 5846, pp. 1921–1926, 2007.

[457] J. J. Morris, R. E. Lenski, and E. R. Zinser, "**The black queen hypothesis: Evolution of dependencies through adaptive gene loss,**" *mBio*, vol. 3, no. 2, pp. 1–7, May 2012.

[458] S. J. Giovannoni, H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappé, J. M. Short, J. C. Carrington, and E. J. Mathur, "**Genome Streamlining in a Cosmopolitan Oceanic Bacterium,**" *Science*, vol. 309, no. 5738, pp. 1242–1245, Aug. 2005.

[459] A. Dufresne, L. Garczarek, and F. Partensky, "**Accelerated evolution associated with genome reduction in a free-living prokaryote.,**" *Genome biology*, vol. 6, no. 2, p. R14, 2005.

[460] V. S. Cooper, D. Schneider, M. Blot, and R. E. Lenski, "**Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of Escherichia coli B,**" *Journal of Bacteriology*, vol. 183, no. 9, pp. 2834–2841, May 2001.

[461] R. Gil, B. Sabater-Munoz, A. Latorre, F. J. Silva, and A. Moya, "**Extreme genome reduction in Buchnera spp.: Toward the minimal genome needed for symbiotic life,**" *Proceedings of the National Academy of Sciences*, vol. 99, no. 7, pp. 4454–4458, 2002.

[462] F. J. Silva, A. Latorre, and A. Moya, "**Genome size reduction through multiple events of gene disintegration in Buchnera APS,**" *Trends in Genetics*, vol. 17, no. 11. pp. 615–618, 2001.

[463] I. J. Tsai, M. Zarowiecki, N. Holroyd, A. Garciarrubio, A. Sanchez-Flores, K. L. Brooks, A. Tracey, R. J. Bobes, G. Fragoso, E. Sciutto, M. Aslett, H. Beasley, H. M. Bennett, J. Cai, F. Camicia, R. Clark, M. Cucher, N. De Silva, T. A. Day, P. Deplazes, K. Estrada, C. Fernández, P. W. H. H. Holland, J. Hou, S. Hu, T. Huckvale, S. S. Hung, L. Kamenetzky, J. A. Keane, F. Kiss, U. Koziol, O. Lambert, K. Liu, X. Luo, Y. Luo, N. MacChiaroli, S. Nichol, J. Paps, J. Parkinson, N. Pouchkina-Stantcheva, N. Riddiford, M. Rosenzvit, G. Salinas, J. D. Wasmuth, M. Zamanian, Y. Zheng, X. Cai, X. Soberon, P. D. Olson, J. P. Laclette, K. Brehm, M. Berriman, E. Morett, T. Portillo, M. V Jose, J. C. Carrero, C. Larralde, J. Morales-Montor, J. Limon-Lason, M. A. Cevallos, V. Gonzalez, A. Ochoa-Leyva, A. Landa, L. Jimenez, V. Valdes, A. Garciarrubio, R. J. Bobes, G. Fragoso, A. Sánchez-Flores, K. Estrada, M. A. Cevallos, E. Morett, V. González, T. Portillo, A. Ochoa-Leyva, M. V José, E. Sciutto, A. Landa, L. Jiménez, V. Valdés, J. C. Carrero, C. Larralde, J. Morales-Montor, J. Limón-Lason, X. Soberón, J. P. Laclette, X. Cai, X.

