Comparative genomic analyses of *Entamoeba* species

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

Ian William Wilson

September 2014

UNIVERSITY OF LIVERPOOL

Acknowledgements

Firstly, I'd like to thank my supervisors Neil and Steve for all of their support and guidance over the past four years. I'm particularly indebted to Neil for giving me the opportunity to do this project and to visit parts of the world I never thought I'd get to see whilst doing so. If the bottom ever falls out of the whole genomics scene, he'd make an outstanding New York City tour guide!

I'd also like to thank everyone in the CGR for their efforts in preparing and sequencing my DNA libraries, as well as anyone in the Institute of Integrative Biology who gave me advice whenever I needed it. Gareth Weedall deserves special mention here for effectively acting as my third supervisor, even after he stopped working on *Entamoeba*! Simply put, I owe that man a lot.

Outside of the Institute, Graham Clark in the London School of Hygiene and Tropical Medicine has given me live cultures, DNA extracts, information and advice, without complaint or expectation of anything in return. For this he has my sincerest gratitude. I'd also like to thank Nancy Guillén, Chung-Chau Hon and Marc Deloger at the Institut Pasteur for contributing toward the *E. moshkovskii* annotation, and Seiki Kobayashi of Tomo Nozaki's group in Tokyo, who provided the *E. dispar* AS16IR sample without which *E. dispar* would not have even featured in Chapter Three!

Special thanks go to a few others too - tea and coffee for keeping me awake during the long days and nights; Joshua Radin, who will never read this but whose music kept me calm; and my amazing friends who kept me as near to sane as I'll ever get with support, advice and unexpected food deliveries!

Finally, I want to express overwhelming gratitude to my parents and sister for their love and support throughout this PhD, not least during the final year when they shared in every stress and struggle to help me across the finish line. I couldn't have conquered this behemoth without you.

Abstract

Amoebiasis is the third-most common cause of mortality worldwide from a disease borne of a parasitic infection. It affects up to 50 million people annually, of which 40,000 to 100,000 cases are fatal. *Entamoeba histolytica* is an obligate protozoon parasite of humans and is the aetiological agent of the disease. Recent suggestions that other members of the *Entamoeba* genus are human-infective, and potentially pathogenic, have been investigated here. A draft assembly and annotation of the 25 Mb genome of *E. moshkovskii* strain Laredo is presented, to which multiple *E. moshkovskii* strains were mapped. The *E. moshkovskii* genome was found to be approximately 200 times more variable than that of *E. histolytica*. Performance of the four-haplotype test revealed that genetic recombination does not seem to occur in *E. moshkovskii*. As such, it is suggested that it be referred to as a 'species complex', rather than an individual species. A comparative genomic analysis of *E. histolytica* HM-1:IMSS, *E. moshkovskii* Laredo, *E. invadens* IP-1 and the avirulent *E. dispar* SAW760 was performed. Subsequent comparative analyses against members of genera representative of the diversity in the Unikonts clade enabled the identification of orthologous gene families unique to the *Entamoeba* genus. Analysis of virulence factors within this set revealed that gene families involved in adhesion of amoebic trophozoites to host cells play a key role in the development of invasive amoebiasis. The Gal/GalNAc lectins and members of the BspA family are of particular interest, being present in all analysed species, except for *E. dispar*. The presence of these key families, plus cysteine proteases, in the *E. moshkovskii* genome suggests that some sequence types within this species complex may be pathogenic. *E. invadens* was found to possess larger numbers of more variable genes within many virulence factor families, including the BspA family and the Gal/GalNAc lectins. This suggests that sequence diversity facilitates *E. invadens*' polyxenous lifestyle. Finally, a novel species recently isolated from a human faecal sample - *E. bangladeshi*, strain 8237 – was sequenced. Its genome was assembled using multiple *de novo* genome assemblers and coding sequences were assembled individually. A combination of all methods tested was found to be beneficial in maximising the number of gene sequences assembled, which is advised as good practice in future similar assemblies. The phylogeny of *E. bangladeshi*, achieved using the combined assemblies' outputs suggested that the novel species is human-infective. The work presented here utilised modern comparative genomic techniques to improve understanding of *Entamoeba* species, their capacity for causing disease and their potential impact upon the epidemiology of amoebiasis.

Table of Contents

Figure Legends

Chapter One

Chapter Two

Chapter Three

Chapter Four

Table Legends

Chapter Two

Chapter Three

Chapter Four

List of abbreviations

ABS	Adult bovine serum
ALA	Amoebic liver abscess
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Aligner
CCD	Charge-coupled device
CDS	Coding sequence
CEG	Core eukaryotic gene
CEGMA	Core Eukaryotic Genes Mapping Approach program
CGR	Centre for Genomic Research
COG	Cluster of Orthologous Groups
CPBP	Cysteine protease binding protein
CRD	Carbohydrate recognition domain
CRT	Cyclic reversible termination
CTAB	Cetyltrimethylammonium bromide
dN	SNPs per non-synonymous site in a DNA coding sequence
dS	SNPs per synonymous site in a DNA coding sequence
dN/dS	Ratio of dN to dS, calculated as dN divided by dS
dN-dS	Difference between dN and dS, calculated as dN minus dS
E-value	Exponent value (in NCBI BLAST search)
EhSTIRP	*Entamoeba histolytica* serine-, threonine- and isoleucine-rich protein
FPKM	Fragments per kilobase per million
GO	Gene ontology
ICDDR	International Centre for Diarrhoeal Disease Research
LINE	Long interspersed elements
LTR	Long terminal repeat
NCBI	National Center for Biotechnology Information
ORF	Open reading frame
PAML	Phylogenetic Analysis Using Maximum Likelihood package
PCR	Polymerase chain reaction
PE	Paired end
PHYLIP	Phylogeny Inference Package
PRANK	Probabilistic Alignment Kit
RBH	Reciprocal best hit (in BLAST searches)
ROS	Reactive oxygen species

SINE	Short interspersed elements
SNP	Single nucleotide polymorphism
SOD	Superoxide dismutase
SREHP	Serine-rich *Entamoeba histolytica* protein
STR	Short tandem repeat
TAGC	Taxon-Annotated-GC-Coverage plot
TE	Transposable element
TLR	Toll-like receptor
TMRCA	Time to most recent common ancestor
TNF-α	Tumour necrosis factor alpha

Chapter One – Introduction

1.1 Amoebiasis and *Entamoeba histolytica*

Amoebiasis is the third-most common cause of mortality worldwide from a disease borne of a parasitic infection. It affects up to 50 million people annually, of which 40,000 to 100,000 cases are fatal [1]. The species defined as the aetiological agent of amoebiasis in humans is the obligate parasite *Entamoeba histolytica*. The *Entamoeba* genus is, however, relatively poorly understood. In recent years, questions have been raised over the contributions made by other species within the genus towards the disease's prevalence [2-4]. In this thesis, genomic analyses have been used to study members of the genus in a bid to improve understanding of their roles as causative agents and proliferators of amoebiasis, as well as the genes that pathogenic *Entamoeba* species and strains require to cause the disease.

1.1.1 Symptoms and prevalence of amoebiasis

E. histolytica is transmitted between human hosts by a faecal-oral route. As such, it is particularly prevalent in areas of poor hygiene where wastewater and drinking or bathing water are not kept separate. Two such regions in which amoebiasis is known to be endemic are the Mirpur slums in Dhaka, Bangladesh and Hué City in Vietnam, as will be discussed in greater detail in Section 1.2. Cases of amoebiasis are also seen in more affluent countries, however. Travellers returning from endemic regions are at heightened risk, as are people who engage in oral or anal sex, most commonly homosexual men [5, 6]. Outbreaks have also been documented in institutionalised individuals in Japan and the Philippines, where occurrences of infections with *E. histolytica* are increasing [7, 8]. Furthermore, accidental contamination of municipal water supplies can result in outbreaks of *E. histolytica* infections as happened in 1998 in the Republic of Georgia [9], highlighting the dangers of contaminated water supplies.

The clinical manifestations of *E. histolytica* infections range in severity, as not all cases progress to a disease state. Asymptomatic infections make up over 90% of cases, although precise numbers are difficult to ascertain for reasons explained in Section 1.4 [1]. The majority of hosts that develop symptomatic infections experience abdominal pains and dysentery as the parasites are contained within the intestinal

tract. However, in relatively rare cases, the disease can progress to an extra-intestinal disease state, most commonly resulting in fatal amoebic liver abscesses (ALA) [1, 10]. The reasons for these symptoms can best be explained by studying the life cycle of the parasite.

1.1.2 Life cycle of *Entamoeba histolytica*

E. histolytica has a two-stage life cycle (Figure 1.1.1; [10]). In the environment it exists as infective quadrinucleate cysts before being ingested by a human host. Upon reaching the ileum, the amoebae undergo excystation, developing into potentially pathogenic trophozoites. In this form, the parasites can inhabit the intestine asymptomatically as allochthonous members of the gastrointestinal tract flora [11], phagocytosing enteric bacteria as a source of nutrients [12] and multiplying by binary fission. Invasive amoebiasis occurs when the trophozoites instead degrade the mucosal layer of the lumen epithelium before attaching to, and invading, the epithelium itself. This causes characteristic flask-shaped ulcers (Figure 1.1.2) and inflammation of the intestinal wall, leading to a loss of absorptive function, and dysenteric symptoms. In either eventuality, some trophozoites will subsequently encyst and pass out in host faeces into the environment, completing the life cycle.

Extra-intestinal infections can occur when trophozoites degrading the intestinal wall enter the bloodstream and are disseminated to other organs, most commonly the liver. Fatal ALAs develop as a result of trophozoites inducing apoptosis in hepatic immune cells, releasing lytic enzymes. The enzymes lyse parenchymal cells and the resulting necrotic foci often coalesce, forming abscesses, destroying the liver's ability to function [13]. Many undefined or unexplained factors contribute towards the extent and severity of infections, including the strain or cell line of the infective cells, with some capable of inducing a more virulent infection than others [14, 15]. These differences are caused, or indicated, by genomic polymorphisms [16, 17].

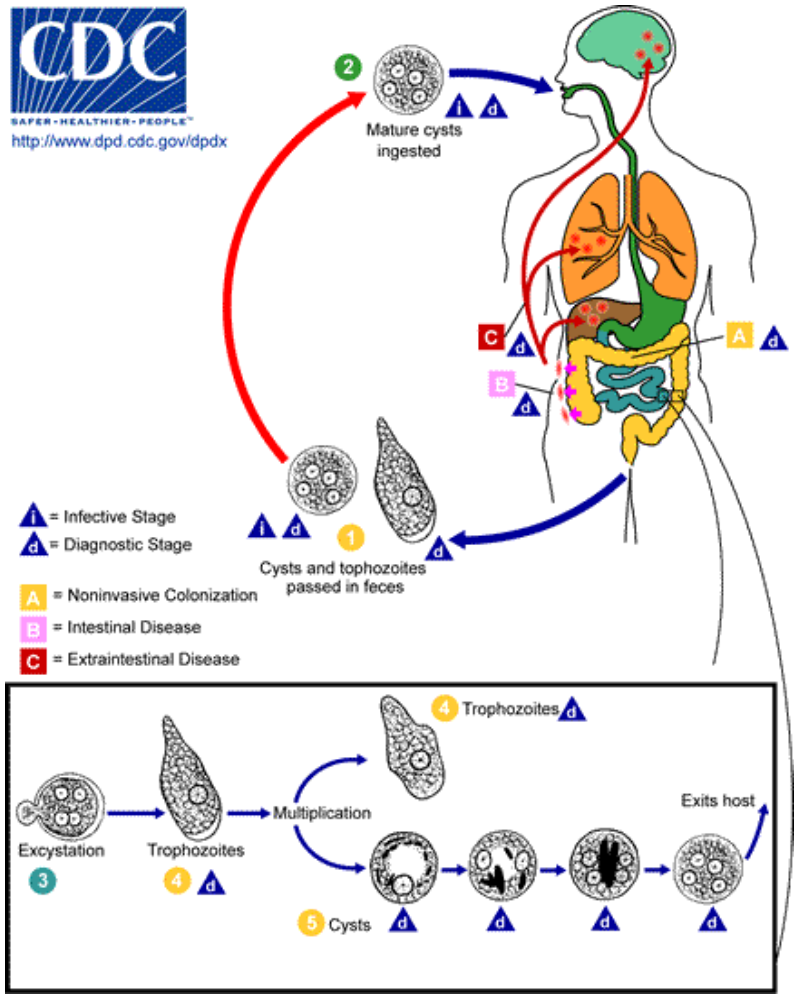


Figure 1.1.1. The life cycle of *Entamoeba histolytica*. The three broadly defined routes an infection can take in a human host are labelled 'A', 'B' and 'C'. Stages of the life cycle common to all three types of infection are labelled in chronological order (1-5). Image provided courtesy of the US Centers for Disease Control and Prevention.

Figure 1.1.2. Flask-shaped ulcer in intestinal mucosa, characteristic of amoebic colitis. Image provided courtesy of the US Centers for Disease Control and Prevention.

1.1.3 Drugs available for treatment of amoebiasis

Whilst prevention is better than cure, the transmission route of *E. histolytica* makes the former near impossible in certain regions of the world. There is a range of amoebicides currently in use to treat infections and a number of regimens by which to administer them. For patients with asymptomatic amoebiasis or mild, non-invasive intestinal amoebiasis, the drug iodoquinol is the most effective treatment. It can also be used in conjunction with other treatments in more severe infections. Iodoquinol is effective against trophozoites and cysts, acting at the point of infection in the large intestine, but by a poorly defined mechanism [18]. Where iodoquinol is unavailable or ineffective, an alternative is diloxanide furoate. Acting via an unknown mechanism, this drug is used as a supplement to others in treating asymptomatic infections or non-invasive intestinal amoebiasis [18, 19].

Metronidazole is the most effective treatment against invasive and extra-intestinal amoebiasis. When followed by a luminal amoebicide such as iodoquinol or diloxanide furoate, complete clearance of an infection is highly probable. Recommended dosages and regimens for the 5'-nitroimidazole antibiotic vary between sources and severity of infection. The World Health Organisation advises a regimen of 10 mg/kg three times a day for 8-10 days [20]. However, for critically ill adults with extensive ALAs, regimens of 500 mg (intravenously) or 800 mg (orally) three times a day for 5 – 10 days have been recommended. Critically ill children can be given 50 mg/kg every day for 10 days, orally [21].

In extreme cases, where metronidazole treatment has been ineffective, other, potentially dangerous, drugs might be prescribed. Paramomycin is an orally delivered aminoglycoside amoebicide effective in treating intestinal amoebiasis [22] and patients in comas resulting from liver damage. However, it is ineffective against extra-intestinal amoebae themselves, and it has many serious side effects [18]. The final alternative is dehydroemetine. Owing to its toxicity and the fact that it is an irritant when taken orally, it is delivered by injections directly into muscle tissue. It inhibits protein synthesis, but can have serious complications in sufferers of cardiac issues. Chloroquine and needle aspiration of abscesses are also recommended in such extreme cases [18, 20].

1.2 Epidemiology of *Entamoeba histolytica* in disease foci

As mentioned in Section 1.1.1, cases of amoebiasis are predominantly seen in areas of developing countries with particularly poor levels of hygiene. Two notable regions of interest include the slums of Dhaka, Bangladesh and a densely populated area of Hué, Vietnam. Large bodies of research have been carried out in recent years in both regions in a bid to understand the causes and dynamics of amoebiasis, generating some important results.

1.2.1 Mirpur thana of Dhaka, Bangladesh

Spearheaded by Dr William Petri Jr's team at the University of Virginia and Dr Rashidul Haque's team at the International Centre for Diarrhoeal Disease Research (ICDDR), research into amoebiasis in the Mirpur slum of Dhaka has been ongoing for over two decades [3, 4, 16, 23-27]. One of the earliest studies from the group demonstrated genetic diversity within the endemic region by studying the serine-rich *Entamoeba histolytica* protein (SREHP) in children [16]. More specifically, they identified polymorphisms between amoebae isolated from intestines and those isolated from extra-intestinal sites, implying that there is a genetic cause to the development of ALAs. A separate series of studies investigated the Gal/GalNAc lectin heavy subunit in *E. histolytica* [23, 24, 27]. Their findings identified sequence conservation in the subunit's carbohydrate recognition domain (CRD) against which IgA antibody responses were directed. This naturally acquired, but incomplete, host immunity reduced colonisation of the intestine by *E. histolytica*, identifying the lectin subunit as a potential vaccine target.

Two other important papers have focused on the little-studied *Entamoeba moshkovskii*. One studied the prevalence of *E. moshkovskii* in a group of children, producing evidence to suggest that the species was human-infective, contradicting existing beliefs [3]. In the second study, mice intra-caecally infected with *E. moshkovskii* trophozoites were found to suffer from amoebiasis, whilst children experiencing diarrhoea tested positive for infection by *E. moshkovskii* [4]. Taken together, these results strongly implied that *E. moshkovskii* is not only human-infective but also pathogenic. Genomic analyses aiming to corroborate or refute these theories make up large proportions of Chapters Two and Three of this thesis. Finally, one of the most noteworthy findings to come from the Bangladeshi team in recent years was the

discovery of a novel human-infective *Entamoeba* species – *Entamoeba bangladeshi* [26]. The genetic analysis of this species forms the basis of the third and final data chapter of this thesis.

1.2.2 Hué, Vietnam

Hué is the third largest city in Vietnam, with a population of approximately 300,000. High rates of invasive amoebiasis cases have been diagnosed in the surrounding province Thua Thien Hué, with a record 21 cases per 100,000 inhabitants per annum diagnosed with ALAs between 1990 and 1998 [28]. A series of collaborations between Prof Egbert Tannich's research group at the Bernhard Nocht Institute for Tropical Medicine and the University of Hué has investigated the cause of this high rate of infection [29]. They found that male adults were at greatest risk of suffering from ALAs, and that the number of extra-intestinal amoebiasis cases increased during the summer season and in more densely populated areas. However, greater infection rates overall were seen in females, and poor access to hygiene facilities and education increased the risk of infection [28]. This supported, to a degree, previous findings that males were more susceptible to development of invasive amoebiasis than females, despite similar incidences of asymptomatic infections [30].

These results suggest that host factors play an important role in defining the course of an infection, with the initiation of invasive or extra-intestinal disease not being exclusively determined by the parasite. This was subsequently corroborated by findings suggesting that, in mice, a small number of host genes determined host resistance to invasive amoebiasis when challenged by pathogenic *E. histolytica* [31]. Furthermore, a 15 month longitudinal study carried out by the group revealed that the endemicity in the Thua Thien Hué province was a result of reinfection, supporting the findings from a longitudinal study in Bangladesh [23]. Taken together, the results from Vietnam highlight the socio-economic factors that need to be accounted for when treating amoebiasis.

1.3 Other *Entamoeba* species important to research into amoebiasis

1.3.1 *Entamoeba dispar*

For decades, *E. histolytica* was thought to be the only human-infective member of the genus. However, the low rate of invasive amoebiasis observed remained unexplained. We know now that this was caused, in part, by the existence of a second, non-invasive, member of the *Entamoeba* genus - *Entamoeba dispar* [2]. Morphologically identical to *E. histolytica*, and closely related (Figure 1.3.1), *E. dispar* is infective to humans, but is thought to be avirulent [2, 32], despite liver-derived clinical isolates of *E. dispar* bringing its avirulence into question recently [33]. The differences in virulence capabilities seen between *E. dispar* and *E. histolytica* have been exploited by various groups attempting to determine which proteins may enable virulence capabilities in *E. histolytica* but not in *E. dispar* [34, 35]. Virulence factors identified in *E. histolytica* are discussed in detail in Section 1.5.

1.3.2 *Entamoeba moshkovskii*

A more distantly related species, *Entamoeba moshkovskii*, was originally thought to be free-living and therefore non-pathogenic [36-38]. However, as with *E. dispar*, human-derived clinical isolates and cases of diarrhoea directly associated with *E. moshkovskii* infection, as described above, have challenged this assumption [3, 4]. So, too, has a strain isolated from a snake in 1948 (unpublished). Since the latter study from Bangladesh, the ability of *E. moshkovskii* to cause invasive and symptomatic amoebiasis has been of great interest. The mounting evidence is leading to calls for the re-examination of *E. moshkovskii*'s pathogenicity, a goal that forms the basis of Chapter Two, and a significant proportion of Chapter Three in this thesis. *E. moshkovskii* is closely related to, and morphologically identical to, *E. histolytica* and *E. dispar* (Figure 1.3.1). Several studies have attempted to design polymerase chain reaction (PCR)-based methods of differentiating between the three species in order to improve accurate detection of the individual species [39-43] and, in some cases, to investigate *E. moshkovskii*'s ability to cause disease [44].

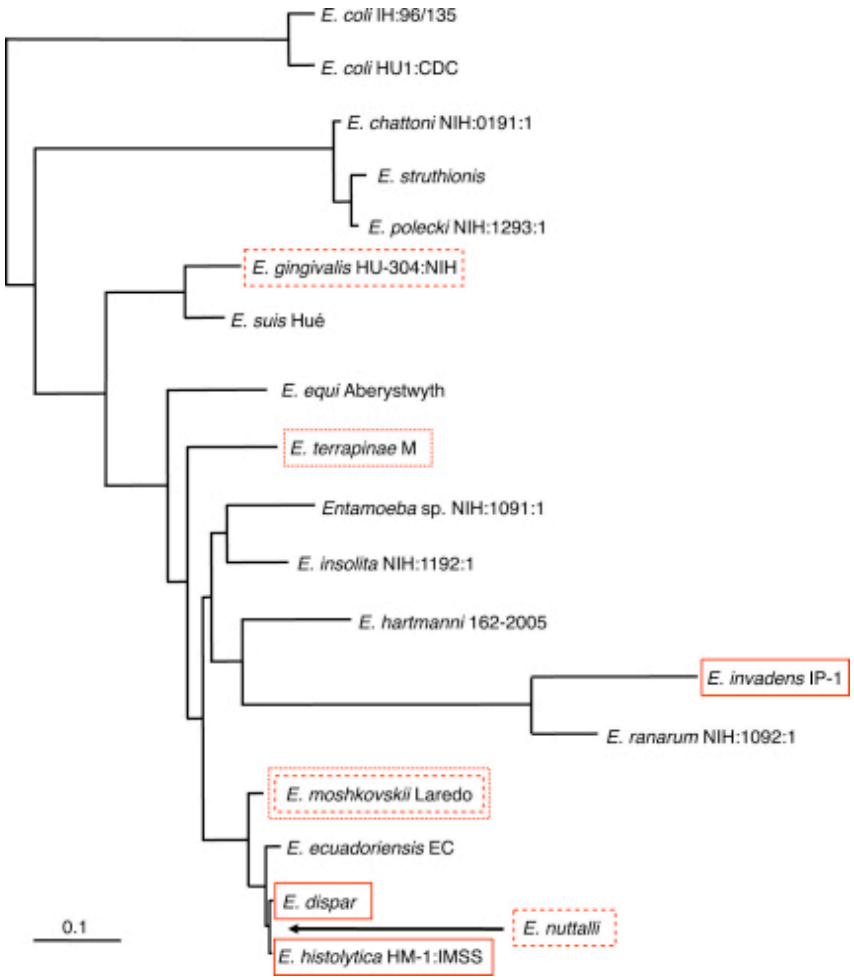


Figure 1.3.1. Phylogeny of the *Entamoeba* genus, based upon small subunit rRNA genes. At the time of its original publication, species surrounded by dashed boxes were due to be sequenced; low coverage shotgun sequencing data existed for those in dotted boxes; and fully sequenced species were in solid boxes [45].

1.3.3 *Entamoeba invadens*

Despite its evolutionary distance from *E. histolytica*, *E. dispar* and *E. moshkovskii*, [46], the reptile-infective *Entamoeba invadens* is of great interest in research into human-infective species' life cycles. This is because *E. invadens* is the only member of the genus whose encystation has been successfully induced *in vitro* [47]. Recently, through genome sequencing, it was found that *E. invadens* has an average sequence identity with *E. histolytica* of 60% [48]. The sequencing of this more distant species is utilised in Chapter Two when identifying sequences required by all *Entamoeba* species.

1.4 The genomic structure of *Entamoeba histolytica*

The genome of *E. histolytica* strain HM-1:IMSS was sequenced and published for the first time in 2005 [49]. The collaborative effort revealed a 23.75 Mb assembly containing 9,938 gene models. The gene content suggested that the parasite possessed a reduced repertoire of metabolic functions, but that this had been bolstered by the horizontal gene transfer of at least 96 genes from bacterial species, more than half of which encoded metabolic enzymes, increasing the range of carbohydrates and amino acids that the eukaryotic parasite could utilise [49]. The assembly was subsequently extensively described and analysed, providing structural and functional details regarding many of the genes identified in the original assembly [50]. In 2010, the genome was reassembled and reannotated, resulting in a shorter 20.80 Mb assembly containing 8,201 gene models [51]. The number of genes has since been increased, through manual annotations, to 8,306 sequences.

The *E. histolytica* genome possesses a number of interesting features that also make it a challenge for sequencing and assembling, including its very low GC content (24%). This prevents traditional Sanger sequencing from exploiting large clone libraries [50]. Furthermore, next-generation sequencing platforms have been reported to demonstrate read coverage bias, favouring balanced GC contents at the expense of more extreme proportions [52-54].

1.4.1 Transposable elements and tRNA arrays

Complicating assemblies further is the high repeat content of the genome. Approximately 20% of the *E. histolytica* genome is comprised of transposable elements (TEs), and repeat regions have been identified in both *E. dispar* and *E. invadens* [55]. TEs are DNA sequences that can 'jump' from one genomic location to another. They typically belong to one of two classes. Class I is comprised of retroelements; that is, genetic material that has been reverse transcribed from RNA, duplicated and transposed as DNA. Class II is made up of transposons, elements that move between loci rather than creating copies of themselves [55, 56].

Class I includes the non-long terminal repeat (LTR) retrotransposons found in large numbers in the *E. histolytica* genome, namely the autonomous long interspersed elements (LINEs) and non-autonomous short interspersed elements (SINEs) [55, 57]. LINEs encode the genetic machinery required to copy themselves, whilst SINEs are dependent upon making use of the LINEs' abilities, being unable to copy themselves [57, 58]. These elements, and other such repeat regions, can affect genome sequencing and gene annotation as their repetitive nature makes discerning one genomic region from another difficult in some cases. If, for example, a gene lies between identical repeat regions, assemblers may stack up reads containing the repeats under the assumption that they represent the same locus. Without both of its flanking regions included in the assembly, reads containing the gene could then be omitted from the assembly [55, 59-61].

Repeat regions are not, however, simply 'junk' DNA that interferes with genome assemblies – they can have both structural and functional roles within the genome. Depending on where they insert into a genome, LINEs and SINEs can affect gene expression by interrupting or inserting promotor regions or splice sites [57]. Additionally, it appears that *Entamoeba* species possess large numbers of subtelomeric tRNA genes, arranged in tandem arrays [62, 63]. Indeed, over 10% of the *E. histolytica* genome consists of approximately 4,500 tRNA genes clustered into 25 arrays, with variable short tandem repeats (STRs) separating the genes from one another [49, 62]. *E. dispar* possesses similarly complex arrays, albeit containing different repeat units, whilst the tRNA genes in other *Entamoeba* species appear to be separated by simple tandem duplications [63]. The function of these arrays is unconfirmed but it is suspected that they play a role in nuclear protein binding [62].

1.4.2 Karyotype and chromosome structure

Although it is suspected that the tRNA arrays described above are subtelomeric, this has yet to be confirmed. This is because the ploidy, as well as the chromosome number and mass of DNA per cell, has not yet been confirmed for *E. histolytica* [64]. Some groups believe the species to be diploid [65], whilst others believe it to be tetraploid [64]. The ploidies of the other *Entamoeba* species featured in this project have not been investigated to date so are not known either. It is possible that the disparities between results from *E. histolytica* studies are artefacts of the different technologies used to test the ploidy of the species; however, the species' ploidy also appears to differ in varying growth conditions and life cycle stages [66, 67].

This uncertainty, combined with the fact that *Entamoeba* chromosomes do not condense, makes it difficult to discern chromosomes from one another and means that the karyotype of *E. histolytica* has not been conclusively determined either [49]. It is suspected that the majority of the species' genomic content is divided across 14 chromosomes [64, 65], with at least 20% of the content existing in separate circular sections of DNA. Whilst chromosome lengths vary between strains, possibly due to regional duplications [68, 69], the plasmid-like DNA molecules vary in size but have been found in all *Entamoeba* species in which they have been looked for [69]. Their presence across the genus and the fact that the ribosomal RNA genes of *E. histolytica* are contained within one of these plasmids suggest that these molecules play important roles that are yet to be defined [62, 69].

1.5 Virulence factors involved in development of amoebiasis

As stated in Section 1.2, it is thought that a combination of parasite and host genes play roles in determining whether *E. histolytica* infections develop into invasive, symptomatic cases. Although the mechanisms and combinations of proteins involved are still relatively poorly understood, much progress has been made in identifying putative virulence factors in *E. histolytica* [70-72]. In this section, major virulence factor families will be discussed with regards to their roles in the progression of invasive amoebiasis, as well as their interactions with the host immune response (Figure 1.5.1). This will inform as to the actions of the gene families of greatest interest as indicators of pathogenicity in Chapter Two of this thesis.

1.5.1 Degradation of intestinal mucosal layer

Whilst the molecular triggers that set *E. histolytica* off down a pathogenic route are unknown, one of the first major families involved in causing invasive amoebiasis is relatively well described. To invade the intestinal epithelium, trophozoites must first degrade and cross the mucosal layer that covers and protects it. To this end, a group of enzymes called cysteine proteases are secreted. The cysteine proteases are a group of at least 50 endopeptidases, 36 of which form three major clades - 'A', 'B' and 'C' [50, 73]. Whilst, collectively, the cysteine proteases are regarded as virulence factors, evidence suggests that approximately 90% of *E. histolytica*'s cysteine protease-derived proteolytic activity is provided by just three proteins – EhCP-A1, EhCP-A2 and EhCP-A5 [74-77]. EhCP-A5 is of particular interest as no orthologue exists in the non-pathogenic *E. dispar* [78]. In concert with amoebic glycosidases, an undefined number of cysteine proteases degrade the MUC2 polymers that constitute much of the mucosal layer [79, 80].

1.5.2 Adherence to, and cytolysis of, host epithelial cells

Trophozoites employ surface-bound proteins to bind to host mucins and, once they have degraded the mucosal layer, epithelial cells. Two major proteins are the Gal/GalNAc lectin and the *Entamoeba histolytica* serine-, threonine-, and isoleucine-rich protein (EhSTIRP). The Gal/GalNAc lectin is a heterodimer, comprising a 170 kDa heavy subunit and a 35 kDa light subunit, associated with a 150 kDa intermediate subunit, as described in a detailed overview of the lectin [81]. The lectin binds to galactose and N-acetyl-D-galactosamine on host cell membranes. Without it, *E. histolytica*'s ability to adhere to host cells is significantly diminished, as is its cytotoxic impact upon the host cells, leading to the understanding that the cytokine cascade by which *E. histolytica* degrades host cells is contact-dependent [10, 82-85]. Downregulation of the EhSTIRP-encoding gene, which is expressed exclusively in virulent strains of *E. histolytica,* was also linked to a reduction in adherence and cytotoxicity in Chinese hamster ovary cells [86], implying that both proteins play a key role in amoebiasis.

1.5.3 Subversion of host immune responses

E. histolytica employs a large number of proteins in evading host immune responses to its assault. The surface protein lipopeptidophosphoglycan (LPPG) sets into motion a series of cellular and molecular interactions that require a number of surface-bound proteins to defend the invading trophozoite. LPPG is recognised by the host proteins Toll-like receptors 2 and 4 (TLR2/4) [87, 88]. This stimulates the release of multiple signalling molecules, including the chemoattractant tumour necrosis factor alpha (TNF-α) [72, 88], which attract macrophages and neutrophils to the site of invasion. The immune cells subsequently attack the trophozoites by releasing reactive oxygen species (ROS), requiring the action of a group of amoebic proteins. A superoxide dismutase (SOD), a NADPH:flavin oxidoreductase and a peroxiredoxin work in sequence to convert the ROS into water, via hydrogen peroxide (Figure 1.5.1; [89-92]). An iron hydrogenase also plays an undefined role in protection against oxidative stress [93, 94]. This challenge, induced by the trophozoites themselves, is thus met with strong resistance by the *E. histolytica* cells. Such protection is not, however, afforded to the enterocytes around the invading trophozoites. These cells are lysed during the immune response, allowing faster progress for the *Entamoeba* through the intestinal mucosa [70].

1.5.4 Avoidance of host immune response

In addition to these defences, *E. histolytica* is able to employ a number of proteins to defend itself against its host's immune system. Firstly, amoebic cysteine proteases cleave the secreted immunoglobulin IgA [95, 96]. IgA antibodies are thought to play a number of roles in defending host cells against invasive trophozoites, including reducing adhesion of trophozoites to host cells [97]. As such, destruction of IgA allows the amoebae to bind to host cells, triggering the destructive immune response described in Section 1.5.3. *E. histolytica* also has a unique strategy that it applies to those antibodies that do successfully bind to its surface antigens. The trophozoites make use of a process termed 'capping and shedding', whereby surface antigens to which host antibodies have bound are relocated, through plasma membrane folding, to appendices at the posteriors of the amoebae, called the uroid, and sloughed off, temporarily concealing the invading pathogen from view of the immune system (Figure 1.5.1; [72, 98-100]). These two methods of immune system evasion contrast drastically with the above descriptions of how *E. histolytica* uses the

immune system against its host. They demonstrate the wide range of abilities that have allowed *E. histolytica* to effectively parasitise human hosts and become a prolific health problem.

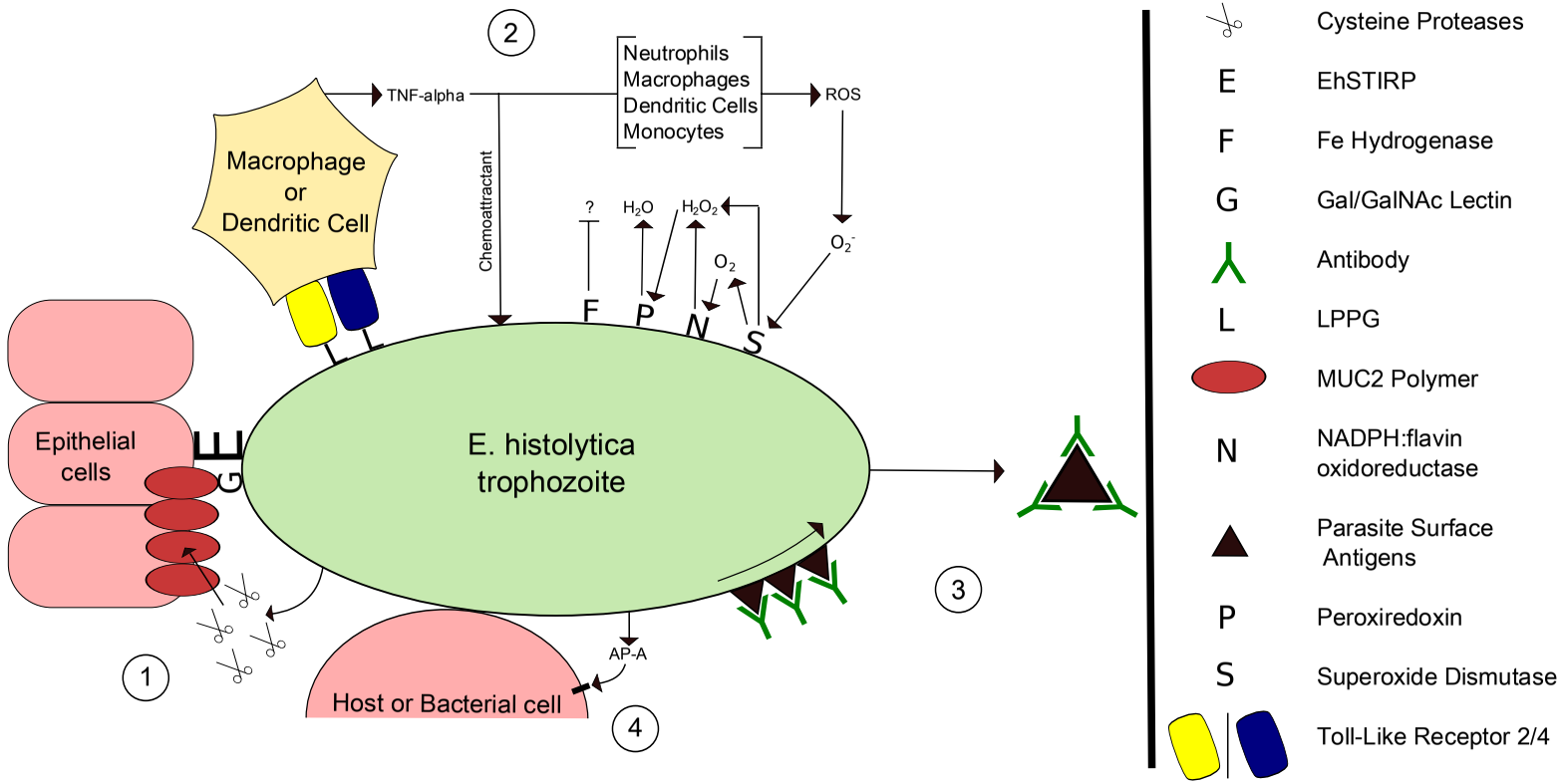


Figure 1.5.1. Actions of key *Entamoeba histolytica* HM-1:IMSS virulence factors identified by genomic studies. 1) Binding to MUC2 polymers by Gal/GalNAc lectin, cleavage of MUC2 by cysteine proteases, and subsequent binding to host epithelial cells by EhSTIRP. 2) Binding to host immune cells via LPPG molecules, prompting release of damaging ROS, which are degraded by a series of membrane-bound proteins. 3) Translocation and shedding of antibody-bound surface antigens. 4) Secretion of amoebapore-A, triggered by direct contact between trophozoite and host or bacterial cells, lysing target cells [72].

1.6 Genome sequencing strategies

1.6.1 Illumina sequencing technology

During this project, I have made use of sequenced datasets generated using all three major second generation sequencing technologies – Illumina, 454 pyrosequencing and SOLiD. The genome sequencing carried out was, however, performed exclusively using Illumina technology [101, 102]. Extensive comparisons and reviews of the available sequencing technologies and platforms, including the Ion Torrent, already exist [103-107], however it is helpful to understand the process by which reads have been generated in this project.

In order to generate a genomic library for Illumina sequencing, short DNA sequences (~500 bp) are generated by mechanical shearing, before universal primers, called adaptors, are ligated to either end of these fragments. These *in vitro* cloned fragments are then bound to the channels of a flow cell amidst oligonucleotides complementary to the adaptor sequences at the unbound ends of the fragments. The adaptors bind to matching sequences in the oligonucleotide field forming bridges consisting of single strands of DNA paired at one end with incomplete complementary strands. Passing nucleotides and polymerase enzymes across the flow cell then results in double stranded sequences as the incomplete strand is extended in a process called 'bridge amplification' [102, 103]. The bridges are subsequently denatured and the amplification process is repeated, resulting in clusters of DNA reads, sequencing of which will be detectable in the next step.

Illumina platforms employ a 'sequencing by synthesis' technique, which makes use of a four-colour cyclic reversible termination (CRT) method, when reading DNA libraries [102, 103, 105]. Over multiple cycles, four different fluorescently labelled individual nucleotides, each containing a removable terminator molecule, are passed over the densely populated flow cell in the presence of a polymerase. During each cycle, the nucleotides will be incorporated into a growing oligonucleotide chain only if they are the next base in the complementary sequence. When bound, laser excitation allows detection of different fluorescent colours by a charge-coupled device (CCD) camera. The combined fluorescent signal from a cluster provides a detectable signal that allows the sequencer to deduce which base was added at each position in the read. After the nucleotides have been detected, the terminators are cleaved from the newly

added nucleotides and the next cycle is performed. These reads can then be entered into genome assemblies, which often take the form of *de novo* assemblies, as featured in Chapter Three of this thesis.

1.6.2 Challenges of *de novo* assemblies

The average cost of genome sequencing has fallen at a rate far beyond most people's reckoning in the past decade, exceeding the long-accepted rate of Moore's Law (Figure 1.6.1) [108]. Whilst this exponential drop, afforded by the introduction of ever cheaper sequencing technologies [103, 105, 109], appears to be levelling out [110], the advances made have left researchers with a tremendous amount of sequencing data to analyse. Unfortunately, even in uncontaminated DNA samples, *de novo* genomic assembly is a complex task with opinions on best practices and the reliability of associated parameters subject to a great deal of debate [111, 112].

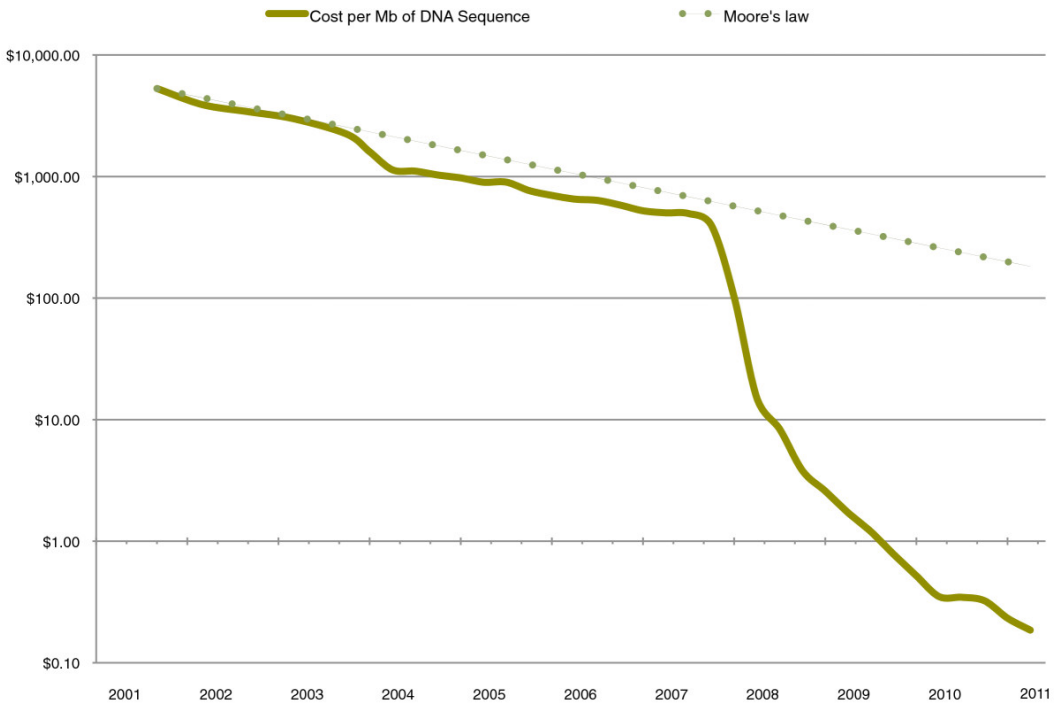


Figure 1.6.1. The falling cost of sequencing 1 Mb of DNA compared with the rate that would have occurred had the decrease followed Moore's Law, as expected. Image originally published by BioMed Central [108].

Second generation sequencing technologies provide users with much larger numbers of reads than could be achieved with traditional Sanger sequencing, at a fraction of the cost [105, 113]. However, in the cases of all but the 454 technology, this has come with the significant caveats that the reads are considerably shorter than those output by Sanger technology and that early platforms were less accurate [105, 113]. Whilst the Pacific Biosciences platform – the first of a new generation of sequencers – offers longer reads than even Sanger sequencing, it is by far the least accurate of the technologies [113] and so requires specialised bioinformatics approaches to be refined before it can be used in assemblies. The short reads with which *de novo* assemblies must, then, be performed present issues. Firstly, as described above, assembling identical repeat regions or duplicated sequences in a genome will likely result in the assembly of one region with an artificially high coverage depth [59-61]. The short reads often do not extend beyond repeat regions so assemblers cannot differentiate between them based upon neighbouring regions. Furthermore, assemblers will often split multi-exon genes across scaffolds, particularly where introns contain duplicate or repetitive sequences, resulting in misleading gene counts and annotations [60].

Despite these potential pitfalls in the *de novo* assembly process, there are many benefits to the technique that, it could be argued, outweigh the possibility of inaccuracies in the assembly. For example, not being reliant upon an existing genome means that an assembly's accuracy is not dependent upon a genome that may or may not be inaccurate itself, having probably been assembled by a *de novo* method. It is also the most effective way of sequencing novel species for which mapping is not an option. Furthermore, any mistakes will likely be rectified over time as more groups re-sequence genomes in order to improve them.

1.6.3 Current strategies for *de novo* genome annotations

Confidently identifying coding regions within *de novo* assembled genomes is often a difficult process. Gene finding using genomic sequences is, broadly speaking, achieved by searching for alignments to known orthologous sequences, properties indicative of a gene sequence, or a combination thereof [114]. Where the genomes of closely related species have been sequenced and annotated, it is possible to align encoded protein sequences to the novel genome, identifying likely sites of gene models. This method forms the basis of the oft-used GeneWise algorithm [115].

Basic *ab initio* gene finding algorithms, such as geneid [116] and Genscan [117], detect genomic regions likely to contain coding sequences based upon detection of both signals and content properties. Signals include transcription start and termination sites, stop codons, and donor and acceptor splice sites [118]. Compositional properties of coding sequences include those shared by exons, introns and intergenic regions [114]. *Ab initio* algorithms, therefore, are capable of identifying novel genes, unlike alignment methods, which are restricted to identifying orthologues. As stated above, many programs, including Ensembl [119] and AUGUSTUS [120], are now capable of making use of orthologue alignments to inform *ab initio* predictions, further improving their accuracy and reliability. The combination of such data types is seen in Chapter Two, where AUGUSTUS is used to predict gene models in the *E. moshkovskii* strain Laredo.

In spite of these advances and improvements in gene prediction, however, *de novo* genome annotations often possess inflated gene counts as a result of gene models being fragmented by assemblies and recorded as multiple genes [121]. An effective method of improving an annotation is to use RNA-Sequencing (RNA-Seq) data [122, 123]. RNA-Seq data consists of sequenced reads of complementary DNA (cDNA), rather than genomic DNA, comprising a library representative of a whole transcriptome. The sequenced reads can either be assembled *de novo* and then aligned to their respective genome or aligned as reads before being assembled based upon their alignments [122].

Such alignments provide accurate annotations of exon boundaries as well as splice junctions, including alternative splice sites that could not be elucidated using genomic data alone [122, 124]. Any genes fragmented by an assembly, being split across contigs, can thus be resolved; and genes that were not originally predicted (perhaps because they are atypical in their content properties) may be detected, thus improving upon gene model predictions based solely upon genomic data. It is generally accepted that most initial annotations of a gene set should only be considered drafts [125], and many have been improved upon by the addition of RNA-Seq data at a later date [126].

1.6.4 Comparative genomics of other parasitic species

Comparative genomic studies such as the ones undertaken in this thesis have been performed for multiple protozoon parasites in recent years [127-129]. Most

recently, a comparison of the genomes and transcriptomes of closely related parasites *Toxoplasma gondii* and *Neospora caninum* was carried out [127]. *T. gondii* is an intracellular parasite of most, if not all, warm-blooded vertebrates, capable of causing blindness, spontaneous abortions and congenital disease [127, 130]. *N. caninum* has a more limited host range and causes spontaneous abortions in cattle [127, 131-133]. Within the genomes of the two species, genes encoding surface antigens and virulence factors were identified that have diverged whilst the majority of the two species' gene sets remain conserved, distinguishing the two species from one another [127]. The comparison also allowed deductions to be made about the species' adaptations that allow them to exploit the niches that they occupy.

Surface antigens were found to be similarly divergent, relative to the rest of the gene sets, in a comparison of the human- and mammal-infective parasites *Trypanosoma cruzi*, *Trypanosoma brucei* and *Leishmania major* [128]. *T. brucei* and *T. cruzi* rely upon a large degree of antigenic diversity to evade host immune responses [134]. Large novel groups of co-localised genes, identified as a result of genomic analyses, were, therefore, implicated in immune evasion, presenting new potential drug targets [128]. The implication of such studies, which is of great interest in Chapter Three of this thesis, is that divergence in an otherwise conserved gene set is indicative of genes that play major roles in a species' ability to parasitise its host.

Furthermore, comparative genomic analyses can also identify genes unique to individual species, which may play important roles specific to their life cycle, as was found in the malaria-causing parasite *Plasmodium vivax* when compared with other members of its genus [129]. Eight novel gene families were discovered in *P. vivax* as a result of this, including one gene known to illicit an immune response, implicating it in pathogenesis. The results of these analyses can, therefore, provide information that might identify key virulence factors involved in diseases.

1.7 Aims of thesis

As a neglected disease, there is still much to learn about amoebiasis and its causative agent or agents. This thesis aims to improve understanding of the genus beyond *E. histolytica*, as well as offering insight into the gene families necessary for survival of the different lifestyles exhibited by members of the genus.

In Chapter Two, the genome of *E. moshkovskii* Laredo is sequenced, assembled and annotated, providing a draft reference genome for the species. The genome is then entered into comparative genomic analyses involving other reference genomes in the genus, namely those of *E. histolytica* HM-1:IMSS, *E. dispar* SAW760 and *E. invadens* IP-1. These analyses allow for the elucidation of a putative core *Entamoeba* gene set consisting of orthologues present in all members of the genus. Gene families thought to have key roles in the development of amoebiasis are identified, whilst the discovery of orthologous sequences in *E. moshkovskii* supports the theory that *E. moshkovskii* is a human-infective strain, rather than a free-living organism. Finally, comparative analyses of *E. histolytica* HM-1:IMSS with representatives of other genera from the Unikonts clade of eukaryotes are performed. Studying species with well annotated genomes – *Dictyostelium discoideum* [135], *Acanthamoeba castellanii* [136, 137] and *Saccharomyces cerevisiae* [138] - allows the generation of an *Entamoeba*-exclusive gene set, defining some of the gene families that distinguish the *Entamoeba* from their close relatives.

Chapter Three shifts the focus from the inter-species comparisons seen in Chapter Two to an intra-species approach. Genomic sequences of multiple strains of the two best-studied species – *E. histolytica*, *E. dispar* – and the little-studied *E. moshkovskii* are compared to identify those genes that demonstrate the greatest genetic variation within each species. Following the Red Queen Hypothesis [139, 140], this was seen as a way to identify the genes most important in survival of the different lifestyles seen in the three species. The chapter also presents comparisons of intra-species diversity seen in *E. histolytica* and *E. moshkovskii*, leading to the conclusion that *E. moshkovskii* is a species complex rather than an individual species comprised of genetically diverse strains.

Finally, Chapter Four offers comparisons between multiple methods of *de novo* genome assembly and gene detection made during efforts to sequence the genome of the recently discovered species *E. bangladeshi*. Four publicly available whole genome assembly programs – Velvet, Ray, SOAPdenovo and ABySS [141-144] – are compared with an alternative method focusing on coding sequences rather than the entire genome. The results of this technical evaluation offer guidance for future similar assemblies, whilst also gleaning information about the gene content of *E. bangladeshi* and its place in the evolutionary history of the *Entamoeba*, thus further adding to our knowledge of this poorly understood genus.

Chapter Two - Genomic annotation of *Entamoeba moshkovskii* strain Laredo and comparative analysis between members of the *Entamoeba* genus

2.1 Introduction

A wealth of knowledge exists regarding the gene families involved in the various stages of parasitic amoebic infections, including those responsible for causing pathogenic amoebiasis. However, much uncertainty remains regarding which of these families play essential roles and what key differences exist between those species and strains capable of causing pathogenesis and those that cannot. It was hoped that comparisons of *Entamoeba histolytica*'s gene complement with those of closely related species might offer greater insight into the genes involved in host-parasite interactions. Annotated genomes already exist for the species *Entamoeba dispar* and *Entamoeba invadens*, although this is the first time they have been presented. However, in order to achieve a more complete comparison, it was first necessary to assemble and annotate a reference genome for the species *Entamoeba moshkovskii*.

2.1.1 Early isolation and identification of *Entamoeba moshkovskii*

E. moshkovskii is a close relative of the causative agent of amoebiasis, *E. histolytica*. It was first identified upon isolation from sewage effluent in Moscow in 1941 [36]. Soon after, as described by Neal, a number of research groups around the world isolated *E. histolytica*-like amoebae from sewage, offering credence to the suggestion that this *Entamoeba* species could survive outside of a host [37]. Strains have subsequently been isolated from freshwater sediments and brackish water pools [38] and, as a result, *E. moshkovskii* has long been considered a free-living organism.

E. moshkovskii is morphologically indistinguishable from both *E. histolytica* and *E. dispar* at all stages in its life cycle. As such, positive identification of isolates from sewage and bodies of water was, for many years, dependent upon physiological and biochemical differences between *E. moshkovskii* and its human-infective relatives. Perhaps the most immediately obvious difference between *E. moshkovskii* and both *E. histolytica* and *E. dispar* is the temperatures at which the species can grow. *E. moshkovskii* can tolerate a greater range of temperatures than *E. histolytica* and *E. dispar*, with an optimal incubation temperature of around 24°C [37]. This is

considerably lower than the 37°C required to imitate the conditions within a human host and successfully culture *E. histolytica* and *E. dispar* [36, 145]. In addition to this, *E. moshkovskii* is both more osmotolerant and more resistant to several drugs, including the anti-protozool drug emetine [3, 146].

2.1.2 Isolation from human hosts

The isolate that this chapter will designate as the species' reference strain – *E. moshkovskii* Laredo – was isolated from a human host in Laredo, Texas in 1956 [147]. Despite the strain's ability to grow both at room temperature and body temperature, its isolation from a human and the symptoms demonstrated by the patient – abdominal pain and diarrhoea – meant it was presumed to be a strain of *E. histolytica* [146]. Following on from this, multiple strains isolated from human hosts were similarly identified as 'atypical *E. histolytica*', or '*E. histolytica*-like' cells. A review of many such cases was carried out by Goldman as several of the distinguishing features of these strains, mentioned above, became better known [148].

In more recent years, such 'atypical *E. histolytica*' strains have been redefined as *E. moshkovskii*, reflecting the defining characteristics they share with environmental isolates [146]. It is important to stress that, despite being isolated from humans, some of whom were ill, the *E. moshkovskii* strains in these instances could not be confirmed as the causative agents of any observed symptoms [148]. Indeed, the symptoms demonstrated by the patient from whom *E. moshkovskii* Laredo was isolated were later judged to be the result of a benign colon tumour [146]. As such, none of these cases definitively demonstrated pathogenicity and parasitism as traits in *E. moshkovskii*.

2.1.3 Evidence of parasitism and pathogenicity in humans

In the last decade, several studies have presented more conclusive evidence of *E. moshkovskii*'s status as a human-infective, and potentially pathogenic, organism. Many of these studies have been made possible through the utilisation of nested, single-round, or real-time polymerase chain reactions (PCR) targeting the small ribosomal subunits of *E. histolytica*, *E. dispar* and *E. moshkovskii*. The three targets all differ in length, meaning the process can differentiate between the species, allowing for confirmation of the presence of *E. moshkovskii* in samples [3, 41, 42, 149-151]. With many such studies detecting *E. moshkovskii* in human stool samples, it is generally

agreed that *E. moshkovskii* should be considered a potential human parasite. Indeed, one study concluded that humans may, in fact, be the primary host of *E. moshkovskii* [3].

The question of *E. moshkovskii*'s pathogenicity is currently of considerable interest. Studies have detected *E. moshkovskii* in stool samples of subjects presenting with diarrhoea, both in mixed amoebic infections and on its own [4, 44], whilst not detecting it in asymptomatic test subjects. *E. moshkovskii* has also been shown to successfully establish symptomatic infections in members of several congenic strains of mice when injected into the caecum. Results of *E. moshkovskii* infections were comparable with those of *E. histolytica*. *E. dispar* was unable to sustain an infection in any of the five strains, suggesting that *E. moshkovskii* and *E. histolytica* share virulence factors that *E. dispar* does not [4]. Taken together, such results imply, whilst not proving, *E. moshkovskii*'s involvement in human amoebiasis.

2.1.4 Indications of a species complex and resulting caveats

It should be noted that not all strains of *E. moshkovskii* are capable of parasitising human hosts. It is likely that *E. moshkovskii* is, in fact, a species complex, comprised of multiple closely related species. Riboprinting has revealed a total of six ribodemes within *E. moshkovskii*, with all human-infective isolates known at the time being non-exclusively present within ribodeme 2 [38, 146, 152]. This suggests that our definition of *E. moshkovskii* may be insufficient and that the parasitic strains of this species may be better defined as their own species, set apart from the others. However, until firm evidence supports such a change, we must assume that *E. moshkovskii* can be regarded as one species, albeit a highly variable one.

2.1.5 Importance of studying genetic content of *Entamoeba moshkovskii*

Despite the first *E. moshkovskii* strain being discovered over 70 years ago, there has been comparatively little research carried out on the organism. This is likely due to the research community being relatively small and the fact that *E. moshkovskii* was not considered a pathogenic, and therefore 'important', organism until recently. As such, no DNA sequencing has ever been carried out to generate a reference genome for the species. If *E. moshkovskii* strains are capable of pathogenic infection of human hosts then it may be of benefit to understand how they do so. Many of the molecular and

chemical pathways employed by *E. histolytica* in pathogenic infections are known but relatively poorly understood. Comparing the genetic content of a parasitic strain of *E. histolytica* with that of a potentially pathogenic strain of *E. moshkovskii* might highlight common gene families not present in non-pathogenic species and strains. Such findings could be used to facilitate research into the pathways essential for causing amoebiasis.

2.1.6 Gene model annotation and validation used in this chapter

i. Annotation

In Chapter One, current *de novo* gene annotation techniques were described. In this project, a combination of alignment-based and *ab initio* methods was used to annotate gene space in the *E. moshkovskii* Laredo genome. Broadly speaking, such an approach involves two steps. Firstly, the chosen annotation pipeline must be provided with a sample of gene models the accuracy and validity of which one can be confident. Functional genes cannot include 'stop' codons, meaning that all predicted gene models exist within Open Reading Frames (ORFs), containing no such codons. As such, the first step is to identify ORFs large enough to contain an individual exon. Subsequently employed are orthologous sequences - that is, genes in another species that share ancestry with suspected gene models and have diverged following a speciation event [153]. It is logical that the largest numbers of orthologous genes will always be found shared by closely related species because evolution from their common ancestor has been relatively recent so the genes will share larger proportions of their protein sequences. As such, the sequences of annotated close relatives (in this case, *E. histolytica* and *E. dispar*) can be searched for in an unannotated genome, such as that of *E. moshkovskii* Laredo, with matches suggesting the presence of an orthologous gene. As explained in Chapter One, transcriptomics are of use here as RNA-Seq data can be used to precisely identify splice sites in multi-exon genes. The RNA-Seq data provided by Dr Guillen's group was used to this effect here, ensuring that accurate exon boundaries were annotated. For the second step, the annotation pipeline attempts to identify regions that possess physicochemical properties common to gene sequences. This project's strategy, in particular, uses the provided sample of gene models to further inform the gene prediction software as to the expected signals and content.

ii. Validation

A common way of determining the quality of an assembly and annotation is to to search for genes orthologous to the 'Core Eukaryotic Genes' (CEG) set [114]. The CEGs are genes suspected of being shared by all eukaryotic species. They were identified by finding orthologous groups present in a wide-ranging group of six species - *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. This approach is imperfect, as is discussed in Section 2.3.2, but it is a widely used rudimentary measure of completeness of a genome so was implemented in this project [154-157]. There are myriad parameters to measure the quality of the assembly itself and these are discussed in Chapter Three.

2.1.7 Aims of chapter

In this chapter I have sequenced, annotated and analysed the genome of the *E. moshkovskii* strain Laredo. I have used these new genomic data in a comparative analysis of four members of the *Entamoeba* genus – *E. histolytica*, *E. dispar*, *E. moshkovskii* and *E. invadens* - in order to identify multiple gene sets. Firstly, I have discovered gene families unique to each species, gaining some insight into the sequences necessary for survival of their different lifestyles. Of particular interest are the genes that shed light on the capacity of *E. moshkovskii* and *E. dispar* to infect and cause disease. Secondly, I have identified a core set of gene families exclusive to *Entamoeba*, thus highlighting the genes that set this parasitic genus apart from other genera. Finally, I have presented detailed analyses of the virulence genes that play crucial roles in the advancement of amoebiasis.

2.2 Materials and Methods

2.2.1 LYI-S-2 growth medium

E. moshkovskii Laredo cultures were kindly provided by Dr Graham Clark (London School of Hygiene and Tropical Medicine) and maintained by Dr Gareth Weedall (University of Liverpool) in LYI-S-2 medium [158]. The complete medium consisted of 880 mL basal medium, 20 mL vitamin mix #18 [159] and 130 mL heat-inactivated adult bovine serum (ABS). 2 mL penicillin-streptomycin solution at a concentration of 5,000 units/mL penicillin, 5 mg/mL streptomycin (Gibco by Life Technologies) was added.

The heat-sterilised basal medium consisted of 1.0 g potassium phosphate dibasic, 0.6 g potassium phosphate monobasic, 1.0 g L-cysteine, 0.2 g ascorbic acid, 2.0 g NaCl, 25.0 g yeast extract, 5.0 g neutralised liver digest, 10.0 g glucose and 1.0 mL ferric ammonium citrate brown form (22.8 mg/mL), dissolved in distilled water and adjusted to pH 6.8. Sigma-Aldrich produced all chemical components used to make the basal medium except for neutralised liver digest (Oxoid), and NaCl and glucose (VWR BDH Prolabo).

Vitamin mix #18 consisted of five solutions. Solution one consisted of 2.5 mg niacinamide, 0.4 mg pyridoxal hydrochloride, 2.3 mg calcium pantothenate, 0.5 mg thiamine hydrochloride and 0.12 mg vitamin B12, dissolved in distilled water up to 2.5 mL. Solutions two, three and four, consisting of 0.7 mg riboflavin dissolved in minimal 0.1 M NaOH; 0.55 mg folic acid dissolved in minimal 0.1 M NaOH; and 0.2 mg biotin, respectively, were each dissolved in distilled water up to 4.5 mL. The four solutions were combined and added to a 4.0 mL solution of 50 mg Tween-80, and 0.1 mg ±-α-lipoic acid in 0.5 ml 95% ethanol, suspended in distilled water. The complete solution was filter-sterilised using a 0.22 μm filter. All chemical components of vitamin mix #18 were produced by Sigma-Aldrich.

Different batches of ABS support growth of *Entamoeba* to varying degrees of efficacy. The batch used throughout this project was recommended by Dr Clark, and produced by Sera Lab (Product ID: S-202-FS; Batch No: M106028). The ABS was heat-inactivated by 30 minutes' incubation at 56°C before being put on ice.

2.2.2 DNA extraction

The method Dr Weedall used to extract DNA from *E. moshkovskii* Laredo cells was a modification of the technique on the '*Entamoeba* Homepage' at http://entamoeba.lshtm.ac.uk/dnaisoln.htm. Chilled late-log/stationary phase cultures were pelleted at 11,000 rpm for 5 minutes at 4°C, then, twice, were re-suspended in 1 mL phosphate-buffered saline solution and pelleted for 1 minute at 6,000 rpm at room temperature. Resulting pellets were re-suspended in 300 µL Qiagen Cell Lysis Solution with 3 µL proteinase K (10 mg/mL) and incubated at 55°C for 120 minutes. 42 µl 10% cetyltrimethylammonium bromide (CTAB)/0.7 M NaCl solution, heated to 65°C, and 75 µL 3.5 M NaCl were added prior to incubation at 65°C for 20 minutes. Following incubation, CTAB solutions were mixed with 400 µL phenol:chloroform:isoamyl alcohol (25:24:1) and centrifuged at 15,000 rpm for 10 minutes at 18°C. Supernatants were chilled with 2 volumes of 100% ethanol and 1/10 volume of 3 M NaOAc (pH 5.2) at -20°C for 12 hours, before being pelleted at 15,000 rpm for 30-45 minutes at 4°C. Final pellets were washed in 70% ethanol, re-suspended in 30-50 µl nuclease-free water and purified using illustra MicroSpin S-200 HR Columns, according to the manufacturer's protocol.

2.2.3 Sequencing and assembling the *Entamoeba moshkovskii* Laredo genome

Dr Weedall prepared four libraries from the DNA, according to the manufacturer's protocol (Roche): two single-end fragment libraries, a 3 kb insert Paired End (PE) library and a 8 kb insert PE library. Sequencing was carried out by the University of Liverpool's Centre for Genomic Research (CGR) using the Roche 454 GS FLX Titanium system. The Newbler Assembler v2.3 [160] was used to carry out a *de novo* assembly of the total 3,812,076 generated reads > 150 bp using default parameters. The resulting scaffolds, and contigs no smaller than 500 bp, were concatenated to produce a disordered draft assembly.

2.2.4 *Entamoeba moshkovskii* Laredo annotation

AUGUSTUS v2.5.5 [120] required a training set of coding sequences, the accuracy and validity of which one could be confident, in order to identify common features of start and stop codons, and splice junctions (http://augustus.gobics.de/binaries/retraining.html). To generate the training set,

Open Reading Frames 150 amino acids or greater in length were cross-referenced with 'hits' generated by entering the first 3.5 Mb of the assembly into a BLASTX search against the *E. histolytica* HM-1:IMSS protein set, with an Exponent Value (E-value) threshold of 1e-10. The Basic Local Alignment Search Tool (BLAST) [161] is a tool made available by the National Center for Biotechnology Information (NCBI).

I also used transcriptomic data kindly generated by Dr Nancy Guillén (personal communication) using a previously published protocol [162]. Briefly, a PE cDNA library was prepared from polyA+ mRNA extracted from log-phase *E. moshkovskii* Laredo trophozoites and sequenced on an Illumina HiSeq2000 instrument, generating 100 bp reads. Reads that Bowtie v0.12.7 [163] failed to align to the assembled *E. moshkovskii* Laredo genome were mapped using HMMSplicer v0.9.5 [164]. Default cutoff scores were used, unlike in the previously published protocol. The *E. moshkovskii* Laredo genome was then entered as a query sequence into a BLASTN search against the 62 bases either side of each unambiguously identified splice junction, with an E-value threshold of 1e-5.

The three datasets were viewed in the Wellcome Trust Sanger Institute's program Artemis v13.0 [165, 166] and used to inform manual gene model curation. AUGUSTUS' training script autoAug was run using a final training set of 197 models, using default parameters. AUGUSTUS was then run using default parameters and a set of 'hints', consisting of weighted intron positions inferred from Dr Guillén's splice junction data (Bonus = '10'; Penalty = 0.7; un-weighted values = 1).

Proteins encoded by putative coding sequences (CDSs) in AUGUSTUS' output were entered into a reciprocal BLASTP search against the protein set of *E. histolytica* HM-1:IMSS, using default parameters. Predicted sequences with a reciprocal best hit (RBH) were included in the final annotation set. Those without a definite orthologue were only included if their total exon length exceeded 350 bp and if they were attributed an AUGUSTUS confidence score of at least 0.75 or 'hit' an *E. histolytica* HM-1:IMSS gene in a one-way BLASTP search using an E-value threshold of 1e-5.

As a rudimentary measure of completeness, the annotated gene set, along with the gene sets of *E. histolytica* HM-1:IMSS, *E. dispar* SAW760 and *E. invadens* IP-1, was compared with a group of sequences theoretically common to all eukaryotes, called the CEG set [114]. The 458 CEG families were downloaded from The Korf Lab's Core

Eukaryotic Genes Mapping Approach (CEGMA) dataset at http://korflab.ucdavis.edu/Datasets/cegma/ [114, 167]. Associated functions were acquired from the NCBI Clusters of Orthologous Groups (COG) database, from which the CEGs were derived (http://www.ncbi.nlm.nih.gov/COG/; [114, 168]). The CEG protein sequences were entered into a BLASTP search against the *Entamoeba* sequences using an E-value threshold of 1e-5.

2.2.5 Reference strain data

Genomic, CDS and protein sequences, as well as genomic feature files, for *E. histolytica* HM-1:IMSS, *E. dispar* SAW760 and *E. invadens* IP-1 were downloaded from AmoebaDB v2.0 [169, 170]. Average fold coverage values were acquired from the NCBI Whole Genome Sequence Project pages. The accession numbers for the versions of the three projects used are as follows (with original project accession numbers in parentheses): *E. histolytica* HM-1:IMSS: AAFB02000000 (AAFB00000000); *E. dispar* SAW760: AANV02000000 (AANV00000000); and *E. invadens* IP-1: AANW03000000 (AANW00000000). All equivalent data regarding *E. moshkovskii* Laredo was derived from the assembly and annotation carried out in this project.

2.2.6 Intra-genus comparative analyses

OrthoMCL v2.0.3 [171] was run to identify gene families with orthologues in *E. histolytica* HM-1:IMSS, *E. dispar* SAW760, *E. invadens* IP-1 and *E. moshkovskii* Laredo. Default parameters were used, though an E-value threshold of 1e-5 was applied to the All-vs-All BLASTP search stage. MySQL served as the relational database. A 50% cutoff value was applied. All proteins from all four species were included in the comparison. MCL was run using a clustering granularity value of 3.0. All clusters containing at least one member from each of the four species were included in the core gene set.

To compare species-specific functions, Gene Ontology (GO) annotations [172] were applied to each of the four *Entamoeba* species' gene sets using BLAST2GO v2.6.4 – v2.7.1 [173]. The sequences were entered into BLASTX searches against the NCBI's nr database, with an E-value threshold of 1e-3 and a limit of 25 hits per query. GO terms associated with hit sequences were pooled and applied to the respective query sequences. InterProScan GO terms derived from all available applications and GO terms attributed to enzymes were also applied. Within each species, all GO annotations

for genes unique to that species were separated into the groups 'Components', 'Functions' and 'Processes'. The same was done for all genes in the core *Entamoeba* gene set that were identified in OrthoMCL results. Each GO group was uploaded to the web-based program, CateGOrizer v3.218 [174], which was used to collate the GO terms into higher-level categories called 'GOA2GO GO Slim' (GOA) terms. Every occurrence of every GO term was included in the counts.

2.2.7 Inter-genus comparative analyses

Protein sequences were downloaded from multiple sources. *E. histolytica* HM-1:IMSS [49, 51] and *Acanthamoeba castellanii* Neff [136] sequences were downloaded from AmoebaDB v2.0. *Dictyostelium discoideum* AX4 sequences and *Saccharomyces cerevisiae* S288c sequences were acquired from the dictyBase [135, 175, 176] and *Saccharomyces* Genome Database [177, 178] websites, respectively. All sequence lists were downloaded on 03/09/13.

OrthoMCL v2.0.3 was used to identify gene families orthologous to *E. histolytica*, *D. discoideum* and *A. castellanii*, as well as those shared by *S. cerevisiae*, as representatives of their genera. Parameters used were as described in Section 2.2.6, save for a 35% cutoff value being applied. All clusters containing at least one member from only *E. histolytica*, *D. discoideum* and *A. castellanii* formed a core Amoebozoa gene set. All clusters containing at least one member from each of the four species were included in a core Unikont gene set. Genes exclusive to *E. histolytica*, and their orthologues from the core *Entamoeba* gene set, formed an *Entamoeba*-exclusive core gene set.

GO annotations were used to compare genus-specific functions. *D. discoideum* GO terms were downloaded from the Gene Ontology website (gene association revision 16105) and attributed to their respective sequences. *Acanthamoeba castellanii* CDSs were downloaded from AmoebaDB v2.0 and *Saccharomyces cerevisiae* CDSs were downloaded from the NCBI Nucleotide database. GO terms were generated for both species using BLAST2GO, as described in Section 2.2.6. All GO annotations for genes unique to each species, including *E. histolytica*, were separated into the groups 'Components', 'Functions' and 'Processes', and collated using CateGOrizer v3.218, as described in Section 2.2.6.

2.2.8 Investigating the *Entamoeba*-exclusive core gene set

Excluding hypothetical proteins, the number of occurrences of every function or domain in the gene set was manually counted. Functions or domains were grouped together when thought to imply similar roles. Where available, *E. histolytica* annotations were used, unless an orthologous *E. dispar* gene's annotations were considered more informative, whilst concurring with the *E. histolytica* annotations. If no *E. histolytica* or *E. dispar* annotations existed, *E. invadens* annotations were used.

2.2.9 Inter-species comparisons of virulence factors

i. Identifying virulence factor families

E. histolytica HM-1:IMSS genes suspected of encoding putative virulence factors were identified using AmoebaDB, NCBI's Gene Database and the scientific literature (for search terms, see Appendix A, Table A.1). Corresponding protein sequences were entered into a TBLASTN search against the complete gene sets of *E. histolytica* HM-1:IMSS, *E. dispar* SAW760, *E. invadens* IP-1 and *E. moshkovskii* Laredo to identify orthologues. An E-value threshold of 1e-5 and a limit of 50 hits per search were applied to limit the number of poor quality hits and computational expense incurred in analysing them.

Where 50% or more of a query sequence's length was cumulatively matched across all hits to a particular reference sequence, that reference sequence, and all genes with which OrthoMCL clustered it, were added to its respective virulence factor family. Clusters or individual genes present in 2 families were manually investigated to determine to which family the gene and their cluster should be added. Any identified orthologues lacking functional annotations on AmoebaDB were entered into a BLASTP search against the NCBI's nr database, using default parameters, to subjectively identify any high-quality hits against a member of the virulence factor family to confirm their function. In addition to this, any informative or requisite domains or functions were identified using the InterPro and ProtoNet subsections of UniProt.

In groups containing a noticeably different number of genes in one species, an *E. histolytica* HM-1:IMSS gene within the clade, or an *E. dispar* SAW760 gene in the absence of an *E. histolytica* gene, was entered into a TBLASTX search against the

genome of the 'missing' species, using default parameters. High-quality hits were determined subjectively, using the E-values of known family members. Non-pseudogenous hits were added to their respective virulence factor family.

ii. Phylogenetic analysis

MUSCLE v3.8.31 [179] was used, with default parameters, to align sequences within each family. Bootstrapped Maximum Likelihood phylograms, were generated for each virulence factor family using the Phylogeny Inference Package (PHYLIP) v3.69 [180]. Default parameters were used unless otherwise stated. Seqboot was run with 1,000 bootstrap replicates. Protdist was then run using the Jones-Taylor-Thornton matrix, set to receive 1,000 datasets. The gamma distribution of evolution rates among amino acid positions, and proportion of invariant sites if greater than 0, were determined using values calculated by MEGA v5.2.1 (Appendix A, Table A.2), using default parameters [181, 182]. Fitch estimated phylogenies with the Fitch-Margoliash criterion for the 1,000 randomised data sets before Consense output bootstrapped trees. To apply branch lengths that represent evolutionary distances to the trees, the first two PHYLIP programs described above were run again, using the same parameters, but for 1 dataset rather than 1,000. Bootstrapped trees were input to Fitch with their respective single data set trees, applying branch lengths to the relationships.

In families containing pseudogenes, all incomplete CDSs were entered into a BLASTN search against their species' complete gene set, with an E-value threshold of 1e-4. Query sequences and sequences hit by them were accepted as members of the family, as were pseudogenous virulence factors identified in Section 2.2.9 Part i. Phylogenetic trees were generated for the nucleotide sequences using a method similar to the one above but implementing PHYLIP's DNAdist as opposed to Protdist and using the F84 distance matrix.

iii. Expression data

Expression data for all *E. histolytica* HM-1:IMSS virulence factor CDSs were downloaded from AmoebaDB v2.0. The RNA-Seq data from which these figures were derived were generated using the protocol described in Section 2.2.4 [162]. Extracted from trophozoites of the *E. histolytica* strains HM-1:IMSS and Rahman in the log phase of growth, RNA was collected, sequenced using the Illumina HiSeq2000 machine and

measured in a comparative study of expression of alternative isoforms. The figures attributed to each gene represented the number of transcript fragments per kilobase per million (FPKM) generated by Cufflinks v.2.0.2 [183] using the RNA-Seq data, thus reflecting expression levels of those genes.

2.3 Results and Discussion

2.3.1 Assembly of the *Entamoeba moshkovskii* Laredo genome

The *E. moshkovskii* Laredo genome was sequenced to provide a reference strain for the species with which comparative analyses with other *Entamoeba* species could be carried out. Four DNA libraries were sequenced on the Roche 454 GS FLX Titanium system. Combined, the two single-end fragment libraries yielded 2,211,151 reads, 86% of which were longer than 150 bp. The 3 kb and 8 kb insert PE libraries generated 743,770 reads (86% > 150 bp) and 857,155 reads (90% > 150 bp), respectively. Assembly of the combined total of 3,812,076 reads generated 12,880 contigs. When assembled into scaffolds, 3,352 contigs were included in 1,147 scaffolds. The scaffolds were concatenated, along with 3,460 contigs of at least 500 bp in length. A total of 6,068 contigs were shorter than 500 bp and were not included in the assembly independently or in any scaffolds.

The *E. moshkovskii* Laredo genome has a lower N50 scaffold length than *E. histolytica* and *E. invadens*, but a greater average scaffold size than *E. histolytica* (Table 2.3.1). It consists of a similar number of scaffolds to these two genomes, whereas *E. dispar* has a higher number of scaffolds and correspondingly lower N50 and average scaffold length values. The fact that the *E. moshkovskii* assembly is comparable to the more complete assembly of *E. histolytica* suggests that it is of a high quality. Additionally, if the unscaffolded contigs are omitted from the assembly, the N50 value increases to 47,690, bringing it closer to the value seen in *E. histolytica*.

The *E. moshkovskii* assembly is the only one of the four reference genomes featured herein to be sequenced using second-generation sequencing technology. As such, higher sequencing depths were achieved than for the other species (Table 2.3.1; [49]). However, the average depth was inflated by a relatively small number of contigs and scaffolds with uncommonly high coverage depths, which skewed the distribution of depths (Figure 2.3.1). Exclusion of those with coverage depths beyond 2 standard deviations of the mean lowered the average depth to 54.41x, whilst the peak alignment depth of the assembly was 27x.

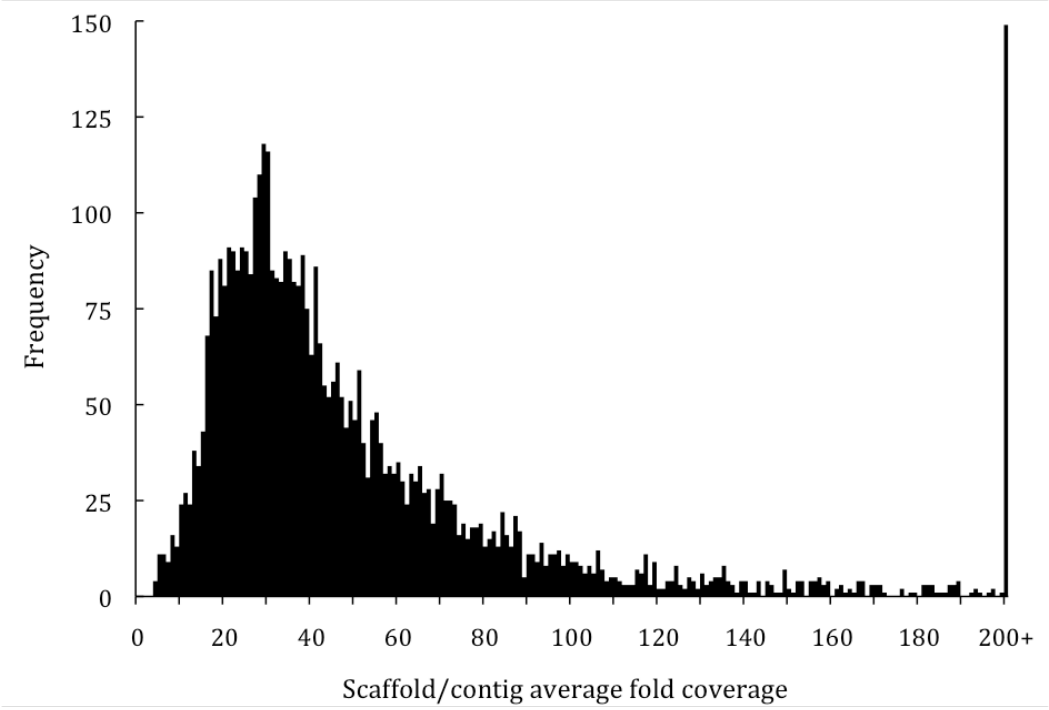


Figure 2.3.1. Frequencies of mean read depths within each scaffold/contig in the *Entamoeba moshkovskii* Laredo genome. The highest fold coverage recorded was 7730.04x.

The GC content of *E. moshkovskii* is similar to those of the other three species, and the range of GC contents seen across the genome is normally distributed (Table 2.3.1; Figure 2.3.2). The relatively narrow range suggests that the GC content of the genome is largely uniform and, more importantly, that the assembly has not been adversely affected by the relatively low GC content, a factor that can introduce bias, and therefore a skewed GC content distribution, into an assembly at certain proportions [52, 54].

It is more likely that, as often occurs in next-generation sequencing and assembly projects, such inflated coverage depths are the result of repeat regions in the genome [184, 185]. Overall, 19.7% of the *E. histolytica* genome is comprised of transposable elements, with two retrotransposon types (LINEs and SINEs) making up 56.9% of that total [55]. Whilst not all such elements are present in both *E. histolytica* and *E. dispar* [186-188], many are, meaning it is highly likely that *E. moshkovskii* also possesses several of them. Indeed, according to AmoebaDB, of the 33 contigs and 1 scaffold with average coverage depths beyond 2 standard deviations from the

assembly mean, the scaffold and 21 of the contigs contain repeat regions or areas of low complexity. Worthy of particular note is contig 01592, 99.88% of which consists of a 361 base tandem repeat.

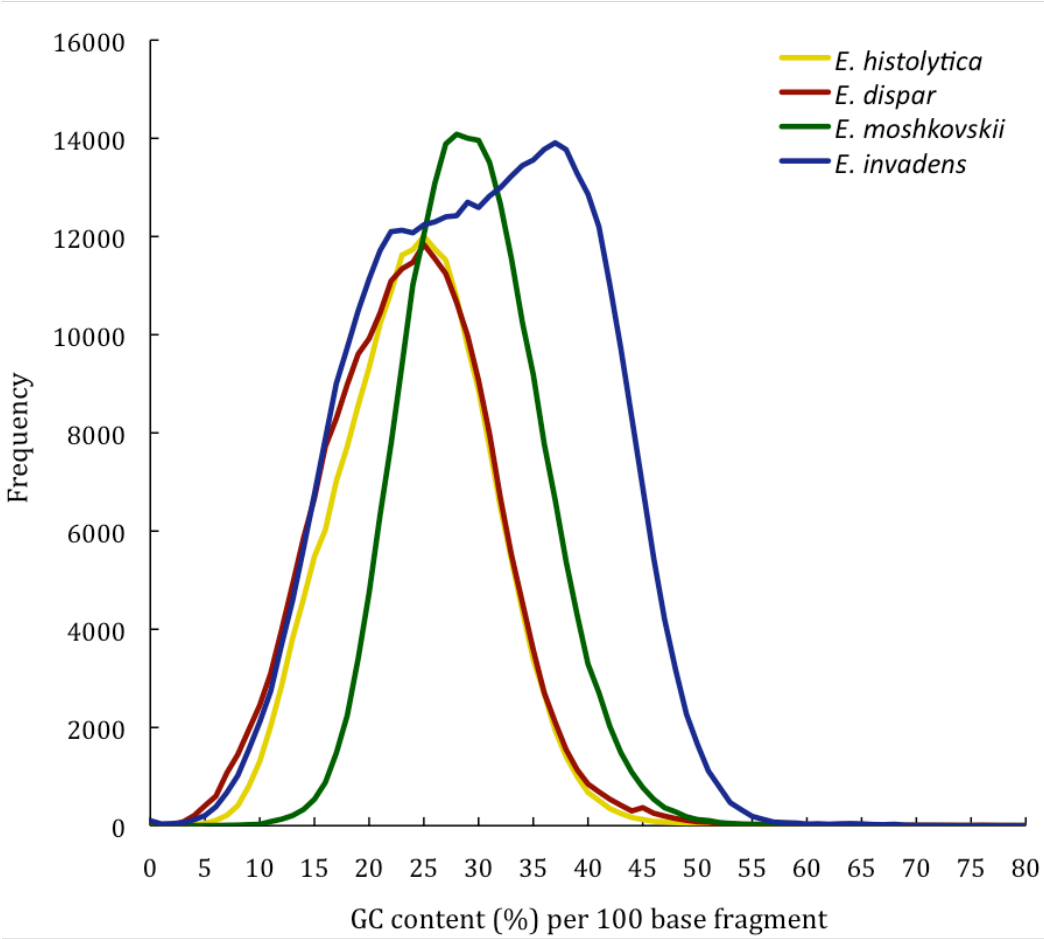


Figure 2.3.2. The range of GC contents in 100 base sections of reference genome assemblies for *Entamoeba histolytica*, *Entamoeba dispar*, *Entamoeba moshkovskii* and *Entamoeba invadens*. In total, 99.19% of the *E. histolytica* assembly was included, as was 98.49% of the *E. dispar* assembly, 88.75% of the *E. moshkovskii* assembly, and 98.47% of the *E. invadens* assembly.

LINEs and some SINEs in *E. histolytica* exceed the length of the reads generated for Laredo [55, 57]. As such, the Newbler genome assembler may have been unable to unambiguously assemble such regions in Laredo. This would have lead to reads erroneously being mapped to the same location, artificially inflating coverage depths at those sites [184, 185]. Despite this common assembly artefact, the high peak coverage depth still means that one can be confident that the sequenced and aligned bases of the genome have been read accurately.

The total concatenated length of the genome was 25,247,493 bp. This is slightly larger than the genomes of the closely related *E. histolytica* and *E. dispar*, but far shorter than that of the more distant *E. invadens* (Table 2.3.1). However, *E. moshkovskii* is the only one of the four *Entamoeba* species featured in this chapter whose genome includes contigs not mapped to scaffolds. It is possible that some of these contigs are positioned in large gaps within scaffolds, such as those that can occur across repeat regions, as well as some being situated between or beyond the regions covered by scaffolds. As such, it is uncertain exactly how many of these contigs actually extend the assembly length beyond that defined by the scaffolds. Without further information, it must be assumed that all of the contigs lie between or beyond the scaffolds and so constitute 12.82% of the total genome, meaning the total length of the *E. moshkovskii* genome represented by scaffolds is actually similar to those of the *E. histolytica* and *E. dispar* genomes. One might expect this given the relatively short evolutionary distance between the three species [189].

Additionally, it must be remembered that the ploidy of *E. moshkovskii* is not known. When assembling genomes with a ploidy of two or greater, assemblers can encounter problems with heterozygous regions. In these cases, assembly algorithms cannot resolve the differences to create one individual contig and, instead, contruct multiple contigs, one for each copy of the region. This makes the two contigs appear to be the result of a segmental duplication, as opposed to an assembly error, artificially lengthening the assembly [190]. Such regions have caused problems in a variety of assemblies [190], from highly polymorphic genomes, including those of *Candida albicans* [191] and *Anopheles gambiae* [192], to the far less variable mouse genome [193, 194], indicating that this is a pervasive problem in *de novo* assemblies.

If *E. moshkovskii* were known to be haploid then there would be no such complication. However, the ploidy of *E. histolytica* is unconfirmed but the species is

thought to be diploid or tetraploid [64, 65]. With such a close phylogenetic relationship to the more studied species, *E. moshkovskii* cannot confidently be thought of as haploid. It is, therefore, possible that sections of the assembly contain multiple copies of heterozygous regions of the genome, resulting in the haploid assembly being larger than the haploid genome itself. It is hoped that the development of a new assembly file format, such as FASTG (http://fastg.sourceforge.net; [112]), will allow for better representation of heterozygous regions, paving the way for more accurate assemblies. However, until assembly algorithms can accurately distinguish between heterozygous regions of a genome and segmental duplications, erroneous segmental duplications are likely to remain an issue in *de novo* genome assemblies.

Furthermore, unsequenced bases, used to fill gaps between contigs in a scaffold, make up 2,509,483 bp of the total length. This accounts for a considerably higher proportion of the genome than the equivalent in the reference strains of *E. histolytica*, *E. dispar* and *E. invadens* (Table 2.3.1). Therefore, whilst it can be considered of high quality, a second mapping-based assembly could still considerably improve the assembly in the future. A second round of sequencing and assembly may also serve to concatenate some of the contigs into scaffolds and resolve potential heterozygous regions misrepresented as segmental duplications.

Table 2.3.1. Statistics relating to the genome assemblies of *Entamoeba histolytica* HM-1:IMSS, *Entamoeba dispar* SAW760, *Entamoeba invadens* IP-1 and *Entamoeba moshkovskii* Laredo. Statistics are derived from AmoebaDB v2.0 data, except for asterisked (*) figures, taken from NCBI WGS Projects AANV02 and AANW03; and the double-asterisked (**) figure, taken from [49].

Statistic	*E. histolytica*	*E. dispar*	*E. invadens*	*E. moshkovskii*
Genome length (bp)	20,799,072	22,955,291	40,888,805	25,247,493
GC content (%)	24.20	23.53	29.91	26.54
Non-ACGT (%)	0.31	0.56	0.93	9.94
Number of scaffolds	1,496	3,312	1,149	1,147
N50 of scaffolds (bp)	49,118	27,840	243,235	40,197
Average scaffold size (bp)	13,903	6,931	35,586	19,190
Number of contigs	-	-	-	3,460
Average contig size (bp)	-	-	-	935
Average coverage depth	12.5x**	4.32x*	4x*	82.65x

Table 2.3.2. Genomic comparison of *Entamoeba histolytica* HM-1:IMSS, *Entamoeba dispar* SAW760, *Entamoeba invadens* IP-1 and *Entamoeba moshkovskii* Laredo. Annotation files upon which statistics are based were obtained from AmoebaDB v2.0. Note, *E. histolytica* statistics only include protein-encoding genes. They do not include the 27 tRNA-encoding genes in the genome that bring the total gene count to 8,333 genes. A total of 458 CEG families exist.

Statistic	*E. histolytica*	*E. dispar*	*E. invadens*	*E. moshkovskii*
No of CDSs	8,306	8,748	11,549	12,449
Avg Gene size (bp)	1,280	1,259	1,401	1,230
% coding DNA	50.12	46.62	38.01	59.04
Avg protein size (aa)	418	408	449	399
Avg Intergenic dist (bp)	1,223	1,365	2,139	798
Proportion of multi-exon genes (%)	24.16	30.73	34.48	26.24
Avg intron size (bp)	74	81	104	89
Avg no of introns per spliced gene	1.27	1.34	1.48	1.31
Number of CEG families	372	358	356	368

2.3.2 Prediction of gene models in the *Entamoeba moshkovskii* Laredo genome

Gene models were predicted for the newly sequenced *E. moshkovskii* Laredo genome so as to establish the genetic content and functions available to this little-studied species. The gene models used to train gene prediction software AUGUSTUS totalled 197, with 57 of these having multiple exons in an approximate estimate of the 3:1 ratio seen in the *E. histolytica* gene set (for full training sequences, see Appendix C, File C.1). The preliminary *E. moshkovskii* Laredo gene set produced by AUGUSTUS totalled 15,711 genes. The 6,092 sequences with RBHs against *E. histolytica* genes were accepted as legitimate models. As determined by Wilcoxon rank sum tests, the 9,619 models without RBHs demonstrated statistically significantly different characteristics to the other group in both model lengths (W = 39453778, p-value < 2.2e-16, two-tailed test) and AUGUSTUS confidence scores (W = 39530924, p-value < 2.2e-16, two-tailed test). As such, these two characteristics were used to discriminate between 'good' and 'poor' gene models (Figure 2.3.3).

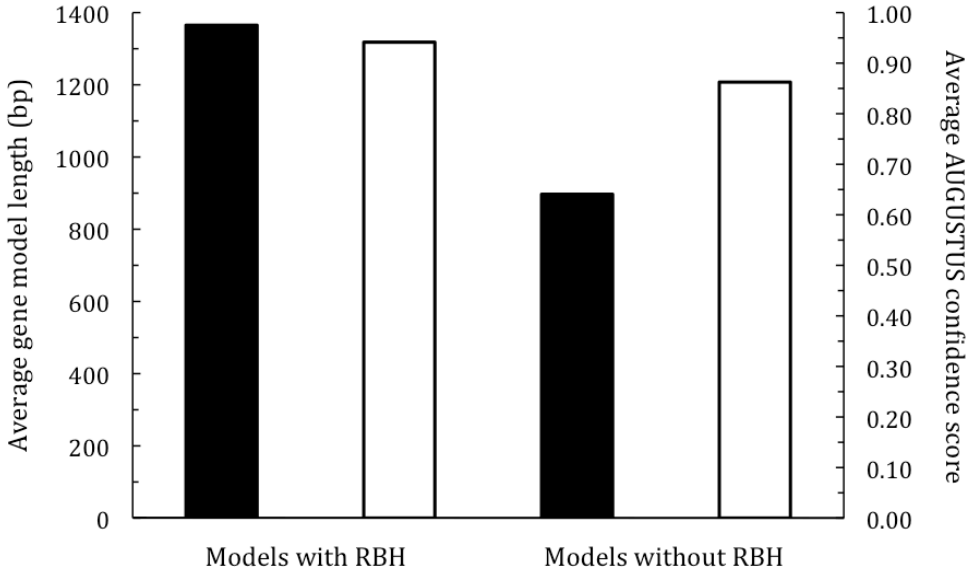


Figure 2.3.3. Average AUGUSTUS confidence scores and average lengths of predicted *Entamoeba moshkovskii* Laredo gene models with and without RBHs against genes in *Entamoeba histolytica* HM-1:IMSS. Black columns show average gene lengths; white columns show average AUGUSTUS confidence scores.

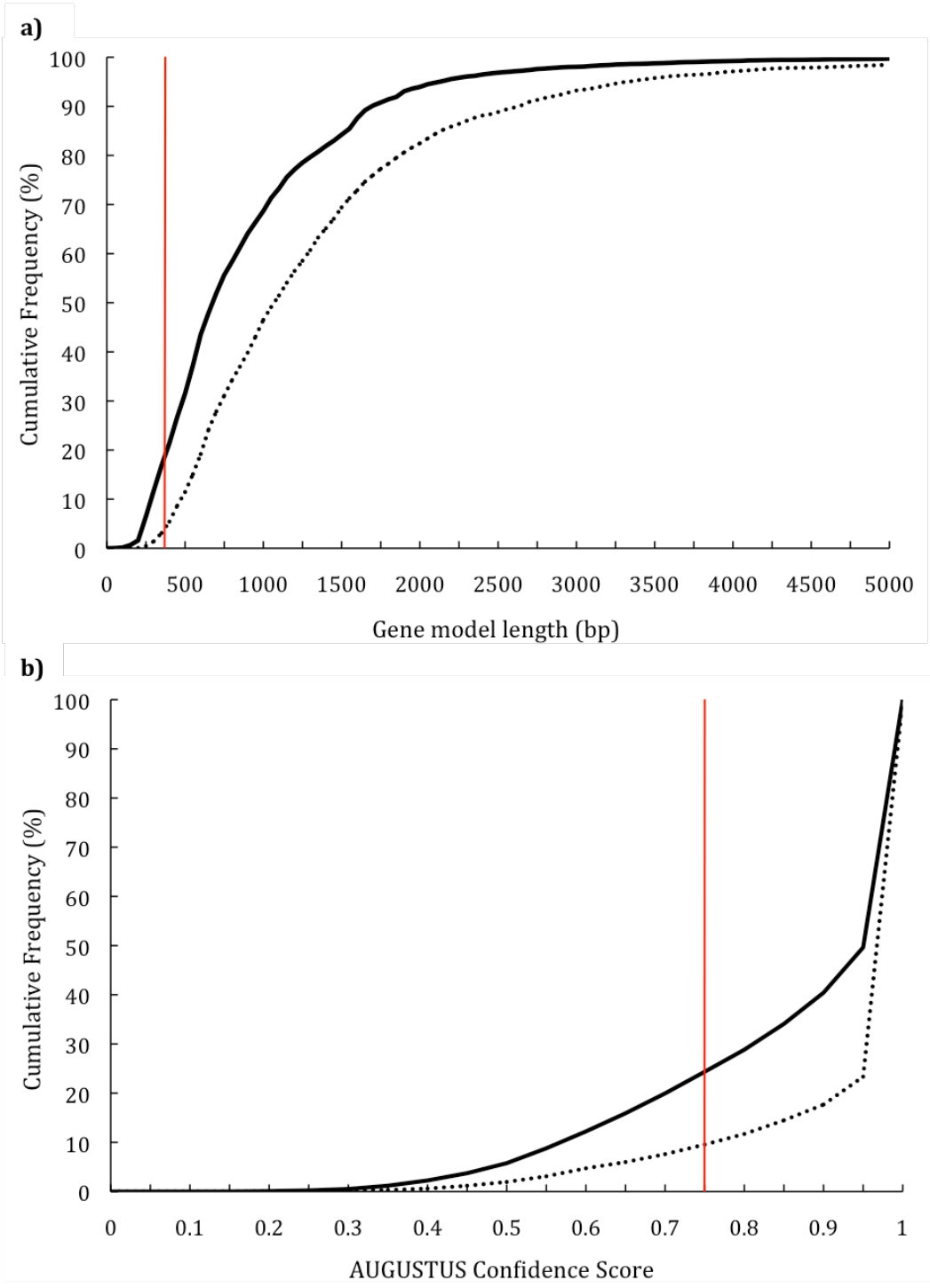


Figure 2.3.4. Cumulative frequency comparisons of a) lengths, and b) AUGUSTUS confidence scores, of those *Entamoeba moshkovskii* gene models with a RBH against an orthologue in *Entamoeba histolytica* and those without a RBH. Dashed lines represent models with an RBH; solid lines represent models without an RBH. Red lines represent cutoff values applied to define 'high-score' and 'long' models.

Approximately 95% of *E. histolytica* HM-1:IMSS CDSs are greater than 350 bp long. This length was chosen as an appropriate cutoff value below which non-RBH *E. moshkovskii* gene models could be omitted. The criterion removed 1,622 gene models. Only 190 sequences with an RBH would have been omitted by this method, had the RBH gene models been included, demonstrating the difference between the two sets of sequences and the appropriateness of 350 bp as a cutoff value (Figure 2.3.4.a). Approximately 90% of genes with RBHs were assigned an AUGUSTUS confidence score of 0.75 or greater. Non-RBH models with an AUGUSTUS score equal to, or greater than, this value and with lengths exceeding 350 bp were, therefore, accepted as legitimate models in the gene set (Figure 2.3.4.b). Of the 2,503 gene models greater than 350 bp in length but possessing relatively poor AUGUSTUS confidence scores, 863 gene models hit *E. histolytica* HM-1:IMSS genes with an E-value of at least 1e-5 when entered into a BLASTP search against them. When added to the models that met the previously described criteria, *E. moshkovskii* was predicted to possess 12,449 genes.

The number of genes attributed to *E. moshkovskii* is greater than those seen in any of the other three species studied in this chapter (Table 2.3.2). The average length of the *E. moshkovskii* predicted gene models is similar to those of *E. histolytica* and *E. dispar*, but slightly shorter than that of *E. invadens*. Taken together, these values mean that *E. moshkovskii* is also predicted to contain the greatest proportion of coding DNA, separated by the shortest average intergenic distances (Table 2.3.2). However, the concatenated contigs not assembled into scaffolds may artificially inflate the gene count in the *E. moshkovskii* Laredo assembly, as is discussed in Section 2.3.3.

The proportion of multi-exon genes in *E. moshkovskii* is predicted to be similar to the proportions seen in *E. histolytica* and *E. dispar*, with a value inbetween those of the two other species. The average number of introns in multi-exon genes is similarly related to the values of the other two species, whilst the average length of introns in *E. moshkovskii* is higher. As is the case with every statistic concerning gene models other than the percentage of the genome made up of coding DNA, *E. invadens* demonstrates greater values regarding multi-exon genes than the other three species (Table 2.3.2).

As the manually curated gene set used to train AUGUSTUS was based upon gene models in *E. histolytica*, it is unsurprising that the statistics relating to the *E. moshkovskii* gene set are similar to those seen in *E. histolytica*. *E. histolytica*

possesses the best studied of the genomes here and is the only one to have a manually curated assembly and gene set. This improves confidence in the gene set predicted for *E. moshkovskii*. However, it does also come with the caveat that gene models unique to *E. moshkovskii* are at greater risk of being omitted and that any mistakes or omissions in the *E. histolytica* gene set could be carried into the *E. moshkovskii* set.

Such limitations were potentially seen in the results of a BLASTP search against the *Entamoeba* species' sequences with the CEG protein set [114, 167, 168]. In total, 368 of the 458 CEG gene families were found to have orthologues in the Laredo gene set, with only 4 more families being present in HM-1:IMSS. This may be indicative of a deficiency in the *E. histolytica* annotation, which led to a similar deficiency in *E. moshkovskii.* However, only 358 CEG families are found in *E. dispar* and only 356 in *E. invadens*, suggesting that this is actually a result of the relative distance between *Entamoeba* and the genera used to generate the CEG database. The BLAST results imply that the CEG protein set should be reduced to account for more distant species. The relative relationships of the *Entamoeba* genus with other eukaryotic genera are investigated further in Section 2.3.5.

The risks inherent in using automated gene prediction software must also be considered. During an update of physical annotations in the gene set to be included in the release of AmoebaDB v4.0, after much of the work for this project had been completed, 6 genes in *E. moshkovskii* Laredo were found to have incorrect coordinates (Appendix A, Table A.3). However, the changes made were minimal and should not have affected the BLAST searches and subsequent generation of orthologous groups carried out in this chapter to a great degree, if at all. This indicates an obvious drawback of the method used, one that is a risk with all annotations of this nature. In addition to the possibility of exon boundaries having been inaccurately modelled, some gene models may have been omitted altogether, whilst others may have been annotated where no CDS actually exists. For example, it is possible that the potential pitfalls associated with assembling a genome of unknown ploidy (see Section 2.3.1) have resulted in individual gene sequences being annotated multiple times. It is impractical to manually curate an entire genome, however, and so the benefits of an expedient method greatly outweigh the relatively small cost in accuracy that may be introduced. The use of transcriptomic data would be likely to identify where CDSs exist, as well as their accurate exon boundaries. Several genome assemblies have subsequently been improved through the use of transcriptomic data [195-197] and it is

likely that the *E. moshkovskii* Laredo assembly would benefit from the application of such data in future. Overall, however, it can be said with confidence that the *E. moshkovskii* Laredo assembly represents a good first draft of the genome, which provided enough data to make the comparative genomic study that follows viable and informative.