Soberón, P. D. Olson, J. P. Laclette, K. Brehm, and M. Berriman, "**The genomes of four tapeworm species reveal adaptations to parasitism,**" *Nature*, vol. 496, no. 7443, pp. 57–63, Mar. 2013.
[464] P. D. Spanu, "**The Genomics of Obligate (and Nonobligate) Biotrophs,**" *Annual Review of Phytopathology*, vol. 50, no. 1, pp. 91–109, Sep. 2012.
[465] C. W. Quist, G. Smant, and J. Helder, "**Evolution of Plant Parasitism in the Phylum Nematoda,**" *Annual Review of Phytopathology*, vol. 53, no. 1, pp. 289–310, Aug. 2015.
[466] L. Mendoza, F. Hernandez, and L. Ajello, "**Life-Cycle of the Human and Animal Oomycete Pathogen Pythium-Insidiosum,**" *Journal of Clinical Microbiology*, vol. 31, no. 11, pp. 2967–2973, 1993.
[467] M. Crook, "**The dauer hypothesis and the evolution of parasitism: 20years on and still going strong,**" *International Journal for Parasitology*, vol. 44, no. 1. pp. 1–8, 2014.
[468] J. Fosu-Nyarko, P. Nicol, F. Naz, R. Gill, and M. G. K. Jones, "**Analysis of the transcriptome of the infective stage of the beet cyst nematode, H. schachtii,**" *PLoS ONE*, vol. 11, no. 1, 2016.
[469] V. L. Hunt, I. J. Tsai, A. Coghlan, A. J. Reid, N. Holroyd, B. J. Foth, A. Tracey, J. A. Cotton, E. J. Stanley, H. Beasley, H. M. Bennett, K. Brooks, B. Harsha, R. Kajitani, A. Kulkarni, D. Harbecke, E. Nagayasu, S. Nichol, Y. Ogura, M. A. Quail, N. Randle, D. Xia, N. W. Brattig, H. Soblik, D. M. Ribeiro, A. Sanchez-flores, T. Hayashi, T. Itoh, D. R. Denver, J. D. Stoltzfus, J. B. Lok, H. Murayama, A. Streit, T. Kikuchi, M. Viney, M. Berriman, W. Grant, J. D. Stoltzfus, J. B. Lok, H. Murayama, J. Wastling, A. Streit, T. Kikuchi, M. Viney, and M. Berriman, "**The Genomic Basis of Parasitism in the Strongyloides Clade of Nematodes,**" *Nature Genetics*, vol. 48, no. 3, pp. 299–307, Feb. 2016.

Supplementary legends

S4.3.1 Output from OrthoMCL clustering of Stramenopile genomes. 'groups.txt' contains the raw output from the programme. Other sheets contain the breakdown of how the raw output was interpreted. The numbers of sequences that contributed to these groups from each genome is included in the 'summary' sheet. Refer to Section 4.2.3 for cluster definitions.

S4.3.2 Output from OrthoFinder clustering of Stramenopile genomes. 'OrthologousGroups.txt' contains the raw output from the programme. Other sheets contain the breakdown of how the raw output was interpreted. The numbers of sequences that contributed to these groups from each genome is included in the 'summary' sheet. Refer to Section 4.2.3 for cluster definitions.

S4.3.3 Full table of results for the enrichment of OrthoMCL *Blastocystis* sp. STs gains. The columns contain the domain annotation, the number of *Blastocystis*-specific sequences annotated with a domain, the total number of sequences in the *Blastocystis* sp. STs genomes annotated with a domain, the total number of genes in the *Blastocystis* sp. STs genomes that are not annotated with a domain, the total number of *Blastocystis*-specific sequences, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *Blastocystis* sp. STs genomes with a domain annotation.

S4.3.4 Full table of results for the enrichment of OrthoFinder *Blastocystis* sp. STs gains. The columns contain the domain annotation, the number of *Blastocystis*-specific sequences annotated with a domain, the total number of sequences in the *Blastocystis* sp. STs genomes annotated with a domain, the total number of genes in the *Blastocystis* sp. STs genomes that are not annotated with a domain, the total number of *Blastocystis*-specific sequences, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *Blastocystis* sp. STs genomes with a domain annotation.

S4.3.5 Full table of results for the enrichment of OrthoMCL *Blastocystis* sp. STs losses. The columns contain the domain annotation, the number of *P. lacertae* sequences with no orthologue in *Blastocystis* annotated with a domain, the total number of sequences in the *P. lacertae* genome annotated with a domain, the total number of genes in the *P. lacertae* genome that are not annotated with a domain, the total number of *P. lacertae* sequences with no orthologue in *Blastocystis*, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *P. lacertae* genome with a domain annotation.

S4.3.6 Full table of results for the enrichment of OrthoFinder *Blastocystis* sp. STs losses. The columns contain the domain annotation, the number of *P. lacertae* sequences with no orthologue in *Blastocystis* annotated with a domain, the total number of sequences in the *P. lacertae* genome annotated with a domain, the total number of genes in the *P. lacertae* genome that are not annotated with a domain, the total number of *P. lacertae* sequences with no orthologue in *Blastocystis*, the P-

value for the enrichment, the significance (* p < 0.1, ** p < 0.05, *** p < 0.01) and the gene identifiers for all the genes in the *P. lacertae* genome with a domain annotation.