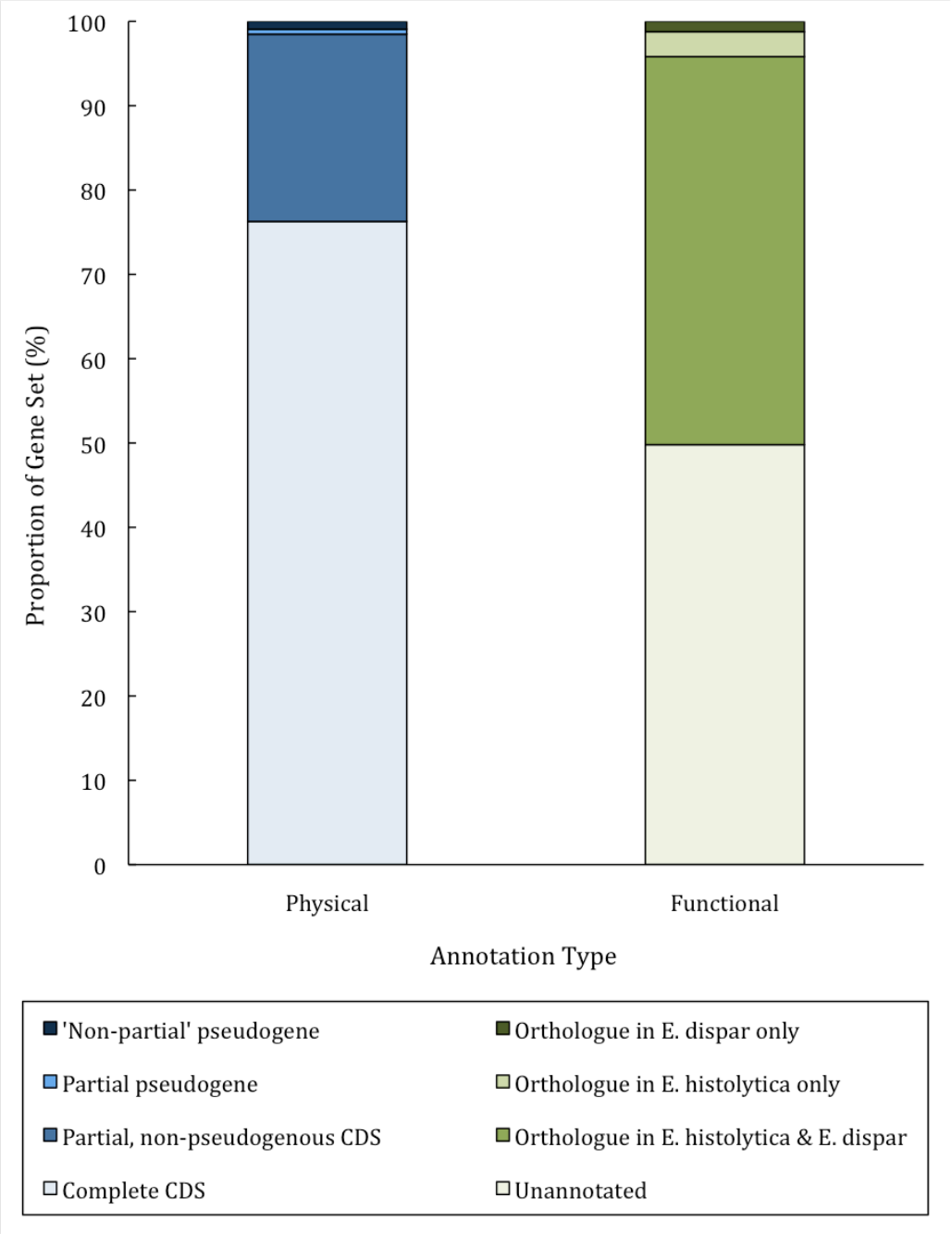


Figure 2.3.5. Proportions of functional and physical annotations in *Entamoeba moshkovskii* Laredo gene models. Of the combined 2,839 partial models: 31.91% are partial at the 5' end only; 39.70% are partial at the 3' only; 20.75% are partial at both ends; 6.90% are partial within the sequence only; 0.32% are partial within the sequence and at the 5' end; and 0.42% are partial within the sequence and at the 3' end. No gene models are partial within the sequence and at both ends.

2.3.3 Adding structural and functional annotations to gene models

Any incomplete gene models, potentially arising as artefacts of the assembly, were identified and annotated accordingly as part of a community effort to improve annotation of *Entamoeba* genomes. Incomplete models were defined as those that began or ended at the last base of a scaffold or contig, were situated within 30 bp of an unsequenced region, or consisted of complete triplet codons yet lacked a methionine start codon or a stop codon. Several sequences predicted by AUGUSTUS contained internal stop codons as a result of splice junctions unevenly splitting triplet codons. These 189 models were annotated as pseudogenes. A total of 9,495 genes in the *E. moshkovskii* gene set are predicted to be complete gene models (Figure 2.3.5).

This left 2,954 incomplete gene models, including both partial genes and pseudogenes. AUGUSTUS was theoretically incapable of predicting pseudogenes, so the presence of at least 189 pseudogenes in its output highlights its limitations in this respect. Unfortunately, as stated above, such relatively limited errors are a necessary, and worthwhile, cost of the speed and overall quality of automated predictions. The accuracy of these predicted pseudogenes cannot be determined without further detailed analysis and transcriptional evidence, which fall beyond the remit of this project.

The contribution of the gene models labelled as partial to the proportion of coding DNA seen in *E. moshkovskii*, as well as the total number of genes predicted for the Laredo genome, is of great interest. Of the 2,839 partial genes, 1,855 are predicted to be present in contigs, rather than scaffolds. The average size of these contigs (935 bp) is lower than the species' average gene size (1,230 bp). As such, it is perhaps unsurprising that 75.68% of the gene models encoded within the contigs are incomplete, being annotated as partial, pseudogenes or both. As is the case for all of these partial genes, it may be that AUGUSTUS predicted gene models in some of these regions where none actually existed, or individual genes may have been split into multiple parts if they spanned an unsequenced region, in which case one gene could have been annotated as multiple separate sequences. Alternatively, the inflated gene count could be indicative of a more diverse 'lifestyle' than is seen in the obligately parasitic relatives of *E. moshkovskii*. Parasites generally evolve to have reduced genetic material as they can manipulate their hosts to provide for them [198]. It is conceivable that *E. histolytica* and *E. dispar* have evolved in such a manner whilst *E. moshkovskii*, if

only facultatively parasitic, would require a larger gene repertoire to survive when outside of a host. Again, further analysis, perhaps in the form of a second round of sequencing, should be considered to improve this draft annotation.

To add functional annotations to gene models, the *E. moshkovskii* Laredo protein set was entered into reciprocal BLASTP searches against the protein sets of *E. histolytica* HM-1:IMSS and *E. dispar* SAW760, using default parameters. Where an *E. moshkovskii* Laredo protein had a RBH against a protein from either of the other species' sets, the gene by which it was encoded was annotated with the same function as its orthologue. CDSs with RBHs in both *E. histolytica* HM-1:IMSS and *E. dispar* SAW760 were thus functionally annotated twice. The majority of genes in the *E. moshkovskii* genome with functional annotations are orthologous to genes in both *E. histolytica* and *E. dispar* (Figure 2.3.5). In total, 50.20% of predicted genes are functionally annotated. Their functions are studied in greater detail in Section 2.3.4. The final set of gene models, and the concatenated assembly upon which they were based, have been made publicly available as part of AmoebaDB v2.0, released on 11th March 2013. Functional and structural annotations were included in AmoebaDB v4.0, released on 8th May 2014.

2.3.4 Species-specific gene sets and a core *Entamoeba* gene set

Strains within a species share physiological characteristics and genetic sequences that define the species. In turn, species within a genus share traits that define the genus. By comparing the genetic content of four *Entamoeba* species – *E. histolytica*, *E. dispar*, *E. moshkovskii* and *E. invadens* – I endeavoured to identify the gene families that are found exclusively in each of the four species as well as those conserved in the genus, which constitute a 'core *Entamoeba* gene set'. Taken together, the generated data have shed more light on the poorly characterised genus.

i. Identifying orthologous clusters

When clustering orthologous sequences from *E. histolytica*, *E. dispar*, *E. invadens* and *E. moshkovskii* using OrthoMCL v2.0.3, a 50% cutoff value was applied. This was based upon the average similarity of 'hits' in a reciprocal BLASTP search, using default parameters, between the protein sets of the two most evolutionarily distant species, *E. histolytica* and *E. invadens* (54.19%). Whilst it will not have included

every sequence in the clustering stage, this method was considered relatively inclusive without being so lenient as to introduce false positive associations between sequences from the more closely related species.

MCL was tested with multiple clustering granularity values (1.2, 2.0, 3.0, 4.0 and 5.0). A value of 3.0 produced the greatest number of direct orthologues (i.e. clusters containing one gene from each species). Given that the aim of the task was to identify orthologous clusters, the output generated using this value was selected. All proteins from all four species were included in the comparison, with none being filtered out. Whilst both the cutoff scores and the granularity values used could have been refined further, the approaches used were deemed an acceptable trade-off between expediency and accuracy.

Overall, 4,704 gene families generated by OrthoMCL were shared by all four species (Figure 2.3.6; Appendix C, File C.2). Comprising 21,741 genes, these families can be tentatively considered a core set that all *Entamoeba* species possess. The number of genes unique to each species positively correlates with the total number of genes in their genomes. A total of 4,624 genes were found to be unique to the newly annotated *E. moshkovskii*, of which 1,394 were 'singletons', belonging to no paralogous family. The remaining 3,230 sequences comprised 605 paralogous clusters.

The total number of gene families shared between pairs of species reflects the phylogenetic relationship between the species: *E. histolytica* and *E. dispar* share the largest number of gene families at 6,134. *E. histolytica* shares slightly more gene families with *E. moshkovskii* than *E. dispar* does (5,831 compared with 5,648). *E. invadens* shares 5,242 families with *E. moshkovskii*, 170 more than it has in common with *E. histolytica* and 303 more than it shares with *E. dispar.* This suggests that OrthoMCL performed a logical analysis of the datasets. There are, however, limitations to the usefulness of the generated core gene set. The comparisons were not wholly inclusive of the *Entamoeba* genus. Of the four species analysed, two are known to be human-infective and one is strongly suspected of being so, suggesting a strong relatedness that is reinforced in their phylogenetic arrangement [45]. The only relatively distant species in the comparison is *E. invadens*, which means that the majority of *Entamoeba* species are not included. As such, it is likely that the core *Entamoeba* gene set is an overestimate of the real number of shared genes in the genus.

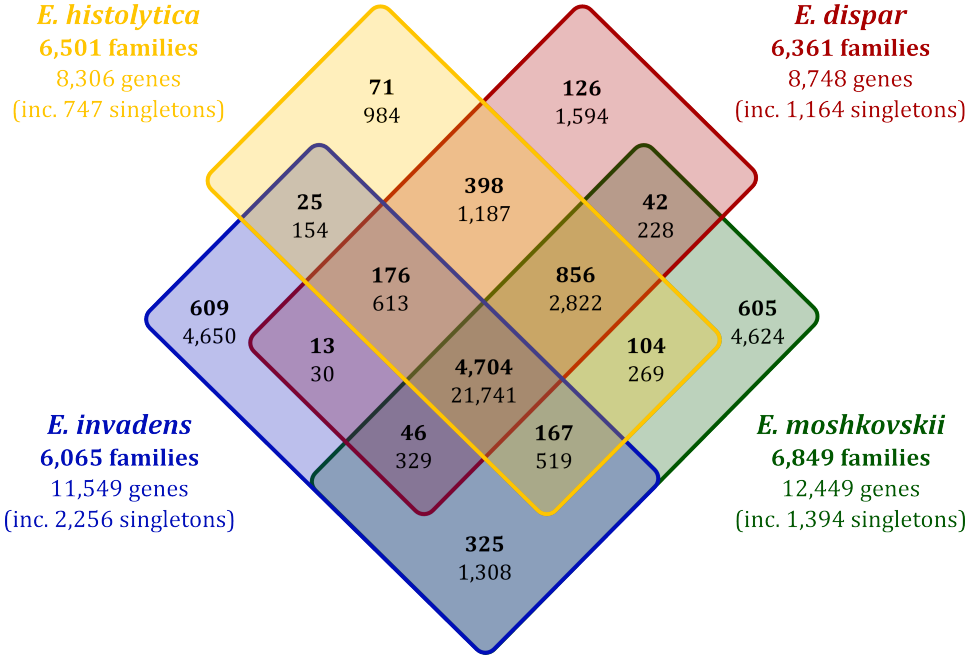


Figure 2.3.6. Venn diagram showing numbers of unique and orthologous genes and families in the genomes of *Entamoeba histolytica, Entamoeba dispar, Entamoeba invadens* and *Entamoeba moshkovskii*. Numbers are based upon OrthoMCL output. Numbers in bold represent gene families; numbers in regular font represent individual gene numbers.

ii. Species-specific gene families

Analysis of the gene families unique to each of the species (except for *E. moshkovskii* for which non-orthologous sequences obviously have no annotations) revealed several interesting characteristics about these members of the *Entamoeba* genus (for full lists of functions, see Appendix A, Table A.4). In *E. histolytica*'s unique gene set, three of the four most prevalent families encode surface proteins. The largest group of genes encodes BspA sequences. The vast BspA family (115 sequences in *E. histolytica* alone) lies within one of seven subfamilies containing leucine-rich repeat regions. At least one BspA protein in *E. histolytica* is known to be located on the plasma membrane of trophozoites [199] and BspA proteins are known to be play roles in adhesion to extracellular membranes in both *Bacteroides forsythus* and *Trichomonas vaginalis* [200-202]. As such, the family has a clear potential role in amoebiasis, which, it is interesting to suggest, could explain why the virulent *E. histolytica* has such a

unique expanded set of BspA genes. Of course, as only a subset of the 115 BspA genes in *E. histolytica* are unique to the species, it is possible that members of the family exist in several of the *Entamoeba* species featured in this project. This supposition is revisited in Section 2.3.6, Part i.

Twelve genes in two families annotated as 'mucin-like protein 1 precursors' and 'mucin-5AC' may also play a role in adhesion, but to the intestinal mucosal layer. This, however, is speculation based upon ambiguous annotation, although it would fit with the theory that *E. histolytica* must possess some genes that allow it to establish infections where *E. dispar* cannot. It would be interesting to study how a 'knock out' of these genes would affect trophozoites' ability to adhere to the mucosal membrane.

Eighteen ariel1 surface proteins are found in the *E. histolytica*-exclusive gene set, as well as 2 orthologous serine-rich antigen proteins. Their presence in this gene set confirms past research, which noted that the ariel1 family was present in *E. histolytica* but not in *E. dispar* [203]. The family belongs to the same larger family as the SREHP protein [204], which has been shown to have some use in immunising against amoebic infection [205]. It is interesting to note that *E. dispar* does not have unique copies of any of these three surface-bound gene families and, according to AmoebaDB, it has no copies of ariel1 whatsoever, giving credence to the theory that proteins involved in adhesion play essential roles in establishing infections that distinguish between *E. histolytica* and *E. dispar*.

Twelve members of the AIG1 family are present only in *E. histolytica*, whilst 13 are found only in *E. dispar*. These GTPases, originally isolated in *Arabidopsis thaliana*, are thought to confer resistance to bacterial infections [206, 207], and have been shown to be more highly expressed in virulent *E. histolytica* cell lines [15]. The presence of commensal gut microbiota in the species' trophozoites' environment means it is logical for them to have a large number of genes encoding AIG1 proteins (49 in total in *E. histolytica*). What is unclear is why the two species have so many unique copies when they are challenged by the same microbiota. Whilst unlikely to be a key factor in explaining how invasive amoebiasis begins, this is an interesting difference between the two species.

Of arguably greater interest are the unique genes encoding 7 cysteine proteases and 3 peroxiredoxins in *E. histolytica*. The occurrence of these sequences is

perhaps to be expected given the families' roles in invasion and survival of ROS, respectively; abilities that are known to be key parts of *E. histolytica*'s pathogenic repertoire [14, 208, 209]. However, all of the peroxiredoxin sequences are pseudogenes, as are 5 of the 7 cysteine proteases. The two complete cysteine protease genes are expressed at relatively low levels with FPKM values of 6.39 and 52.06, compared with the most highly expressed cysteine protease's value of 7,729.12. As such, it would appear that these sequences confer no advantage upon *E. histolytica* and offer no explanation as to why this species is pathogenic in humans whilst the others are not. It is possible that these genes are not as significant as once thought. Their evolutionary relationships to other sequences in their families are studied in Section 2.3.7, Parts i and iv.

There are many more unique genes and families in *E. invadens* than in *E. histolytica* and *E. dispar*, the latter of which possesses a relative paucity of unique genes. Much of this is likely a direct result of its larger gene complement. *E. invadens* possesses genes that encode a number of cysteine proteases, thioredoxin proteins, heat shock proteins and lysozymes. Without greater analysis of the variation between the members of these families, it is impossible to say whether or not this large number of unique virulence factors is required to allow infection of *E. invadens*' range of hosts. Conspicuous by its absence, however, is the large Gal/GalNAc lectin subunit family. Given that *E. invadens* is capable of causing amoebic infections in a variety of reptilian hosts, it might be expected that it would possess a number of host-binding lectin subunits not seen in its human-infective relatives. It would appear that, regardless of target host, these proteins share enough similarities to be considered orthologous. This is considered in greater depth in Section 2.3.8 Part iii.

As in all three species, the majority of the genes unique to *E. invadens* are housekeeping genes, not typically associated with virulence. For example, *E. invadens* possesses 214 unique protein kinases and a large collective number of genes involved in cytoskeletal rearrangement. The existence of such paralogous gene clusters is indicative of duplication events occurring independently within each species' genome [153]. What cannot be gleaned simply from their existence, however, is how these paralogues have impacted upon diversification within the *Entamoeba* species.

Whilst the duplicates have obviously become fixed in the populations, they could have evolved since their duplication in a number of ways. Firstly, they may have

not accrued many, or any, mutations if their duplication of an existing gene function was advantageous to the organism. This is rare but it is possible that a capacity to produce a greater number of transcripts could be selected for and the two gene sequences conserved [210, 211]. More often, one of the duplicates will accrue mutations, as it is functionally redundant. This can result in a slight alteration to the function of the original gene, or it can cause changes that confer a distinctly different function upon the duplicated sequence [210-212]. Alternatively, as the duplicate is not under any selective pressures, it can accumulate mutations that alter the structure or functional capacity of the gene, which would otherwise be selected against in non-redundant genes. This can result in pseudogenisation of the gene [211], as is seen in a number of paralogous clusters in *E. histolytica* (Table A.4). Whilst their exact functions and the roles they play in distinguishing between lifestyles cannot be known without further investigation, it is important to note that these genes, in addition to the suspected virulence factors mentioned above, play important roles unique to the *Entamoeba* species investigated here.

iii. Gene ontologies within the species-specific gene sets

Whilst studying the individual gene families and the proteins they encoded was highly informative with regards to certain families, it revealed nothing about unannotated ones. As such, I chose to attribute GO terms to each of the genes unique to each species and count the occurrences of each GO term in the gene sets. It was hoped that this would give a more general overview of the gene functions specific to the different *Entamoeba* species. However, the GO terms' annotations were found to be too specific to the genes for which they were originally generated, resulting in several nonsensical annotations, including 'compound eye development (GO:0048749)' and 'embryo development (GO:0009790)'.

To attain a more general, and therefore useful, set of annotations, these GO terms were divided into the subgroups 'Components', 'Functions' and 'Processes' within each species, and entered into the online program CateGOrizer to be grouped into more general GOA Slim categories. An individual GO term can be associated with multiple GOA Slim categories and can be attributed to a GOA Slim category multiple times, along different paths of a directed acyclic graph [213]. As such, the total numbers of associations of GO terms with GOA Slim categories can be higher than the number of GO terms input. Of the 4 generic GO Slim classifications offered by

CateGOrizer, the GOA Slim annotations were chosen as they grouped GO terms together into a meaningful, but limited, set of categories. Two general GOA Slim categories, which included the majority of the GO terms, were omitted from each of the three subgroups: in Components, 'GO:0005575: Cellular Component' and 'GO:0005623: Cell'; in Functions, 'GO:0003674: 'Molecular Function' and 'GO:0003824: Catalytic Activity'; and in Processes, 'GO:0008150: Biological Process' and 'GO:0009987: Cellular Process'.

Pairwise and multi-sample Pearson's Chi-squared tests were used (with the Monte Carlo simulation where counts were less than 5) to compare the proportions of GO terms allocated to each GOA Slim category for the species-specific gene sets. An alpha level of 0.05 was used for all tests. Statistically significant differences were seen in 7 of the 10 Components categories, 20 of the 21 Functions and 13 of the 17 Processes (Table 2.3.3). Three common patterns accounted for over half of these cases (Figure 2.3.7; Appendix A, Table A.5). In 7 GOA Slim categories, all proportions were similar to one another except for those seen in *E. invadens*. In a further 11 cases only the proportions in *E. histolytica* and *E. dispar* were similar. Finally, in 4 categories, similarities were seen between *E. histolytica* and *E. dispar*, and between *E. invadens* and *E. moshkovskii*.

Starting with the categories exhibiting these three patterns, the proportions of *E. histolytica* and *E. dispar* genes associated with macromolecule metabolism are significantly different to the proportions in the other two species. A similar trend is seen in the 'Metabolism' category in which the proportions differ significantly between all 4 species. The highest proportion of GO terms is seen in *E. histolytica*, followed by *E. dispar,* then *E. moshkovskii*, with *E. invadens* providing the lowest proportion. These differences may, however, be less indicative of a large number of metabolism-related genes in the two human infective species, and more a sign of losses of other types of genes. They both have lower proportions of genes associated with many different processes including cell communication, differentiation and motility, development, and regulation of biological processes. As is often seen in obligate parasites, it is possible that *E. histolytica* and *E. dispar* have lost the ability to perform certain tasks without their hosts [198, 214, 215], but have retained the ability to carry out metabolic tasks.

Secondly, a similar proportion of genes unique to *E. moshkovskii* are involved in antioxidant behaviour as in *E. histolytica* and *E. dispar*, whilst a comparative paucity of

such genes is seen in *E. invadens*. In addition to this, the two obligate human parasites' genes have significantly higher proportions of associations with oxidoreductase activity than *E. invadens* and *E. moshkovskii.* Whilst it should be remembered that these activities are also employed in a non-pathogenic capacity, this is possibly indicative of lower resistance to ROS in *E. invadens*.

One can also see that a greater percentage of *E. moshkovskii's* unique genes are involved in signal transducer activity, required to trigger changes in cell function, than in the other species. A relatively greater number of its genes are involved in cell communication than in the two human-infective species too, although *E. invadens* has a similarly high proportion involved in this process. It is interesting to speculate as to the reason behind these differences. Certainly, it is feasible that *E. moshkovskii* needs to be able to alter cellular functions in order to survive in different environments with different temperatures and acidity levels. However, such suggestions could only be verified by transcriptomic data, which is lacking for *E. moshkovskii* at this stage.

Looking beyond the three main patterns of association, *E. dispar* lacks unique genes associated with components of the extracellular matrix and extracellular space genes, and possesses only 2 extracellular region associations and less than half of the number of protein binding associations as *E. histolytica* (Table 2.3.3). Whilst it is true that the species may possess some genes that have orthologues in one or two of the other species but not all three, the fact that the other species contain greater proportions of unique genes with these associations implies that they have more genes with these functions overall. This supports the findings in Section 2.3.4, Part ii that *E. dispar* perhaps lacks some surface-bound proteins. It may be that its surface composition differs in more ways than just the Gal/GalNAc lectin, which also suggests that there are other genes involved in adhesion that may play a similarly important role to the lectin. This is supported by the fact that a greater proportion of genes unique to *E. invadens* are involved in selective molecular binding than in the other three species despite the lack of unique Gal lectin heavy subunits. A wider range of binding molecules makes sense considering the host range of *E. invadens*.

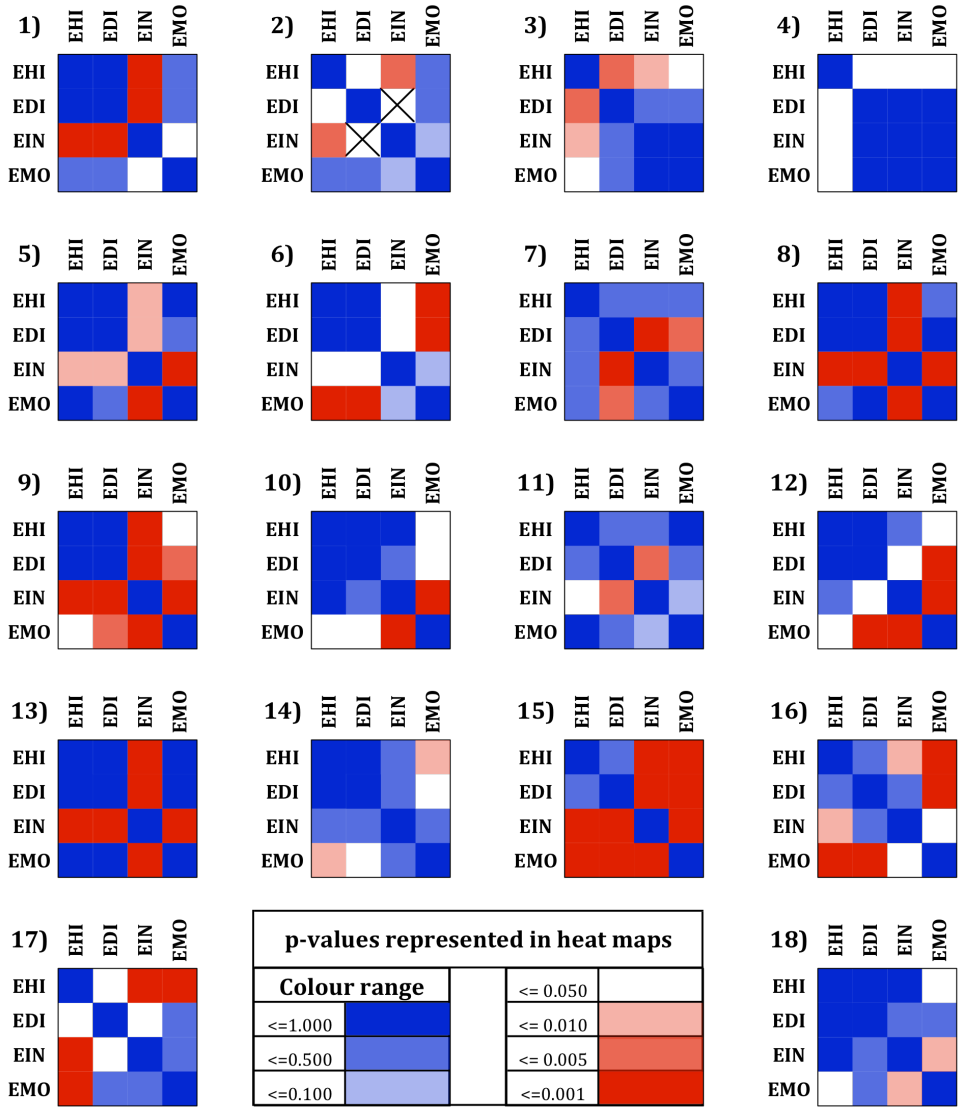
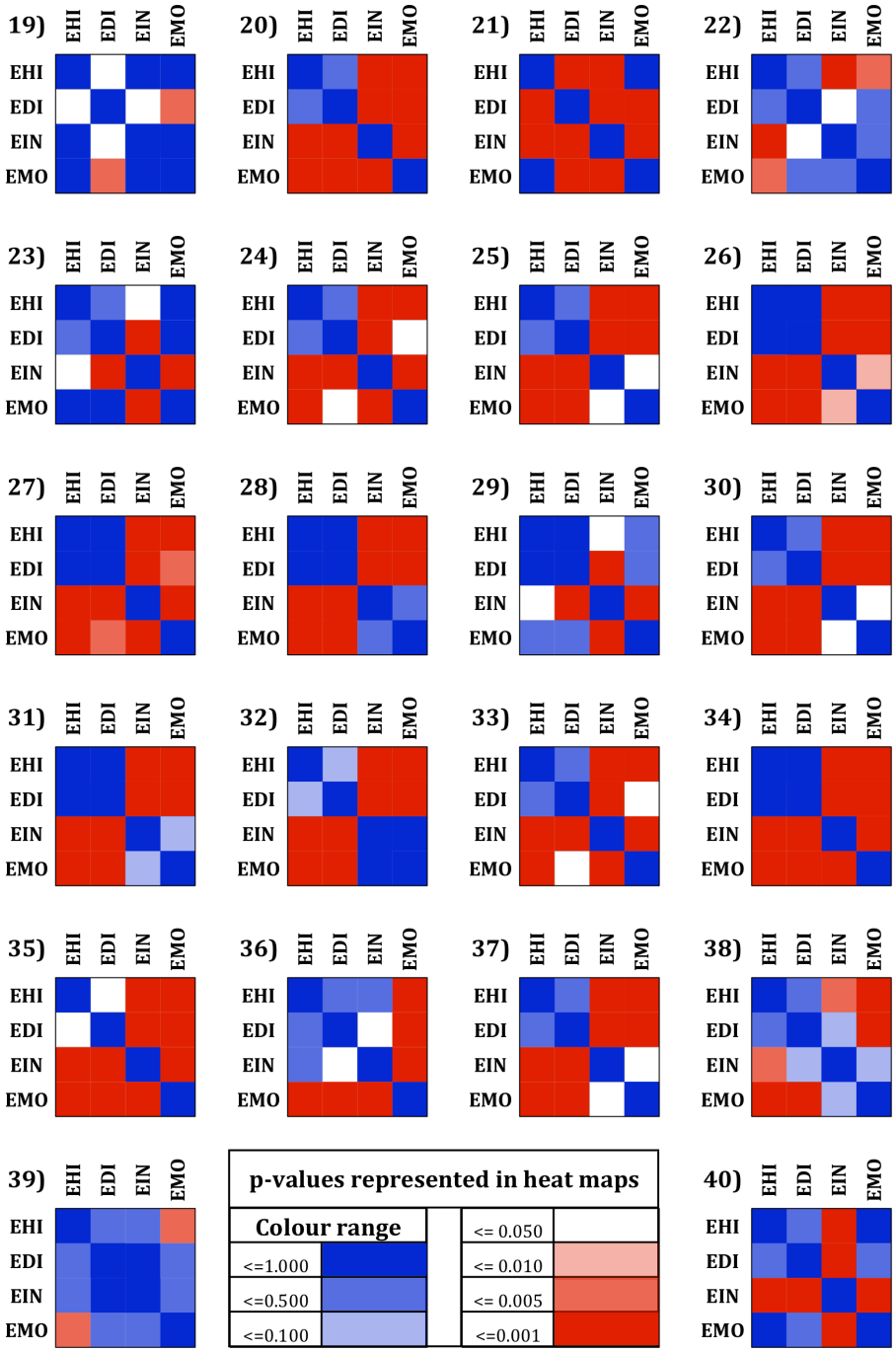


Figure 2.3.7. Significance of Chi-squared test results between proportions of GO terms from *Entamoeba histolytica*, *Entamoeba dispar*, *Entamoeba invadens* and *Entamoeba moshkovskii* in 40 GOA Slim categories. Crosses indicate comparisons in which proportions in both species equalled 0%, so p-values could not be generated. GOA Slim categories:

1) Cytoplasm; 2) Extracellular matrix (sensu Metazoa); 3) Extracellular region; 4) Extracellular space; 5) Intracellular; 6) Membrane; 7) Nucleus; 8) Antioxidant activity; 9) Binding; 10) Enzyme regulator activity; 11) Helicase activity; 12) Hydrolase activity; 13) Ion transporter activity; 14) Isomerase activity; 15) Kinase activity; 16) Ligase activity; 17) Lyase activity; 18 Motor activity (**Continued overleaf)**



(Continued from previous page): 19) Nucleic acid binding; 20) Oxidoreductase activity; 21) Protein binding; 22) Protein transporter activity; 23) Receptor activity; 24) Signal transducer activity; 25) Structural molecule activity; 26) Transferase activity; 27) Transporter activity; 28) Biosynthesis; 29) Catabolism; 30) Cell communication; 31) Cell differentiation; 32) Cell motility; 33) Development; 34) Macromolecule metabolism; 35) Metabolism; 36) Nucleobase, nucleoside, nucleotide and nucleic acid metabolism; 37) Regulation of biological process; 38) Response to stimulus; 39) Secretion; 40) Transport.

Table 2.3.3. GOA Slim categories with which GO terms have been associated in significantly different proportions in *Entamoeba histolytica* HM-1:IMSS, *Entamoeba dispar* SAW760, *Entamoeba invadens* IP-1 and *Entamoeba moshkovskii* Laredo. Chi-squared tests with 3 degrees of freedom were performed with alpha values of 0.05. Asterisked comparisons required the Monte Carlo simulation.

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics
Components					
GO:0005737: Cytoplasm	20.81	21.37	14.62	17.90	χ^2 = 16.9944, p-value = < 0.001
GO:0005578: Extracellular matrix	0.00	0.95	0.00	0.41	χ^2 = 14.6668, p-value = 0.002*
GO:0005576: Extracellular region	0.30	2.48	0.75	1.07	χ^2 = 15.3001, p-value = 0.001*
GO:0005615: Extracellular space	0.00	0.95	0.17	0.08	χ^2 = 15.4265, p-value = 0.002*
GO:0005622: Intracellular	44.49	43.89	51.25	42.61	χ^2 = 20.5095, p-value < 0.001
GO:0016020: Membrane	22.02	19.08	26.66	30.13	χ^2 = 23.2269, p-value < 0.001
GO:0005634: Nucleus	10.11	7.25	5.15	6.24	χ^2 = 17.5126, p-value < 0.001
Functions					
GO:0016209: Antioxidant activity	1.55	1.26	0.05	1.58	χ^2 = 57.9926, p-value < 0.001*
GO:0005488: Binding	30.98	30.39	37.91	27.42	χ^2 = 121.8086, p-value < 0.001
GO:0030234: Enzyme regulator activity	0.85	0.98	1.10	0.43	χ^2 = 15.4211, p-value = 0.001
GO:0004386: Helicase activity	0.60	0.35	0.12	0.35	χ^2 = 9.9565, p-value = 0.019
GO:0016787: Hydrolase activity	14.39	13.66	12.01	10.41	χ^2 = 28.3263, p-value < 0.001

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics
GO:0015075: Ion transporter activity	2.41	2.52	0.62	2.31	χ^2 = 47.0643, p-value < 0.001
GO:0016853: Isomerase activity	0.55	0.70	0.32	0.21	χ^2 = 10.7821, p-value = 0.013
GO:0016301: Kinase activity	7.17	6.23	12.01	14.97	χ^2 = 137.7319, p-value < 0.001
GO:0016874: Ligase activity	0.90	1.33	0.57	0.29	χ^2 = 137.7319, p-value < 0.001
GO:0016829: Lyase activity	0.60	1.47	0.17	0.35	χ^2 = 41.9066, p-value < 0.001
GO:0003774: Motor activity	0.05	0.14	0.15	0.00	χ^2 = 9.0255, p-value = 0.029*
GO:0003676: Nucleic acid binding	6.32	4.62	4.72	4.44	χ^2 = 11.7252, p-value = 0.008
GO:0016491: Oxidoreductase activity	9.12	7.91	2.77	5.26	χ^2 = 126.6868, p-value < 0.001
GO:0005515: Protein binding	2.76	5.88	9.49	5.76	χ^2 = 111.7518, p-value < 0.001
GO:0008565: Protein transporter activity	0.35	0.63	0.07	0.16	χ^2 = 18.0134, p-value < 0.001*
GO:0004872: Receptor activity	0.85	0.56	0.15	0.73	χ^2 = 18.2987, p-value < 0.001
GO:0004871: Signal transducer activity	3.06	2.38	0.50	4.35	χ^2 = 131.5486, p-value < 0.001
GO:0005198: Structural molecule activity	1.85	2.45	0.27	0.62	χ^2 = 81.1735, p-value < 0.001
GO:0016740: Transferase activity	11.23	11.48	15.36	17.43	χ^2 = 62.0115, p-value < 0.001
GO:0005215: Transporter activity	4.41	4.90	1.45	2.86	χ^2 = 66.7445, p-value < 0.001

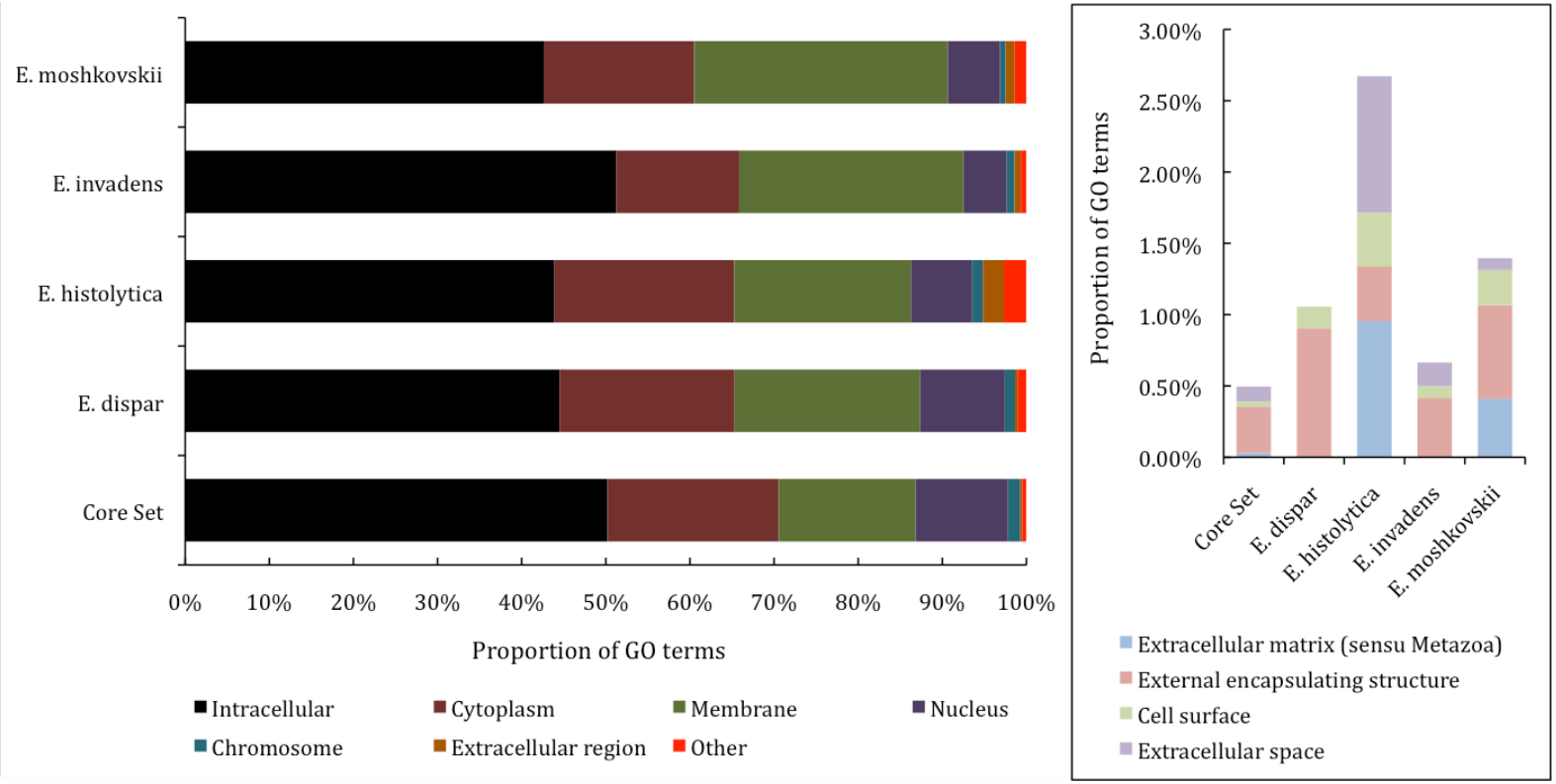
GOA Slim ID	EDI	EHI	EIN	EMO	Statistics
Processes					
GO:0009058: Biosynthesis	5.68	5.36	2.34	2.10	$\chi^2 = 125.4503$, p-value < 0.001
GO:0009056: Catabolism	2.79	3.05	4.35	2.40	$\chi^2 = 37.6463$, p-value < 0.001
GO:0007154: Cell communication	5.89	5.20	8.51	9.70	$\chi^2 = 64.2153$, p-value < 0.001
GO:0030154: Cell differentiation	0.42	0.32	2.67	2.19	$\chi^2 = 79.417$, p-value < 0.001
GO:0006928: Cell motility	0.14	0.43	1.28	1.42	$\chi^2 = 41.4464$, p-value < 0.001*
GO:0007275: Development	2.05	1.55	4.14	2.99	$\chi^2 = 43.8864$, p-value < 0.001
GO:0043170: Macromolecule metabolism	19.13	19.08	12.70	14.88	$\chi^2 = 77.9659$, p-value < 0.001
GO:0008152: Metabolism	30.53	33.44	22.87	25.60	$\chi^2 = 105.3392$, p-value < 0.001
GO:0006139: Nucleobase, nucleoside, nucleotide and nucleic acid metabolism	7.17	6.43	5.84	3.98	$\chi^2 = 51.5633$, p-value < 0.001
GO:0050789: Regulation of biological process	12.00	11.09	16.18	17.73	$\chi^2 = 82.4045$, p-value < 0.001
GO:0050896: Response to stimulus	10.48	9.32	11.95	13.12	$\chi^2 = 27.2197$, p-value < 0.001
GO:0046903: Secretion	0.11	0.27	0.08	0.03	$\chi^2 = 10.2852$, p-value = 0.015*
GO:0006810: Transport	3.11	3.86	6.61	3.54	$\chi^2 = 81.1489$, p-value < 0.001

iv. Gene ontologies within the core *Entamoeba* gene set

Whilst it was practical to manually count the numbers of gene families unique to the individual species, the number of gene families in the core *Entamoeba* set was too large. It was only practical to attribute GO terms to each of the genes and calculate the proportions of these GO terms that were, in turn, associated with the more general GOA Slim categories. It was hoped that this would give an overview of the gene functions found in all *Entamoeba* and reveal some information about the gene set. This was carried out using the same method as used in Section 2.3.4, Part iii.

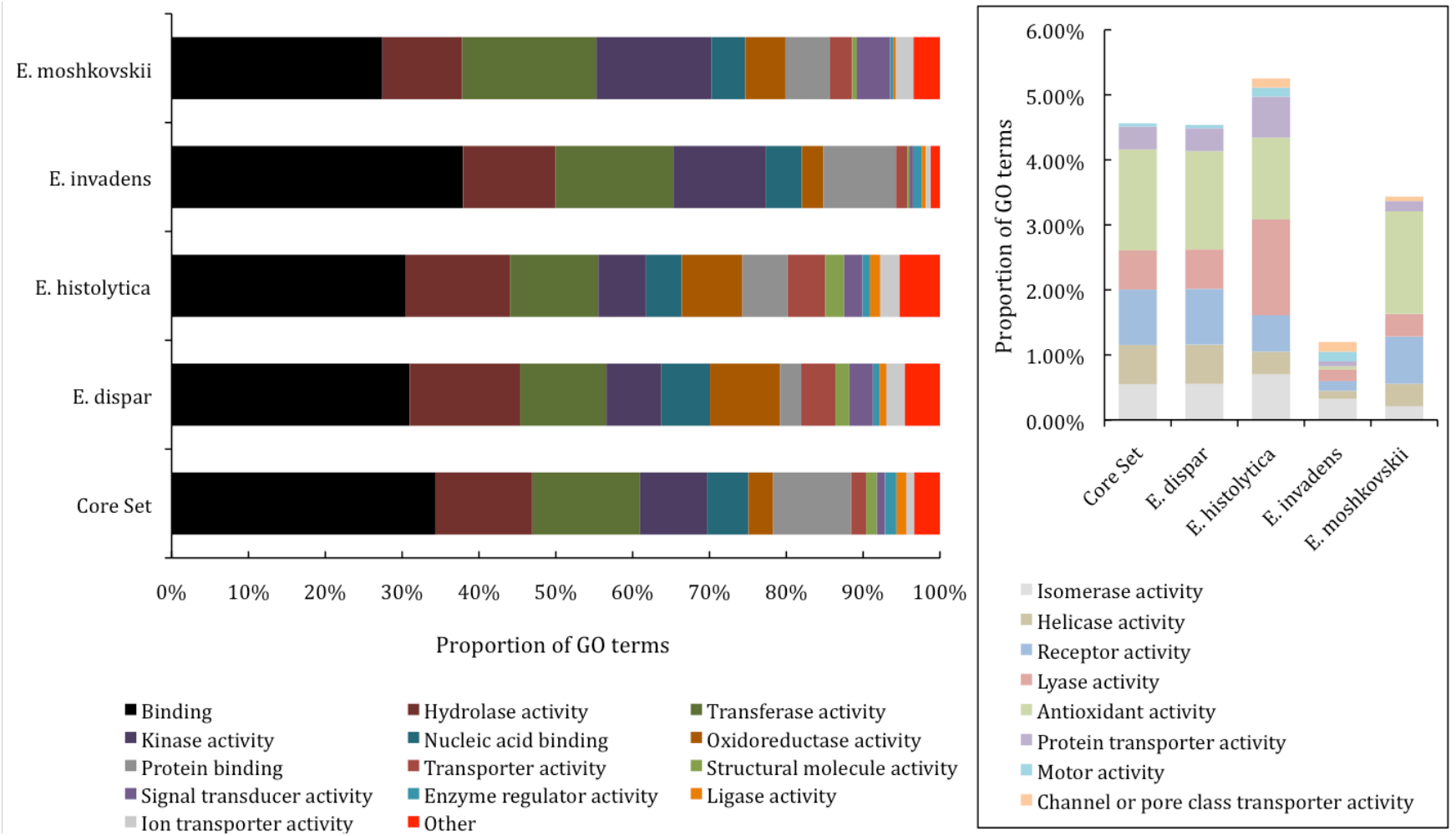
A total of 3,261 core families (14,278 genes) were annotated with at least one GO term. When the core gene set's GO terms were associated with GOA Slim categories, 238,821 associations were made. This was reduced to 108,674 once the 6 broad GOA Slim categories were removed. Eight GO terms were not associated with a GOA category in Components. Five GO terms were similarly excluded from Functions, with eighteen excluded from Processes. GOA Slim terms in the 'Components' subgroup accounted for 14.86% of the associations, 'Functions' accounted for 34.19%, and 'Processes' 50.95%.

Unfortunately, the GO associations revealed little about the core gene set, simply suggesting that the species-specific gene sets do not encode any components, functions or processes that the core gene set itself does not (Figure 2.3.8). One key observation of interest was made however. Firstly, 83% of associations between Component GO terms and GOA Slim categories in the core gene set represent intracellular components. This is hardly surprising given that the vast majority of cells' components are intracellular; however, the species-specific sets all contain higher percentages of membrane components and extracellular components than the core gene set. This reinforces the suggestion that a large proportion of each species' membrane-bound proteins are species-specific and of paramount importance in distinguishing between *Entamoeba* species.

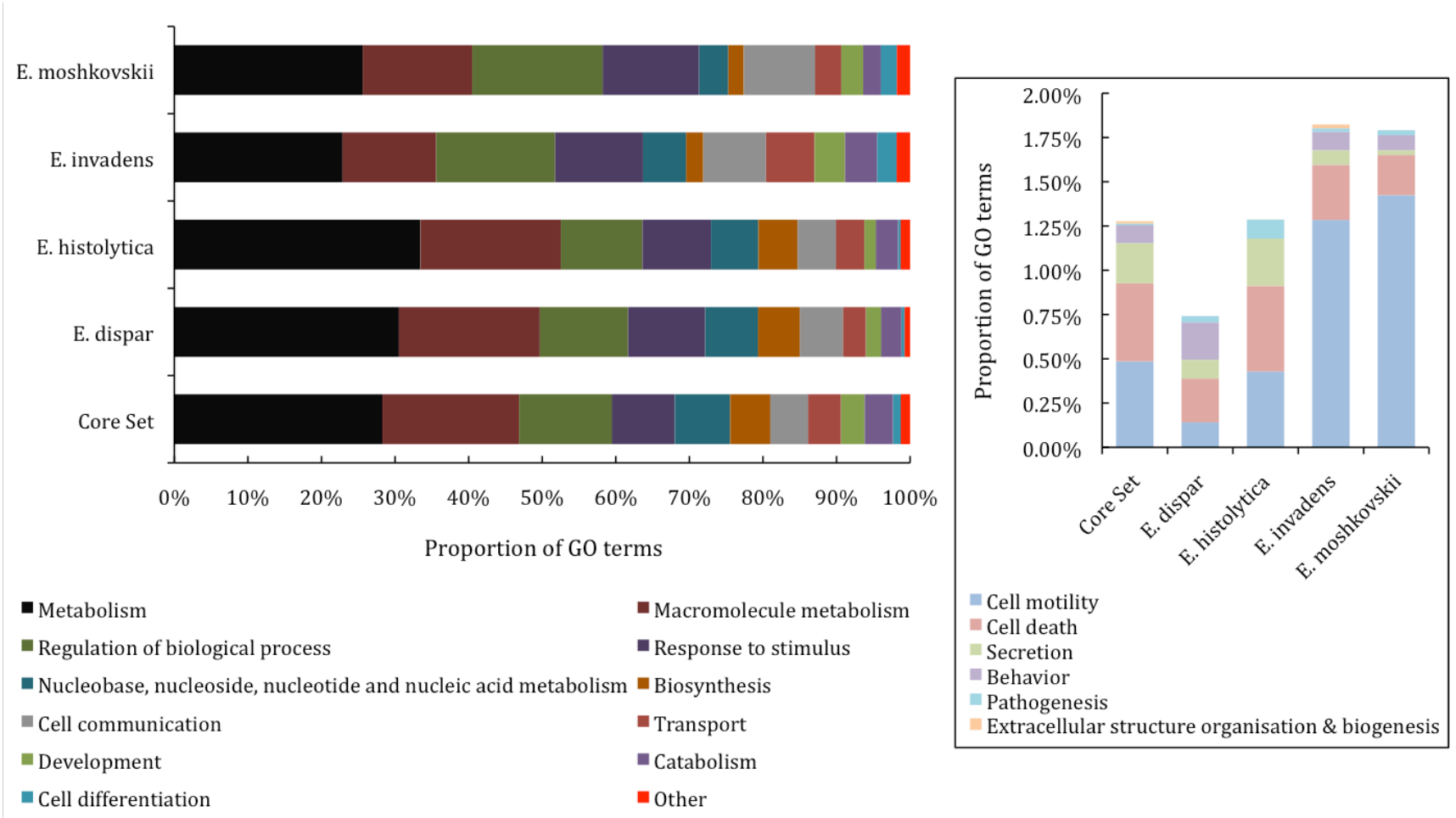


a) Proportions of 'Components' GO terms assigned to GOA Slim categories

Figure 2.3.8. Proportions of GO terms assigned to GOA Slim categories for the *Entamoeba* core gene set and genes unique to *Entamoeba histolytica*, *Entamoeba dispar*, *Entamoeba invadens* and *Entamoeba moshkovskii*. GOA Slim categories are divided into a) Components; b) Functions; c) Processes. Inset boxes contain an expanded view of the contents of the 'Other' category from the main chart.



b) Proportions of 'Functions' GO terms assigned to GOA Slim categories



c) Proportions of 'Processes' GO terms assigned to GOA Slim categories

2.3.5 Generation of *Entamoeba*-exclusive gene set

Analysis of the GO terms and their associations with GOA Slim categories revealed disappointingly little about the gene families found in all *Entamoeba* species. The core gene families identified thus far were also somewhat limited in use as they did not reveal which gene families were found exclusively in *Entamoeba* species, thus were important in the parasitic lifestyles of members of the genus. As such, it was necessary to compare the core *Entamoeba* gene set with genomes from other genera to identify genes present in all *Entamoeba* that are also unique to the genus. This would also reduce the number of families to one that it was practical to manually count. Comparisons were made with other species representative of the Amoebozoa, and again with an additional species that allowed the group to represent the more evolutionarily distant Unikonts superfamily.

i. Orthologous clusters

The comparisons involved representatives of the *Dictyostelium* slime mold genus, the free-living *Acanthamoeba* and the *Saccharomyces* yeast genus. Respectively, these representatives were *D. discoideum*, *A. castellanii* and *S. cerevisiae*. Rather than running OrthoMCL independently for the Amoebozoa and Unikont gene sets, clusters were generated for the Unikonts, from which Amoeboza clusters were extracted. This ensured that the numbers of orthologous gene families were directly comparable between the two cluster sets. When clustering orthologous sequences from the four species, using OrthoMCL v2.0.3, a 35% cutoff value was applied. This was based upon the average percent similarity of 'hits' in a reciprocal BLASTP search using default parameters between the protein sets of the two most evolutionarily distant species, *E. histolytica* and *S. cerevisiae* (35.25%). As in Section 2.3.4, an optimal granularity value of 3.0 was applied to MCL. One sequence from *A. castellanii* 5 amino acids in length was filtered out prior to clustering.

Firstly, let us consider the three species representing the Amoebozoa – *E. histolytica*, *D. discoideum* and *A. castellanii* (Figure 2.3.9.a). In the same way that shared genes between the *Entamoeba* species were regarded as a putative core *Entamoeba* gene set, the 885 families shared between the three genera can be considered a core Amoebozoa gene set. It contained 1,472 *Entamoeba* genes (17.72% of *E. histolytica*'s total gene complement), 1,079 *Dictyostelium* genes (8.76% of *D. discoideum*'s total) and

1,058 *Acanthamoeba* genes (6.74% of *A. castellanii*'s total). As expected, the number of genes exclusive to each genus was positively correlated with the number of genes in each representative species' genome, with *E. histolytica* possessing 921 unique gene families. *E. histolytica* shared more gene families with the more closely related *D. discoideum* than with *A. castellanii*, reflecting the evolutionary relatedness of the three species [216].

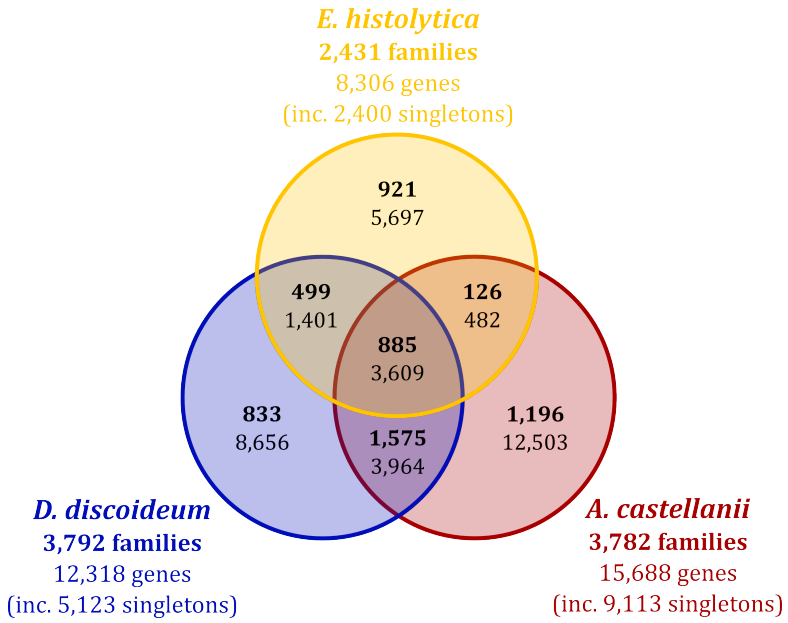
When expanding the comparison to include the Unikonts, it was expected that the number of core gene families would be reduced and that genus-specific family counts would, conversely, increase. Whist this occurred, it was to a relatively small degree as only 42.28% of *Saccharomyces* genes were shared with the Amoebozoa (Figure 2.3.9.b; Appendix C, File C.3). The Unikonts core gene set included 655 shared families, 230 fewer common families than were seen between the Amoebozoa. The core set comprised 1,082 *Entamoeba* genes, 780 *Dictyostelium* genes, 751 from *Acanthamoeba* and 798 from *Saccharomyces*. The effect on the number of gene families unique to *E. histolytica* was minimal, with the count declining by just 10, resulting in a total of 67.96% of the *Entamoeba* representative's gene set making up the genus' exclusive gene set. This was comparable to the proportion of *D. dictyostelium*'s gene set that was genus-specific (67.00%) but considerably lower than that of the *Acanthamoeba* representative (78.33%) and higher than that of *S. cerevisiae* (57.22%).

The comparison involving genera representative of the Unikonts also generated gene clusters thought to be shared exclusively by members of the Amoebozoa. The number of clusters consisting of genes present in species from the *Entamoeba*, *Acanthamoeba* and *Dictyostelium* genera, but not the *Saccharomyces* genus, totalled 230 (Figure 2.3.9.b). This count is lower than a 297-strong Amoebozoa-specific gene set previously generated, however the methodology and set of genes used in the earlier study were considerably different to those used in this project [217].

Firstly, the earlier work predicted the presence of approximately 21,000 gene models using an incomplete assembly of the *A. castellanii* genome. Conversely, I used a considerably reduced set of genes, predicted using RNA-Seq data and based upon a whole genome assembly [136]. Secondly, rather than the OrthoMCL-based approach employed in this project, which made use of the *E. histolytica* gene set as well as those of *D. discoideum* and *A. castellanii*, the 297-gene set was generated solely using the results of BLAST searches, which did not involve the *Entamoeba* representative. The

predicted *A. castellanii* sequences were compared with the gene sets of *D. discoideum* and a number of non-Amoebozoan species, with stringent parameters applied to ensure that only those predicted genes with strong hits to *D. discoideum* and relatively poor hits to non-Amoebozoan species were included in the final data set [217]. Given that the *A. castellanii* gene set used to predict the previously generated Amoebozoa-specific gene set is now known to be inaccurate, and the methodology only made use of two species in the clade, the difference between the numbers of Amoebozoa-exclusive genes output by the two methods was not considered an issue.

a)



b)

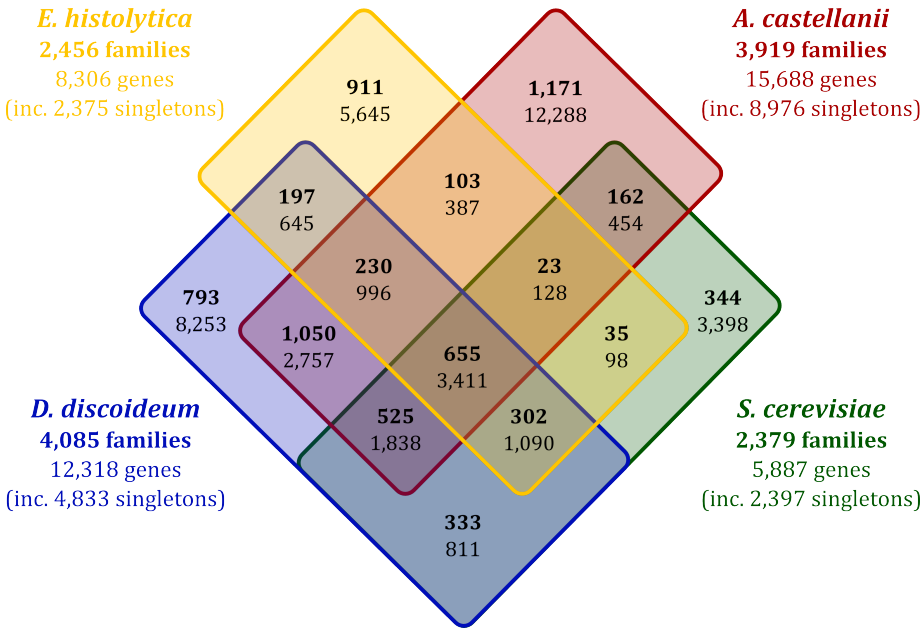
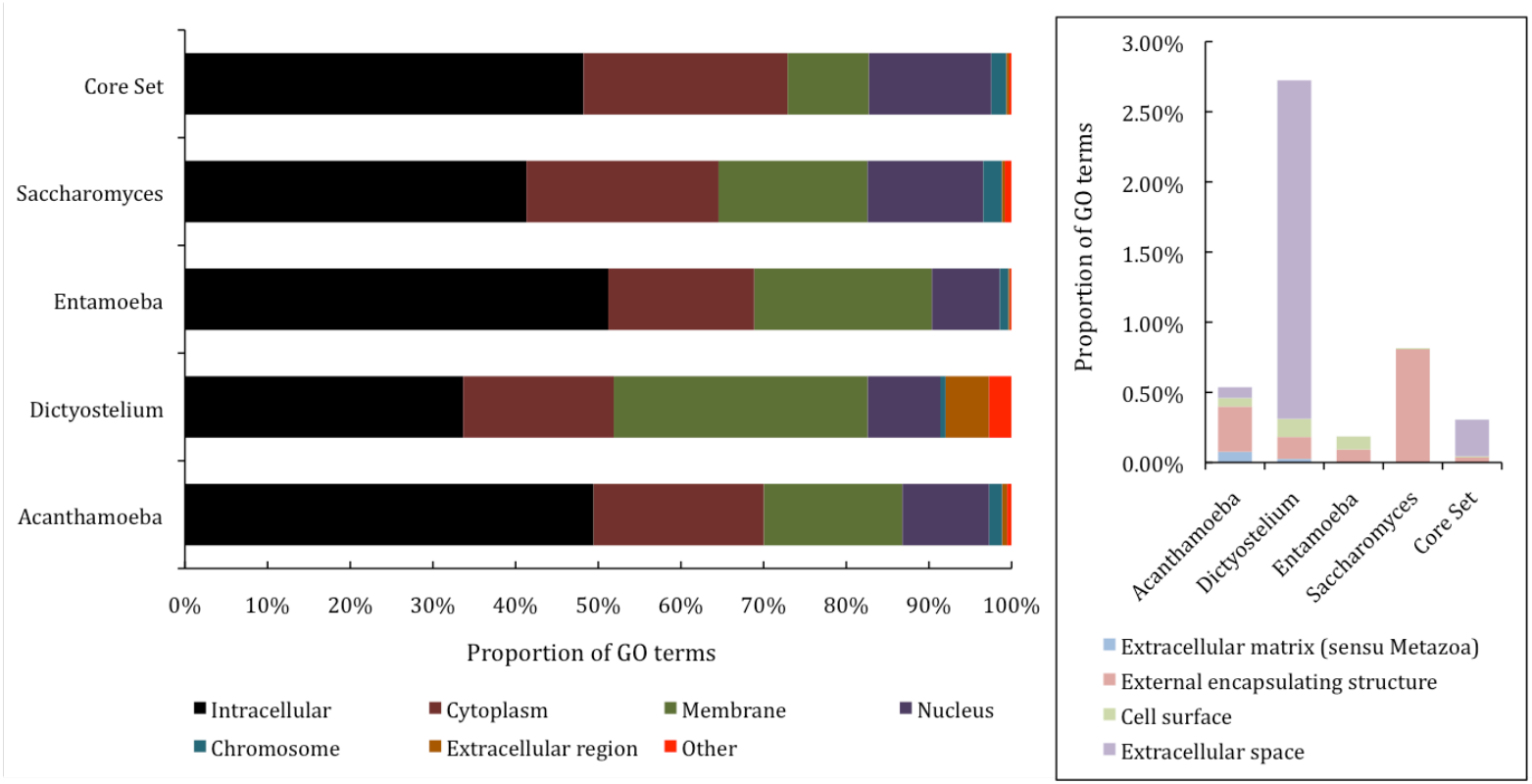


Figure 2.3.9. Venn diagrams showing numbers of unique and orthologous genes and families in a) the genera *Entamoeba*, *Acanthamoeba* and *Dictyostelium*, representing the Amoebozoa; and b) the genera *Entamoeba*, *Acanthamoeba*, *Dictyostelium* and *Saccharomyces*, representing the Unikonts. Numbers are based upon OrthoMCL output. Numbers in bold represent gene families; numbers in regular font represent individual gene numbers.

ii. Gene ontologies within the *Entamoeba*-exclusive gene set

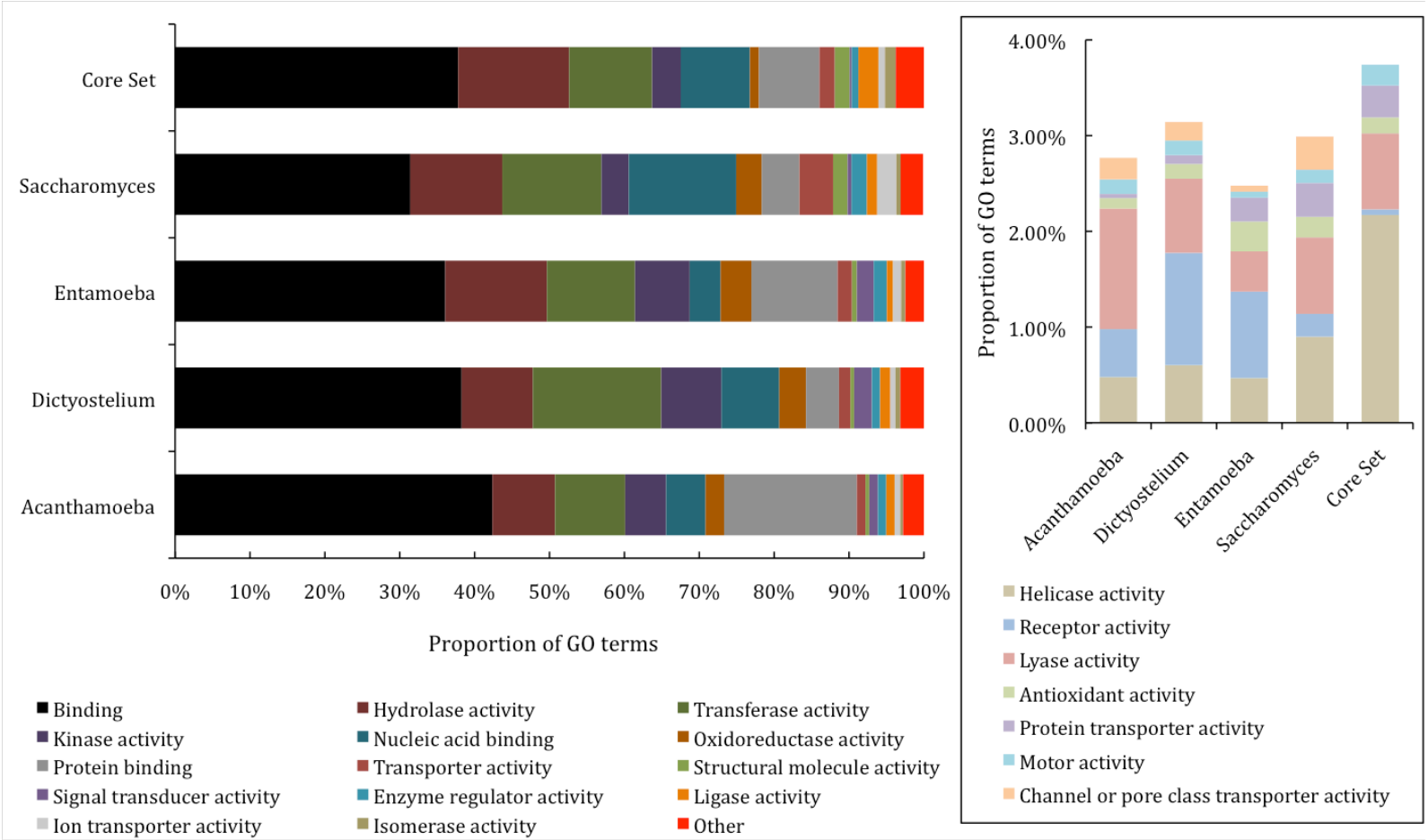
Given the similarity between the Amoebozoa and Unikont comparisons, it seemed unlikely that studying the GO associations in both sets would yield more information. As such, only the Unikont-based gene sets were analysed. It was hoped that an overview of the GO associations within each genus-specific gene set and the core Unikonts gene set might offer some broad impressions of the differences between the genera. As in Section 2.3.4, BLAST2GO was used to attribute GO terms to the *E. histolytica* and *A. castellanii* genus-specific gene sets. *D. discoideum* GO terms were downloaded from the DictyBase web resource. *S. cerevisiae* GO terms were downloaded from the Yeast Genome web resource. Both downloaded sets required modification as several GO terms were wrongly labelled as 'Component', 'Function' or 'Process'. The species-specific gene sets' GO terms, as well as the core Unikonts gene set's GO terms, were uploaded to CateGOrizer and grouped into GOA Slim categories (Figure 2.3.10). Within both the 'Components' and 'Processes' subcategories, one GO term was excluded because it could not be associated with a GOA Slim category.

Relatively few GO associations were made for the *Entamoeba*-exclusive genes. Only 46.18% of the genus-specific genes were annotated with GO terms. As such, the information that can be gained from the associations of the GO terms with GOA Slim categories is limited. A total of 39,148 associations were made for the genes unique to *Entamoeba*. Once reduced, as above, to 17,803 associations, 'Components' made up 12.11%, 'Functions' 36.06%, and 'Processes' 51.83%. Overall, 22 GO terms were not associated with any GOA Slim categories – 5 'Components', 2 'Functions' and 15 'Processes'. Pairwise Pearson's Chi-squared tests were used (with the Monte Carlo simulation where counts were less than 5), with alpha values of 0.05, to compare the proportions of GO terms allocated to each GOA Slim category in *E. histolytica* with each of the other 3 genus-specific gene sets (Appendix A, Table A.6). When compared with *Dictyostelium*, the most closely related genus, significant differences were seen in 28 of the 49 GOA Slim categories. This increased slightly to 29 differences with the more distant *Acanthamoeba*, with comparisons with *Saccharomyces* producing the largest number of significant differences at 39, as might be expected. Unfortunately, these differences clarify little about the general components, functions and processes that distinguish *Entamoeba* from other genera in the Unikonts.

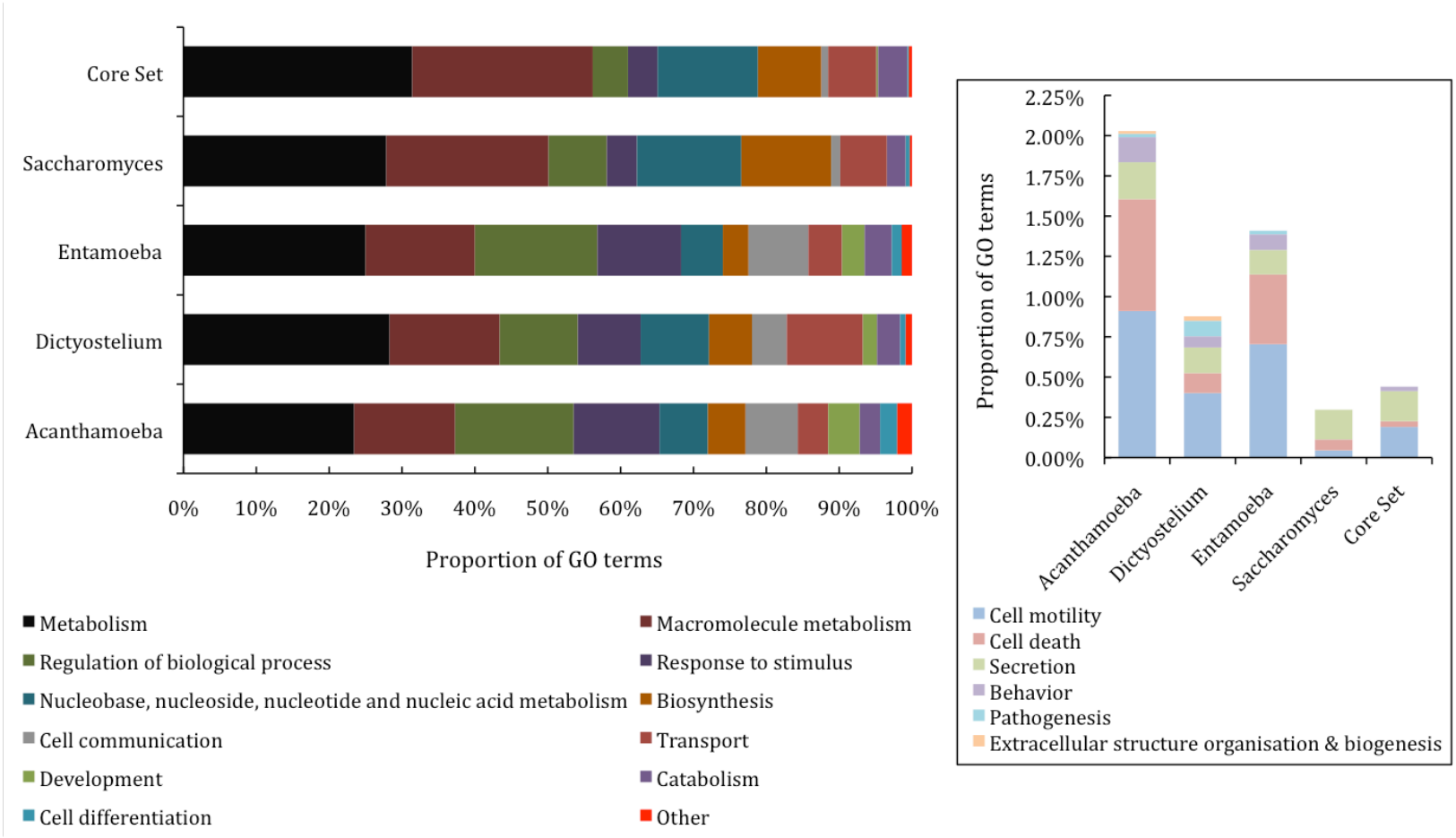


a) Proportions of 'Components' GO terms assigned to GOA Slim categories

Figure 2.3.10. Proportions of GO terms assigned to GOA Slim categories for the Unikonts core gene set and genes unique to *Entamoeba*, *Acanthamoeba*, *Dictyostelium* and *Saccharomyces*. GOA Slim categories are divided into a) Components; b) Functions; c) Processes. Inset boxes contain an expanded view of the contents of the 'Other' category from the main chart.



b) Proportions of 'Functions' GO terms assigned to GOA Slim categories



c) Proportions of 'Processes' GO terms assigned to GOA Slim categories

2.3.6 Functions of *Entamoeba*-exclusive gene set

By identifying the number of core *Entamoeba* families identified in Section 2.3.4 with at least one member present in the gene families exclusive to the *Entamoeba* genus, it was possible to generate a list of gene families theoretically present in all *Entamoeba* species, but not in any other Unikont genus. This reduced the number of gene families to one for which it was practical to manually count occurrences of functions. This allowed the identification of gene families that distinguish *Entamoeba* from other genera, potentially highlighting families directly responsible for facilitating amoebic infections.

It is important to note that this should not be considered a definitive list of *Entamoeba*-exclusive gene families. It is unlikely that all of the families included here are present in all *Entamoeba* species and no other genera. As such, this gene set should be edited and reduced over time if further studies seek to improve it. Currently, there are 2,981 such *Entamoeba*-exclusive gene families, comprising 13,672 genes. Of these families, 1,798 comprise gene models with unconfirmed functions and, as such, have no functional annotations. The remaining clusters were grouped by function to provide an overview of the families that define the *Entamoeba* genus.

i. Most prevalent families and functions

The majority of the identified families perform diverse or poorly defined functions so could not easily be grouped together. However, a number of families shared functions that, together, fulfil particular cellular processes, such as DNA repair and ubiquitination of proteins. These groups of families were counted alongside frequently occurring individual functions (Figure 2.3.11). The most prevalent function is derived from the highly diverse serine/threonine/tyrosine protein kinase superfamily. The Ras GTPase superfamily, and its associated guanine nucleotide exchange factors and GTPase-activating proteins, also form a significant proportion of the core *Entamoeba*-exclusive gene set, making up 6.37% combined. The Ras superfamily is an evolutionarily conserved group of proteins that perform a wide range of cellular functions [218]. Additionally, a large number of the families and functions most often seen in the gene set, including helicases and chaperones, are linked to DNA and RNA assembly, maintenance and repair.

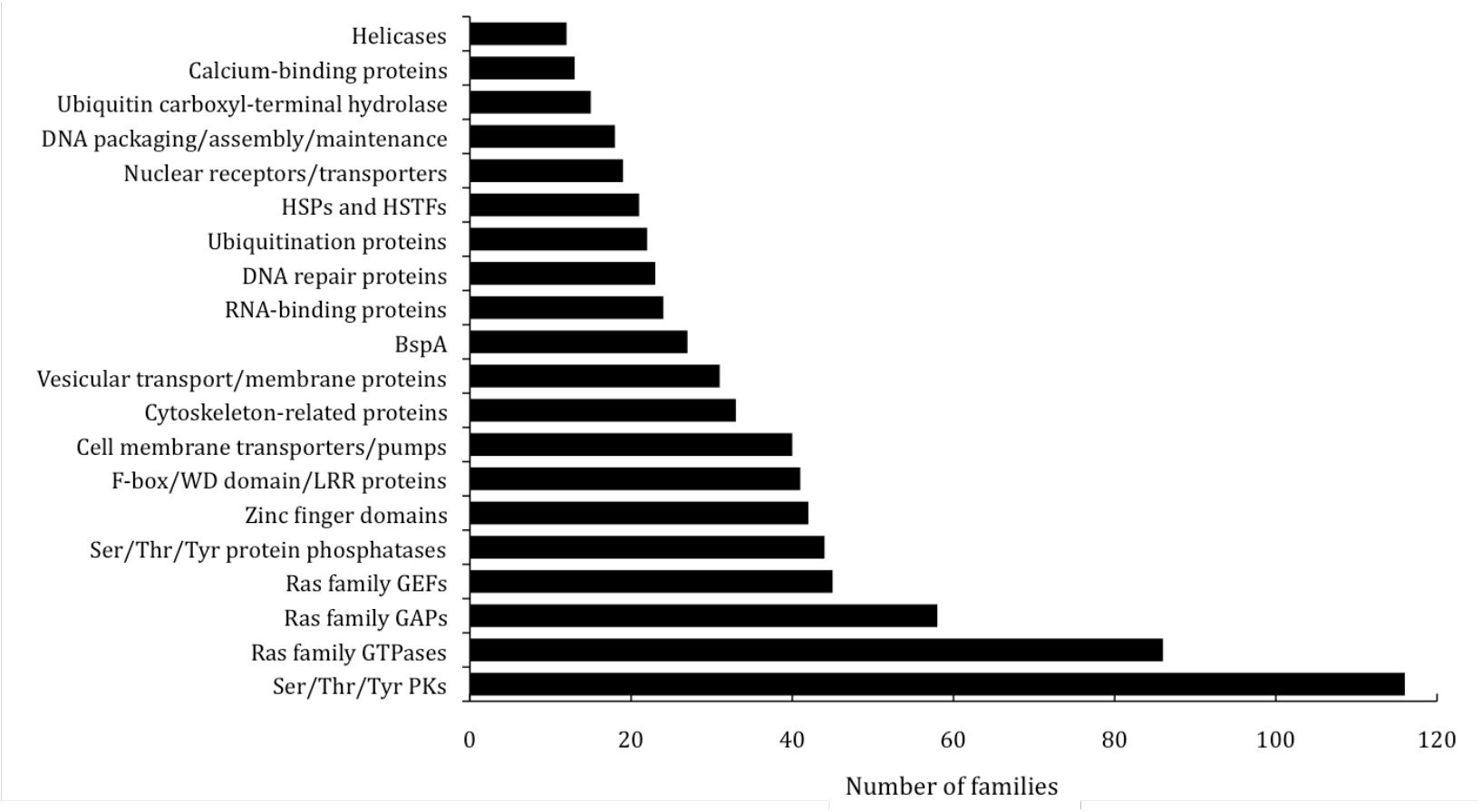


Figure 2.3.11. The most prevalent gene families/functions in the core *Entamoeba*-exclusive gene set. PK = Protein kinase; GAP = GTPase-activating protein; GEF = Guanine nucleotide exchange factor; LRR = Leucine-rich repeat; HSP = Heat shock protein; HSTF = Heat shock transcription factor.

Also prevalent in large numbers were several families whose functions, when performed by so many genes, are indicative of a genus that has evolved to improve survival in a particular niche. This includes cytoskeleton-related families and vesicular and vacuolar protein families. Whilst it is likely that the majority of these proteins play 'housekeeping' roles, it is possible that *Entamoeba* species require unique families of these genes for engulfing and lysing bacteria [12, 219]. If these families played such a role it would certainly be necessary for allowing the *Entamoeba* to survive in a host environment. Furthermore, a large number of families encoding cell membrane transporters and pumps are also exclusively present in *Entamoeba* species. It is probable that these transporters are required for, amongst other functions, acquiring nutrients from the trophozoites' colonic environment. Proteins with a more direct link to the genus' parasitic nature are members of the BspA family. This family is associated with adhesion to extracellular membranes [200-202]. It is interesting to suggest that these family members may play a major role throughout the *Entamoeba* genus in enabling a parasitic 'lifestyle', as was discussed earlier.

ii. Virulence factors

Whilst no known virulence factor families, other than the relatively unstudied BspA family, were present in large numbers in the *Entamoeba*-exclusive gene set, 15 putative virulence factor groups were identified in smaller numbers (Table 2.3.4). One might consider that some, if not all of, these families are essential for establishing parasitic amoebic infections in a range of hosts. The cysteine proteases and Gal/GalNAc lectins have long been considered key virulence factors in the genus [74, 81, 220-222]. Interestingly, 8 thioredoxin families, responsible for reducing oxidative stress, also exist in the gene set [209, 223]. Thioredoxins are ubiquitous [224], so it is probable that there are simply a greater number of them in *Entamoeba* species than in the other Unikonts studied here. However, it is not unreasonable to propose that some of them play roles in facilitating a parasitic lifestyle. The roles of the known virulence factors seen exclusively in *Entamoeba*, as well as several others, are discussed in greater detail in Section 2.3.7.

Two families have not been previously discussed in the literature as putative virulence factors in *Entamoeba*, but have distant orthologues with potentially pathogenic properties. The bullous pemphigoid antigen triggers an autoimmune response, leading to blistering of skin and mucous membranes [225, 226]. M proteins,

meanwhile, are surface-bound proteins, some of which act as virulence factors in *Streptococcus pyogenes* [227]. Those M proteins in which a particular signal sequence is found are capable of stimulating production of opsonic antibodies or increasing the bacterium's resistance to phagocytosis.

It is highly improbable that the proteins annotated as bullous pemphigoid antigens actually fulfil a similar pathogenic role, however it is interesting to suggest that they may contain domains and motifs similar to those seen in their namesake that confer similar properties. This is, however, conjecture as no information regarding domains exists to confirm or deny this family's role in virulence. Functional domains are, however, recognised in the families annotated as M proteins. The functions carried out by some M proteins in Streptococci could conceivably be of benefit in *Entamoeba* infections. However, not all M proteins are capable of such virulence-inducing properties [228] and further work would be encouraged to discover whether the M proteins seen in *Entamoeba* have the domains required to make them pathogenically important.

Table 2.3.4. Frequencies with which gene families, whose members or orthologues are known to act as virulence factors, exist in the core *Entamoeba*-exclusive gene set.

Family name/function	Count	Reference
Alcohol Dehydrogenase	1	[229]
Amoebapore C	1	[230]
Bullous pemphigoid antigen	1	[225]
Cysteine proteases	6	[74, 220]
Gal/GalNAc lectin subunits	5	[222]
Immuno-dominant variable surface antigen	1	[231]
Iron hydrogenase	2	[93, 94]
Iron-sulphur flavoprotein	2	[232]
M protein	2	[227]
Peroxiredoxin	1	[209]
Rhomboid	1	[98, 233]
Rubrerythrin	1	[234]
Serine-Threonine-Isoleucine-Rich Protein	2	[86]
Thioredoxin	8	[209, 223]
Type A flavoprotein	2	[94, 235]

2.3.7 Comparison of all virulence factor families in the *Entamoeba* genus

Having identified putative virulence factor families belonging exclusively to the *Entamoeba* genus, I next chose to study all virulence factors present in *E. histolytica*, *E. dispar*, *E. invadens* and *E. moshkovskii*. Once the ubiquitous families were identified (Table 2.3.5), phylogenetic analyses were carried out to identify if, and when in the evolutionary past, expansions or redactions of orthologue families occurred. Ultimately, it was hoped that this would highlight particular genes that have been lost or gained, or whose sequences have diverged or converged over time. Greater understanding of the genus' virulence factors may help identify the families that are crucial in the development of amoebiasis.

It has been proposed that *E. histolytica* demonstrates long branch attraction (LBA) in phylogenetic relationships with members of other genera [217]. LBA is

characterised by a significantly different evolutionary rate to the expected rate. It is a result of reduced genome size in obligate parasites. It is assumed in this work that, if any *Entamoeba* species demonstrate LBA, they are all likely to and so phylogenetic comparisons within the genus are a viable form of comparison. Comparisons of branch lengths, representative of evolutionary distances between genes, were manually calculated. Mann-Whitney-Wilcoxon tests (with continuity correction) were performed for each data set using alpha values of 0.05.

A total of 29 virulence factor gene families were studied (Table 2.3.5). Members of 12 of those families interact directly with host proteins and cells. Fourteen other families' members are involved in metabolic or enzymatic processes, rather than outright defensive or offensive actions. In these families, having an impact upon virulence could almost be considered a convenient coincidence. Furthermore, the host range of a particular species would not be expected to influence gene diversity. The final three families have unconfirmed roles in pathogenicity.

Of the 'directly virulent families', three contained similar numbers of genes from each species in a phylogenetic arrangement that approximately reflected the evolutionary relationships between them [189]. Eight of the 'metabolic families' and two of the unconfirmed virulence factors demonstrated this typical phylogeny and possessed similar numbers of coding sequences from all 4 species.

Within many of the gene families it was observed that *E. invadens* possessed greater numbers of genes than the other species. It was suspected that this implied greater variation that may be necessary for parasitising, and causing disease in, its wider range of hosts. This theory was investigated further as each gene family was studied in turn.

Table 2.3.5. Virulence factors that directly interact with host proteins and cells present in *Entamoeba histolytica*, *Entamoeba dispar*, *Entamoeba invadens* and *Entamoeba moshkovskii*. Only coding sequences included in AmoebaDB v2.0 are included. Shaded rows indicate families that demonstrate atypical phylogeny. NO = Nitric oxide.

Virulence Factor Family	Protein Function	Species			
		Entamoeba histolytica	*Entamoeba dispar*	*Entamoeba invadens*	*Entamoeba moshkovskii*
Directly interact with host proteins and cells					
C2 protein kinase [236]	Involved in initiation of phagocytosis	2	2	2	2
Cysteine protease Family A [74, 220]	Cleave host proteins, inc. MUC2 mucin barrier	14	10	18	14
Cysteine protease Family B [74, 220]		15	9	22	17
Cysteine protease Family C [74, 220]		13	14	11	15
Heavy Gal/GalNAc lectin [222]	Allow adhesion to host cells and subsequent invasion	5	2	9	4
Intermediate Gal/GalNAc lectin [222]		3	3	13	4
Light Gal/GalNAc lectin [222]		7	6	13	5
KERP [237]	Involved in adherence to host cells	3	1	1	2
Poreformers [50]	Disrupt host cell membranes	12	10	9	11
Serpin [238]	Inhibit host proteins	1	2	8	22
Sphingomyelinase C [239]	Disrupt phagosome membranes	10	3	4	6
STIRP [86]	Involved in cytotoxicity and adhesion to host cells	4	4	6	2

Virulence Factor Family	Protein Function	Species			
		Entamoeba histolytica	***Entamoeba dispar***	***Entamoeba invadens***	***Entamoeba moshkovskii***
Involved in metabolic processes					
NADPH:flavin oxidoreductase [92, 240]	Involved in response to oxidative stress	19	12	12	11
Peroxiredoxin [209]		10	26	13	3
Rubredoxin [234]		1	1	1	1
Rubrerythrin [234]		1	1	2	1
SOD [90]		1	1	4	4
Fe hydrogenase [93, 94]		4	4	5	4
Thioredoxin [209, 223]		4	4	5	8
Thioredoxin reductase [224, 241]		2	1	1	1
CPBP [242]	Traffic, process and activate cysteine proteases	12	13	13	22
Lysozyme [243]	Form acidic lysozyme to degrade host proteins	7	8	14	10
P21-activated kinase [244]	Involved in capping process	7	7	8	8
Rhomboid [98, 233]		5	5	3	4
Arginase [245]	Prevents macrophages synthesising NO	1	1	1	1
Type A flavoprotein [94, 235]	Detoxifies NO	5	5	4	4

Virulence Factor Family	Protein Function	Species			
		Entamoeba histolytica	***Entamoeba dispar***	***Entamoeba invadens***	***Entamoeba moshkovskii***
Undetermined function					
Adhesin [246]	N/A	2	2	2	2
Grainins [247]	N/A	7	4	5	9
Phospholipases [248]	N/A	21	22	20	22

i. Indirect virulence factor families – involved in ROS protection

Four of the six remaining metabolic families are involved in protecting trophozoites against oxidative stress, a challenge most notably, but not exclusively, faced by *Entamoeba* in the host bloodstream. The gene numbers and arrangements within these families vary but they all lead to similar conclusions. Members of the SOD family convert ROS to O_2 and H_2O_2, acting as the first in a chain of proteins that render ROS harmless. The family conforms to the typical phylogeny but contains expanded sets of *E. invadens* and *E. moshkovskii* sequences, resulting in four of each, compared with one sequence each from the other two species. Three of the *E. invadens* sequences were nearly identical (mean branch length: 1.33e-5; s = 5.77e-06), as were two *E. moshkovskii* sequences (branch length: 1.00e-5). Overall there was no significant difference in variability between the two species' expansions.

A similar situation was seen in the NADPH:flavin oxidoreductase family, the members of which convert the O_2 produced by SOD into H_2O_2. For the most part, the family conforms to the typical phylogenetic arrangement expected; however, an expansion containing five *E. histolytica* genes has contributed to there being considerably more genes from that species than from any of the others. Branches between the members of this expanded clade are very short (mean length: 0.010318; s = 0.006502), indicating relatively few differences between the sequences. RNA-Seq expression data downloaded from AmoebaDB v2.0 was used to assess the functionality of these genes. The FPKM values for the expanded clade ranged from 0.727 to 27.226, which is particularly low when compared with the maximum known value in this family of 466.989. Taken together, these observations suggest that *E. histolytica* does not gain any additional functionality or variability from its additional genes.

Peroxiredoxin forms the final part in the chain of proteins designed to neutralise ROS, converting the H_2O_2 produced by the other two proteins into water. The family consists of two major clades, one of which is made up of three large expansions. One expansion contains 22 near-identical *E. dispar* sequences (mean branch length: 0.027049; s = 0.028944); another contains 8 closely related *E. histolytica* sequences (mean branch length: 0.009314; s = 0.007994); and the third contains 10 *E. invadens* sequences (mean branch length: 0.612409; s = 0.454135). Variation between the *E. invadens* sequences is significantly greater than that seen between the *E. histolytica* sequences (p-value < 0.001), which is, in turn, greater than

that seen between the *E. dispar* sequences (p-value = 0.012). According to RNA-Seq data, only three of the *E. histolytica* peroxiredoxin genes are expressed at a relatively high level in trophozoites (FPKM values between 56.456 and 704.433). The five remaining genes are expressed very little (values range from 0.292 to 12.676), suggesting that these sequences lack functionality.

The fourth and final atypical family involved in countering oxidative stress is the thioredoxin gene family, which forms a paired system with the typical thioredoxin reductase family. It has a greater number of genes from *E. moshkovskii* than from any other species, due to the presence of an expanded clade including four sequences, which also includes an *E. moshkovskii* pseudogene. The *E. moshkovskii* sequences demonstrate relatively high similarity (mean branch length: 0.087867; s = 0.063244). The fact that one sequence in the closely related clade has been pseudogenised implies that these additional sequences offer no advantage to *E. moshkovskii.*

Considering the above observations as one, it seems unlikely that gene duplications and clade expansions have conferred any greater functionality or variability on the abilities of the different *Entamoeba* species to counteract ROS using this chain. Given the relative lack of variability between the members of the expanded clades, the differences in gene numbers in the families appear simply to be the result of multiple gene duplications. As such, it seems unlikely that survival of ROS is a key differential between the four *Entamoeba* species studied here. This is, perhaps, unsurprising, given that entrance into the host bloodstream commences extraintestinal infections, which are counterproductive to the parasite's life cycle, and are, in fact, not seen in *E. dispar* infections.

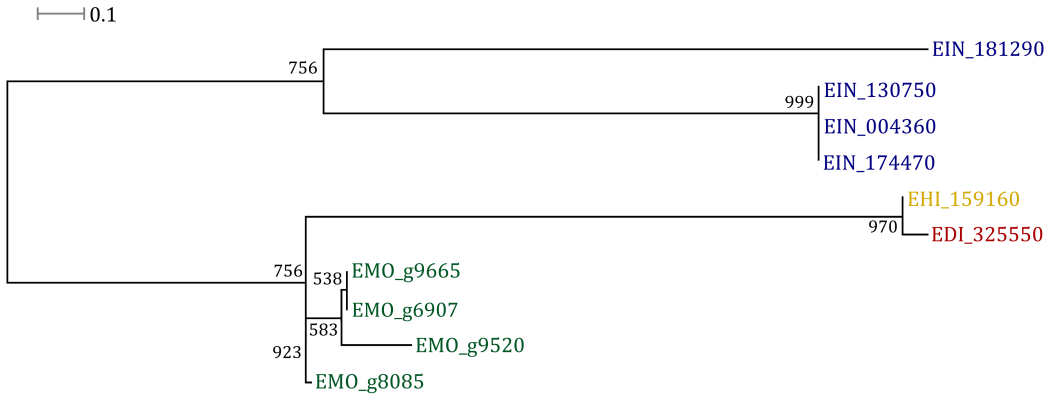
ii. Indirect virulence factor families – other families of interest

E. invadens and *E. moshkovskii* both have greater numbers of lysozyme genes than their counterparts due to multiple duplications in each, plus a small expansion in *E. invadens*. The two sets of genes are similarly variable, with both species demonstrating significantly greater variability in these expansions than they do in the families discussed above (*E. moshkovskii* sequences when compared with both SOD and thioredoxin sequences: p-value < 0.001; *E. invadens* sequences compared with peroxiredoxin sequences: p-value < 0.001), with the exception of the *E. invadens* expansion in the SOD family. The variation seen in this family is unexpected given that

lysozymes are internal structures so one would imagine that there would be little requirement for the two species exposed to the most variable environments to have additional members of the virulence factor family. It is possible that this is indicative of losses in the obligately parasitic *E. histolytica* and *E. dispar*, though further investigation would be required to confirm this.

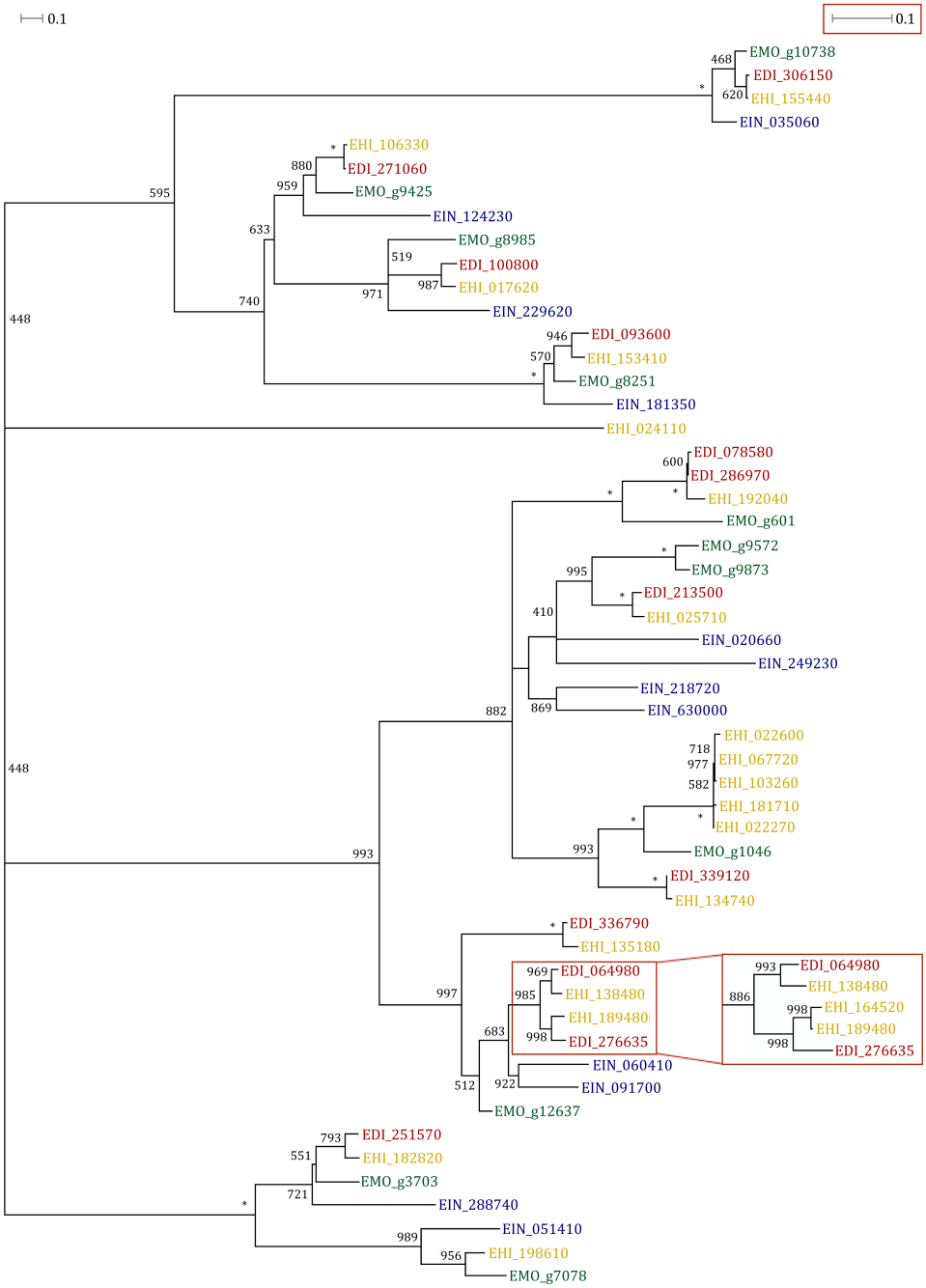
The cysteine protease binding protein (CPBP) family contains several typical clades. There is, however, a clade containing an eight-gene expansion of *E. moshkovskii* sequences. There are two *E. histolytica* genes in this clade. Both are expressed at relatively low levels (9.693 and 0.013), compared with the highest FPKM value in the family of 240.537, suggesting a lack of improved functionality resulting from the expansion, at least in *E. histolytica*.

Overall, these results suggest that there are few important differences between the metabolic, or indirect, virulence factor families of the four *Entamoeba* species. Increased gene numbers do not appear to be indicative of greater functionality and they are rarely a cause of evolved variability. Whilst the literature leaves little doubt as to their classification as virulence factors, it would appear that differences in proteins that do not interact directly with host cells cannot account for differences in pathogenicity.

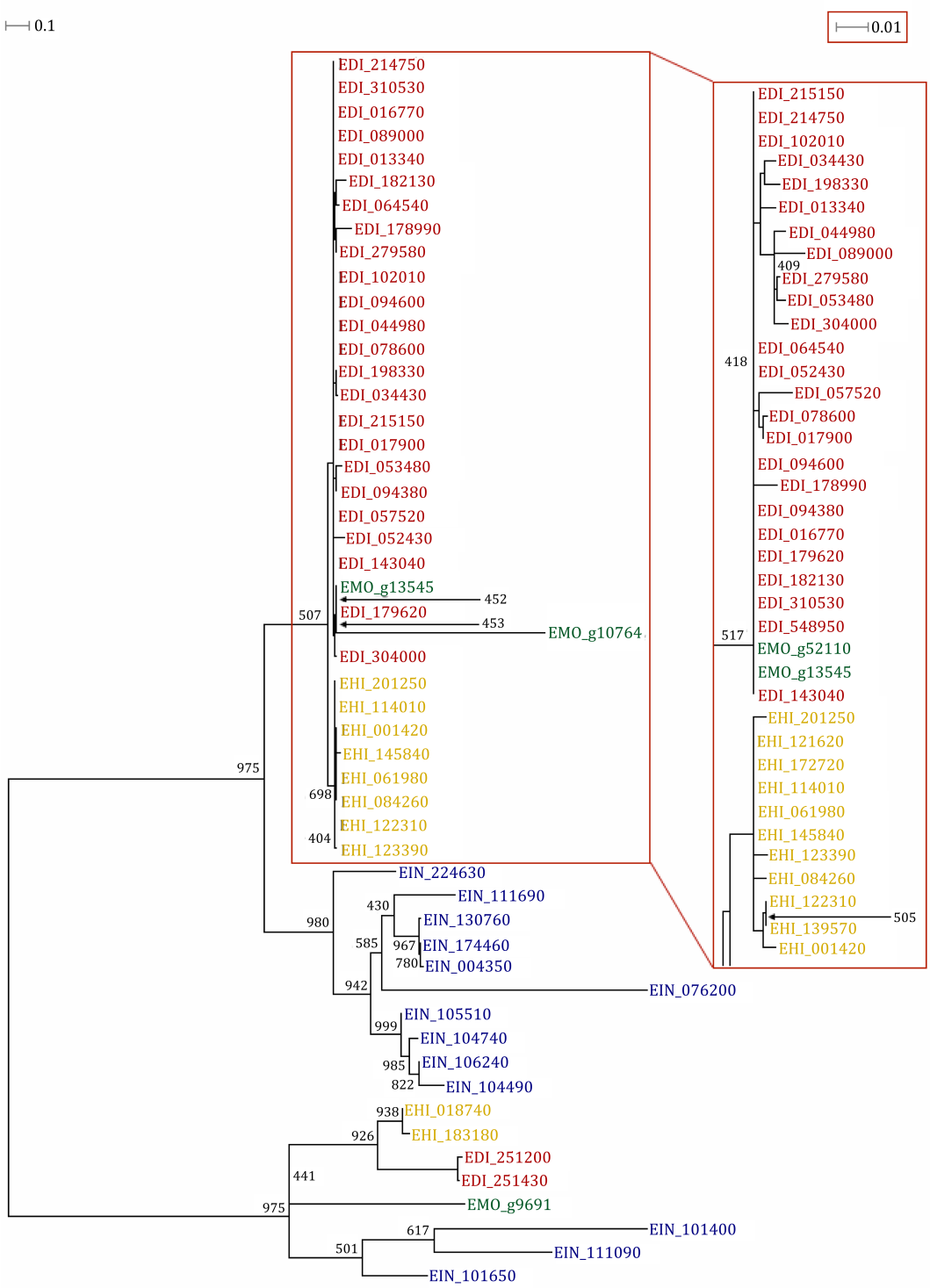


a) Superoxide dismutase

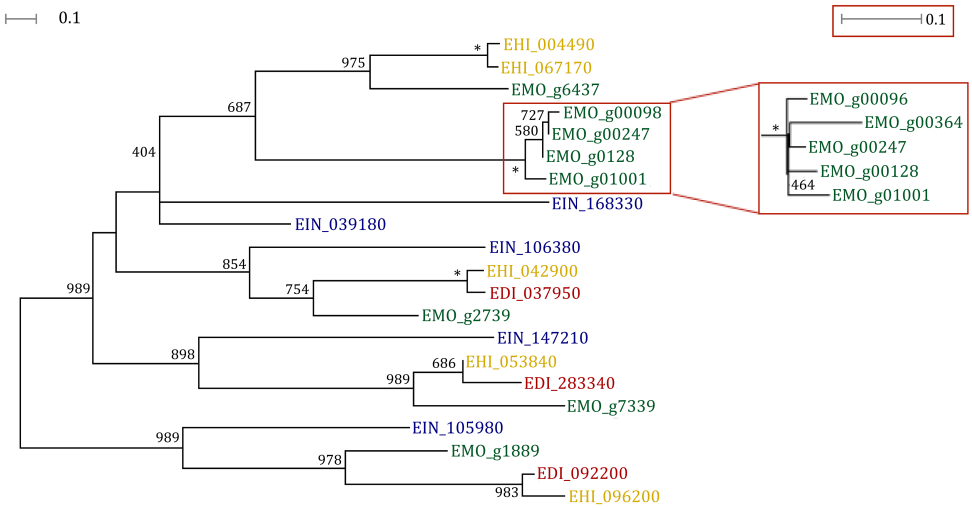
Figure 2.3.12. Phylograms of the *Entamoeba* virulence factor families indirectly involved in virulence that demonstrate atypical phylogeny. *E. histolytica* sequences are represented by yellow text, *E. dispar* sequences by red text, *E. moshkovskii* sequences by green text, and *E. invadens* sequences by blue text. Red boxes highlight clades in which pseudogenes were identified. They are linked to red boxes showing the same clades when phylogeny was calculated using nucleotide sequences, including the pseudogenes. Scale bars in red boxes represent nucleotide phylograms. All phylograms are midpoint rooted. Bootstrapping was performed for 1,000 replicates. Bootstrap values of 1,000 are represented by asterisks (*). Bootstrap values below 400 are not shown.