S4.3.7 Full table of results for the enrichment of OrthoMCL *P. lacertae* gains. The columns contain the domain annotation, the number of *P. lacertae*-specific sequences annotated with a domain, the total number of sequences in the *P. lacertae* genome annotated with a domain, the total number of genes in the *P. lacertae* genome that are not annotated with a domain, the total number of *P. lacertae*-specific sequences, the P-value for the enrichment, the significance (* p < 0.1, ** p < 0.05, *** p < 0.01) and the gene identifiers for all the genes in the *P. lacertae* genome with a domain annotation.

S4.3.8 Full table of results for the enrichment of OrthoFinder *P. lacertae* gains. The columns contain the domain annotation, the number of *P. lacertae*-specific sequences annotated with a domain, the total number of sequences in the *P. lacertae* genome annotated with a domain, the total number of genes in the *P. lacertae* genome that are not annotated with a domain, the total number of *P. lacertae*-specific sequences, the P-value for the enrichment, the significance (* p < 0.1, ** p < 0.05, *** p < 0.01) and the gene identifiers for all the genes in the *P. lacertae* genome with a domain annotation.

S4.3.9 Full table of results for the enrichment of OrthoMCL *P. lacertae* losses. The columns contain the domain annotation, the number of *Blastocystis* sp. STs sequences with no orthologue in *P. lacertae* annotated with a domain, the total number of sequences in the *Blastocystis* sp. STs genomes annotated with a domain, the total number of genes in the *Blastocystis* sp. STs genomes that are not annotated with a domain, the total number of *Blastocystis* sp. STs sequences with no orthologue in *P. lacertae*, the P-value for the enrichment, the significance (* p < 0.1, ** p < 0.05, *** p < 0.01) and the gene identifiers for all the genes in the *Blastocystis* sp. STs genomes with a domain annotation.

S4.3.10 Full table of results for the enrichment of OrthoFinder *P. lacertae* losses. The columns contain the domain annotation, the number of *Blastocystis* sp. STs sequences with no orthologue in *P. lacertae* annotated with a domain, the total number of sequences in the *Blastocystis* sp. STs genomes annotated with a domain, the total number of genes in the *Blastocystis* sp. STs genomes that are not annotated with a domain, the total number of *Blastocystis* sp. STs sequences with no orthologue in *P. lacertae*, the P-value for the enrichment, the significance (* p < 0.1, ** p < 0.05, *** p < 0.01) and the gene identifiers for all the genes in the *Blastocystis* sp. STs genomes with a domain annotation.

S4.3.11 Full table of gene numbers mapped to metabolic (KEGG) pathways in *Blastocystis* sp. STs, *P. lacertae* and *C. roenbergensis.* Conserved are those found in all genomes, specific are those found in only one genome. ‘Blast sp’ are those only found in two or more *Blastocystis* genomes and ‘Blast loss’ are those found in only in *P. lacertae* and *C. roenbergensis*.

S4.3.12 List of gene names and species from 4.3.8.

S4.3.13 Output from OrthoMCL clustering of Stramenopile putative secretome. Sheets contain the breakdown of how the raw output was interpreted. The numbers of sequences that contributed to these groups from each genome is included in the 'summary' sheet.

S4.3.14 Full table of results for the enrichment of OrthoMCL *Blastocystis* sp. ST4 putative secretome gains. The columns contain the domain annotation, the number of *Blastocystis* sp. ST4-specific sequences annotated with a domain, the total number of sequences in the *Blastocystis* sp. ST4 putative secretome annotated with a domain, the total number of genes in the *Blastocystis* sp. ST4 putative secretome that are not annotated with a domain, the total number of *Blastocystis* sp. ST4-specific sequences, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *Blastocystis* sp. ST4 putative secretome with a domain annotation.