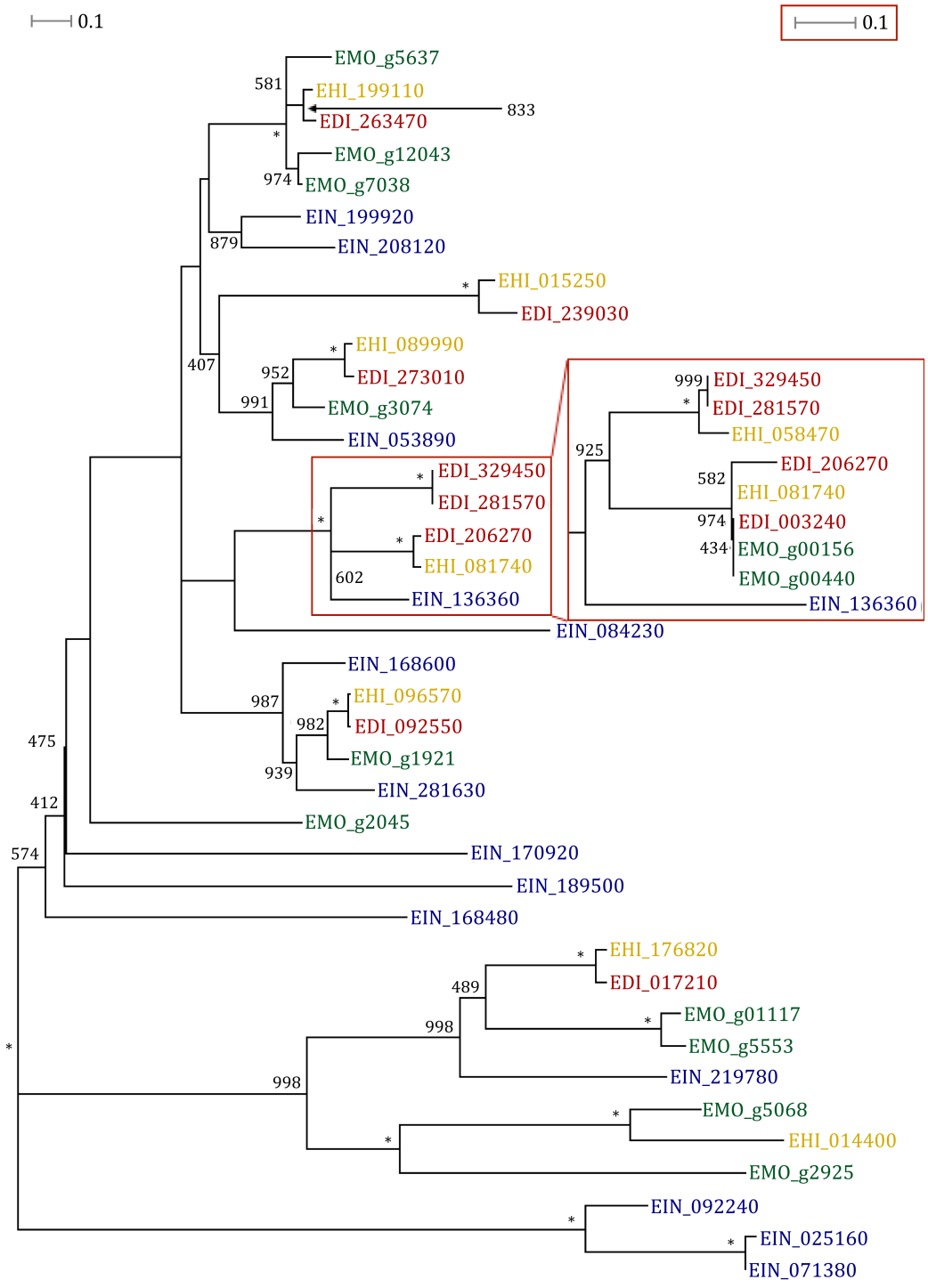
b) NADPH:flavin oxidoreductase



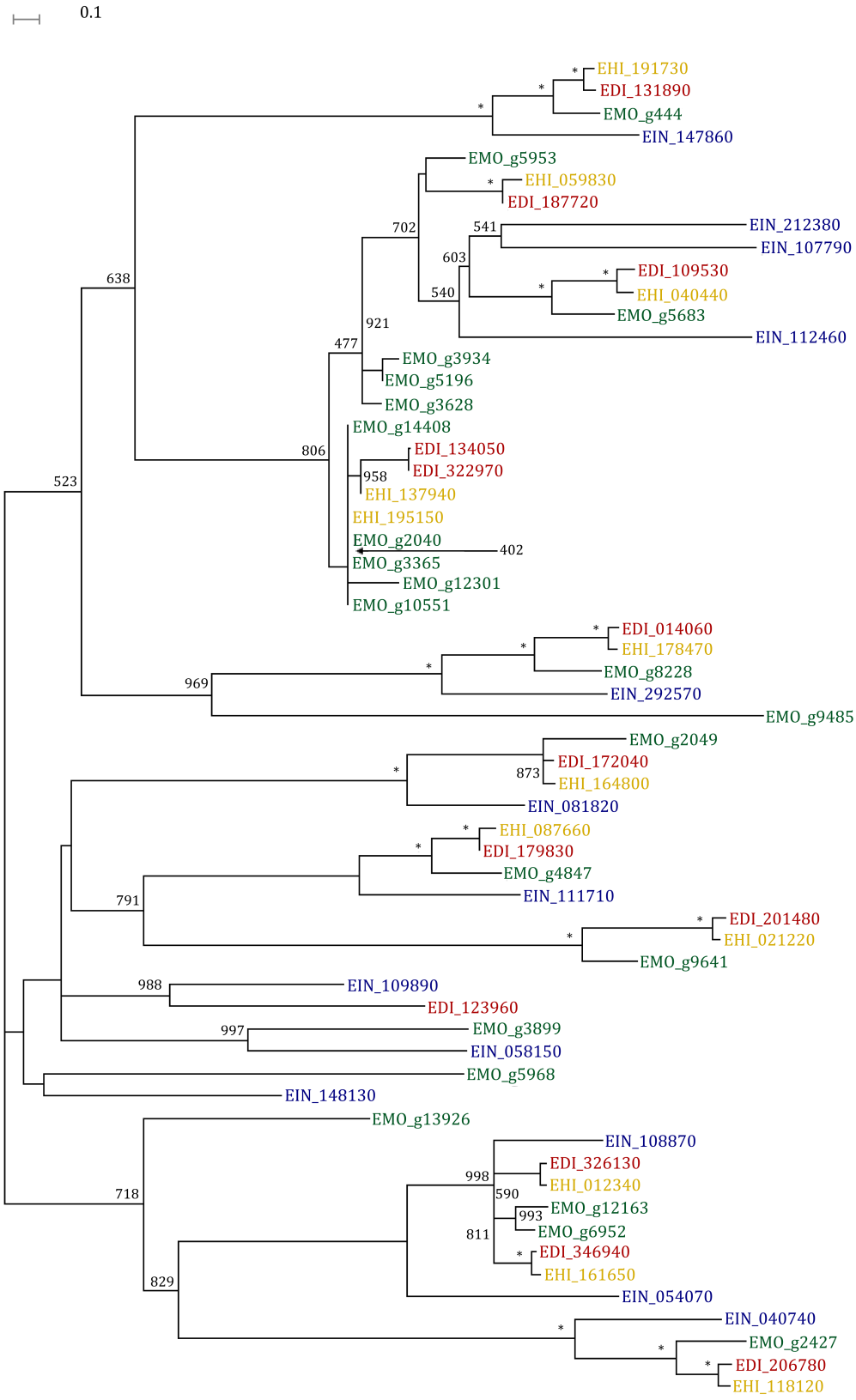
c) Peroxiredoxin



d) Thioredoxin


0.1
0.1
EMO_g5637
581
EHI_199110
833
EDI_263470
*
EMO_g12043
974
EMO_g7038
EIN_199920
879
EIN_208120
*
EHI_015250
EDI_239030
*
EHI_089990
407
952
EDI_273010
991
EMO_g3074
EIN_053890
999
EDI_329450
*
EDI_281570
EHI_058470
925
EDI_206270
582
EHI_081740
974
EDI_003240
434
EMO_g00156
EMO_g00440
EIN_136360
*
EDI_329450
EDI_281570
*
*
EDI_206270
EHI_081740
602
EIN_136360
EIN_084230
EIN_168600
987
*
EHI_096570
982
EDI_092550
475
EMO_g1921
939
EIN_281630
412
EMO_g2045
574
EIN_170920
EIN_189500
EIN_168480
*
EHI_176820
EDI_017210
489
*
EMO_g01117
998
EMO_g5553
EIN_219780
998
*
EMO_g5068
*
EHI_014400
EMO_g2925
*
EIN_092240
*
EIN_025160
EIN_071380

e) Lysozyme



f) Cysteine protease binding proteins

iii. Direct virulence factor families – Gal/GalNAc lectins

The heavy Gal/GalNAc lectin subunit allows *Entamoeba* species to adhere to host cells by binding to Galactose (Gal) and N-acetyl-D-galactosamine (GalNAc) on their membranes [222]. It is known to be heavily conserved in *E. histolytica*, with each protein containing a 104 amino acid CRD [27, 249]. Whilst comparing expansions and redactions of gene families between *Entamoeba* species it was interesting to investigate how well conserved the CRD was in other members of the genus. Greater understanding of this potential vaccine target could help inform treatments of amoebic infections.

The heavy Gal/GalNAc lectin subunit family contains expansions in *E. histolytica* and *E. moshkovskii*, as well as an expansion in *E. invadens* containing approximately twice as many sequences (Figure 2.3.14.a). The genes in the expanded *E. invadens* clade are significantly more diverse than the genes in the other two expanded species (based upon branch lengths compared with *E. histolytica*, p-value < 0.001; compared with E*. moshkovskii*, p-value = 0.045). Importantly, there are only two *E. dispar* genes in the family. A lack of proteins required for adherence to host cells may explain why invasive disease is seen so infrequently in this species [2, 33]. This could be a crucial characteristic of *E. dispar* that distinguishes it from its pathogenic relatives.

A multiple sequence alignment was performed for the heavy Gal/GalNAc lectin subunits in each species to compare and contrast CRDs (Figure 2.3.13). In *E. histolytica*'s heavy Gal/GalNAc lectin 1 sequence the CRD lies between residues 899 and 1002 [27]. Twenty-four residues, 10 of which were cysteine residues, were conserved between all five *E. histolytica* sequences when the alignment of the shorter EHI_046650 was manually edited. Without that outlying sequence, all but 6 residues were conserved. In this case, every difference was seen in the sequence Hgl2 (EHI_012279). This is very much in line with the conservation described in the literature [249]. The two *E. dispar* sequences' CRDs share 29 conserved residues. Twelve of these are cysteine residues, including the 10 seen in the conserved motif. These sequences clearly demonstrate greater variability than those in *E. histolytica* but, crucially, they both contain the cysteine-rich motif seen in their close relative. In the expanded set of *E. invadens* heavy Gal/GalNAc lectin paralogues, every sequence contained the same cysteine-rich motif within the CRD, though the sequences between the cysteine residues varied more than in *E. histolytica*. Even omitting a significantly

shorter sequence (EIN_018210) and the member of the distant clade (EIN_149390), only 25 amino acids, 10 of which were cysteine residues, were conserved within the CRD.

In *E. moshkovskii* only the sequence EMO_g1010 appears to be complete. EMO_g1855 contains an internal gap, whilst the other two sequences are suspected of lacking 3' ends. These are artefacts of the *E. moshkovskii* Laredo genome assembly. The conserved CRD motif of 10 cysteine residues is complete in the annotations of EMO_g6677 and EMO_g1010 with manual refinement of the assembly; however, the two remaining sequences are both incomplete beyond the fourth cysteine residue. The three *E. moshkovskii* sequences that are clustered together in Figure 2.3.14.a share 68.57% of their first 926 alignment positions (excluding the first 12 amino acids of EMO_g1855, which have no homologous residues, suggesting the sequence should actually begin at position 13), implying that the sequences are highly conserved across their entire lengths.

These data show that the highly conserved motif of 10 cysteine residues within the CRD is present in every species. This motif is most likely required to create a definitive structural arrangement within the lectin subunits. Conversely, the amino acid sequences between these residues vary between and within the four species, an observation supported by their phylogenetic relationships. Vaccines targeting the CRD of *E. histolytica* sequences have met with some success in the past, so it would be interesting to revisit this potential vaccine candidate in light of this new information [250-252].

The intermediate and light subunits of the Gal/GalNAc lectin offer considerably fewer differences than the heavy subunit (Figures 2.3.14.b and c). The intermediate chain group shows some similarities to the heavy chain group in that there is a large, relatively varied, *E. invadens* expansion raising the number of *E. invadens* genes above the number of genes seen in the other species (mean branch length: 2.962247; s = 1.610896). The light chain family, meanwhile, contains 2 *E. invadens* expansions and a slightly smaller expansion in both *E. dispar* and *E. histolytica*, but none in *E. moshkovskii.* Again, the *E. invadens* expansions are more variable than those of *E. histolytica* (p-value = 0.002) and *E. dispar* (p-value = 0.002). The *E. histolytica* expansion contains a poorly expressed gene FPKM value of 4.091, suggesting a relative lack of importance in terms of functionality. The relative lack of variability in the light

and intermediate lectin subunits, when compared with the heavy subunit family, suggests that the smaller subunits are less crucial to the success of amoebic infections than the heavy chain subfamily. This is, perhaps, unsurprising given their roles in the lectin [81] and the fact that we have seen that proteins indirectly involved with the host do not appear to demonstrate great variability between and species.

The Gal/GalNAc lectin subunits of *E. invadens*, in particular the heavy subunits, are of considerable interest. It was proposed above that the greater numbers of genes in virulence factor families seen in *E. invadens*, when compared with the other three species featured herein, are linked to the species' capacity for infecting a range of reptilian hosts. The significantly greater diversity seen in *E. invadens* heavy Gal/GalNAc lectin subunits suggests that these proteins in particular may be of importance. It is likely that *E. invadens* would require a range of proteins to adhere to, and parasitise, a range of hosts. As such, one can theorise that the variable heavy Gal/GalNAc lectin subunits, suggested above as key determinants in the development of invasive amoebiasis, are a key family in allowing *E. invadens* to do so.

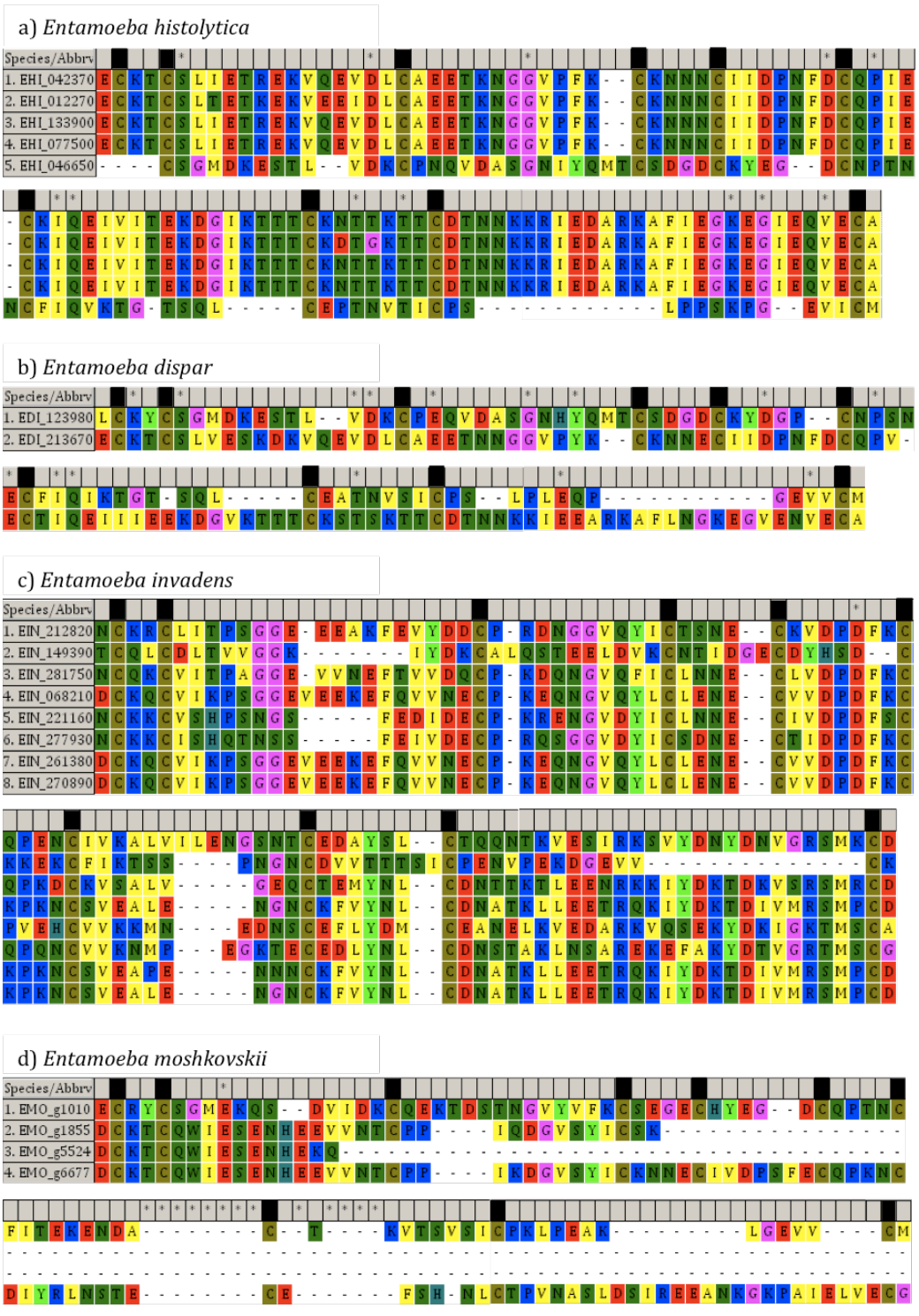


Figure 2.3.13. Alignments within species of Carbohydrate Recognition Domains within heavy Gal/GalNAc lectin subunits. Conserved cysteine residues are highlighted by black boxes. *E. moshkovskii* sequences were aligned to Hgl1 in *E. histolytica* to improve alignment.

iv. Direct virulence factor families - Cysteine proteases

The cysteine proteases can be divided into 3 subfamilies – A, B and C. When considered together, these three subfamilies highlight interesting differences between the *Entamoeba* species. In Family A (Figure 2.3.14.d), there are more *E. invadens* genes than there are genes from the other species, and a notably lower number of *E. dispar* sequences. The higher number of *E. invadens* genes is due to a largely variable expansion (mean branch length: 0.814269; s = 0.266459), which forms part of the clade containing the two most highly expressed *E. histolytica* genes. *E. dispar*, meanwhile, has a pseudogene in a region syntenic to *E. histolytica*'s CP-A5. There are no other *E. dispar* pseudogenes, whereas there are nine *E. histolytica* pseudogenes.

In Family B (Figure 2.3.14.e), *E. dispar* has considerably fewer genes than the other three species, as it is the only species whose genes have not been subject to expansion. However, RNA expression data shows that none of the 6 near-identical *E. histolytica* sequences are actually expressed under the studied conditions. The *E. moshkovskii* gene expansion is significantly more varied (p-value < 0.001) but relatively closely related to the *E. histolytica* expansion, being part of the same clade. Conversely, the expanded *E. invadens* genes are more varied than both the *E. moshkovskii* sequences (p-value < 0.001) and the *E. histolytica* sequences (p-value < 0.001) and have expanded in an independent event. As was seen in Family A, *E. invadens* appears to possess a larger, more variable set of cysteine proteases than the other three species.

In Family C (Figure 2.3.14.f), the numbers are much more even, though there are fewer *E. invadens* genes than there are of the other three species. This is due to a large clade consisting mostly of very similar sequences across those three species (mean branch length: 0.266321; s = 0.351707). Of the five *E. histolytica* sequences present in that clade, three have FPKM values of 35.711 and higher, whilst the remaining two are barely expressed (FPKM values of 0.233 and 1.788). As such, it may be that Family C demonstrates no major differences in numbers of functional proteins between the four species, although expression data for *E. dispar* and *E. moshkovskii* would offer some clarity here.

When one considers all three families together, the relative paucity of *E. dispar* sequences (33 CDSs, compared with 42 in *E. histolytica*, 51 in *E. invadens*, and 46 in

E. moshkovskii) is of interest as it suggests a diminished requirement for these virulence factors. Taken alongside the fact that *E. dispar* is the only one of the four species to have a pseudogenised orthologue of the pathogenically important CP-A5 gene [74, 75], it appears that *E. dispar* has experienced a general reduction in genes involved in host attachment and invasion. This reduction is likely to be at least partly responsible for its reduced virulence, which has long been recognised in the literature [2]. Conversely, this argument implicates *E. moshkovskii* in a pathogenic lifestyle, supporting the findings of an increasing number of studies [4, 44].

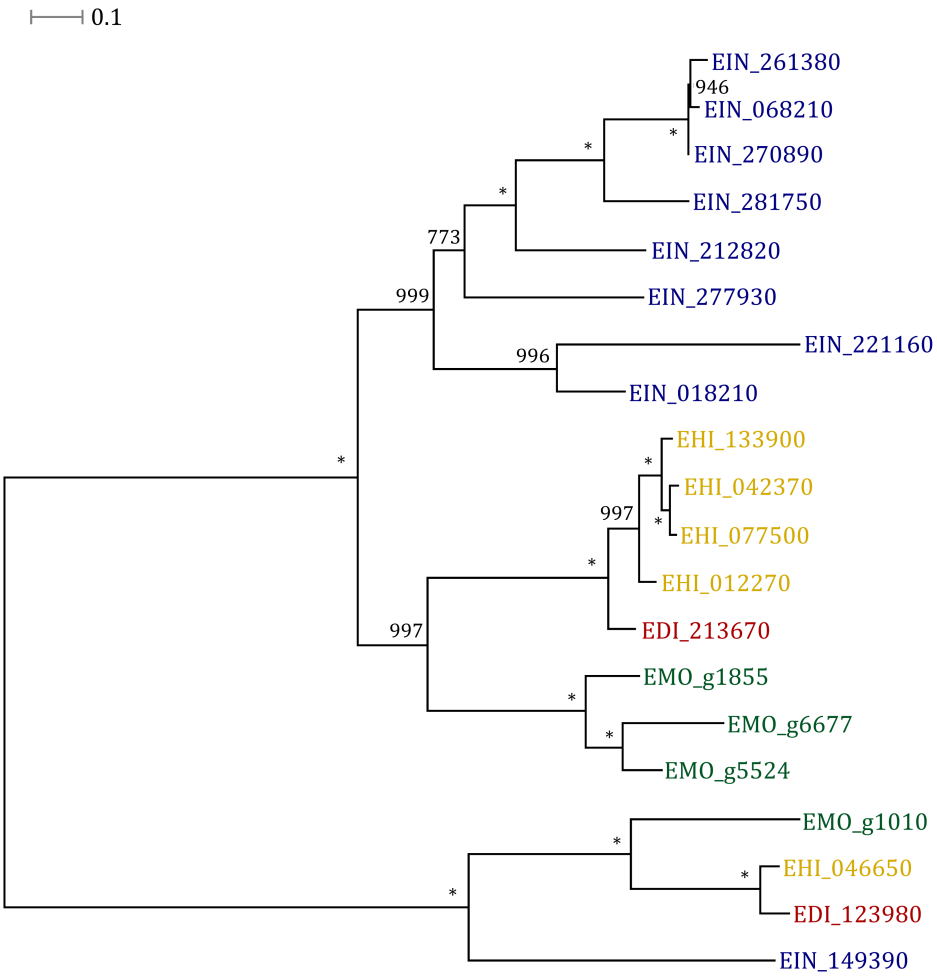
v. Direct virulence factor families - other families of interest

The serpin family (Figure 2.1.14.g) contains a large expansion of *E. moshkovskii* genes, with little variability between the sequences overall, although two sequences are relatively distant (mean branch length: 0.458415; s = 0.458697). The family also contains a smaller *E. invadens* expansion, the sequences in which are significantly more variable than the *E. moshkovskii* sequences (p-value < 0.001). There is only one *E. histolytica* sequence and two *E. dispar* sequences in the family. This family appears to have expanded relatively recently. Given the serpins' role in inhibiting host proteins [238], the lack of variability in the *E. moshkovskii* genes is surprising as it suggests there is little functional diversity in the family. It may be that *E. moshkovskii* utilises protein inhibitors that we are not yet aware of.

Finally, there are considerably more *E. histolytica* genes than any of the other species' genes in the sphingomyelinase C family. This is entirely due to an expansion of seven near-identical sequences (mean branch length: 2.28571e-5; s = 7.17137e-6). Only two of these genes are well expressed (all others have a FPKM value lower than 1.000), and one of these two only has an FPKM value of 4.803. There are four *E. dispar* pseudogenes, which, when included in a nucleotide-based phylogenetic analysis, were found to be most closely related to the sequences in the *E. histolytica* expansion. Therefore, these sequences have been lost in the *E. dispar* lineage and could be important candidates to explain the different phenotypes of the two species.

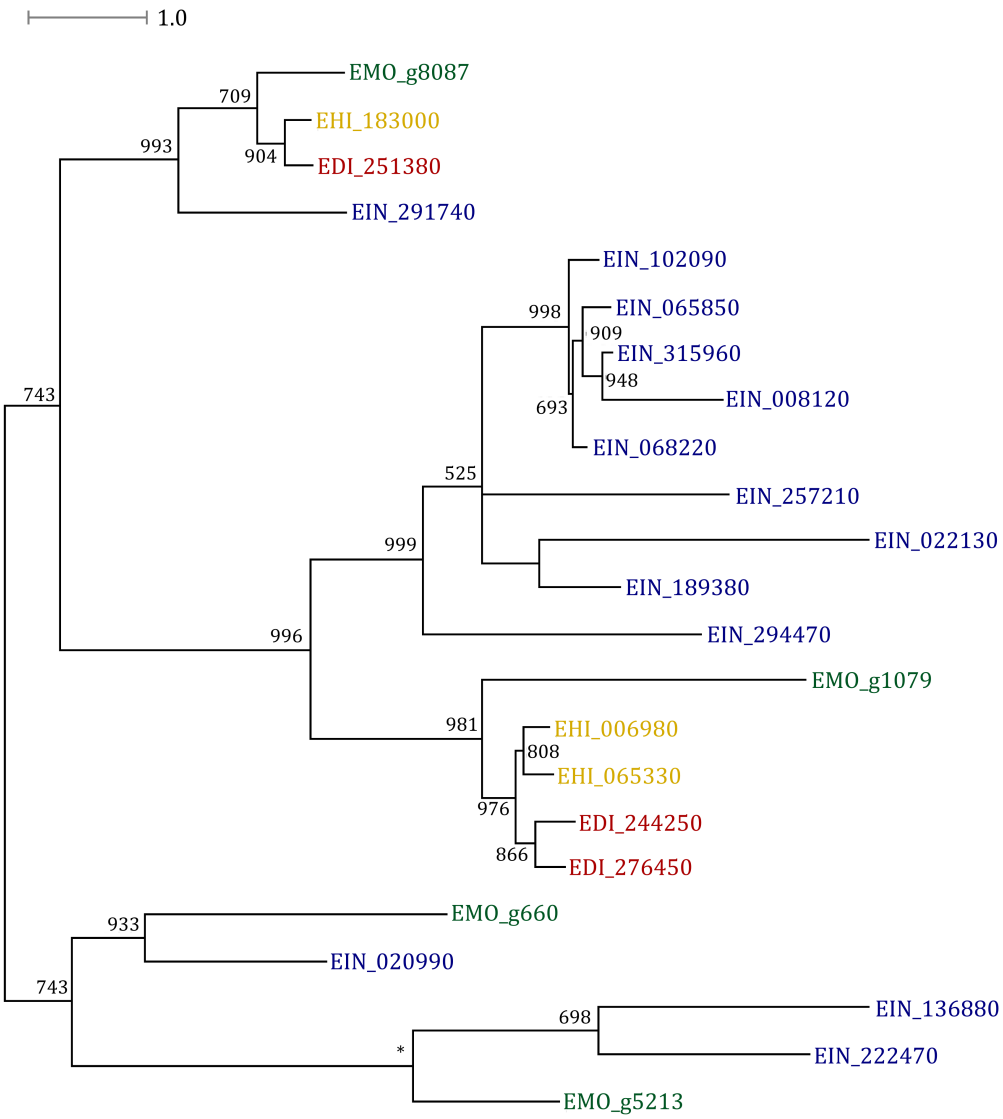
Overall, and in contrast to the 'indirect' virulence factor families, the directly virulent families demonstrated informative patterns of diversity. It is likely that these genes are under greater selective pressure to evolve, as per the Red Queen Hypothesis (discussed in depth in Chapter Three), allowing the parasites to manipulate a greater

range of hosts [140]. The direct virulence families of *E. invadens* proved particularly interesting and informative. In addition to the already discussed lectins, *E. invadens* appears to require a variety of cysteine proteases and serpins. Not only does this add credence to the established belief that the cysteine proteases are major facilitators of amoebiasis, it also supports the theory that *E. invadens* requires a greater number of important virulence factors to allow it to effectively parasitise its wide range of hosts. It is reasonable to conclude that a proportion of the enlarged gene set seen in *E. invadens* (relative to *E. histolytica* and *E. dispar*) consists of genes required to facilitate the amoeba's polyxenous lifestyle.

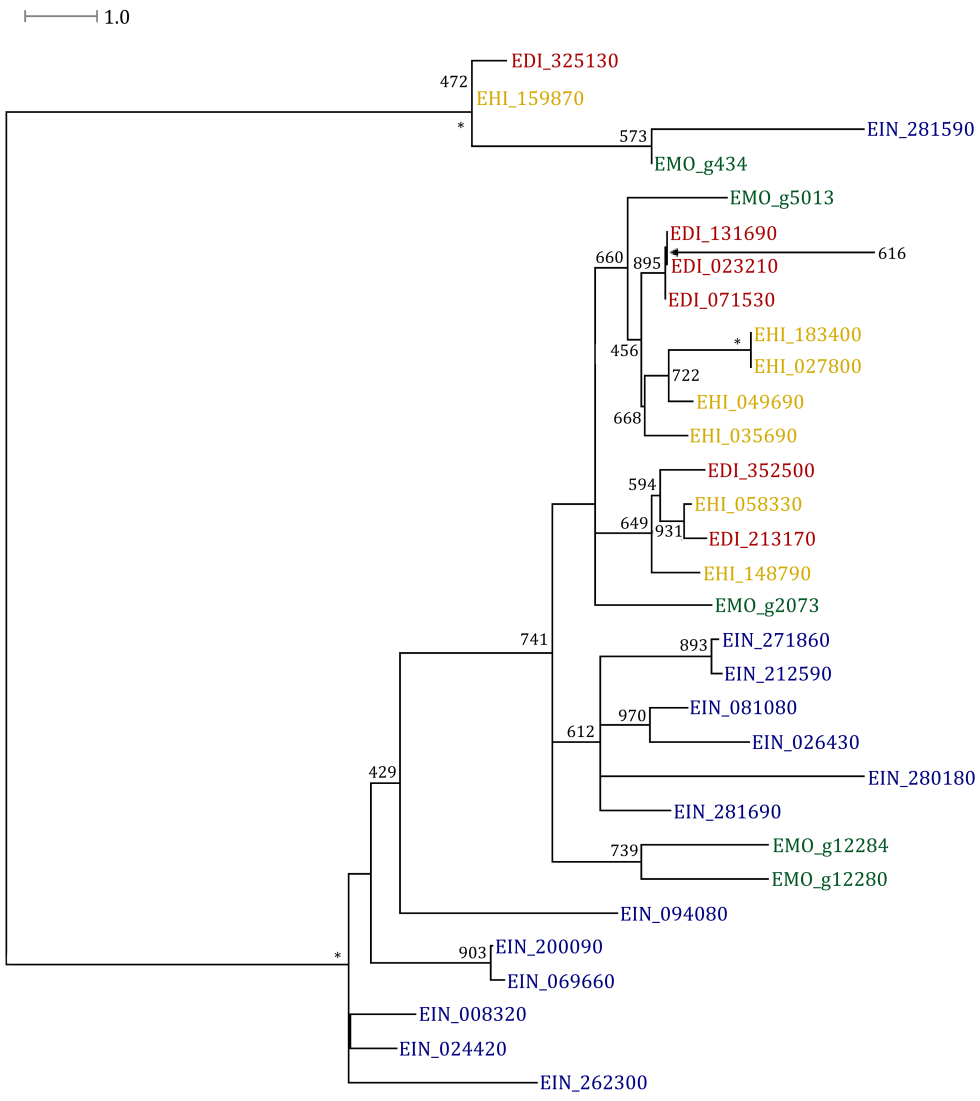


a) Heavy Gal/GalNAc lectin subunits

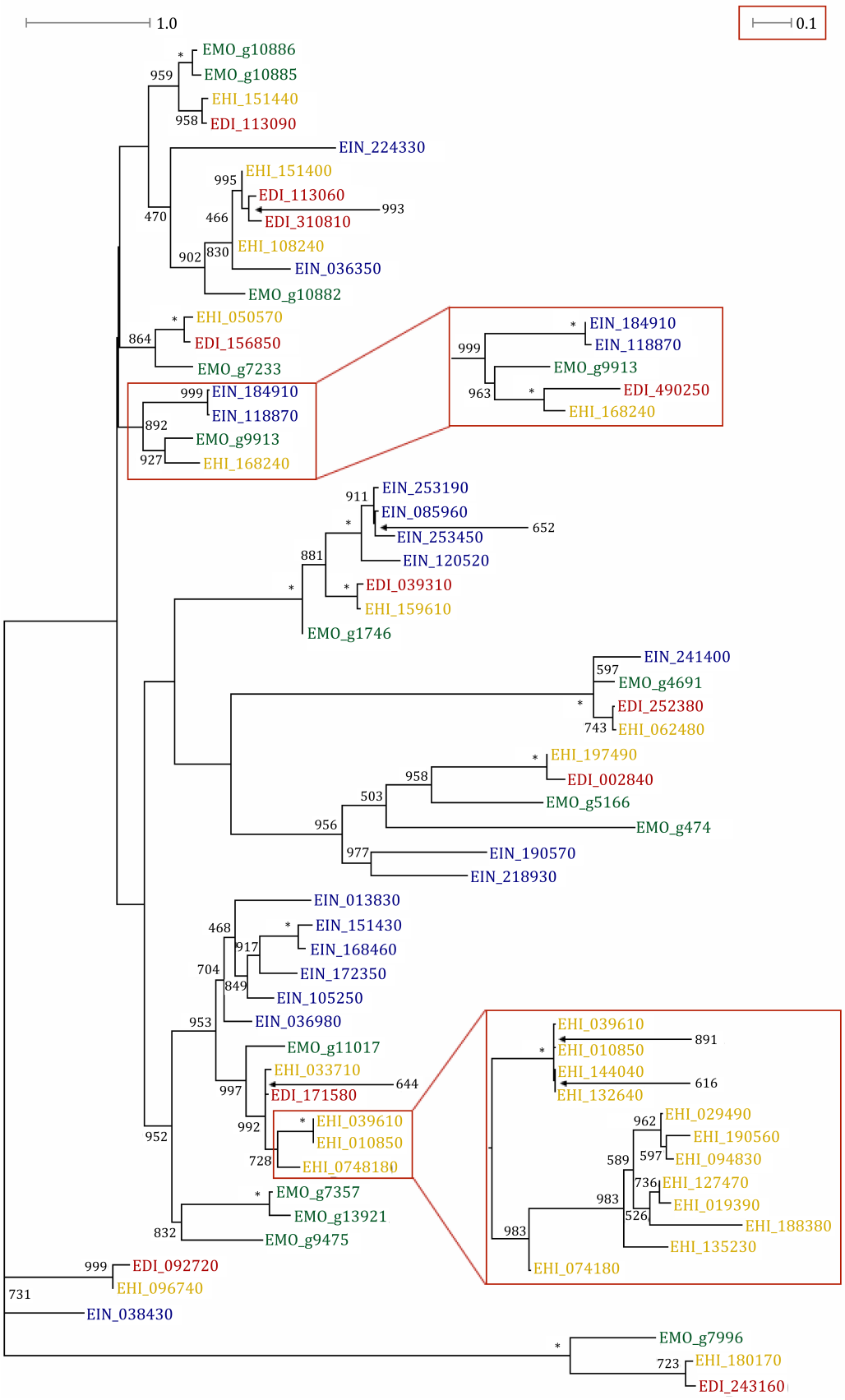
Figure 2.3.14. Phylograms of the *Entamoeba* gene families directly involved in virulence that demonstrate atypical phylogeny. Red boxes highlight clades in which pseudogenes were identified. They are linked to red boxes showing the same clades when phylogeny was calculated using nucleotide sequences, including the pseudogenes. Scale bars in red boxes represent nucleotide phylograms. All phylograms are midpoint rooted. Bootstrapping was performed for 1,000 replicates. Bootstrap values of 1,000 are represented by asterisks (*). Bootstrap values below 400 are not shown.



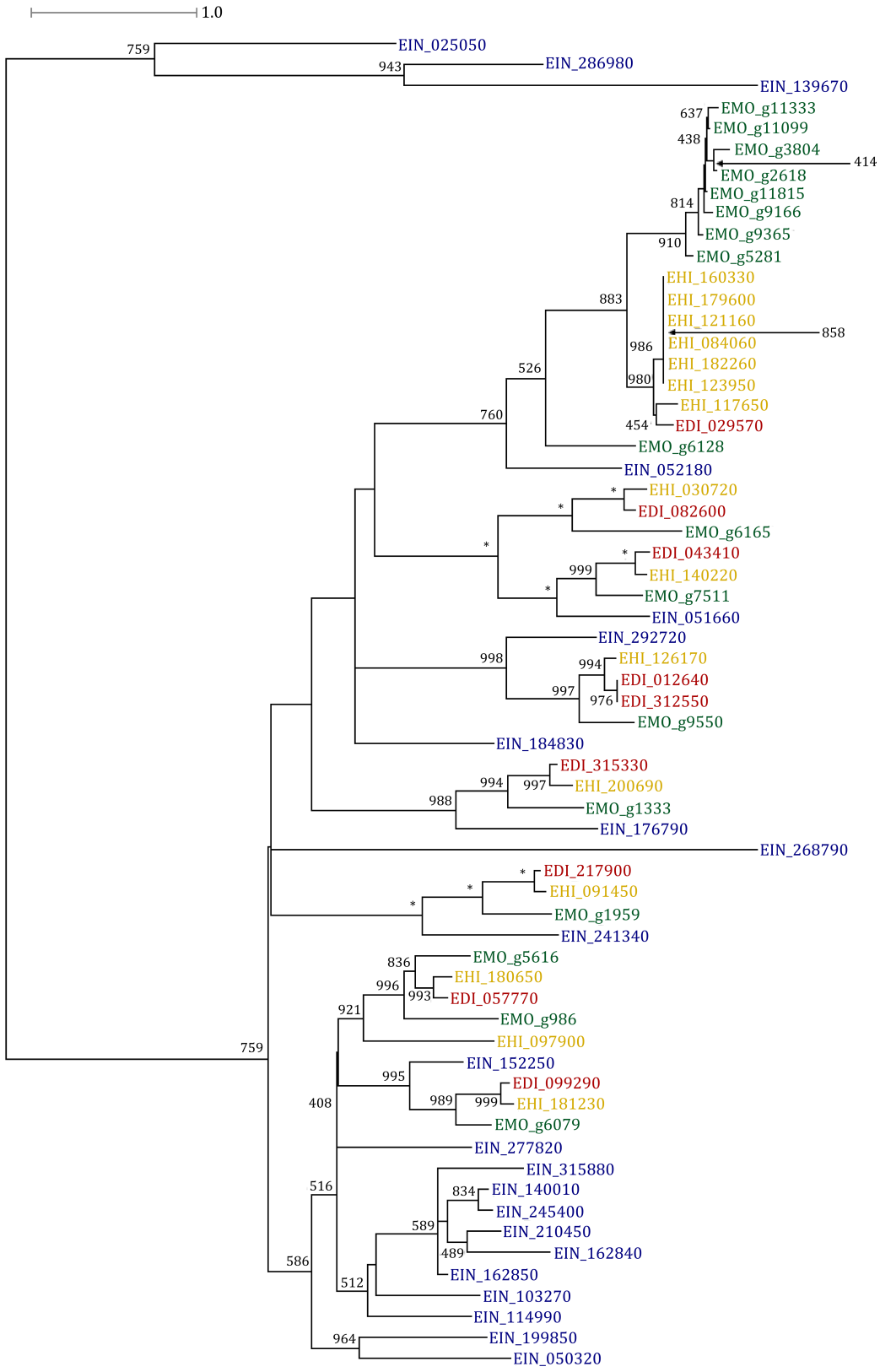
b) Intermediate Gal/GalNAc lectin subunits



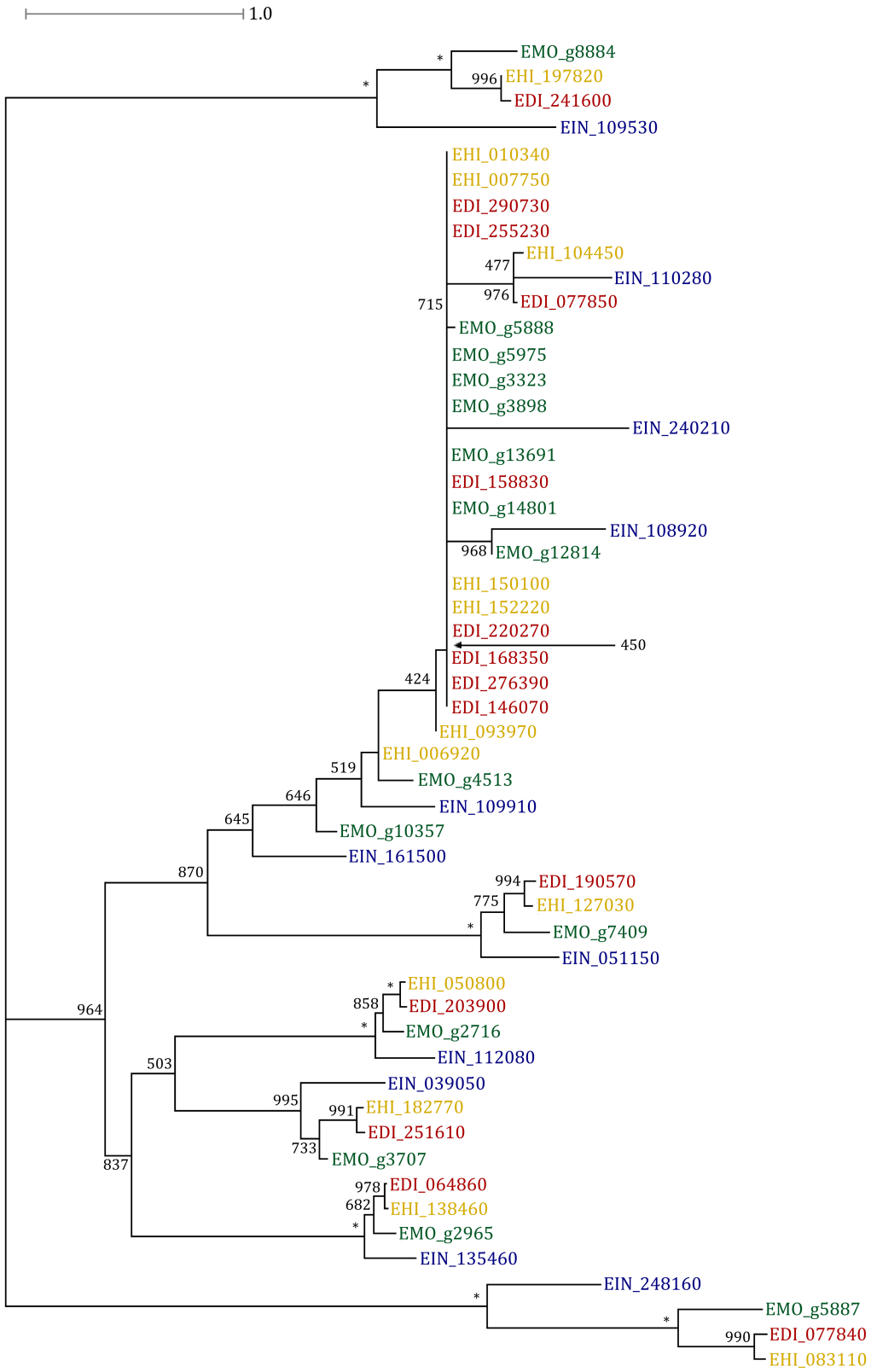
c) Light Gal/GalNAc lectin subunits



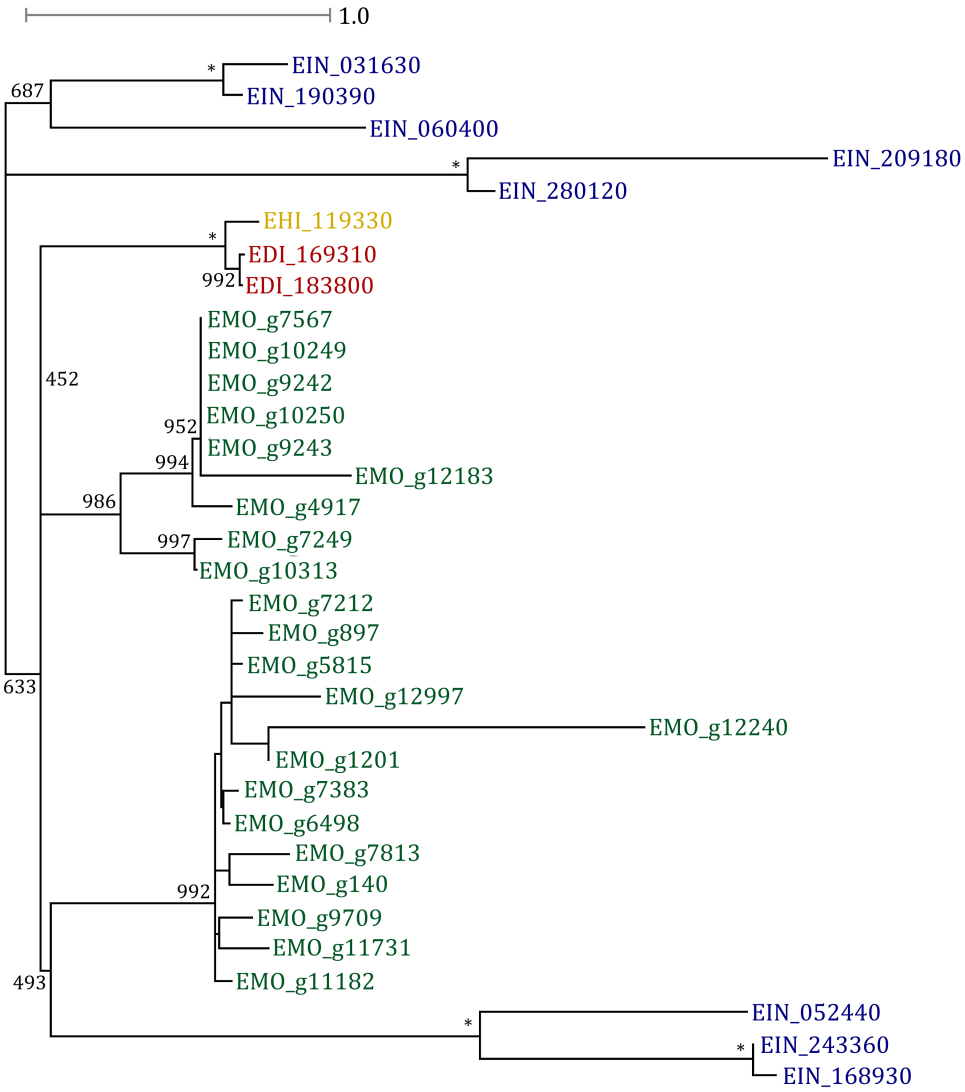
d) Cysteine protease Family A



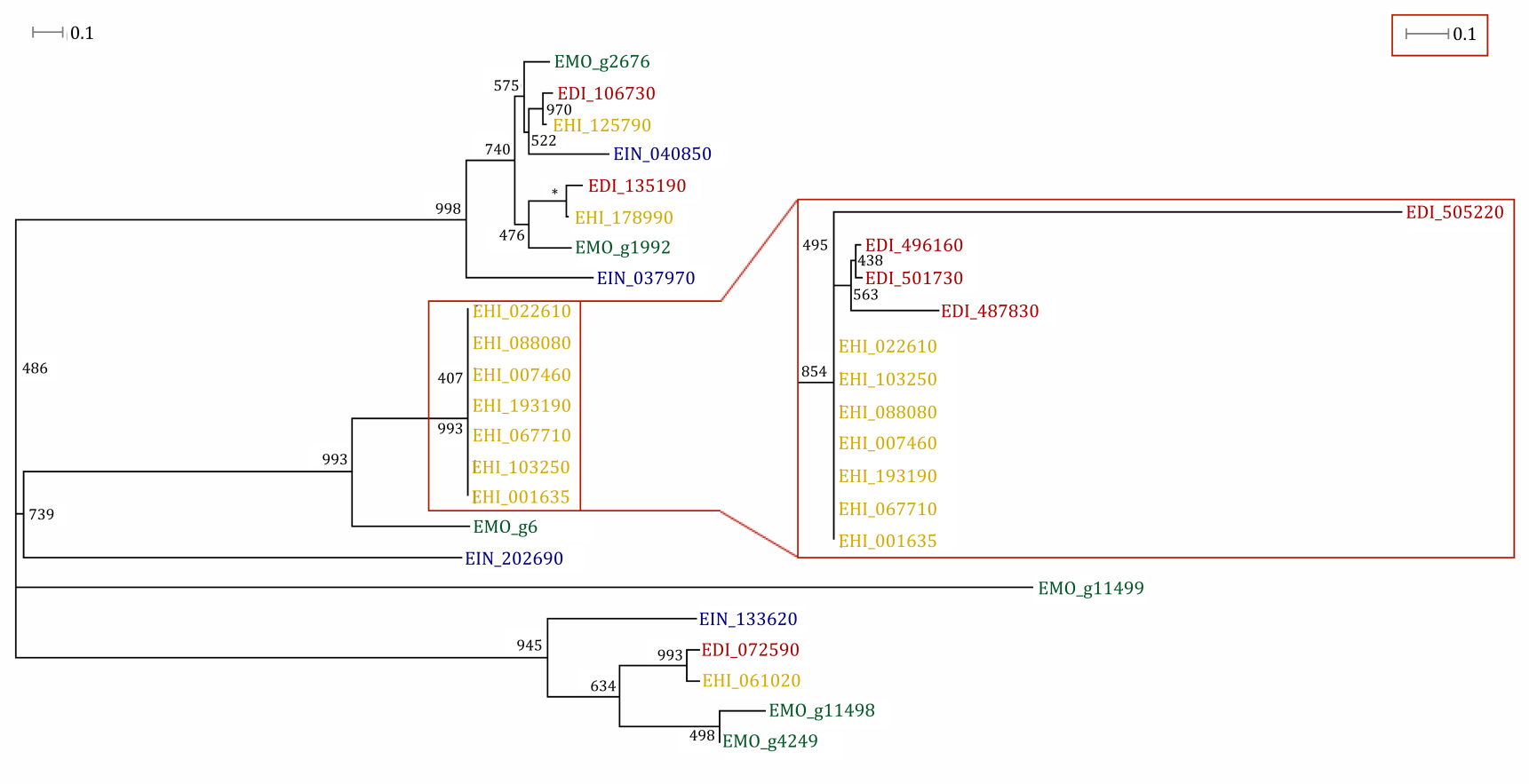
e) Cysteine protease Family B



f) Cysteine protease Family C


1.0
EIN_031630
EIN_190390
EIN_060400
EIN_209180
EIN_280120
EHI_119330
EDI_169310
EDI_183800
EMO_g7567
EMO_g10249
EMO_g9242
EMO_g10250
EMO_g9243
EMO_g12183
EMO_g4917
EMO_g7249
EMO_g10313
EMO_g7212
EMO_g897
EMO_g5815
EMO_g12997
EMO_g12240
EMO_g1201
EMO_g7383
EMO_g6498
EMO_g7813
EMO_g140
EMO_g9709
EMO_g11731
EMO_g11182
EIN_052440
EIN_243360
EIN_168930
687
992
452
952
994
986
997
633
992
493

g) Serpin



h) Sphingomyelinase C

2.4 Concluding remarks

In this chapter I have presented an assembled and annotated genome for the *Entamoeba moshkovskii* reference strain Laredo. It is the first of the *Entamoeba* reference strains to be generated using next-generation sequencing technology and is of the high quality that this technology makes possible. The genome contains 12,449 coding sequences, although many of these are incomplete, suggesting that this number is inflated above the actual gene count. Nevertheless, the assembly and annotation comprise a good quality first draft of the genome, which allowed comparisons with other members of the genus for the first time.

Through comparison of the *Entamoeba* genus with genera representing the diversity of the Unikonts, I have identified gene families present, and ubiquitous, in *Entamoeba* species only. The gene families present in the largest numbers were, as one might expect, largely housekeeping genes. However, they were families that allow *Entamoeba* species to exploit their environmental niche, including genes potentially involved in phagocytosis of bacteria. Additionally, the large BspA gene family has been identified as a potentially genus-wide facilitator of parasitism, allowing adhesion to host cells. This work also identified putative virulence factors theoretically present in all *Entamoeba* species, including two families not previously mooted as virulence factors in the genus.

Surface-bound proteins play a major role in the development of amoebiasis. The pathogenic *E. histolytica* possesses a large number of unique surface proteins, which contrasts starkly with the absence of genes encoding such proteins in the non-pathogenic *E. dispar*. Analysis of the heavy subunits of the Gal/GalNAc lectin revealed that every such gene in every species should possess a complete CRD, which is variable around a conserved motif of 10 cysteine residues. I propose that this information might help inform research into the utilisation of the Gal/GalNAc lectins as potential vaccine targets. This study also suggested that other surface-bound proteins might play roles similar in importance to the Gal/GalNAc lectins with regards to enabling pathogenic infections in the genus.

Furthermore, *E. invadens* was found to possess a greater number of genes in multiple gene families, including the Gal/GalNAc lectin heavy subunits and the cysteine proteases (families A and B). The genes comprising the expansions in these families

were also often significantly more variable than those genes seen in the other *Entamoeba* species. Taken together, these results suggest that *E. invadens* requires a greater variety of certain virulence factors to allow it to infect the large host range it is known to exploit. In contrast, the low numbers of surface proteins seen in *E. dispar* are also seen in its cysteine protease virulence factor gene families. Taken together, these results support the literature in implying that *E. dispar* is non-pathogenic. A further suggestion of counts of these virulence factors, however, is that current changing opinions about the pathogenicity of *E. moshkovskii* are correct – the species does indeed appear to be virulent. With an annotated genome of *E. moshkovskii* now in the public domain it is likely that research into this interesting species will pick up pace and further confirmation of its pathogenicity status will follow soon thereafter.

Chapter Three – Intra-species comparative analyses of *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii*

3.1 Introduction

3.1.1 Measures of genomic diversity

Genomic diversity defines, and distinguishes between, strains and individuals of a species. This variation can be attributed to three sources – horizontal gene transfer, recombination and mutations. Mutations are caused by errors made during the replication of genomic sequences, both coding and non-coding. They inherently cause differences in the nucleotide sequences in which they lie and, often, any amino acids encoded by those nucleotides. This introduces differences between the sequences and their ancestors, which can be used as a measure of evolutionary distance. For example, microsatellites – concatenated STR sequences - are commonly used to generate linkage maps, which offer a measure of recombination rates and, therefore, diversity between sequences [253-256]. They are used because highly repetitive sequences are difficult even for cellular DNA polymerases to replicate, resulting in the frequent addition or removal of repeat units as a polymerase 'slips', dissociating from, and annealing to, the template of the sequence erroneously [257]. This high mutation rate can be used to generate a relatively high-resolution inference of evolutionary distances.

This chapter is concerned with a particular form of mutation, not restricted to microsatellites, for which evidence of superiority when determining evolutionary distance is growing [258, 259]. Single nucleotide polymorphisms (SNPs) are defined as point mutations found in at least 1% of a given population [260]. SNPs are becoming increasingly favoured as a measure of diversity over microsatellites for a number of reasons [258, 259]. Firstly, although they are less abundant than STRs, they are widespread and less prone to further mutations, making them both reliable and representative of the entire genome [261]. It is also argued by several research groups that the semi-automated protocol used to study microsatellites, along with the high number of alleles possible in such regions, leads to a greater error rate than when analysing SNPs [262, 263]. An additional reason, specific to the focus of this study, is that microsatellites in *Entamoeba histolytica* are thought to be far less variable than

would be expected in a eukaryote [49, 50] making traditional marker based approaches impossible.

3.1.2 Diversity within members of the *Entamoeba* genus

The differences in reported diversity that can arise when comparing SNP-based studies and those that use microsatellites have been seen in members of the *Entamoeba* genus. Several reports studying STRs in particular genes [16, 264, 265] and loci [266, 267] have found diversity within *E. histolytica* to be relatively high. As stated, however, these studies were limited to individual genes or a small number of repeat regions and they were also subject to the same limitations as all microsatellite-based studies, as described in Section 3.1.1. Several reports, focusing on SNPs, have found contradictory evidence to support the theory of limited genetic diversity amongst *E. histolytica* strains [27, 68, 268]. Initially, this was thought to indicate a clonal species, however evidence of meiotic recombination has been discovered recently suggesting that *E. histolytica* actually reproduces sexually [68].

There is a relative paucity of studies into diversity in *E. histolytica*'s close relatives, *Entamoeba dispar* and *Entamoeba moshkovskii.* In the case of *E. moshkovskii*, this is, of course, because there was previously no reference genome with which to compare different strains. In spite of this, there is support for the theory that *E. moshkovskii* is, in fact, highly variable and may be a species complex, rather than an individual species [44, 146]. The few studies that have analysed diversity in the non-pathogenic *E. dispar* have focused on individual genes or loci containing STRs and, as might be expected given the results of similar work in *E. histolytica*, they found the species to be highly diverse [269, 270].

It should be noted that, whilst SNPs are ubiquitous throughout genomes, they occur at greater frequencies in certain regions. For example, in *E. histolytica*, as in humans [271], it has been reported that non-coding regions of the genome mutate at a faster rate than coding regions [268]. As such, it is possible for limited investigations to identify different mutation rates dependent upon the sequences they study. The most inclusive alternative is, of course, to perform genome-wide studies. In light of the increasing support for diversity studies using SNPs rather than microsatellites, as well as the shortcomings of limited-scale studies, it was decided that this study would use SNPs to compare genome-wide diversity levels between and within the species *E.*

histolytica, *E. dispar* and *E. moshkovskii*, thus going some way to filling the current deficit in such studies in the genus. Only human-infective species were compared, as these are the only species for which there exist numerous axenic isolates.

The strains studied in this chapter were acquired from a variety of sources, presenting interesting challenges. Many of the strains used were isolated in different years, with the range covering decades. Those strains that were isolated longer ago are likely to have gone through more generations *in vitro* than more recently isolated strains, meaning it is less likely that their genomes resemble the original isolated samples. However, the efforts that would be required to isolate and axenise a sufficient number of strains for this study meant that isolating strains from new samples was an unviable option.

As such, this study made use of existing axenised strains and, where available, sources of genomic data. The genomes of *E. moshkovskii* strains and *E. dispar* strains were sequenced specifically for this project. For *E. histolytica*, existing genomic sequencing data, generated in two individual studies, were utilised. Eight of the ten strains were sequenced on the SOLiD 4 platform in a study of the genomic diversity seen in *E. histolytica* [68]. The original analysis identified SNPs across the genomes of the strains as a measure of diversity within the species. With the exception of the reference strain, the genomic data for these strains were also featured in a study that combined and compared *E. histolytica* strains' genomes sequenced using SOLiD or 454 technologies [272]. That piece of research also investigated SNP content of *E. histolytica*, but focused on identifying differences between virulent and avirulent strains. This latter study demonstrated that it is possible to gain meaningful results from data gained from multiple sequencing platforms. This point was explored in this chapter as the strains sequenced in the latter diversity study were also sequenced using Illumina technology, which could be mapped using the same commands as those used for SOLiD data.

3.1.3 Selective pressures and the Red Queen hypothesis

Mutations at individual sites in a coding sequence can have a multitude of effects upon the translated amino acid sequence. At some sites (always the final base in a triplet codon), no mutation is capable of causing a change to the amino acid sequence. Such sites are described as fourfold degenerate (4D) synonymous sites.

Mutations at other nucleotide positions will always lead to amino acid sequence alterations. These sites are termed non-synonymous sites. Still other sites can be a fraction non-synonymous and a fraction synonymous, depending upon the number of changes to the site that, respectively, will and will not cause changes to the encoded amino acid.

The average number of nucleotide differences between a coding sequence and its corresponding reference sequence per non-synonymous site is described as the sequence's 'dN' value. The equivalent value for synonymous sites is represented as 'dS'. Sequences across which the average dN value is greater than the average dS value are described as being under diversifying selective pressure, whereby mutations that alter the amino acid sequence encoded by the gene are accrued, resulting in a population with multiple alleles. Over time this can result in the evolution of new strains or, ultimately, speciation, theoretically improving the fitness of the organism to a particular environment. Conversely, sequences in which the average dS value exceeds the average dN value are considered to be under purifying selection, whereby only mutations that do not alter the amino acid sequence of the encoded protein are permitted to accrue, preserving the structure and function of the protein. Where more than two sequences are concerned, this principle is usually extended to measure sequence diversity (π) as the mean fraction of nucleotide differences between all possible pairwise comparisons of the sequences [273].

The Red Queen hypothesis states that coevolving species are constantly driven to adapt to changes, both in the environment and one another, in order to survive [139]. In host-parasite pairings, such as those seen in the *Entamoeba* genus [274, 275], the parasite evolves to overcome host defences, thus necessitating the evolution of countermeasures in the host. As such, evolution of these species is driven forward faster than if they were to evolve independently of one another [140]. It is anticipated that such antagonistic co-evolution has advanced the adaptation of host-infective *Entamoeba* species. Of particular interest in this chapter is the fact that, in host-parasite relationships, those genes directly involved in survival evolve faster as they are under stronger diversifying selective pressures than others as the molecular 'arms race' escalates [140, 274, 276]. Informative results of projects investigating these selective pressures have the potential to identify genes directly involved in host-parasite interactions, such as those seen between certain *Entamoeba* species and their human hosts.

3.1.4 Aims of chapter

In this chapter, I have sequenced the genomes of multiple strains of *E. moshkovskii* and two strains of *E. dispar*. I have mapped the reads generated for each of the strains, along with publicly available reads of multiple *E. histolytica* strains, to their respective reference genomes, identifying SNPs based upon pairwise comparisons. Divergence between the genomes and evidence of selective pressures acting upon genes within them were analysed using the generated data. This work had multiple goals. Firstly, it was hoped that the data could be used to compare the divergence levels seen in each of the three species, particularly with a view to gaining evidence that *E. moshkovskii* has a highly variable genome and is, potentially, a species complex [38, 146, 152]. Secondly, the data was used to compare variability between different sequence classes within each genome so as to identify in which genomic regions the greatest variability occurs. Finally, I hoped to identify the genes that were under diversifying selective pressures within each species, indicating those sequences that were perhaps of greatest importance in survival of the different lifestyles exhibited by the various strains and species.

3.2 Materials and Methods

3.2.1 Acquisition of *Entamoeba dispar* extracts and *Entamoeba moshkovskii* cultures

Cultures of *E. moshkovskii* strains Laredo, FIC, 15114 and Snake were grown and maintained using the medium described in Section 2.2.1. Cultures were grown in volumes of 50 - 60 mL in 65 mL sterile filter-capped plastic flasks (Corning), the lids of which were wrapped in Parafilm (Sigma-Aldrich), thus restricting oxygen availability. Inoculum volumes were varied based upon subjective microscopic observation of growth in seeding cultures. All *E. moshkovskii* cultures were incubated at room temperature for 7 days. All cultures were incubated in darkness to prevent a reduction in efficacy of the photosensitive Vitamin Mix #18. Dr Graham Clark provided lysates of *E. dispar* strain SAW760. Dr Tomoyoshi Nozaki's research group (University of Tsukuba, Japan) extracted, and provided, genomic DNA from the *E. dispar* strain AS16IR.

3.2.2 DNA extraction

The method described in Section 2.2.2 was used to extract DNA from the *E. moshkovskii* cultures and the *E. dispar* SAW760 lysates. Purity and concentration of DNA was measured using the NanoDrop 1000 Spectrophotometer (Thermo Scientific) and the Qubit Fluorometer (Invitrogen by Life Technologies), respectively, according to the manufacturers' protocols (Appendix B, Table B.1). The Qubit Fluorometer's dsDNA Broad Range assays were utilised unless low DNA concentrations required the use of the dsDNA High Specificity assay.

3.2.3 Library preparation and sequencing

The CGR prepared, pooled and sequenced 100 bp PE libraries of the purified extracted DNA from *E. moshkovskii* strains FIC, 15114 and Snake using their standard protocol. I prepared 150 bp PE libraries for *E. dispar* strains SAW760 and AS16IR and for *E. moshkovskii* strain Laredo. The details that follow in this section relate to the libraries I prepared only.

The preparation technique used was an adaptation of the 'TruSeq DNA sample prep low throughput protocol' (Illumina). Where necessary, the DNA was diluted or concentrated to achieve the recommended concentration (20 ng/μL) before being sheared using an S220 Focused-Ultrasonicator (Covaris). Covaris-specific tubes were used, rather than the recommended plates so TE buffer was used to increase the sample volume to 130 μL. After shearing, the AMPure XP beads step was used to reduce the sample volume to 50 μL. During the End Repair clean-up step no dilution was carried out. Wherever they were stated as options, the in-line control reagent and gel-free method were used. The three libraries I prepared in this chapter were pooled with the library described in Chapter Four – that of *Entamoeba bangladeshi* strain 8237. Libraries for *E. dispar* SAW760 and *E. bangladeshi* were subsequently pooled together again, without the other two libraries, as they had not formed an acceptable proportion of the initial pooled library. The adapters used in all libraries were the Set B adapters recommended in Option 1 for a plexity of 4 in the TruSeq pooling guidelines. These adapters were applied to the four libraries as follows: AD001: *E. moshkovskii* Laredo; AD008: *E. dispar* SAW760; AD010: *E. dispar* AS16IR; and AD011: *E. bangladeshi* 8237.

TruSeq output libraries were diluted to within the range of the Agilent 2100 Bioanalyzer's (Agilent Technologies) High Sensitivity DNA analysis kit (5 - 500 pg/μL) and measured using the Qubit Fluorometer, as above. The Agilent 2100 Bioanalyzer was used to test the library fragment sizes. As each read in a pair was 150 bp in length, and a 65 bp adapter was annealed to each read, a total fragment length of between 400 and 600 bp was required to limit selection to pairs with up to ~170 bp of intervening sequence. A gap any greater than 200 bp between a pair was thought likely to be so big as to prevent the BWA assembler (Section 3.2.5) from assembling the reads as pairs. A Pippin Prep machine (Sage Science) was used to select such required fragments using a 1.5% agarose gel cassette (range of 250 bp – 1.5 kb). Residual ethidium bromide was cleaned from Pippin Prep eluates using the AMPure XP beads technique described in the TruSeq protocol. Cleaned libraries, measured using the Qubit Fluorometer and Agilent 2100 Bioanalyzer (Appendix B, Table B.1), were submitted to the CGR for sequencing. Each library was sequenced on 1 lane of the Illumina MiSeq platform, according to the optimised protocol used by the CGR.

3.2.4 Acquisition of *Entamoeba histolytica* strain data

Previously sequenced read data for *E. histolytica* strains were used [68, 272]. Strains MS96-3382 and DS4-868, sequenced using Illumina technology, were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena). Their Run accession numbers are SRR368631 and SRR369427, respectively. Dr Gareth Weedall provided SOLiD-derived read data for *E. histolytica* strains Rahman, 2592100, PVB-M08B, PVB-M08F, HK-9, MS27-5030, MS84-1373 and HM-1:IMSS-A.

3.2.5 Variant calling and analysis

Reads from the reference strains were aligned to the existing assembled reference sequences, downloaded from AmoebaDB v2.0 [169, 170], using the Burrows-Wheeler Aligner (BWA) v0.5.9 [277]. Default parameters were applied to the 'aln' command except in two cases. Firstly, suboptimal alignments were permitted for reads that could be mapped to multiple sites provided that there were no more than 10 equally best potential sites. Secondly, maximum edit distances of 4 and 12 were applied to the SOLiD reads and longer Illumina reads, respectively. The 'samse' and 'sampe' commands were used to align the SOLiD and Illumina reads, respectively, using default parameters except for limiting the maximum number of alignments output in the XA tag to 2. Unmapped and non-uniquely mapped reads were filtered out.

SNPs in the aligned reference strains' reads were called using the SAMtools v0.1.18 [278] mpileup command (default parameters were used apart from forcing the output of per-sample read depths) and bcftools view command (default parameters were used except for setting it to output both bases and variants). High quality SNPs were defined as those that met the following parameters: Phred quality score >= 20; read depth >=5 and <= 95th percentile of all depths seen in assembly; and farther than 5 bp from a gap, using a window of 30 bp. High quality homozygous SNPs were inserted in place of their respective original bases within the original reference sequences (see Appendix D, File D.1 for details). The updated reference sequences were then used in place of the original genomes when reads from non-reference strains were mapped, and SNPs called, using the method outlined above.

Total counts of SNPs per gene, excluding pseudogenes and sequences with an incomplete triplet codon, were calculated per strain. A Perl script, provided by Dr

Weedall, was used to distinguish between synonymous and non-synonymous SNPs in coding regions and SNPs in non-coding regions. Programs from the Phylogenetic Analysis Using Maximum Likelihood (PAML) package v4.5 [279, 280] were used to calculate dN and dS values for each gene. The Probabilistic Alignment Kit (PRANK) v.111130 was run using an empirical codon model with other parameters set to default values, followed by codeml, run using default parameters. For SNP counts and respective dN/dS ratios, see Appendix D, Files D.2-D.4.

3.2.6 Phylogenetic analysis

PHYLIP v3.69 [180] was used to generate a neighbour-joining phylogram for nucleotide positions of common 4D sites in *E. moshkovskii* strains, using the additive tree model. Default parameters were used unless otherwise stated. Seqboot was run with 1,000 bootstrap replicates. DNAdist was then run using the Jukes-Cantor model, which does not take codon position into account [281]. Neighbor was subsequently run for the 1,000 data sets, the output of which was processed by Consense. To apply branch lengths that represented evolutionary distances to trees, a distance matrix (Table 3.3.3), consisting of differences between pairs of strains per 4D site, was submitted to Neighbor for a single data set. Branch lengths were manually added to Consense output files.

3.2.7 Expression data

As in Chapter Two, expression data for all *E. histolytica* HM-1:IMSS virulence factor CDSs were downloaded from AmoebaDB v2.0. The figures attributed to each gene represented the FPKM values generated from RNA-Seq data.

3.2.8 Four-haplotype test in *Entamoeba moshkovskii*

The four-haplotype test was used to test for evidence of meiotic recombination between the four *E. moshkovskii* strains featured in this chapter. High quality sites were used for this. These were defined as nucleotide positions called in every strain and existing as homozygotes in every strain, but varying between them. One million pairs of these high quality SNPs were randomly sampled. Within groups of 10,000 pairs, proportions of SNP pairs existing as four haplotypes were calculated and the

group's average distance between pairs of sites was calculated. This test was carried out using a Perl script written by Dr Weedall [68].

3.3 Results and Discussion

3.3.1 Mapping to reference strains

As a method of comparing variation within *Entamoeba* species and identifying the most variable gene sequences in each genome, counts of SNPs across multiple strains in each species were required. The first step in generating them was to map reference strains' reads to their respective genomes in order to identify errors in the existing sequences. This would reduce the number of errors carried forward into subsequent analyses when comparing strains with their respective reference genomes. The reference strains were sequenced to relatively high average coverage depths in this study (Table 3.3.2). As such, it was likely that any bases identified as high quality homozygous SNPs here could be assumed to represent bases that were incorrectly called in the original genome assemblies (Table 3.3.1).

A much greater number of homozygous SNPS were identified in the *E. dispar* SAW760 genome than in the *E. moshkovskii* and *E. histolytica* reference strains. This was most likely a result of the relatively low average coverage depth achieved during the initial assembly of the *E. dispar* reference genome (Table 2.3.1). Assemblies performed at low coverage depths are, naturally, prone to more frequent errors. Interestingly, however, it was the *E. moshkovskii* Laredo genome that demonstrated the greatest number of heterozygous differences and insertions/deletions (indels). This was suggestive of *E. moshkovskii* possessing a more polymorphic genome than the other two species. Existing bases at homozygous positions were replaced with the newly called nucleotides to generate updated and improved sequences. Multiple alleles of heterozygous SNPs, by their very nature, are present in the genomes, making it impossible to define a 'correct' base at those sites. As such, neither heterozygous SNPs, nor indels, were considered high quality calls and so were not used to alter the reference genomes. The number of heterozygous SNPs in *E. histolytica* was extremely low; this may be, in part, due to the sequencing technology used but also because the HM-1:IMSS strain has been kept in culture for a long time and is likely to have been subject to loss of heterozygosity.

Table 3.3.1 Counts and proportions of variants detected when reads from the reference strains were mapped to existing reference genomes

Species	Homozygous SNPs		Heterozygous SNPs		Indels
	Count	SNPs/Kb	Count	SNPs/Kb	Count
E. histolytica	120	5.77 e-3	1,361	6.54 e-2	176
E. dispar	741	3.32 e-2	13,960	6.08 e-1	1,968
E. moshkovskii	165	6.53 e-3	33,527	1.33	2,446

Reads from non-reference strains were mapped to these updated reference genomes, rather than the original assemblies, and SNPs were called within them. Nine non-reference strains of *E. histolytica* were compared along with three strains of *E. moshkovskii* and one strain of *E. dispar* (Table 3.3.2). Acquisition of cultures and lysates of *E. dispar* strains proved problematic, hence the lower than ideal number of strains included in this study. The AS16IR strain of *E. dispar* mapped poorly to the SAW760 reference genome, achieving a very low average coverage depth. It did, however, map to approximately 80% of the reference, meaning that it was still informative, though conclusions drawn from it should be treated with caution, as many SNPs will have been omitted.

As might be expected, the proportion of reads that successfully mapped to a reference genome differed between strains sequenced using different sequencing platforms (Table 3.3.2). By comparing the *E. histolytica* strains alone, to achieve direct comparisons, one can see that greater proportions of reads were successfully mapped to greater proportions of the reference sequence in those strains sequenced using Illumina technology. Mapping of such strains also achieved considerably higher average coverage depths than seen in those sequenced using the SOLiD 4 platform. This is likely to have been caused by a combination of the greater length of the Illumina reads (150bp compared with 50bp) and the fact that the Illumina reads formed PE libraries, so could be mapped across regions 400-600 bp in length. It is feasible that these characteristics allowed the reads to map to regions that single SOLiD reads could not; for example, *Entamoeba* repetitive elements and LINES and SINES, which are known to be prevalent throughout *Entamoeba* genomes [55, 57, 186, 187, 282]. It is also possible that this improved mapping subsequently resulted in more SNPs being detectable. As such, there are likely to be large differences between those *E. histolytica*

strains sequenced using SOLiD and Illumina platforms. Whilst this is a disadvantage of comparing genomes sequenced by multiple technologies, it is one that was unavoidable in this instance. Its affect on SNP counts is considered throughout this chapter so as to avoid drawing inaccurate conclusions from data that are not directly comparable.

In recent years, much has been made of the limitations of individual sequencing platforms and the need to combine technologies to accurately sequence genomes and call SNPs [106, 283, 284]. Indeed, some mutations can only be identified by one of the three second generation sequencing technologies. It is, therefore, probable that greater coverage and, consequently, greater accuracy of SNP detection could have been achieved by combining reads generated by multiple sequencing technologies in individual strains. This is certainly a limitation of this work but, as reads generated using different platforms are made publicly available, it would be interesting to see how their combined application affects mapping quality and variant calling in the genus. It should be noted, however, that both the SOLiD and Illumina platforms demonstrate coverage bias relative to GC content, favouring low GC contents below approximately 40% [284], making them the ideal platforms for single-technology SNP calling in *Entamoeba*.

The non-reference *E. moshkovskii* strains mapped to coverage depths equivalent to, or higher than, those achieved with the *E. histolytica* strains sequenced on the Illumina platform. Despite this, the reads covered a lower proportion of the reference genome than their *E. histolytica* equivalents. This provided an early indicator of the expected disparity between the *E. moshkovskii* strains, relative to *E histolytica*. Strain FIC exhibited the poorest mapping of the *E. moshkovskii* strains, whilst strain 15114 mapped the most successfully. This suggested that FIC was more distant from Laredo than the other two strains, with strain 15114 being most closely related to Laredo; a notion that was reinforced by SNP calling statistics, discussed later in this chapter.

Table 3.3.2. Mapping and coverage statistics for each strain studied in this project. Grey rows represent reference strains, reads from which were mapped to their existing respective reference genome. Positions at which high quality homozygous SNP calls were made in the reads were replaced in the original reference sequence. All other strains were mapped to the updated versions of their respective reference strains. Underlined sections of strain names represent the shortened versions of the names that will be used in this project. References: a) [285]; b) [15]; c) [68]; d) [286]; e) [2]; f) [272]; g) [17]; h) [147].

Strain	Country of origin	Sequencing platform	Year of isolation	Average coverage depth (x)	No of mapped reads	Coverage of reference (%)
Entamoeba histolytica						
HM-1:IMSS-A [a,b]	Mexico	SOLiD 4	1967	43.53	13,743,197 (29.42%)	61.03
2592100 [c]	Bangladesh	SOLiD 4	2005	41.50	13,618,188 (21.97%)	68.83
HK-9 [d]	Korea	SOLiD 4	1951	57.41	21,217,510 (24.61%)	71.86
PVBM08B [c]	Italy	SOLiD 4	2007	50.02	17,688,152 (21.37%)	70.88
PVBM08F [c]	Italy	SOLiD 4	2007	29.61	8,506,016 (10.85%)	71.88
Rahman [e]	UK	SOLiD 4	1964	49.43	19,534,522 (29.84%)	67.78
MS27-5030 [c]	Bangladesh	SOLiD 4	2006	59.97	20,419,790 (24.76%)	63.27
MS84-1373 [c]	Bangladesh	SOLiD 4	2006	63.01	21,499,758 (23.28%)	69.57
MS96-3382 [f]	Bangladesh	Illumina GA II	2007	114.03	20,527,917 (61.17%)	89.00
DS4-868 [g]	Bangladesh	Illumina GA II	2006	72.15	13,361,613 (61.32%)	88.36

Strain	Country of origin	Sequencing platform	Year of isolation	Average coverage depth (x)	No of mapped reads	Coverage of reference (%)
Entamoeba dispar						
SAW760	England	Illumina MiSeq	1979	43.78	9,475,439 (37.21%)	88.54
AS16IR	Iran	Illumina MiSeq	1997	12.77	1,939,915 (32.87%)	81.14
Entamoeba moshkovskii						
Laredo [h]	America	Illumina MiSeq	1956	97.61	8,833,683 (80.65%)	89.91
FIC	Canada	Illumina MiSeq	1959	162.27	19,750,749 (30.98%)	61.58
Snake	France*	Illumina MiSeq	1948*	209.10	25,655,106 (51.40)	76.96
15114	Bangladesh	Illumina MiSeq	1999	265.55	35,292,777 (71.53%)	85.24

* Sent from Institut Pasteur, Paris to Charles University, Prague in 1948. Institut Pasteur has no record of origin (personal communication with Dr Graham Clark).

3.3.2 Comparison of SNP rates in strains of the three *Entamoeba* species

To understand how much variation exists within the three species, SNP counts were made for each strain. The average coverage depth necessary to reliably detect 95% of SNPs in a genetic sequence is 35x [287, 288]. With that in mind, it is important to note that *E. histolytica* strain PVF and *E. dispar* strain AS16IR were sequenced and mapped to lower average coverage depths than this (Table 3.3.2). As such, it is likely that many SNPs were omitted from analyses of these two sequences. Indeed, it should be appreciated that, without 100% coverage of the reference sequence, it is probable that less than 95% of SNPs will have been called in all of the strains, though the likelihood of this is reduced with greater coverage. The average coverage depth across the *E. histolytica* non-reference strains is 59.68x. In *E. dispar* AS16IR the average depth was very poor at 12.77x, whilst the average in *E. moshkovskii* non-reference strains was considerably higher than in both of the other species at 212.31x.

Pairwise SNP rates were calculated across all genotype quality scores for each non-reference strain relative to its respective reference genome as a measure of divergence (Figure 3.3.1). Both homozygous and heterozygous SNPs were included. The large number of bases attributed a genotype quality score of 99 (Figure 3.3.1b) offers some insight into the generally high quality of the SNP calling. Two-tailed Wilcoxon signed-rank tests, with alpha levels of 0.05, were used to compare the statistical significance of any differences in divergence between sets of strains across all genotype quality scores, split evenly into 20 bins.

The average divergence of all *E. moshkovskii* strains from the reference was greater than that demonstrated by *E. histolytica* strains when compared with HM-1:IMSS, a difference apparently independent of genotype quality (p-value < 0.01). The differences between *E. histolytica* and *E. moshkovskii* strains robustly demonstrate a greater evolutionary distance between *E. moshkovskii* Laredo and the species' other strains than exists between HM-1:IMSS and other *E. histolytica* strains. Within *E. moshkovskii*, the three non-reference strains' divergence from Laredo supported the figures seen in Table 3.3.1, suggesting that strain 15114 is the least divergent from Laredo (compared with Snake: p-value < 0.01; compared with FIC: p-value < 0.01). Strain FIC is significantly more divergent than both 15114 and Snake (compared with Snake: p-value < 0.01), supporting the early deductions made in Section 3.3.1.

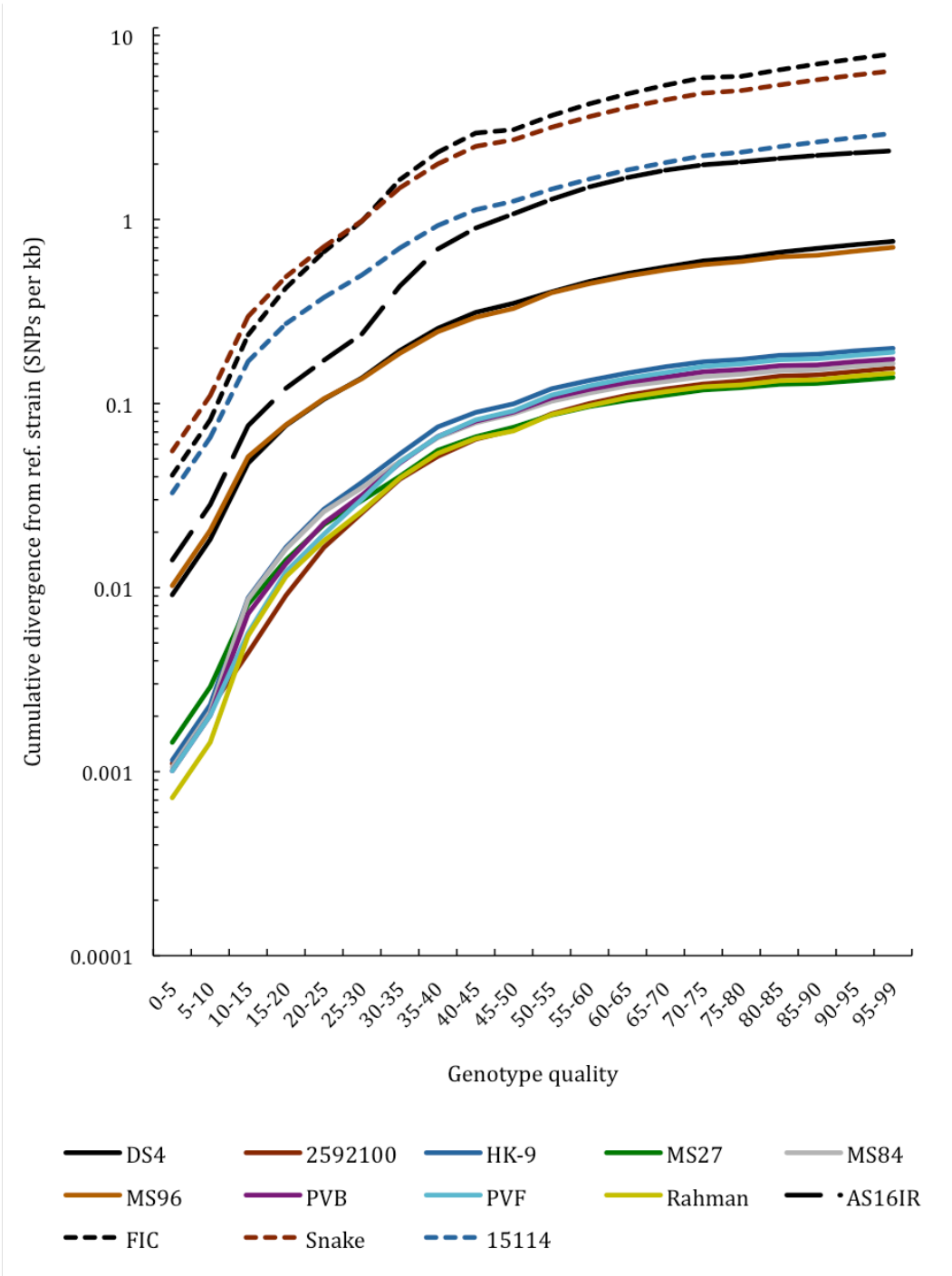


Figure 3.3.1a. Cumulative divergence of *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* strains, relative to their reference strains, as a function of genotype quality up to values of '99'. *E. histolytica* strains are denoted by solid lines, *E. dispar* AS16IR by a dashed line, and *E. moshkovskii* strains by dotted lines.

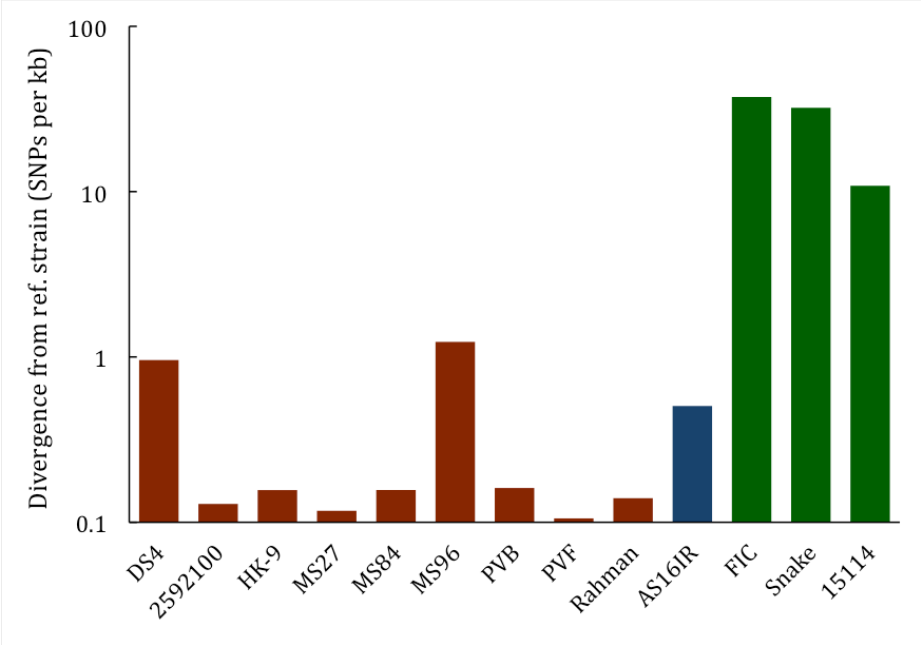


Figure 3.3.1b. Divergence of *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* strains, relative to their reference strains, as a function of genotype quality values of '99'. *E. histolytica* strains are denoted by red columns, *E. dispar* AS16IR by blue, and *E. moshkovskii* strains by green.

Within *E. histolytica*, strains DS4 and MS96 demonstrated consistently greater divergence from HM-1:IMSS than the other *E. histolytica* strains (p-value < 0.01), albeit still at low levels relative to those seen in the *E. moshkovskii* strains. Whilst it is possible that this is indicative of a higher proportion of SNPs in these two strains, it is more likely that this is a consequence of the higher quality mapping seen in Section 3.3.1. Reads for DS4 and MS96 mapped to a larger proportion of the reference genome and to a higher average depth than the other strains so SNPs may have been more confidently called and in loci to which the other strains did not map. Divergence in *E. dispar* strain AS16IR was found to be significantly more diverse than the average values seen in *E. histolytica* (p-value < 0.001). However, the limited number of *E. dispar* strains means no more than a tentative suggestion can be made as to the species' relative diversity. Future projects would benefit from the inclusion of a larger number of *E. dispar* strains in such a study so as to gain a clearer understanding of its diversity relative to other members of the genus.

At this stage, it is important to discuss the differences between the strains and species sequenced in this chapter. Firstly, as was mentioned in Section 3.1.2, the various strains were isolated in different years, ranging from 1948 to 2007. It is possible that the *E. histolytica* reference strain might have lost heterozygosity over the years in which it has been cultured by researchers, making it a less direct comparison with *E. dispar* and *E. moshkovskii*. The degree to which *in vitro* culturing of strains has resulted in reduction of heterozygosity cannot be known for the strains used here. Naturally, strains isolated longer ago that have been frequently cultured *in vitro* are likely to have undergone the greatest losses; however, it is unknown how long each strain has been subcultured for over the years. Furthermore, the longer they have been removed from the environment from which they were isolated, the greater the chances that there are differences between the isolates and their contemporary descendants. As such, one cannot consider the isolates used in this study to be directly comparable. Unfortunately, it was impractical to isolate, axenise and sequence enough strains to perform this comparison using new isolates. The use of the various *E. histolytica* strains in similar studies in recent years [68, 272] demonstrates that one can still acquire meaningful results from the comparisons, albeit with the small caveat that one cannot guarantee that the differences between the isolates would also be seen *in vivo* today.

3.3.3 Investigating genomic diversity within different sequence classes

To better understand in which regions of the genomes differences in diversity between *E. histolytica* and *E. moshkovskii* occur, SNP rates in a range of sequence classes were studied in more detail (Figure 3.3.2; Table 3.3.3). Both homozygous and heterozygous SNPs were included in this analysis. Mann-Whitney statistical tests were used to compare the average divergence between sequence classes between the species. An alpha level of 0.05 was used for all tests. Statistically significant differences in divergence were found between the *E. histolytica* and *E. moshkovskii* strains in all sequence classes (for 4D sites and intronic regions, p-value = 0.02; for all other classes, p-value < 0.01). This confirms that the greater divergence in *E. moshkovskii* suggested in Section 3.3.2 is ubiquitous across the genome.

Within *E. moshkovskii* and *E. histolytica*, occurrences of polymorphisms in coding regions were compared with those in a variety of non-coding regions. The significance of any differences between classes was tested using a two-tailed Welch's

paired t-test, with an alpha level of 0.05. There were no significant differences between the divergence seen in coding regions and those values recorded for the non-coding regions in *E. moshkovskii*. Conversely, coding regions of *E. histolytica* genomes were, overall, significantly more divergent than intronic regions (t = 15.0988, d.f = 7, p-value = 1.34 e-6) and 3' flanking regions (t = 2.5806, d.f = 7, p-value = 0.036). This suggests that polymorphisms occur at different rates in these regions of *E. histolytica*. This could not be proven convincingly in *E. moshkovskii*, possibly implying a greater importance of some non-coding sequences in *E. moshkovskii*.

Once thought to be non-functional 'junk', non-coding DNA is now known to contain a wealth of regulatory elements involved in the control of such important processes as DNA replication and gene expression [289-295]. As intergenic regions in *Entamoeba* genomes are very short it may be that they are very densely packed with regulatory regions. These findings do contradict a previous study that focused on individual genes and associated non-coding regions, which suggested that the latter were more divergent than coding regions due to their being under less selective pressure [268]. However, it is likely that the difference between the two conclusions is because the analyses featured here were performed across the entire genome, as opposed to selected regions, and so are based upon more data.

The 5'- and 3'-flanking regions of a sequence typically contain promoter and enhancer regions, to which transcription factors sometimes bind [296]. SNPs in 5'-flanking regions are known to affect regulation and expression levels [297-299]. It is interesting to hypothesise that the greater relative divergence seen in the *E. moshkovskii* 3' flanking regions is indicative of variable expression levels necessitated by a range of subtly different 'lifestyles'. The effects of promoter-based SNPs on stress resistance have previously been reported, so it is conceivable that SNPs in 5'- and 3'-flanking regions could facilitate, as an example, survival outside of a human host [300]. It was shown in Chapter Two that *E. moshkovskii* possessed a greater number of genes involved in cell signalling and communication than *E. histolytica*. Expression of such genes would likely be highly regulated, by non-coding elements. This might account for some of the difference between the two species. Such appealing conjecture could be supported by future transcriptional studies of *E. moshkovskii* strains and could, indeed, be applied to other *Entamoeba* species.

It is interesting to note that, in all sequence classes, *E. histolytica* strains DS4 and MS96 were more divergent than other *E. histolytica* strains, and that *E. dispar* AS16IR was even more divergent from its reference, though less so than the *E. moshkovskii* strains. The two *E. histolytica* strains' relatively high divergence was likely due to greater sequencing depth as discussed earlier. The higher SNP rate in the *E. dispar* strain, despite a poor average sequencing depth, is interesting as it suggests slightly higher diversity in *E. dispar* than *E. histolytica*. It is important to remember, however, that this observation is based upon a comparison of a limited number of strains and further work would be required to conclusively compare diversity in *E. histolytica* and *E. dispar*.

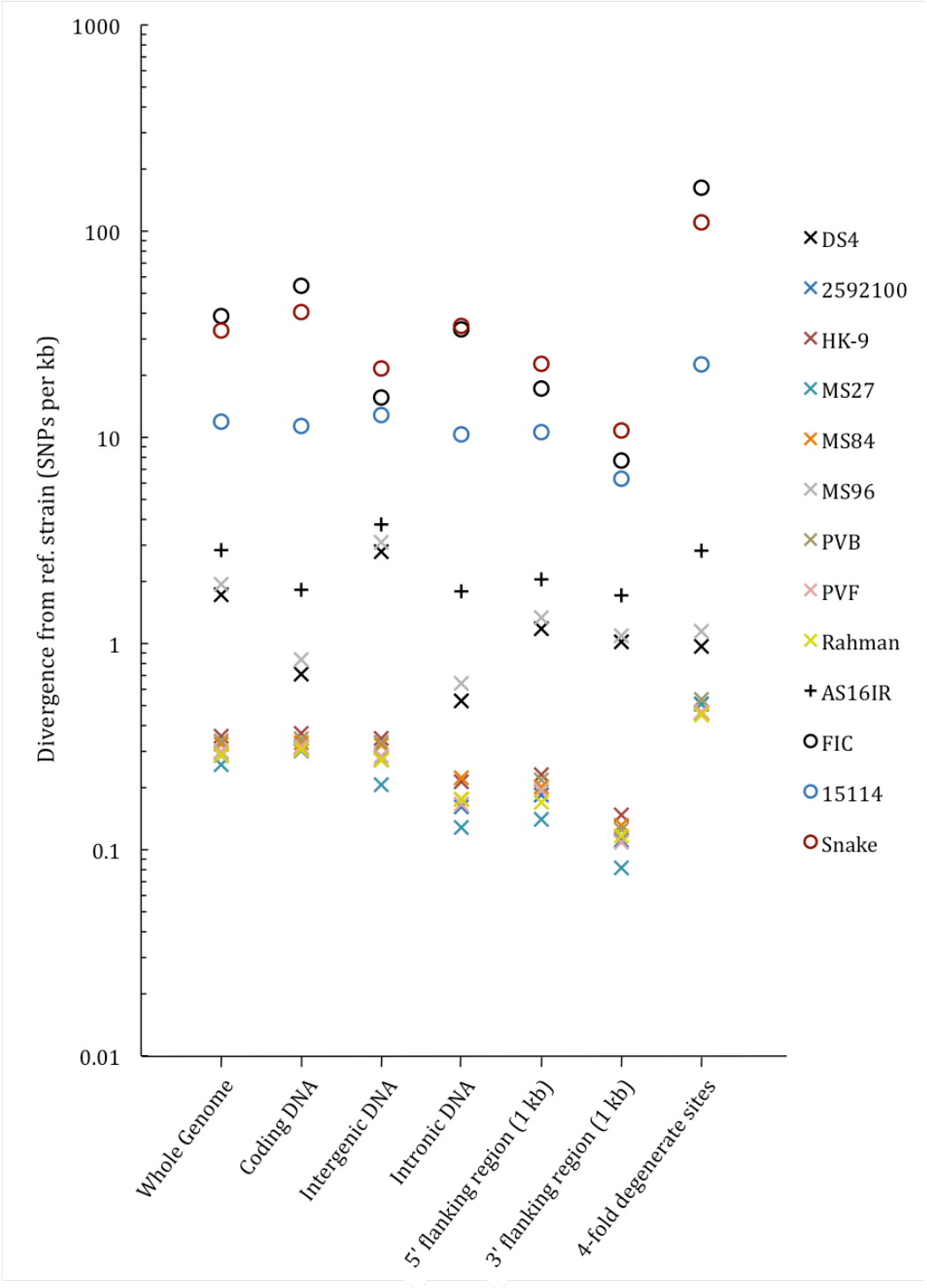


Figure 3.3.2. Divergence of *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* strains, relative to their reference strains, within different sequence classes. SNPs occurring in regions classified as both flanking regions and coding regions were considered to occur in coding regions only. Rates are relative to sites within their respective sequence classes.

Table 3.3.3. SNPs in *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* strains, relative to their respective reference genomes. Divergence values (SNPs/kb) calculated using both homozygous and heterozygous SNPs, and, in parentheses, just homozygous SNPs, are presented as lower and upper limits of potential divergence. Synonymous and non-synonymous counts refer to homozygous SNPs only.

Species	Strain	Hom	Het	Divergence (Hom only)	Synonymous (Hom)	Non-synonymous (Hom)
E. histolytica	DS4	19,964	15,907	1.72 (0.96)	1,673	2,770
E. histolytica	2592100	2,961	3,003	0.29 (0.14)	771	1,116
E. histolytica	HK-9	3,987	3,403	0.36 (0.19)	965	1,448
E. histolytica	MS27	2,909	2,468	0.26 (0.14)	833	1,136
E. histolytica	MS84	3,648	3,112	0.33 (0.18)	910	1,325
E. histolytica	MS96	23,993	16,335	1.94 (1.15)	2,047	3,594
E. histolytica	PVB	3,914	3,088	0.34 (0.19)	1,001	1,431
E. histolytica	PVF	3,426	2,800	0.30 (0.16)	896	1,305
E. histolytica	Rahman	3,532	2,448	0.29 (0.17)	917	1,329
E. dispar	AS16IR	18,720	46,471	2.84 (0.82)	4,309	4,084
E. moshkovskii	FIC	915,287	63,052	38.75 (36.25)	559,771	202,768
E. moshkovskii	15114	213,261	87,605	11.92 (8.45)	345,222	173,196
E. moshkovskii	Snake	707,613	124,789	32.97 (28.03)	65,233	55,120

As stated above, divergence across strains' 4D sites was greater in *E. moshkovskii* than in *E. histolytica* (Figure 3.3.2). Such sites have long been thought to be under neutral selective pressure, given that mutations in them do not affect the amino acid that their triplet encodes [301, 302]. As such, they provide an opportunity to evaluate the overall differences in diversity between species without the added complication of selective pressures influencing results. With this in mind, the 4D sites present in *E. histolytica, E. moshkovskii* and *E. dispar* were employed to approximately calculate the age of each species' strains' most common ancestor (Time to Most Recent Common Ancestor, or TMRCA) to further evaluate relatedness between strains of the species.

All homozygous 4D sites to which reads were mapped at a depth of 35x or greater in all strains of each species were identified and concatenated. This amounted to 339,091 bases in *E. histolytica*, 641,223 bases in *E. moshkovskii*, and 53,216 bases in *E. dispar.* The uncertain impact of heterozygous SNPs upon diversity meant that they could not be considered high quality calls, so could not be included from this point onwards. SNPs between each pair of strains within a species were counted and calculated as fractions of the total number of 4D sites covered to a sufficient depth. The pairwise SNP rates in *E. dispar, E. histolytica*, and *E. moshkovskii* were used to calculate TMRCAs, as well as to visualise, for the first time, the phylogenetic relationships between the strains of *E. moshkovskii* (Table 3.3.4; Figure 3.3.3; and Appendix B, Table B.2).