S4.3.15 Full table of results for the enrichment of OrthoMCL *Blastocystis* sp. ST7 putative secretome gains. The columns contain the domain annotation, the number of *Blastocystis* sp. ST7-specific sequences annotated with a domain, the total number of sequences in the *Blastocystis* sp. ST7 putative secretome annotated with a domain, the total number of genes in the *Blastocystis* sp. ST7 putative secretome that are not annotated with a domain, the total number of *Blastocystis* sp. ST7-specific sequences, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *Blastocystis* sp. ST7 putative secretome with a domain annotation.

S5.3.1 Full table of gene numbers mapped to metabolic (KEGG) pathways in *E. histolytica, M. balamuthi, D. discoideum* and the inclusive dataset for *Mastigamoeba* sp. 'Conserved' are those found in all genomes, 'specific' are those found in only one genome. 'Mastig sp' contains pathways found only in the mastigamoebids. Pairwise sheets contain pathways in only two genomes and 'loss' sheets contain pathways present in all genomes but one.

S5.3.2 Full table of gene numbers mapped to metabolic (KEGG) pathways in *E. histolytica, M. balamuthi, D. discoideum* and the exclusive dataset for *Mastigamoeba* sp. 'Conserved' are those found in all genomes, 'specific' are those found in only one genome. 'Mastig sp' contains pathways found only in the mastigamoebids. Pairwise sheets contain pathways in only two genomes and 'loss' sheets contain pathways present in all genomes but one.

S5.3.3 Output from OrthoMCL clustering of Archamoebae genomes, using the inclusive dataset for *Mastigamoeba* sp. 'MCL_Inclusive.txt' contains the raw output from the programme. Other sheets contain the breakdown of how the raw output was interpreted. The numbers of sequences that contributed to these groups from each genome is included in the 'summary' sheet. Refer to Section 4.2.3 for cluster definitions.

S5.3.4 Output from OrthoFinder clustering of Archamoebae genomes, using the inclusive dataset for *Mastigamoeba* sp. 'FDR_Inclusive.txt' contains the raw output from the programme. Other sheets contain the breakdown of how the raw output was interpreted. The numbers of sequences that contributed to these groups from each genome is included in the 'summary' sheet. Refer to Section 4.2.3 for cluster definitions.

S5.3.5 Full table of results for the enrichment of OrthoMCL *E. histolytica* gains. The columns contain the domain annotation, the number of *E. histolytica*-specific sequences annotated with a domain, the total number of sequences in the *E. histolytica* genome annotated with a domain, the total number of genes in the *E. histolytica* genome that are not annotated with a domain, the total number of *E. histolytica*-specific sequences, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *E. histolytica* genome with a domain annotation.

S5.3.6 Full table of results for the enrichment of OrthoFinder *E. histolytica* gains. The columns contain the domain annotation, the number of *E. histolytica*-specific sequences annotated with a domain, the total number of sequences in the *E. histolytica* genome annotated with a domain, the total number of genes in the *E. histolytica* genome that are not annotated with a domain, the total number of *E. histolytica*-specific sequences, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *E. histolytica* genome with a domain annotation.

S5.3.7 Full table of results for the enrichment of OrthoMCL *E. histolytica* losses. The columns contain the domain annotation, the number of *Mastigamoeba s*p. sequences with no orthologue in *E. histolytica* annotated with a domain, the total number of sequences in the *Mastigamoeba* sp. transcriptome annotated with a domain, the total number of genes in the *Mastigamoeba* sp. transcriptome that are not annotated with a domain, the total number of *Mastigamoeba* sp. sequences with no orthologue in *E. histolytica*, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *Mastigamoeba* sp. transcriptome with a domain annotation.

S5.3.8 Full table of results for the enrichment of OrthoFinder *E. histolytica* losses. The columns contain the domain annotation, the number of *Mastigamoeba s*p. sequences with no orthologue in *E. histolytica* annotated with a domain, the total number of sequences in the *Mastigamoeba* sp. transcriptome annotated with a domain, the total number of genes in the *Mastigamoeba* sp. transcriptome that are not annotated with a domain, the total number of *Mastigamoeba* sp. sequences with no orthologue in *E. histolytica*, the P-value for the enrichment, the significance (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$) and the gene identifiers for all the genes in the *Mastigamoeba* sp. transcriptome with a domain annotation.