The accuracy of the calculation of a species' TMRCA is dependent upon the accuracy of the assumed mutation rate. Whilst the exact mutation rates of the *Entamoeba* species are unknown, the generic eukaryotic rate of 2.2 e-9 substitutions per base per annum was considered an acceptable approximation given its use in a similar previous study [303, 304]. Based upon this, the TMRCA for the two *E. dispar* strains is approximately 208,000 years. These are unlikely to be the two most distant *E. dispar* strains though, so it is probable that the TMRCA of the species will increase if more strains are included in future calculations. The TMRCA for *E. histolytica* is 165,000 years, whilst the TMRCA for the *E. moshkovskii* strains is approximately 81,590,000 years. As such, it can be concluded that *E. moshkovskii* first evolved approximately 494 times longer ago than *E. histolytica*. As mutations accrue at a particular rate over time, as stated above, *E. moshkovskii* would thus have accrued a

greater number of mutations than *E. histolytica*, explaining why greater variation is seen in *E. moshkovskii.*

The limitations of analysing evolution using 4D sites should, however, be discussed. Firstly, as stated earlier, the assumed mutation rate is not specific to the *Entamoeba* species. As such, it is likely to be a slightly inaccurate estimate. Compounding this is the potential for 4D sites to, contrary to long-held beliefs, be subjected to selective pressures that vary between species and even between genes, as has been shown in other species [305, 306]. Of particular interest was the discovery that 4D sites within certain genes in the *Drosophila melanogaster* genome are under strong purifying selective pressures [305]. However, until such research can be corroborated, particularly with respect to *Entamoeba* genomes, it was deemed reasonable to accept the traditional view of selective pressures on 4D sites. The results presented here provide evidence that the apparent higher diversity in *E. moshkovskii* relative to *E. histolytica* exists between every strain, and not just between each non-reference strain and Laredo.

Whilst calculating the TMRCA in *E. moshkovskii*, it was possible to assess the phylogenetic relationships between the four sequenced strains (Figure 3.3.3). This provided confirmation that strain FIC is the most distantly related to Laredo, whilst strain 15114 was the closest sequenced relative of the reference strain. A previous study focusing specifically on *E. histolytica* isolates of Asian origin suggested that there was a geographic link to the disparity between *E. histolytica* strains [307]. It would be interesting to compare more *E. moshkovskii* strains to ascertain whether this is the case in other *Entamoeba* species. The strains compared herein all have different countries of origin (Table 3.3.2).

Table 3.3.4. SNP rates in 4D synonymous sites common to four strains of *Entamoeba moshkovskii*. Rates are given as the number of SNPs per 4D site across the reference genome. The greatest SNP rate between a pair of strains is in bold.

	Laredo	**FIC**	**15114**
FIC	0.104839		
15114	0.014817	0.176561	
Snake	0.095302	**0.179496**	0.096913

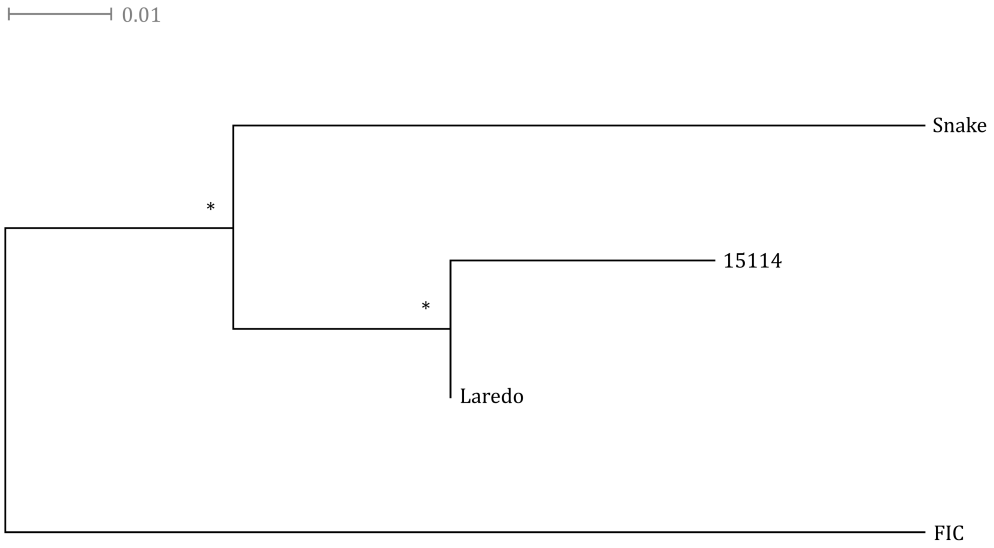


Figure 3.3.3. Phylogeny of *Entamoeba moshkovskii* strains based upon diversity in 4D synonymous sites. The tree was generated using a Neighbour-Joining method and is unrooted. All bootstrap values are 1,000 out of 1,000, indicated by asterisks at all branching points.

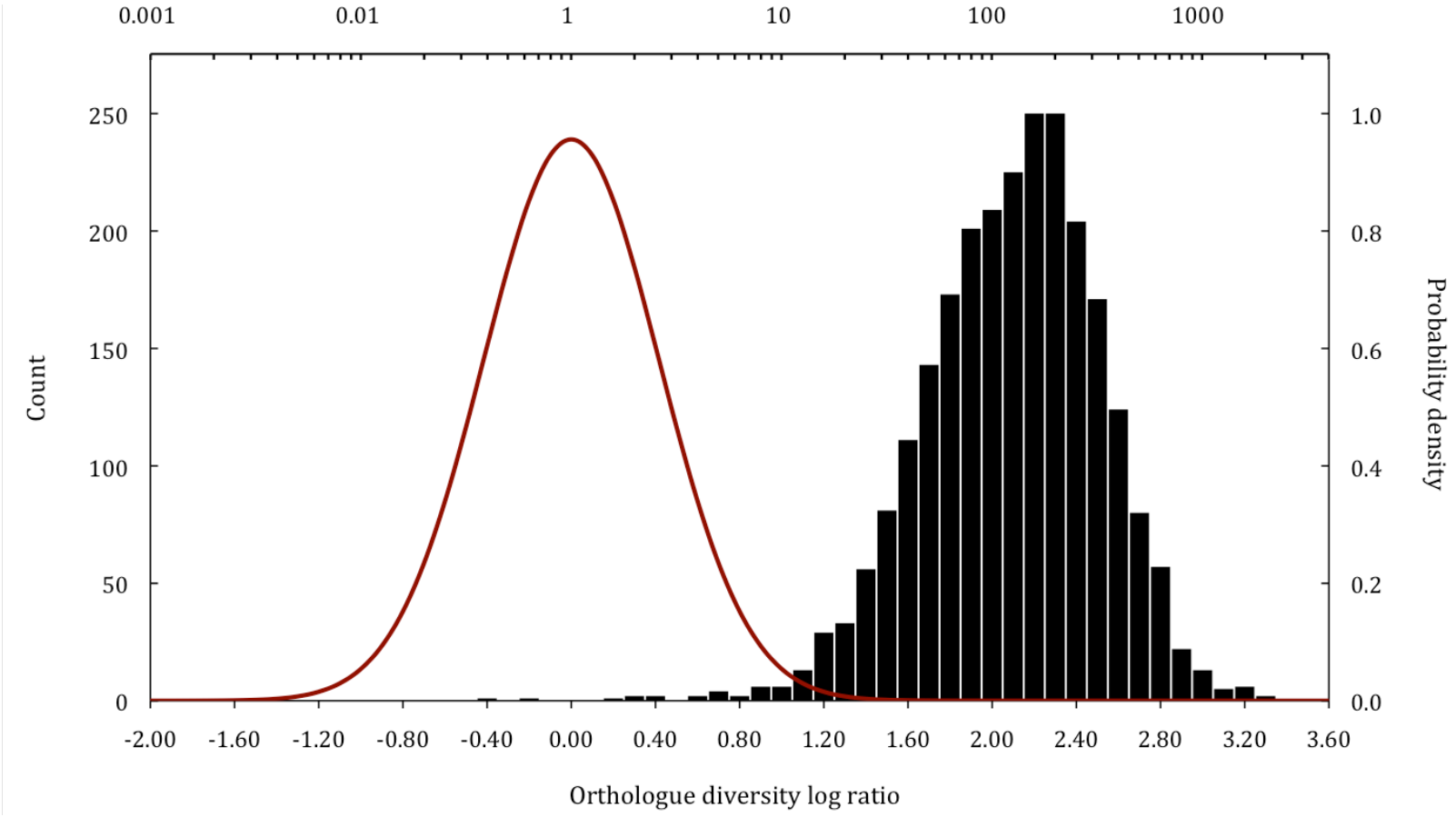


Figure 3.3.4. Probability-distributed log ratios of diversity in 2,485 *Entamoeba histolytica* and *Entamoeba moshkovskii* orthologue pairs. Ten *E. histolytica* and four *E. moshkovskii* strains were compared. Columns indicate observed counts of pairs at different diversity ratios. Red line represents normal distribution expected in the case of equal diversity of values within 3 standard deviations of the mean.

3.3.4 Comparisons of pairwise divergence between orthologues in *Entamoeba histolytica* and *Entamoeba moshkovskii*

Whilst differences in diversity in coding regions between *E. histolytica* and *E. moshkovskii* thus far appeared to be statistically significant, the different gene sets did not offer a direct comparison and, unlike the analysis of 4D synonymous sites, they did not take into account variation between the non-reference strains. As such, diversity (π) in 6,095 pairs of directly orthologous genes (as identified by a reciprocal BLASTP search between the two species' reference strains using default parameters) was measured in pairwise comparisons between all strains in each species. Statistical significance of differences between diversity values in these RBH pairs was tested using a two-tailed Welch's paired t-test. An alpha value of 0.05 was used. Again, to ensure high confidence SNP calls were used, only homozygous SNPs were included.

The comparison between the species was reduced due to the need to omit certain genes from the set of 6,095 RBH pairs. In only 2,485 pairs did both members contain SNPs, meaning that only 2,485 ratios of diversity values could be calculated. Of the pairs that were omitted, 2 had 0 SNPs in both species, whilst 3,606 had 0 SNPs in *E. histolytica* only, and 4 had 0 SNPs in *E. moshkovskii* only. Although these pairs could not be included in the analysis, these numbers do reinforce the following findings. The comparison between the 2,485 RBH pairs demonstrated that diversity in *E. moshkovskii* coding regions is approximately 200 times greater than in *E. histolytica* coding regions ($t = 213.0161$, $d.f = 2484$, $p < 0.01$; Figure 3.3.4). This suggests that the evolutionary distances between strains FIC, 15114 and Snake are, on average, greater than the mean distance between the three non-reference strains and Laredo. It also, importantly, offers more conclusive evidence of the differences in diversity between *E. histolytica* and *E. moshkovskii*

Of the 2,485 RBH pairs, only two contained an *E. histolytica* gene that was more diverse than its orthologue in *E. moshkovskii*. The members of one of these pairs encoded the translation initiation factor eIF-5A (EHI_151540, g12742), whilst the other pair consisted of orthologues containing a ricin B lectin domain (EHI_164470, g12207), found in carbohydrate recognition proteins [308-310]. Whilst the differences in divergence between the members of the pairs were negligible compared with the average difference between orthologues where diversity in *E. moshkovskii* is higher, it is still interesting to suggest that these proteins may play an important role in *E.*

histolytica. Certainly, it is logical that a protein involved in carbohydrate recognition and binding could be of importance in the lifestyle of the pathogenic *E. histolytica*.

3.3.5 Identification of most diverse genes and functions within each species

In the previous section, diversity was studied only in genes with direct orthologues in *E. histolytica* and *E. moshkovskii*, thus omitting the majority of genes in the two genomes. Here, all genes within all three species were studied, allowing for the identification of the most diverse sequences within each species, thus identifying genes that are theoretically specific and important to the lifestyle of the species in question. In order to identify these genes, diversity (π) values were calculated for all genes. In *E. histolytica*, SNPs were identified between at least one pair of strains in 40.57% of the total gene set. As might be expected, the proportion seen in *E. moshkovskii* was considerably higher at 99.03%. The proportion of genes found to vary between the two *E. dispar* strains was considerably lower than both of these values (32.30%) as a result of the limited number of strains used and the relatively limited mapping of AS16IR reads. However, it was still interesting to analyse which genes proved most divergent between these two strains as a preliminary analysis of diversity within the genome.

Those sequences demonstrating the highest levels of diversity in each species were identified, and functional or physical annotations were retrieved from AmoebaDB and UniProt, respectively (Figures 3.3.5, 3.3.6 and 3.3.7). Whilst not a direct comparison, it is interesting to note that the range of diversity values in the three species positively correlates with the number of genes containing SNPs. The highest diversity value seen in *E. moshkovskii* is considerably higher than the equivalent values in *E. histolytica* and *E. dispar*, which are, themselves, relatively similar. Meanwhile, the lowest diversity value in *E. histolytica* is 1.27 e-5 (EHI_114220) and the lowest value in *E. moshkovskii* is 9.70 e-5 (g12207).

Of the 20 most diverse coding sequences seen in *E. histolytica*, the functions of, or functional domains contained within, 12 of them are known (Figure 3.3.5). One of the most diverse of this subset is a sequence annotated as encoding a viral replication-associated protein (EHI_145460). The viral protein to which the *Entamoeba* gene appears to bear similarity is essential to the replication of geminivirus DNA [311] and shares great similarity with bacterial proteins involved in genome replication [312]. This protein has an orthologue in *E. dispar* but not in any other species, according to

AmoebaDB. It would be very interesting for further studies to investigate the role played by this protein in these two human-infective *Entamoeba* species.

Of particular interest amongst the high diversity *E. histolytica* sequences was the large number of AIG1-encoding genes. The AIG1 family may be involved in resistance to bacterial infections [206, 207], suggesting that the ability to resist infections and improve chances of survival in the host intestinal lumen is a key point in the life cycle of *E. histolytica*. Three genes that encode BspA proteins were also present in the 20 most diverse sequences in *E. histolytica*. It is possible that the diversity seen in them is exaggerated by sequencing errors of their leucine-rich repeat (LRR) region; however this observation supports the findings of Chapter Two that this vast protein family is of importance in establishing infections in *E. histolytica*.

As would be expected, the majority of genes found to be highly diverse in *E. moshkovskii* (Figure 3.3.6) were unannotated, either due to a lack of direct orthologues during the annotation of the Laredo genome or because the functions of any orthologues were also unknown. Of those to which functional annotations could be attributed, the majority were identified as housekeeping genes, that is, genes that play no obvious role in pathogenesis but which are expected to be vital to the survival of the niches into which the *Entamoeba* species have evolved to fit. It is helpful to distinguish between genes that are known to contribute towards pathogenesis and those required for maintaining cell functions and viability. This is because it separates those functions specific to the invasive species that can be targeted with a view to decreasing the development of disease states from those genes' functions that are, perhaps, ubiquitous or else could not be targeted.

The highly diverse Ras GTPase in *E. moshkovskii* belongs to a superfamily that, as stated in Chapter Two, performs a variety of cellular functions [218]. The functions of the three diverse enzymes in *E. moshkovskii* have not been well characterised, however it is possible to speculate as to the importance of endo-1,4-beta-xylanase. It is known to be involved in degradation of plant cell walls [313], potentially demonstrating the ability of the *Entamoeba* species to exploit the diet of its host. The only annotated sequence encoding a potential virulence factor is the gene exhibiting the greatest amount of diversity, which encodes a light Gal/GalNAc lectin subunit. Whilst it does not directly interact with host cells or proteins, evidence has been found that light Gal/GalNAc lectin subunits play a role in virulence, potentially being involved

in the gathering of the lectin complexes [314-316]. Unfortunately, it cannot be known at this point whether the protein encoded by this particular gene is involved in such a process and whether it would be of particular importance in the life cycle of *E. moshkovskii*. Future studies would be encouraged to look at this in greater depth. As most of the genes identified as highly diverse in *E. moshkovskii* have unknown functions it is tempting to speculate that this is because they are specific to *E. moshkovskii* and, therefore, unstudied. It is quite possible that these genes are involved in host interaction and, therefore, they are also promising targets for future study.

The majority of the most diverse sequences in *E. dispar*, according to this study, are unannotated, much like in *E. moshkovskii* (Figure 3.3.7). Of the seven that are annotated, three lack annotations but contain LRR regions. Again, whilst it is possible that the repeat regions caused sequencing errors, making the genes appear to be more diverse than they actually are, the regions do appear to play a structural role, allowing protein-protein interactions [317, 318]. However, sequences containing LRRs are numerous and perform a variety of functions, meaning it is not possible to know what roles these genes may play. That being said, given that *E. dispar* is not known to possess any members of the LRR-containing BspA family other than one pseudogene, it is unlikely that these proteins are involved in cell-cell adhesion as they are in *E. histolytica*. Whilst no certain conclusions can be drawn without more complete functional annotations of the *E. dispar* gene set, it would appear that survival of *E. dispar* is less dependent upon an ability to adhere to host cells. This would suggest, as was seen in Chapter Two, that adherence to the intestinal wall of hosts is a crucial difference in the life cycles of *E. dispar* and *E. histolytica*.

The other annotated sequences found to be most diverse in *E dispar* play housekeeping roles, as was seen in the majority of annotated diverse *E. moshkovskii* sequences. Again, a member of the Ras superfamily is seen, however the other annotated genes perform functions not seen in the equivalent *E. histolytica* and *E. moshkovskii* gene sets. Whilst it must be remembered that only two *E. dispar* strains have been used to identify the most variable sequences in the species, these differences between the three species support the suggestion that different genes, beyond virulence factors, are important to their survival. As is the case with the other two species, it is likely that improved functional annotation of the *E. dispar* genome and inclusion of a greater number of strains would reveal more about the genes most important to the survival of this human-infective species.

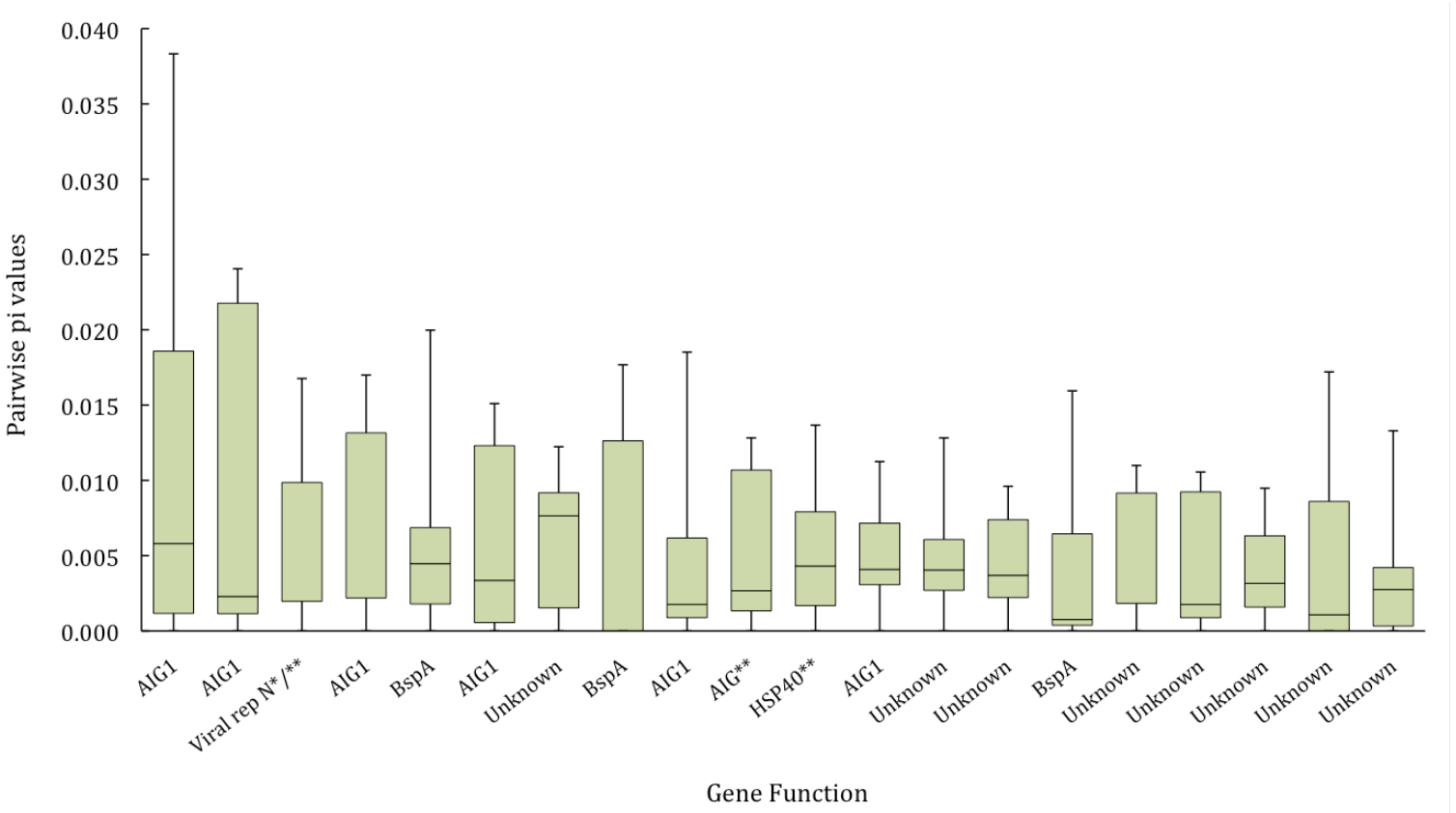


Figure 3.3.5. *Entamoeba histolytica* genes with the highest mean pairwise diversity (pi) [left-to right]. Functions derived from AmoebaDB unless denoted by: * = function derived from orthologues on AmoebaDB; ** = function derived from UniProt. 'Viral rep N' = Viral replication-associated protein, N terminal.

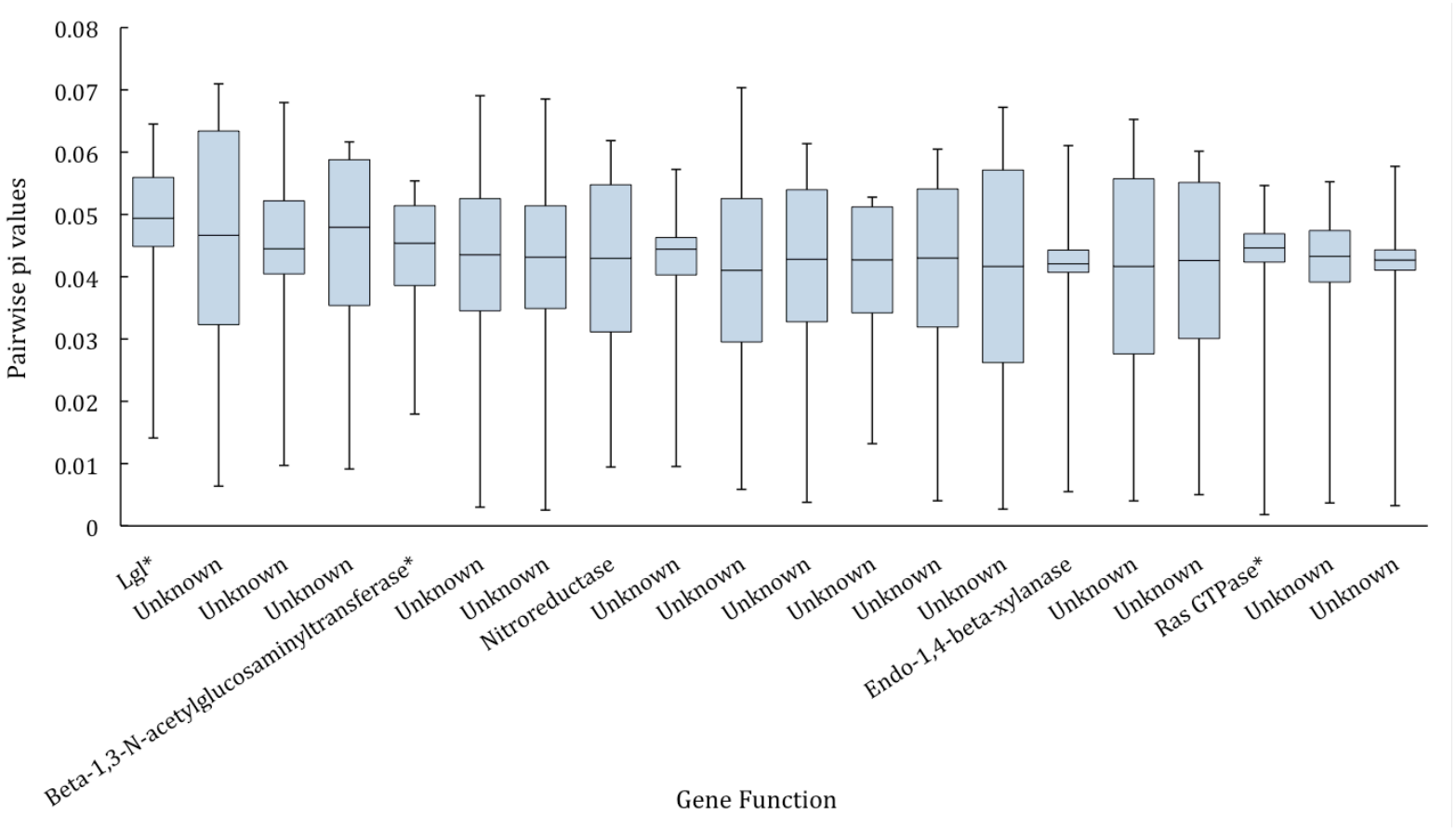


Figure 3.3.6. *Entamoeba moshkovskii* genes with the highest mean pairwise diversity (pi) [left-to right]. Functions derived from AmoebaDB unless denoted by: * = function derived from orthologues on AmoebaDB; ** = function derived from UniProt. Lgl = Light Gal/GalNAc lectin subunit.

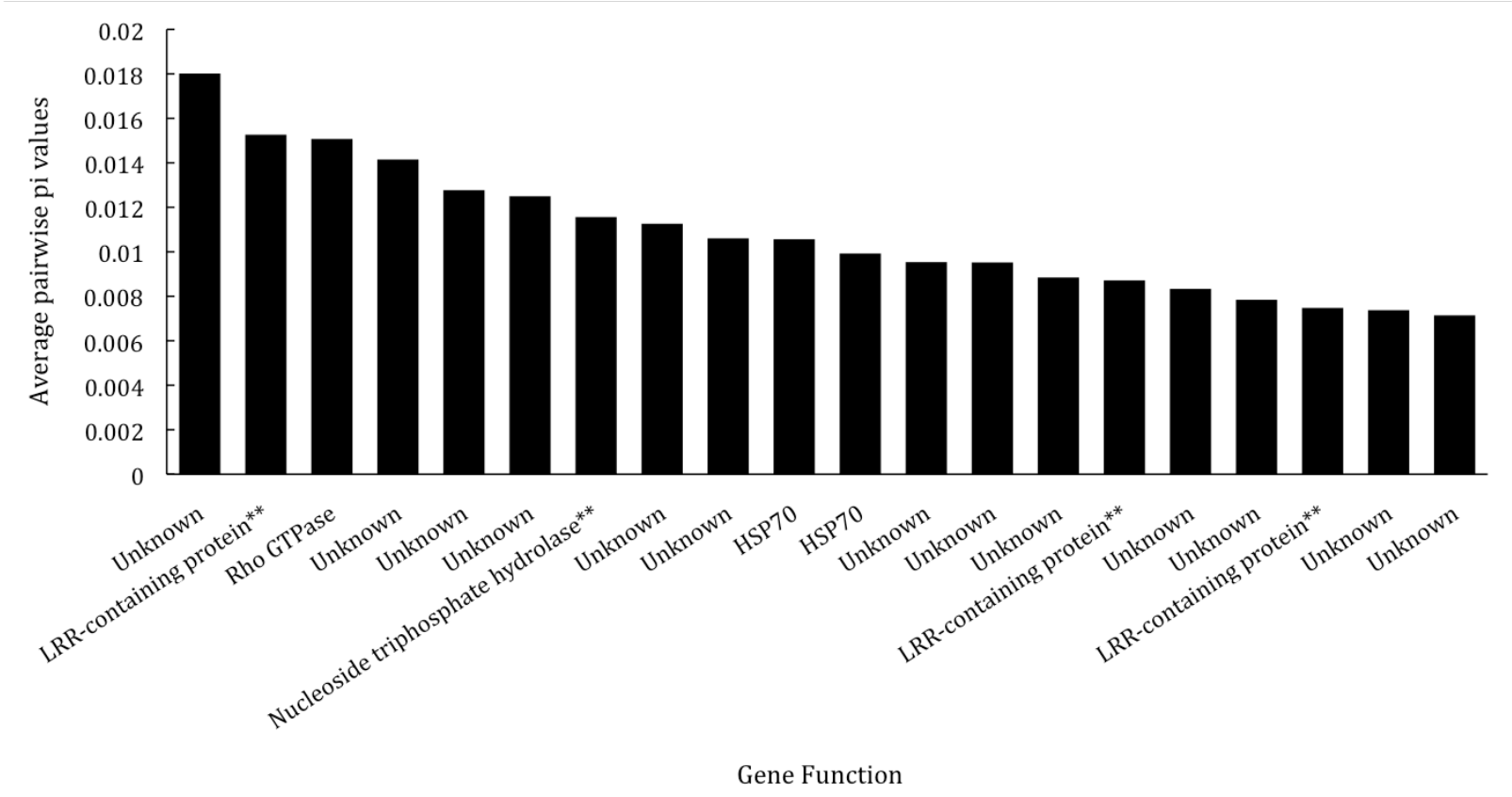


Figure 3.3.7. *Entamoeba dispar* genes with the highest mean pairwise diversity (pi) [left-to right]. Functions derived from AmoebaDB unless denoted by: * = function derived from orthologues on AmoebaDB; ** = function derived from UniProt. LRR = Leucine-rich repeat domain.

3.3.6 Identification of genes and functions under diversifying selective pressure within each species using dN/dS ratios

The above analyses based upon SNP counts revealed less than was hoped about diversity and the functions of most importance in *E. histolytica*, *E. moshkovskii* and *E. dispar*. As such, a more detailed analysis of polymorphisms was carried out to identify the types of selective pressures active upon genes. Ratios of dN to dS values (dN/dS) were calculated for each coding sequence in each strain of *E. histolytica*, *E. moshkovskii* and *E. dispar*, using homologous SNPs only. The ratios were calculated relative to each strain's respective reference genome. Attempting a pairwise calculation would have required one to accept the assumption that all SNPs were called in each strain, which was almost certainly not the case, and that any base not called as a SNP was definitely the same as in the reference strain. Furthermore, heterozygous SNPs were once more omitted, as calculation of the impact they have upon a sequence's dN/dS ratio would have been impractical as they may cause individual positions to be considered as being the site of both synonymous and non-synonymous changes.

In *E. moshkovskii*, the majority of genes identified as being under diversifying selective pressure lacked annotations or known domains (Table 3.3.5). However, a number of sequences, the functions of which are known, were identified, including a relatively large number of BspA family members. It is interesting that members of the BspA family appear to be important to the survival of all three *E. moshkovskii* strains, whilst *E. dispar* possesses no complete BspA coding sequences. Whilst there are numerous *Entamoeba* proteins involved in cell adherence, including the ariel1 surface antigen seen to be under diversifying selection in strain 15114, it may be that *E. moshkovskii*, like *E. histolytica,* utilises BspA proteins to adhere to host cells. In this case, *E. moshkovskii* could conceivably be a more probable cause of symptomatic amoebiasis than *E. dispar.* It would be of great interest to study how crucial the BspA family is in allowing adherence to host cells and promoting invasive amoebiasis.

In addition to the BspA family proteins, all three *E. moshkovskii* strains were found to possess genes with similar housekeeping functions in the form of protein kinases, DNA repair proteins and Ras family GTPases. No further putative virulence factors were seen to have dN/dS ratios above 1 and to, therefore, be of particular importance in survival. These data suggest that the strains of *E. moshkovskii* studied

here are human-infective, though they are inconclusive with regards to the strains' pathogenicity.

Table 3.3.5. Functions under diversifying selective pressure in *Entamoeba moshkovskii* strains, according to dN/dS ratios, relative to the reference strain Laredo. Numbers represent number of genes to which each annotation was attributed within each strain. Annotations were taken from orthologous sequences where none was available for a sequence itself.

Function	Strain		
	FIC	Snake	15114
BspA family	20	33	27
Serine/threonine/tyrosine? protein kinase	5	5	7
Surface antigen ariel1	-	-	2
Serine protease inhibitor/leukocyte elastase inhibitor	1	1	2
DNA double-strand break repair Rad50 ATPase	6	7	8
Heat shock protein 70	-	-	2
Cullin family protein	-	1	-
Ras family GTPase	1	2	2
PQ loop repeat protein	-	1	-
DEAD/DEAH box helicase	-	2	-
Nucleoside diphosphate kinase	1	-	1
Caldesmon	-	-	1
CAAX amino terminal protease family protein	-	-	1
Methyltransferase trm13 protein*	-	-	1
AAA family ATPase	-	-	1
Actin	-	2	2
Transitional endoplasmic reticulum ATPase	-	1	-
Mucin-2	-	-	1
Myosin heavy chain	-	-	1
Proline synthetase associated protein	-	-	1
Acetyltransferase, GNAT family	-	-	1
Hypothetical protein	19	35	70

Table 3.3.6. Functions under diversifying selection in *Entamoeba histolytica* strains and the *Entamoeba dispar* strain AS16IR, according to dN/dS ratios, relative to their respective reference strain. Numbers represent number of genes to which each annotation was attributed within each strain. Annotations were taken from orthologous sequences where none was available for a sequence itself.

Strain	Function	Count	Gene IDs
Entamoeba histolytica			
DS4	Ser/Thr protein kinase	1	EHI_059040
	Hypothetical protein	3	EHI_020930, EHI_013900 EHI_025850
2592100	Hypothetical protein	1	EHI_180400
MS27	Hypothetical protein	1	EHI_013900
MS84	AIG1 family protein	1	EHI_176580
	Hypothetical protein	1	EHI_174560
MS96	BspA	3	EHI_190890, EHI_049160 EHI_123820
	DNA polymerase	1	EHI_132860
	Ser/Thr protein kinase	1	EHI_059040
	Regulator of nonsense transcripts	1	EHI_148970
	DEAD/DEAH box helicase	1	EHI_119920
	AIG1 family protein	1	EHI_115150
	Hypothetical protein	6	EHI_101400, EHI_160470 EHI_083760, EHI_109010 EHI_013900, EHI_154760
PVB	Hypothetical protein	2	EHI_032470, EHI_111770
PVF	Hypothetical protein	2	EHI_032470, EHI_111770
Rahman	Hypothetical protein	1	EHI_180400
Entamoeba dispar			
AS16IR	Ser/Thr protein kinase	3	EDI_334450, EDI_188560, EDI_239440
	Ubiquitin protein ligase	1	EDI_305290
	Protein phosphatase 2C	1	EDI_247520
	Hypothetical protein	8	EDI_353340, EDI_203280 EDI_113920, EDI_195230 EDI_082470, EDI_348320 EDI_313320, EDI_185210

dN/dS ratios indicating diversifying selective pressures acting upon genes were present in eight of the nine *E. histolytica* strains, with only HK-9 appearing to lack sequences under such pressures (Table 3.3.6). However, the numbers recorded in each strain were, compared with counts in *E. moshkovskii*, very low, with only MS96 featuring more than five such diversified genes. Of those genes identified as being under diversifying selection, one can see that, as with the comparison of pi values, the majority are unannotated, but there appear again BspA and AIG1 family proteins. This reinforces the theory that these families play roles that are important to the survival of *E. histolytica*. The number of genes shown to be under diversifying selection in *E. dispar* strain AS16IR is similar to that seen in *E. histolytica* MS96, although only three functions were identified. Members of the protein phosphatase 2C family are known to be involved in regulation of signalling pathways in eukaryotes [319] so it is possible that they play a similar role in *E. dispar*. This enzyme, as well as the protein ligase and protein kinase found to be under diversifying selection in *E. dispar*, and those housekeeping genes under similar selective pressures in the other two species, would need to be studied in greater detail to determine their precise functions and roles in helping the species survive their environmental niches. The relatively low numbers of genes under diversifying selection in *E. histolytica* and *E. dispar* are likely the result of a combination of factors. Firstly, every *E. histolytica* strain excluding DS4 and MS96 were, as was discussed earlier, sequenced to a relatively low depth. As such, fewer SNPs were likely to have been detected, thus profoundly affecting the calculation of dN/dS ratios. Secondly, dN/dS ratios can only be accurately calculated where a sequence contains both synonymous and non-synonymous SNPs. It was likely that many genes containing only non-synonymous SNPs, but which would still certainly be classed as being under diversifying selection, would have been omitted. This would, of course, have also affected the dN/dS ratios in *E. moshkovskii*. In light of this, an alternative method of identifying genes under diversifying selection was devised. Coding sequences containing equal dN and dS values must have a dN/dS ratio of '1'. As such, any sequence containing a greater proportion of non-synonymous SNPs per non-synonymous site than synonymous SNPs per synonymous site must have a value greater than '1'. Similarly, when the difference between a sequence's dN and dS values (dN-dS) is calculated, equal values must give a value of '0', meaning that any gene with a difference value greater than '0' can be considered to be under diversifying selection. This is a slightly crude method but it is far more inclusive, allowing for identification of sequences only possessing one group of SNPs (Figures 3.3.8 and 3.3.9).

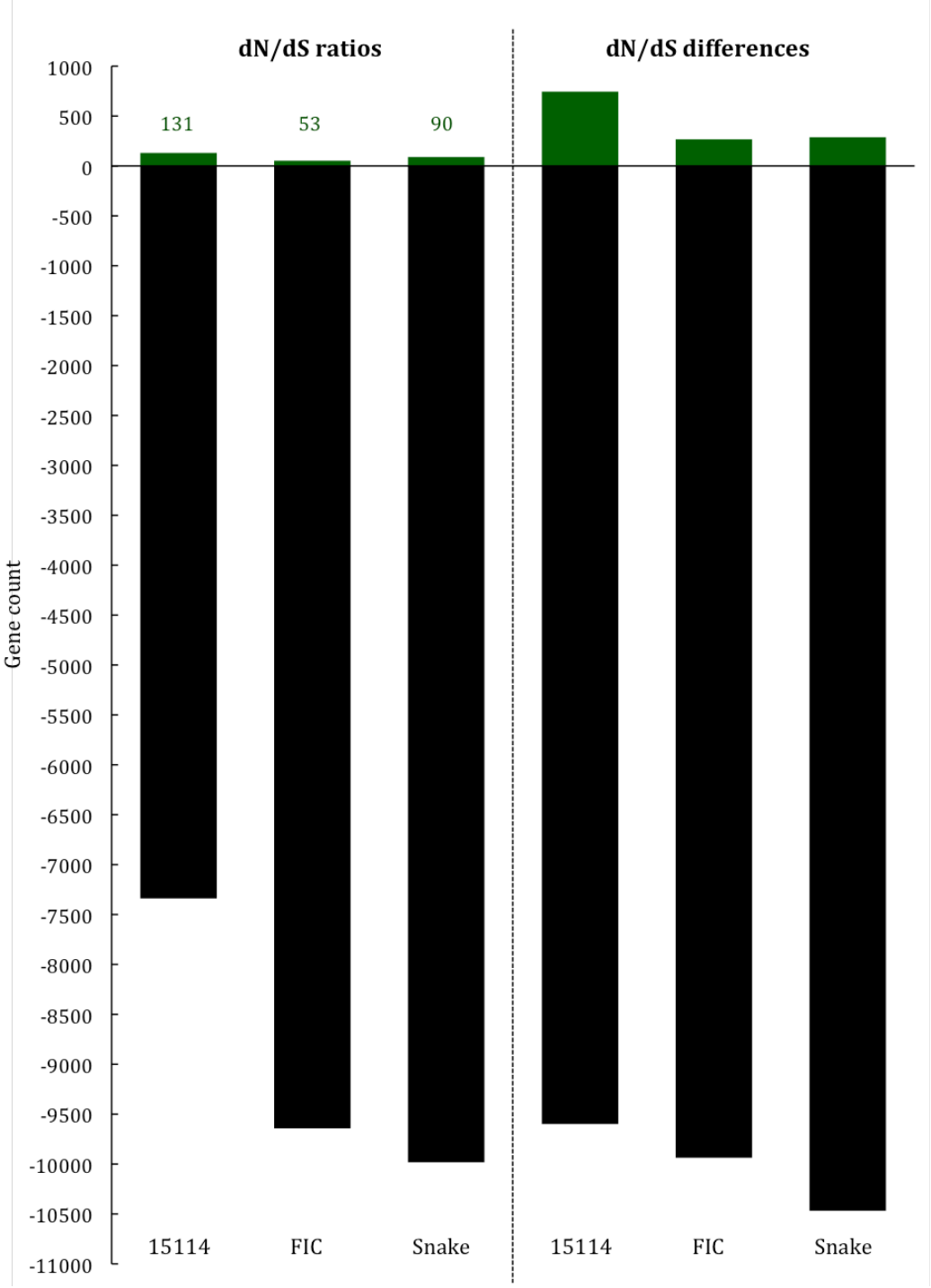


Figure 3.3.8. Counts of genes under diversifying (green bars) or purifying (black bars) selective pressures in *Entamoeba moshkovskii* strains. Genes with dN/dS ratios >1 are under diversifying selective pressures. Those with ratios < 1 are under purifying selection. In differences between dN and dS, the cutoff value is 0. Numbers in green show values of columns that are difficult to see.

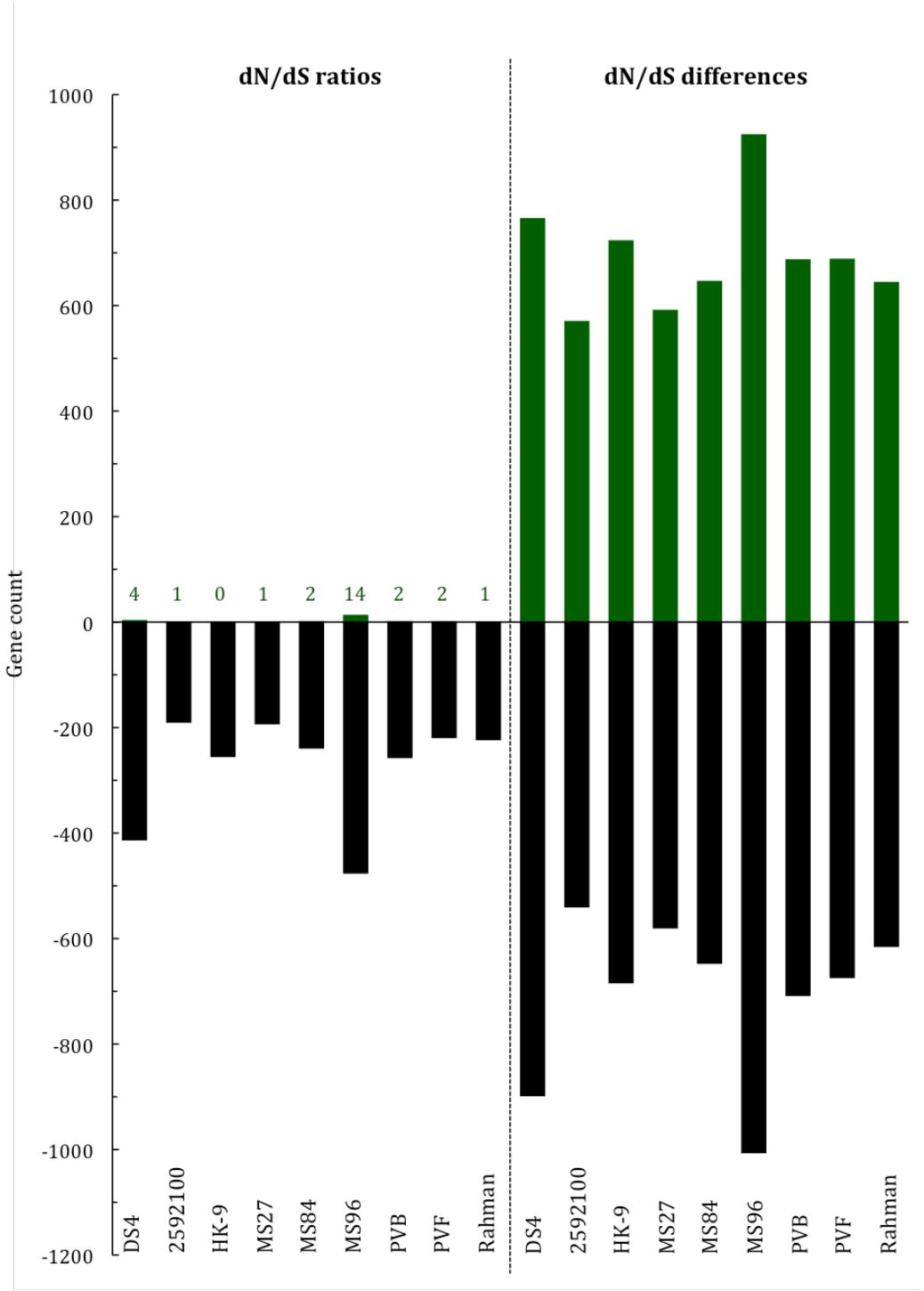


Figure 3.3.9. Counts of genes under diversifying (green bars) or purifying (black bars) selective pressures in *Entamoeba histolytica* strains. Genes with dN/dS ratios >1 are under diversifying selective pressures. Those with ratios < 1 are under purifying selection. In differences between dN and dS, the cutoff value is 0. Numbers in green show values of columns that are difficult to see.

3.3.7 Identification of genes and functions under diversifying selection in *Entamoeba dispar* and *Entamoeba histolytica* using dN-dS differences

Using a positive dN-dS difference value as an indicator of genes under diversifying selection resulted in considerably greater numbers of genes being identified in the *E. histolytica* strains (Figure 3.3.9). In total, 1,998 genes were identified as being under diversifying selective pressure in one or more strains, of which 44.24% were functionally annotated. The annotated sequences encoded a large variety of proteins, predominantly housekeeping genes. However, in each strain were sequences that encoded members of several of the virulence factor families identified in Chapter Two (Figure 3.3.10). These included relatively large numbers of BspA and AIG1 family proteins, as might be expected given the previous results in this chapter. Lesser numbers of other virulence factor family members, including the cysteine proteases and Gal/GalNAc lectin subunits, were also present. It is important to remember here, however, that numbers of members alone are not enough to determine the relative importance of a family to a species. For example, six cysteine proteases are seen to be under diversifying selective pressures in varying numbers of *E. histolytica* strains, yet the gene known to encode the pathogenically important cysteine protease EhCP-A5 (EHI_168240) is not amongst them [74, 75]. As such, whilst one can suggest that these virulence factors play important roles in survival of *E. histolytica*, it is not possible to conclude that they are important in pathogenesis; they may simply have housekeeping roles.

In order to reduce the data set to a practicable size, only those genes with dN-dS difference values above an arbitrary value of 0.005 were compared (Table 3.3.7). This offered a cutoff point and a way of identifying the individual genes thought to be of most importance to the survival of the species, rather than being grouped by function as was necessary in previous sections. As can be seen, the majority of sequences are hypothetical and, of all of the virulence factors identified in the full set, only a subset of the BspA, AIG1 and ariel1 families were present in the reduced list. This suggests that the virulence factors of *E. histolytica* are not under such great diversifying pressures as it first appeared.

Reducing the gene set to only include those with dN-dS values greater than 0.005 was also necessary in *E. dispar* AS16IR. Of the 844 sequences under diversifying selective pressure according to their dN-dS difference values, 268 were functionally

annotated (Appendix D, File D.3) However, in the reduced set of 30 sequences, only a DnaK chaperone protein and a multi-drug resistance protein were annotated, rendering the results rather uninformative until further work has been done to improve the *E. dispar* gene set annotation.

It is worth noting that the only sequence theoretically under diversifying selective pressure in all *E. histolytica* strains is a hypothetical protein the majority of the sequence of which consists of repeat units. Whilst the gene is relatively highly expressed (FPKM value of 98) and so appears to serve a function, it is interesting to note that dN and dS values are heavily influenced by the presence of STRs in coding sequences. This sequence may not, therefore, be of great importance to these eight strains of *E. histolytica*. This caveat must be applied to all dN-dS and dN/dS results. In light of this, and the limitations of dN/dS ratios described earlier, I would suggest that future studies would benefit from a test combining dN-dS difference values and analyses of sites within STRs, which would provide a more inclusive data set whilst compensating for bias introduced by STRs.

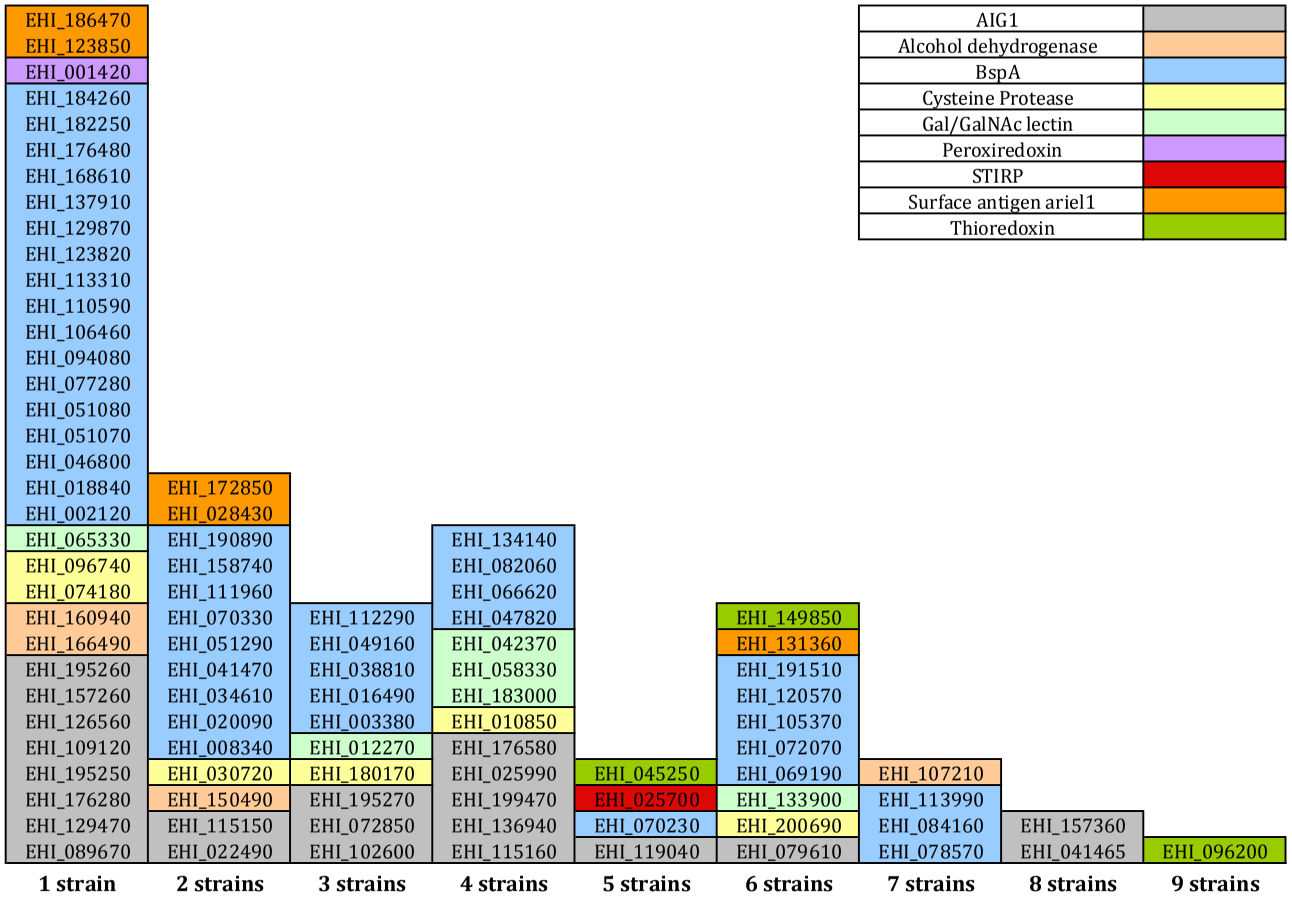


Figure 3.3.10. Graphical representation of the gene IDs encoding virulence factors under diversifying selective pressure in the 9 non-reference strains of *Entamoeba histolytica*. Columns for each group represent the numbers of species in each group in which each gene ID was present.

Table 3.3.7. The prevalence of *Entamoeba histolytica* sequences with dN-dS values above 0.005 in at least 1 strain. Single asterisks (*) indicate functions derived from orthologous sequences listed on AmoebaDB. Double asterisks (**) indicate functions or domains described in UniProt.

Gene ID	Function	No of strains
EHI_013900	Hypothetical protein Highly repetitive	9
EHI_152820	ATP-binding cassette transporter*	5
EHI_191400	Hypothetical protein	4
EHI_132380	Zinc finger RING domain**	4
EHI_079610	AIG1 family protein	4
EHI_072520	Hypothetical protein	4
EHI_176580	AIG1 family protein	3
EHI_004060	Hypothetical protein	3
EHI_190890	BspA family protein	2
EHI_182790	Nucleotide-binding, alpha-beta plait**	2
EHI_170960	Hypothetical protein	2
EHI_156430	NUF1 protein	2
EHI_142160	Hypothetical protein	2
EHI_120570	BspA family protein	2
EHI_084160	BspA family protein	2
EHI_071340	Hypothetical protein	2
EHI_025850	Midasin**	2
EHI_025310	Hypothetical protein	2
EHI_020860	Hypothetical protein	2
EHI_002220	Hypothetical protein	2
EHI_201420	ERE1 repetitive element*/TLDc domain**	1
EHI_196710	Hypothetical protein	1
EHI_193790	Biotin-acetyl-CoA-carboxylase ligase*/**	1
EHI_186870	Hypothetical protein	1
EHI_186470	Surface antigen ariel1	1
EHI_184470	60S ribosomal protein L14	1
EHI_175720	Hypothetical protein	1
EHI_175710	Orthologues with mixed functions*	1
EHI_175700	Orthologues with mixed functions*	1

Gene ID	Function	No of strains
EHI_172850	Surface antigen ariel1	1
EHI_172730	Orthologues with mixed functions*	1
EHI_157360	AIG1 family protein	1
EHI_155330	Prefoldin, alpha subunit	1
EHI_154760	Hypothetical protein	1
EHI_154590	Leucine-rich repeat-containing protein	1
EHI_136940	AIG1 family protein	1
EHI_136840	Hypothetical protein	1
EHI_132250	Orthologues with mixed functions*	1
EHI_130540	Hypothetical protein	1
EHI_126560	AIG1 family protein	1
EHI_123820	BspA family protein	1
EHI_111590	Hypothetical protein	1
EHI_109120	AIG1 family protein	1
EHI_106460	BspA family protein	1
EHI_102490	Hypothetical protein	1
EHI_096690	Hypothetical protein	1
EHI_092340	Sec61 protein	1
EHI_083760	AIG1 family protein**	1
EHI_072850	AIG1 family protein	1
EHI_062960	Hypothetical protein	1
EHI_058330	Gal/GalNAc lectin subunit	1
EHI_049160	BspA family protein	1
EHI_029510	Hypothetical protein	1
EHI_029350	40S ribosomal protein S8*/**	1
EHI_022490	AIG1 family protein	1
EHI_021550	Hypothetical protein	1
EHI_020950	Hypothetical protein	1
EHI_012920	Hypothetical protein	1
EHI_004950	Hypothetical protein	1
EHI_004930	Hypothetical protein	1

3.3.8 Identification of genes and functions under diversifying selection in *Entamoeba moshkovskii* using dN-dS differences

In the case of *E. moshkovskii*, studying every sequence with a dN-dS difference value greater than 0 revealed little because, once again, the majority of sequences lacked functional annotations (Figures 3.3.8 and 3.3.11) and manual attainment of orthologues' functions was impractical. Complete lists of sequences under diversifying selective pressure in each strain are available in Appendix D, File D.4. As was the case in Section 3.3.6, a large difference was observed in the number of sequences and functions demonstrating diversifying selection in the three non-reference strains (Figure 3.3.11). Strain FIC possessed 267 genes with positive dN-dS values, 263 of which were unannotated, whilst strain Snake possessed a similar 282 unannotated genes, alongside 4 annotated sequences. The strain most closely related to Laredo – strain 15114 - possessed a much larger number of genes with dN-dS > 0 (745) than its relatives, though only 80 were functionally annotated. However, this is not indicative of greater pressures acting upon 15114's gene sequences. The strain possesses a comparable number of SNPS to FIC and Snake, however they reside within a greater number of coding sequences (Figure 3.3.12). As such, strains FIC and Snake may possess fewer sequences with dN-dS values < 0 when compared to each other, but those they do possess may be under greater diversifying pressures than those in strain 15114, depending upon proportions of synonymous and non-synonymous SNPS.

The four annotated sequences under diversifying selective pressure in strain FIC encoded a BspA family protein, a prefoldin subunit, a 40S ribosomal protein and a sequence shared by strain Snake that contained the common zinc finger domain. Those also seen in Snake encoded an alcohol dehydrogenase, a 60S ribosomal protein, a small nuclear ribonucleoprotein, a mannose-1-phosphate guanylytransferase and a sequence containing a PQ loop repeat. The BspA protein and alcohol dehydrogenase sequences are of potential interest for further studies, as are three putative virulence factors found to be under selective pressure solely in strain 15114, one encoding an alcohol dehydrogenase, one a peroxiredoxin and the other a thioredoxin. Whilst all are regarded as virulence factors, they do also play roles not associated with virulence [209, 223, 229]. It is interesting to speculate that thioredoxin might be of importance to this strain in surviving environmental oxidative stresses. It is expected that the information and gene IDs acquired here will be more informative once more annotations exist for *E. moshkovskii*.

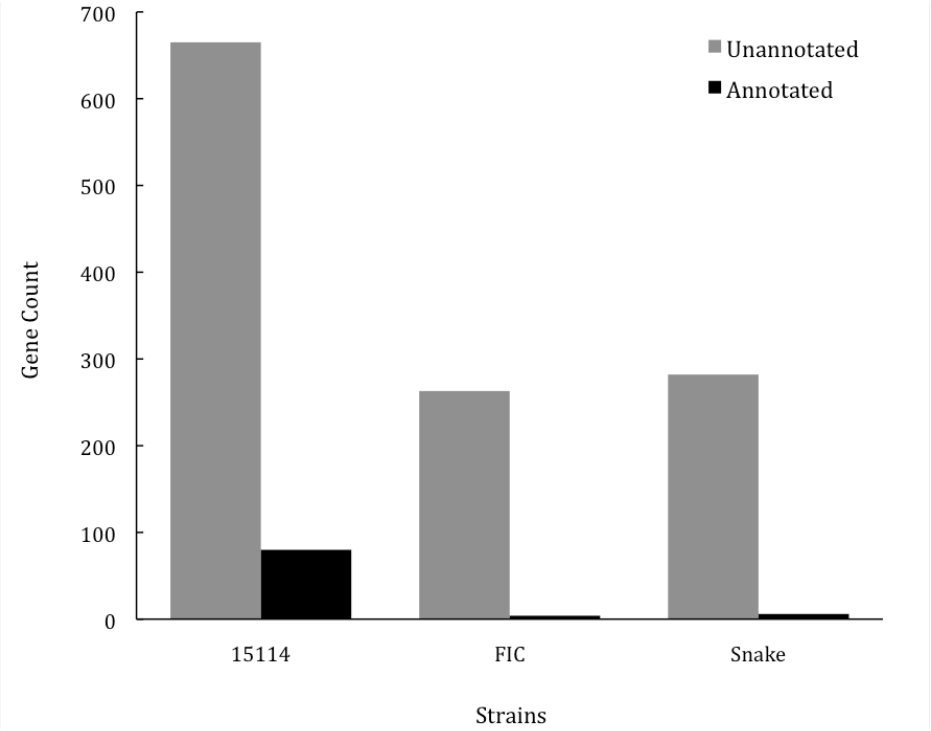


Figure 3.3.11. Counts of sequences under diversifying selection in *Entamoeba moshkovskii* strains with or without functional annotations

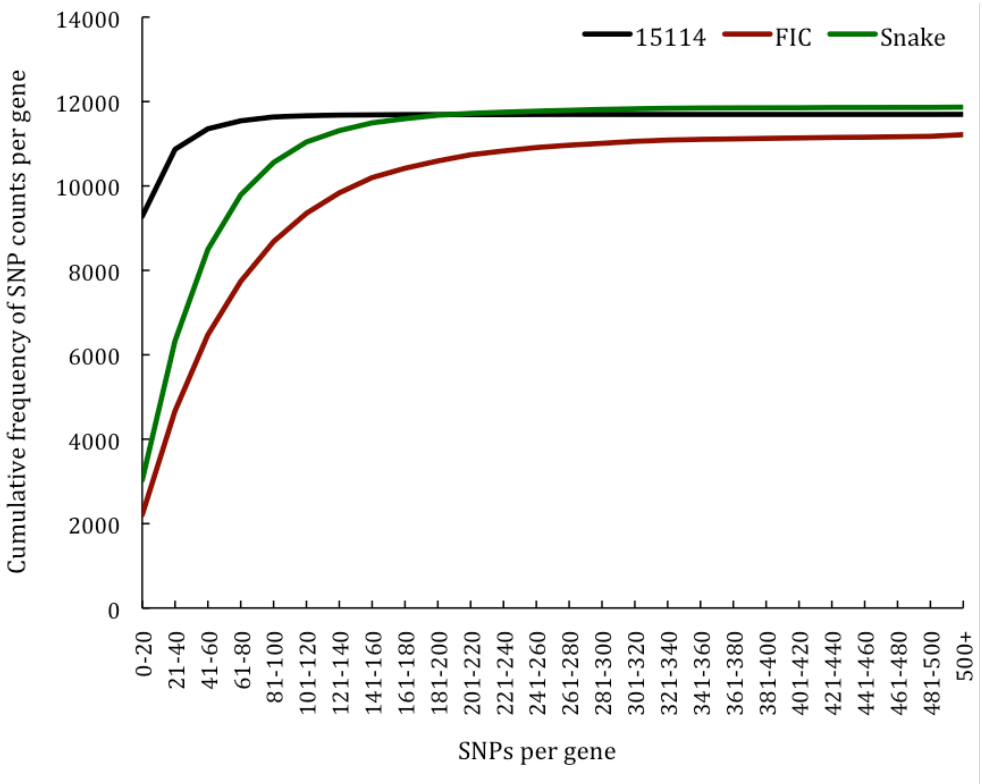


Figure 3.3.12. Cumulative frequencies of SNP counts per gene in the three *Entamoeba moshkovskii* strains

3.3.9 Testing for evidence of meiotic recombination in *Entamoeba moshkovskii*

During the course of this chapter, it has been evident that genomic sequence variation between *E. moshkovskii* strains is significantly greater than that seen in *E. histolytica*. I wished to determine whether this was due to *E. moshkovskii* strains simply having a more distant most recent common ancestor than *E. histolytica* strains or whether it is because *E. moshkovskii* actually consists of multiple individual species that have been grouped together erroneously, i.e. a species complex. The four-haplotype test was used to check for evidence of meiotic recombination between the four *E. moshkovskii* strains. According to the infinite sites model of evolution, individual nucleotide positions can only mutate once, meaning that the maximum possible number of haplotypes between two physically linked sites is three, unless recombination between genomes is possible. Furthermore, recombination is more likely to occur between sites the greater the distance between them. As such, the occurrence of four haplotypes within a species, combined with a greater prevalence of such haplotypes over greater genomic distances act as reliable indicators of meiotic recombination. Evidence of meiotic recombination has previously been reported in *E. histolytica*, demonstrating that it can occur in members of the *Entamoeba* genus [68].

A Spearman's correlation coefficient was applied to test whether there was any significant correlation between the proportions of physically linked SNP pairs that exist as four haplotypes and the distance between members of those pairs (Figure 3.3.13). There was no significant correlation, meaning that four-haplotype SNP pairs are not more prevalent over greater distances. This suggests that recombination has not occurred in the evolutionary history of the strains designated as *E. moshkovskii* and that gene conversion may, in fact, be the cause of the four-haplotype SNP pairs that are present. The lack of evidence of recombination strongly suggests that not all *E. moshkovskii* strains studied here belong to the same species. Further investigation will need to be carried out to devise a method of differentiating between the members of what can confidently now be referred to as a species complex, rather than an individual species.

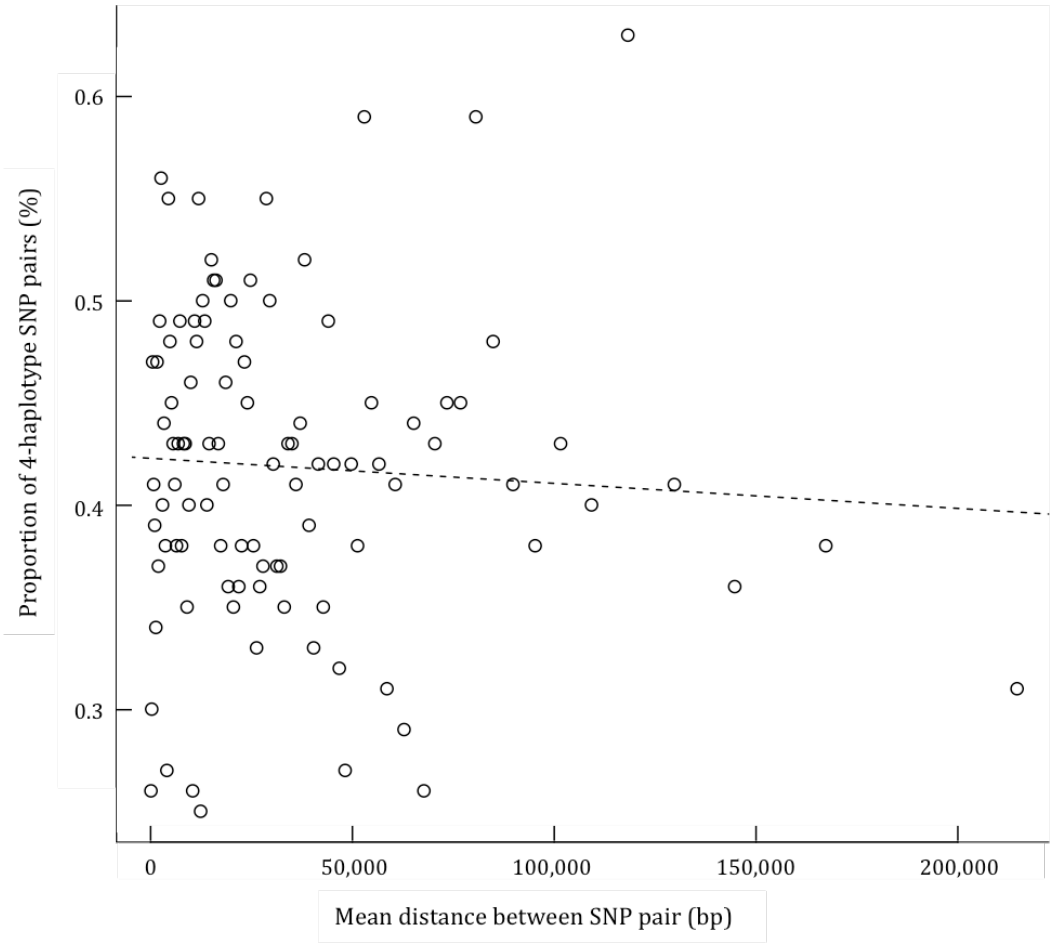


Figure 3.3.13. The proportion of 4-haplotype SNP pairs in *Entamoeba moshkovskii* as a function of the distance between the pairs. SNP pairs were physically linked (i.e. they were on the same scaffold/contig). Each point represents 1,000 randomly selected SNP pairs. The line represents the correlation between distance between pairs and proportions of 4-haplotype SNP pairs. The correlation was statistically insignificant.

3.4 Concluding remarks

In this chapter, I have sequenced for the first time the genomes of three strains of *E. moshkovskii* and one strain of *E. dispar.* I have mapped the reads generated for these strains, along with reads of *E. histolytica* strains present in the public domain, to their respective reference genomes and compared SNP counts and rates between species and strains as a measure of diversity. Overall, the genomes of *E. moshkovskii* strains were found to be more diverse than those of *E. histolytica* strains, with the former species approximately 200 times as diverse as the latter. This greater diversity was found to be the case across multiple sequence classes, demonstrating that it is not restricted to individual regions of the genome. Furthermore, *E. moshkovskii* was found to have diverged from its strains' most recent common ancestor nearly 500 times longer ago than *E. histolytica*'s strains did from theirs. It is likely, therefore, that the reason for the greater diversity within *E. moshkovskii* is that its genome has accrued mutations over a longer period of time than that of *E*. *histolytica*. These results suggest that genetic diversity is very low in *E. histolytica* and this may be one reason why resistance to commonly used drugs such as metronidazole has not been observed. However the greater genomic diversity of *E. moshkovskii* suggests that, if it is a potential emerging pathogen, then it may be more likely to accrue resistance.

I have also identified those functions believed to be under the greatest diversifying selective pressures in *E. histolytica*, *E. moshkovskii* and *E. dispar*. Sequences under such pressures are thought to be of importance in survival of host-parasite interactions. Whilst results were inconclusive for *E. moshkovskii* and *E. dispar* due to lack of functional annotations and lack of sequencing data, respectively, there did appear to be a clear difference between those two species and *E. histolytica*. The latter possesses numerous putative virulence factors, which were not under such selective pressures in *E. moshkovskii* and *E. dispar*. Of particular interest are members of the BspA family and members of Gal/GalNAc lectin complex, which play roles in adherence to host cells, amongst others. Whilst it must be stressed that further work must be done to ascertain the roles of the individual genes identified, taken together these findings appear to confirm that adhesion to host cells is one of the most important steps in *E. histolytica* establishing infections, echoing the results of Chapter Two.

Finally, the four-haplotype test was applied to demonstrate that *E. moshkovskii* should not be considered an individual species, as has been suspected for some time. Recombination does not occur between all of the 'strains' of *E. moshkovskii* featured in this chapter. As such, they can no longer be thought of as belonging to the same species and a new classification for the members of this species complex will be required in the future. The fact that these *E. moshkovskii* strains are probably not the same species is important for understanding the recent reports that it has been associated with human infection. It may be that only one of these sequence types can be infective and, therefore, to understand the epidemiology of this emerging disease we need to develop better diagnostics that can differentiate between the different sequence types. Also, if there are pathogenic and non-pathogenic types of *E. moshkovskii*, they could act as a useful system for studying the emergence of pathogenicity.

Chapter Four – Comparison of *de novo* assemblers and assembly methodology using *Entamoeba bangladeshi* genome

4.1 Introduction

4.1.1 *Entamoeba bangladeshi*

Entamoeba bangladeshi is the most recent species in the *Entamoeba* genus to be discovered. It was isolated in the Mirpur slum region of Dhaka, Bangladesh as part of a long-term study into the epidemiology of amoebiasis [25]. It was identified when PCR tests of faecal samples known to contain *Entamoeba* cells returned negative results for the known human-infective species *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* [26]. Morphologically identical to *E. histolytica* [26] and, by association, *E. dispar* and *E. moshkovskii* [2, 37], *E. bangladeshi* is suggested to be a close relative of these human-infective species. Certainly, it can be cultured at both 25°C and 37°C, much like *E. moshkovskii*, suggesting physiological similarities between the two [26]. The epidemiological implications of this discovery will now be of great interest to the amoebiasis research community. The global prevalence of *E. bangladeshi* is currently unknown, as is its virulence status.

Entamoeba organisms isolated from hosts are unable to be cultured *in vitro* in sterile conditions [158]. They are, instead, cultured with a mixture of bacterial species in a xenic culture. Over time, certain strains and species can adapt to survive in the absence of these additional species [158, 320]. However, sometimes strains simply cannot survive axenically *in vitro*. The process of adapting to sterile conditions (axenisation) is also time-intensive and, therefore, impractical for many studies. The bacterial contaminants pose problems for genomic assemblies, particularly of novel species such as *E. bangladeshi*. There are no axenic cultures of *E. bangladeshi* and, therefore, new methods are required to undertake genomic analyses of this species. This chapter investigates the effects multiple methods can have upon genome assembly quality in xenic cultures of *Entamoeba*, using *E. bangladeshi* as an illustration.

4.1.2 Challenges of *de novo* genome assemblers

Since gaining the ability to generate shorter reads, groups have been competing to generate *de novo* assemblers that can construct scaffolds using them. Included are

the programs employed in this chapter - ABySS [144], Velvet [141], SOAPdenovo [143] and Ray [142]. Whilst all four of these assemblers offer different assembly methods, they all use de Bruijn graphs based upon the Euler algorithm [321, 322]. Many comparisons have been made between the vast number of assemblers available, however the overall conclusion appears to be that the properties and quality of the genome itself are more important factors in determining which assembler performs best than the assemblers themselves [111, 112, 323].

There is, however, an ongoing debate as to which parameter should be used to determine the quality of an assembly. Despite the common use of the N50 statistic (the smallest scaffold length of a set of the largest scaffolds in an assembly the combined length of which contains at least 50% of the assembly [324]), it is not perfect and it is generally accepted that, as one parameter is improved, others might be worsened [112, 125]. The N50 benefits those assemblers that will assemble reads into scaffolds in spite of errors, making longer scaffolds but not necessarily better assemblies [61]. As such, it is not suitable for analysis of every assembly, and nor is any other individual parameter. No comparisons have been made for *Entamoeba* and so it is currently impossible to say which assembler is best suited to constructing an *Entamoeba* genome *de novo* and finding genes, or indeed which parameter should be used to determine this.

4.1.3 Aims of chapter

In this chapter I have sequenced the genome of the recently discovered species *E. bangladeshi*, strain 8237. I have used the generated libraries of reads to compare the efficacy of four genome assembler programs in performing a *de novo* assembly of an *Entamoeba* genome extracted from a xenic culture. Furthermore, as a major aspect of much genome annotation is gene discovery, the outputs of these assemblers were compared with the results of an approach that involved assembling individual CDSs from the *E. bangladeshi* genome to identify the most effective method of gene discovery for the species. The outputs of these various methods have been combined to produce a collaborative summary of the core *Entamoeba* gene sequences known to be present in the *E. bangladeshi* genome. The work described here offers a first glimpse of the genome of a human parasite, and potential pathogen, the discovery of which may have important consequences for our understanding of the epidemiology of amoebiasis.

4.2 Materials and Methods

4.2.1 Acquisition and sequencing of DNA

Lyophilised DNA extracted from a xenic culture of *E. bangladeshi* strain 8237 was kindly provided by Dr Rashidul Haque (ICDDR). The protozoon had been cultured with an undefined mixture of bacterial species. The DNA was rehydrated in ultra-pure water and a 150 bp PE library was prepared for the solution, as described in Section 3.2.3. In total, 8,689,302 pairs of reads were generated.

4.2.2 *Entamoeba bangladeshi* read identification and isolation

To distinguish *E. bangladeshi* reads in the libraries from the bacterial reads, an adapted version of the publicly available 'Blobology' (aka. Taxon-Annotated-GC-Coverage (TAGC) plot) protocol was utilised [325]. A k-mer size of 91 was used. This was proportionately identical to the k-mer size of 61 used in the Blobology paper, in which PE reads of 101 bp were used. A k-mer size of 91 permitted the sensitivity to distinguish between reads from different species without being restrictively specific regarding the positioning and assembly of those reads.

Briefly, the Blobology script (https://github.com/blaxterlab/blobology) used the abyss-pe program of ABySS v1.3.4 [144] to assemble submitted PE reads, before entering all resulting scaffolds >=200 bp into a BLASTN search (BLAST v2.2.29) against the NCBI nt database (May 2013 update). The output was limited to one hit per query sequence with an E-value threshold of 1e-5. Taxonomic annotations of hits were attributed to their respective scaffolds. The script then used Bowtie 2 v2.1.0 [326] to map the initial reads to the generated scaffolds to calculate coverage values for each scaffold, before combining the generated data with calculations of GC content (%) for each scaffold. This penultimate step was edited so that the output included the taxonomic Order of each scaffold, if known. Prior to the final step being run, where species or genera were known but the Order of a scaffold was not annotated, the output file was manually edited to include the respective Order. The final step of the script was then run to generate the TAGC plot. The script was run iteratively, with the number of reads entered into the script being further reduced each time using various criteria, discussed in full in Section 4.3. Within each reduced read set, only those reads

whose paired read also mapped to a scaffold were included in the next round of the assembly.

4.2.3 Assembler comparisons

Multiple assembly approaches were compared to identify the optimal method when performing a *de novo* assembly for members of the *Entamoeba* genus with limited genomic data. Firstly, four PE *de novo* assemblies, each run by a different program, were performed for the 1,950,947 pairs of reads found to be either unannotated at the taxonomic Order level, or to be annotated as Amoebida. In all cases, a k-mer size of 91 was used. The four programs used were ABySS v1.3.4 [144], SOAPdenovo (SOAP) v2.04 [143], Velvet v1.2.02 [141] and Ray v2.3.1 [142]. ABySS's 'abyss-pe' and Ray were run using default parameters. When running Velvet, 'velveth' was run using default settings. At the 'velvetg' stage, an insert length (average length of paired reads plus intervening distance) of 325, with a standard deviation of 66, was specified. This was calculated using a Python script created, and made available on GitHub's Gist site, by Mr Wei Li (https://gist.github.com/davidliwei/2323462). When running SOAP, a maximal read length of 151 bp was specified with no 3' trimming requested. An average insert size of 325 bp was used. The minimal values for the map length and pair number cutoff parameters were applied.

An alternative approach was tested whereby the *E. histolytica* HM-1:IMSS gene set, acquired from AmoebaDB v2.0 [169, 170], was entered into TBLASTN searches against the unfiltered set of forward reads, as well as the unfiltered reverse reads (i.e. including both *Entamoeba* and bacterial reads). TBLASTN was run using default settings, save for limiting the E-value to 1e-5. Every read hit by a gene was entered into a *de novo* assembly run for that gene. The individual gene assemblies were performed using Velvet v1.2.02 [141], using default settings and a k-mer value of 51 (one-third of the common read length). Reads were assembled as singlets.

4.2.4 Comparisons of assemblies

To determine the quality of the assemblies produced, three methods were used. Firstly, statistics for the assemblies were generated using a Perl script made publicly available by Ian Korf's group at the Genome Centre, UC Davis (http://korflab.ucdavis.edu/datasets/Assemblathon/Assemblathon2/Basic_metrics/

assemblathon_stats.pl). Secondly, scaffolds were searched using Ian Korf's group's CEGMA v2.4 program [114, 167] to identify complete and partial CEG models. Finally, core *Entamoeba*-exclusive protein sequences described in Chapter Two, and all protein sequences for *E. histolytica* strain HM-1:IMSS downloaded from AmoebaDB v2.0 were entered as query terms in a TBLASTN search against each assembly's scaffolds, using default terms except for defining the E-value threshold as 1e-5. CEG IDs, core *Entamoeba* cluster IDs and *E. histolytica* genes present in assemblies are listed in Appendix E, File E.1.

4.2.5 Phylogenetic analyses

To identify the phylogenetic relationship of *E. bangladeshi* with other members of the *Entamoeba* genus, a conserved group of orthologous genes encoding the 60S acidic ribosomal protein P2 were studied. The gene IDs of the *E. histolytica*, *E. dispar*, *E. invadens* and *E. moshkovskii* sequences are EHI_186830, EDI_310180, EIN_035540 and EMO_087250, respectively. MUSCLE v3.8.31 [179] was used, with default parameters, to align the sequences. A bootstrapped phylogram, built using the additive tree model, was generated using PHYLIP v3.69 [180]. Default parameters were used unless otherwise stated. Seqboot was run with 1,000 bootstrap replicates. DNAdist was then run using the F84 distance matrix, set to receive 1,000 datasets. The coefficient of variation (1.3608) was calculated using the shape parameter generated by MEGA v5.2.1, using default parameters [181, 182]. Fitch estimated phylogenies with the Fitch-Margoliash criterion for the 1,000 randomised data sets before Consense output bootstrapped trees. To apply branch lengths that represent evolutionary distances to the trees, the first two PHYLIP programs described above were run again, using the same parameters, but for 1 dataset rather than 1,000. Bootstrapped trees were input to Fitch with their respective single data set trees, applying branch lengths to the relationships.

4.3 Results and Discussion

4.3.1 Identification and isolation of *Entamoeba bangladeshi* reads

In this chapter, I aimed to compare different methods of gene detection in the novel species *E. bangladeshi*, strain 8237. The goal of this was to provide advice for future studies involving newly discovered members of the genus. *E. bangladeshi* DNA was extracted from a xenic culture and so the extract contained a mixture of amoebic and bacterial DNA. As there exists no reference genome for *E. bangladeshi*, mapping of the reads to an existing sequence was not possible. Instead, a *de novo* approach was required. Most assemblers are not well suited to metagenomic samples generated using next generation sequencers, as the short reads do not allow great enough distinction between the various species within such populations [327, 328]. As such, an alternative method was tested whereby I attempted to distinguish between *Entamoeba* reads and non-*Entamoeba* reads prior to a final assembly of the former.

Multiple groups have used such *in silico* methods to distinguish between reads from different species within a sample and to improve assembly of contigs from individual species. To date, sources of such samples have varied from contaminants within a DNA sample [329] to symbionts or parasites within a host [330, 331] to bacteria associated with plants [332]. Parameters used to determine how sequenced contigs should be clustered vary between methodologies, though, generally, contigs are entered into BLAST searches against a non-redundant protein database of existing genomes. Contigs may also be separated based upon their GC content and read depth [329, 331]. Such an approach was utilised in this chapter.

To achieve this, the publicly available Blobplot protocol was used, which assembles reads and allows the user to distinguish between scaffolds by taxonomic annotations, GC content and coverage depth [325]. Here, annotations at the taxonomic Order level were applied. Initially, all reads were entered into the protocol. The relationship between coverage and GC content of each resulting scaffold at least 200 bp in length is shown in Figure 4.3.1. As can be seen, there were a relatively large number of scaffolds (4,852 scaffolds; 3,497,519 bp total) belonging to the *E. bangladeshi* genome, annotated using the Order Amoebida. Coverage depths in these scaffolds ranged from 2 - 2,054x, whilst GC contents existed between 13.94 and 56.25%. The majority of these scaffolds, however, existed within a smaller range of each statistic,

forming the large collection of dark blue points in the TAGC plot. Coverage was, in the majority of cases, similar between scaffolds from all genomes, suggesting largely uniform sequencing. It is likely that those with exceptionally high coverage depths contained repeat regions or very common motifs, causing mapping of reads that belonged in different regions of a genome to be collapsed and assembled together [60]. Whilst there are some scaffolds annotated as belonging to bacterial Orders that have similar GC contents to the Amoebida scaffolds, many of the bacterial scaffolds have higher GC contents than the majority of those from the *E. bangladeshi* genome. This is to be expected given that *Entamoeba* species typically have low GC contents (Table 2.3.1), whilst the GC contents of bacterial genera tend to be higher, yet can vary drastically, with extreme known values at 16.5 and 75.0% [333, 334].

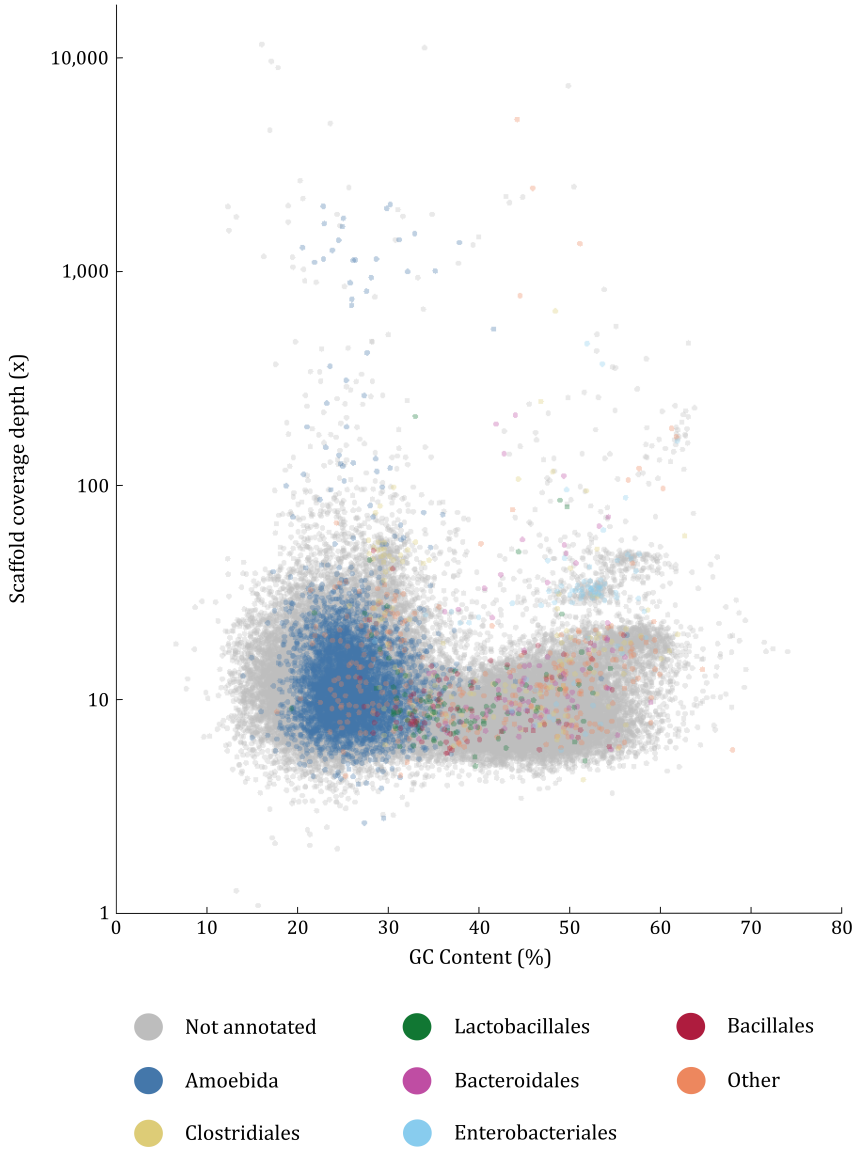


Figure 4.3.1. Scaffolds assembled by ABySS, using all reads generated for a xenic *Entamoeba bangladeshi* culture, plotted as a function of their GC contents and average coverage depths. Taxonomic Orders are listed in order of decreasing prevalence from top to bottom, then left to right, across the key. If >1% of scaffolds shared the same Order annotation, all those scaffolds with that annotation were labelled as 'Other'.

GC content was used to distinguish between *Entamoeba* reads and those of bacterial species so as to reduce the number of reads entered into the second round of this iterative process. There was a clear difference between the range of GC contents seen in Amoebida scaffolds and the range in non-Amoebida scaffolds. Note that the latter included those scaffolds lacking annotation so it may have included unconfirmed *E. bangladeshi* sequences (Figure 4.3.2). There was one roughly Gaussian curve of GC values seen in annotated *E. bangladeshi* scaffolds, peaking at 25%. However, there were two distinct normally distributed curves in the frequency of GC contents occurring amongst the non-Amoebida scaffolds, one peaking at 24% and the other at 46%. Aside from a small number of exceptions, no Amoebida scaffolds have GC contents as great as the scaffolds in the higher curve. It was likely that those in the higher peak in Figure 4.3.2, unless explicitly annotated as Amoebida, were bacterial sequences. The much greater peak frequency seen at 24% GC content was suspected of being due to a large number of unannotated *E. bangladeshi* sequences, as well as a number of bacterial sequences. As no distinctions could be made between them at this stage, all scaffolds with GC contents lower than 37% were included in the next round of assembly. All reads that were assembled into scaffolds with GC contents above this, except those annotated as Amoebida, were removed from the data set. After this first step, the number of reads had been reduced from 8,689,302 pairs to 2,294,270 pairs.

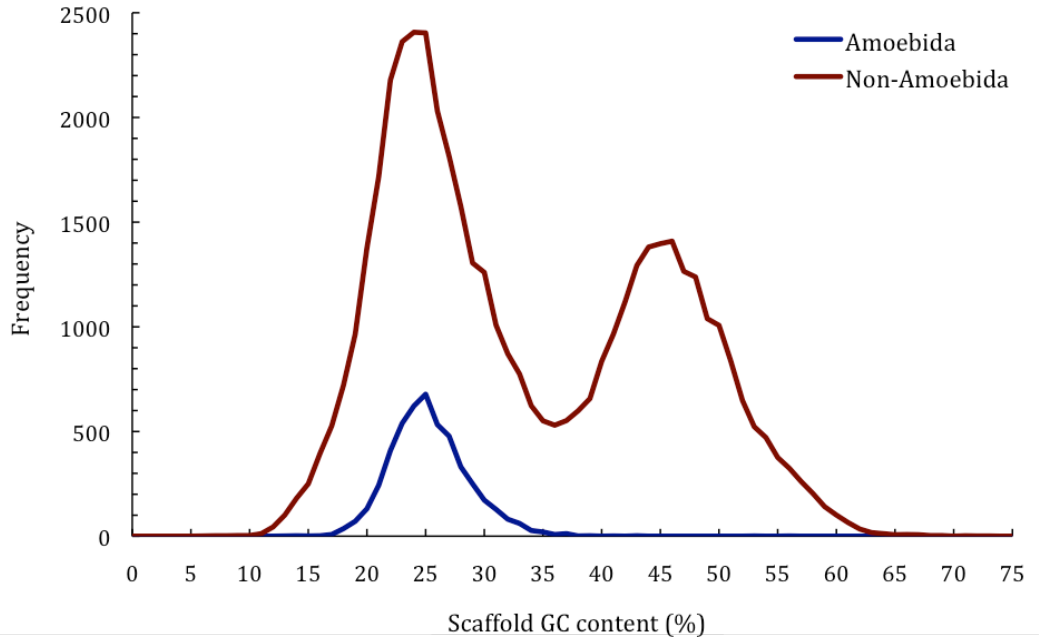


Figure 4.3.2. GC content of scaffolds assembled by ABySS using all generated reads. 'Non-Amoebida' includes unannotated scaffolds that may be unconfirmed Amoebida scaffolds.

The Blobplot protocol was run for a second time, using the reduced set of reads (Figure 4.3.3). The total number of resultant scaffolds annotated as Amoebida was 5,087, with a total combined length of 4,387,155 bp. This was a slight improvement over the assembly of Amoebida reads seen in the previous Blobology round (235 more scaffolds, 889,636 bp longer). Whilst only a modest improvement, this does demonstrate the benefit of reducing the sequencing data, rather than attempting to assemble 'through' the contaminants. ABySS will have been less likely to assemble chimaeric sequences, thus reducing the interference non-Amoebida scaffolds could introduce when ABySS attempted to assemble regions from *E. bangladeshi*. To further reduce the number of bacterial contaminants in the sequencing data, all scaffolds annotated as anything other than Amoebida were analysed. Annotations were derived from BLAST hits against the NCBI nucleotide database. It was deemed that one could be confident that all BLAST hits with E-values of 1e-50 or less were of a high quality. As such, all annotated non-Amoebida scaffolds with such low E-values were removed from the read set. This second round reduced the number of reads to 2,088,206 pairs.

When the Blobplot script was run a third and final time (Figure 4.3.4), for the reduced read set output by the previous step, more scaffolds consisted of *E. bangladeshi* sequences than before, however they comprised a slightly smaller total length (5,092 scaffolds, with a combined length of 4,383,265 bp). There was a slight increase in average coverage across the Amoebida scaffolds compared with those generated in the previous round (16.82 compared with 15.51, respectively). It is likely that the small decrease in total length was a direct result of the slight increase in coverage, with more reads being correctly mapped to the same genomic regions, so the assembly was not adversely affected by the previous filtering step. At this stage, the only non-Amoebida taxonomic Order making up at least 1% of the total number of scaffolds was the Clostridiales. It was deemed that these scaffolds, plus those grouped together in the 'Other' category, were few enough (341 in total) that the omission of their component reads from the final read set purely on the grounds of their annotation would have little impact upon the quality of the final *E. bangladeshi* assembly, even if any had been incorrectly annotated as non-Amoebida sequences. Given the improvement seen in the previous rounds' assemblies when contaminants were removed, it was likely that such a strategy would slightly improve the final assembly. As such, only reads that belonged to scaffolds that were either unannotated, or annotated as Amoebida, were included in the final set to be used in the assembler comparison study in Section 4.3.2. This final set consisted of 1,950,947 pairs of reads.

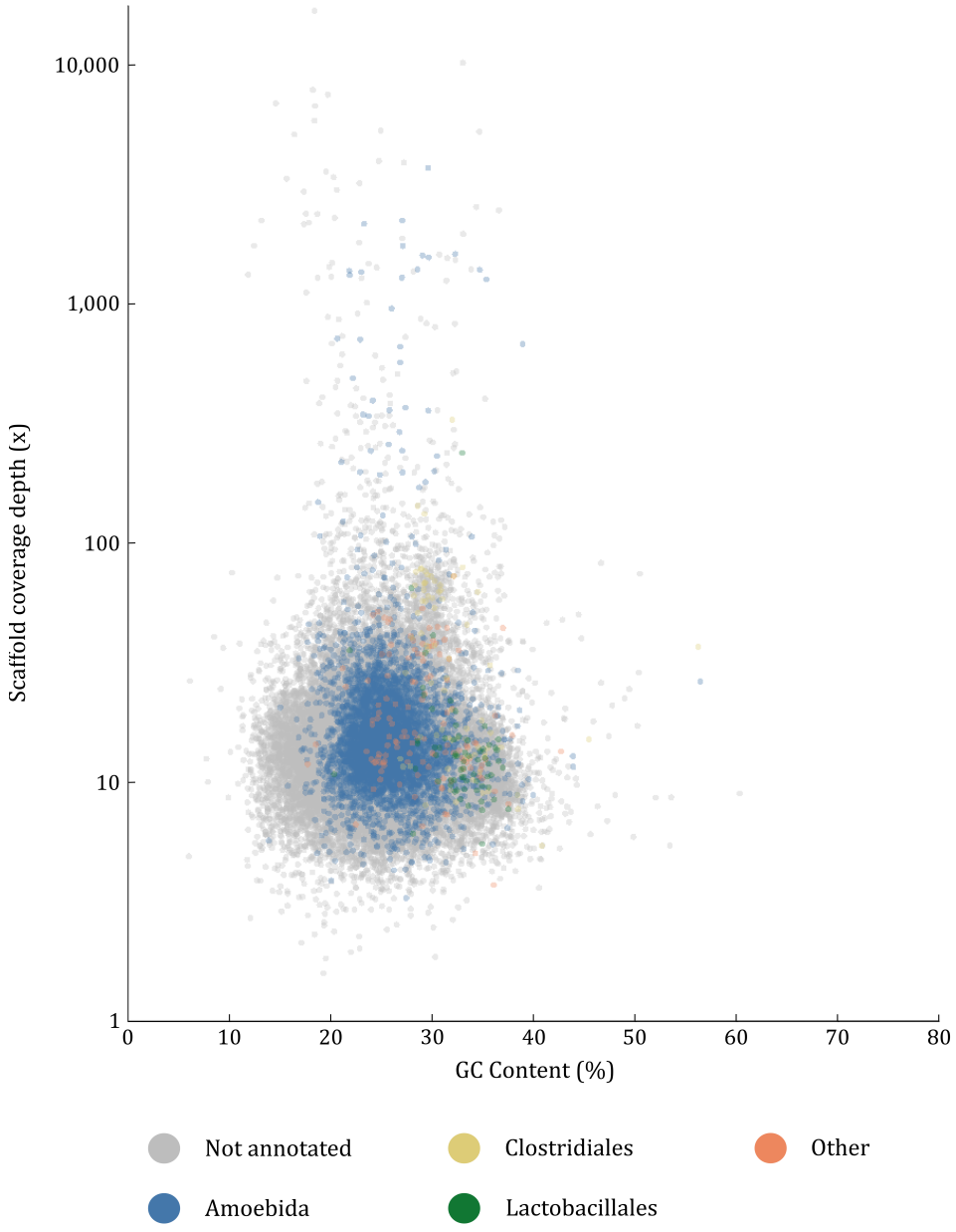


Figure 4.3.3. Scaffolds assembled by ABySS using reads annotated as Amoebida, or seen to have GC contents below 37%, in Blobology Round 1. Scaffolds are plotted as a function of their GC contents and average coverage depths. Taxonomic Orders are listed in order of decreasing prevalence from top to bottom, then left to right, across the key. If >1% of scaffolds shared the same Order annotation, all those with that annotation were labelled as 'Other'.

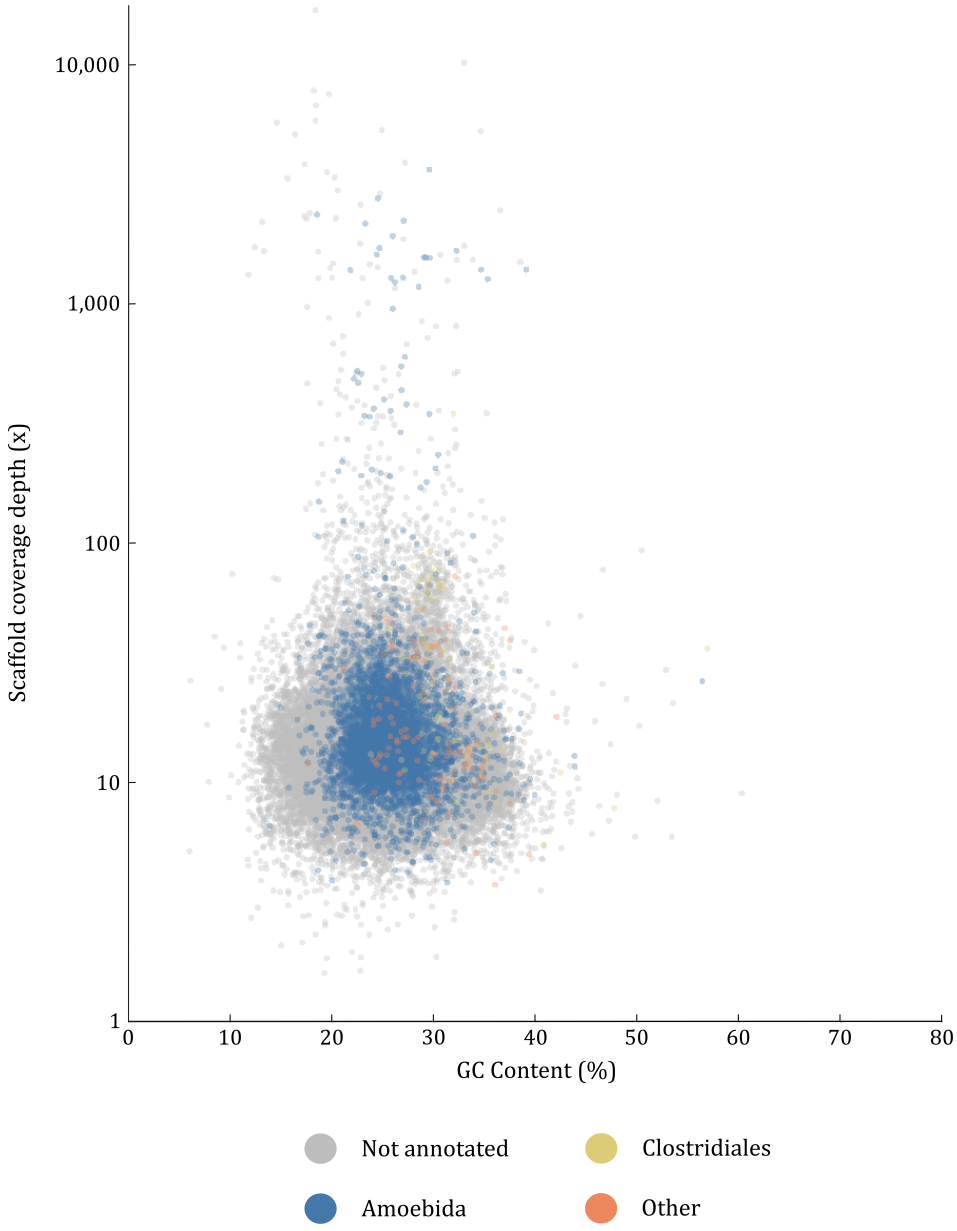


Figure 4.3.4. Scaffolds assembled by ABySS using reads output by Blobology Round 2. Scaffolds are plotted as a function of their GC contents and average coverage depths. Taxonomic Orders are listed in order of decreasing prevalence from top to bottom, then left to right, across the key. If >1% of scaffolds shared the same Order annotation, all those with that annotation were labelled as 'Other'.

4.3.2 Comparison of assemblers

Having identified the reads most likely to be derived from the *E. bangladeshi* genome, I wished to identify the genome assembler that generated the 'best' assembly from uncontaminated read libraries. Dozens of genomic *de novo* assembly programs exist, the majority of which accept Illumina reads as input. Practically, it was not possible to compare all of these programs in this chapter. Instead, a selection of four maintained, oft-cited assemblers were chosen: ABySS v1.3.4 [144], Velvet v1.2.02 [141], SOAPdenovo v2.04 [143] and Ray v2.3.1 [142]. All four are de Bruijn assemblers. They were run using the same set of reads and a k-mer size of '91'. It is not a simple matter to characterise a 'best' assembly, however. A number of publications have considered the difficulties associated with determining the best parameters for a *de novo* assembly [61, 112, 323]. As of yet, no single parameter has been agreed upon universally as an accurate measure. Complicating matters is that assembly quality is dependent upon the genome in question, as some are inherently more difficult to assemble due to their highly repetitive nature [323]. As such, the first step in determining the best assembly was to choose appropriate parameters for the *E. bangladeshi* genome.

i. N50, NG50 and scaffold length

Of the parameters mooted as potential measures of a good assembly, the N50 statistic is one of the most commonly used. The suitability of the N50 statistic has, however, come into question in recent years. Firstly, it does not offer any measure of accuracy or coverage depth within an assembly [111] and, secondly, it is possible for a small number of large scaffolds to result in a large N50, indicating a good quality assembly, masking a poorly assembled and fragmented remainder [61, 112, 323, 335]. An alternative to the N50 – the NG50 - was introduced in the first Assemblathon paper [111]. The NG50 statistic is a variation on N50 whereby the median scaffold length is determined using the expected genome size, rather than the assembly size. This effectively normalises comparisons of assemblies, although the genome size is, as it is here, usually an estimate, rather than a definite value. The NG50 statistic does not eliminate the issue that coverage depth and accuracy are not considered. As such, it was decided that either the N50 or the NG50 would be used, in conjunction with a number of other statistics, as suggested by the Assemblathon 2 team [112].

The first of these additional statistics directly reflects the primary goal of this genome assembly and many others besides – to detect CDSs. Theoretically, in multiple assemblies of similar lengths, the assembly consisting of a greater proportion of scaffolds at least as large as the average predicted gene sequence will contain the greatest number of complete CDS. The average gene size used in this chapter was 1,280 bp. This is the average gene size in *E. histolytica* HM-1:IMSS. The proportions of each assembly made up of gene-sized scaffolds were compared, along with each assembly's N50 and NG50 values, to demonstrate how these statistics compared with one another (Figure 4.3.5).

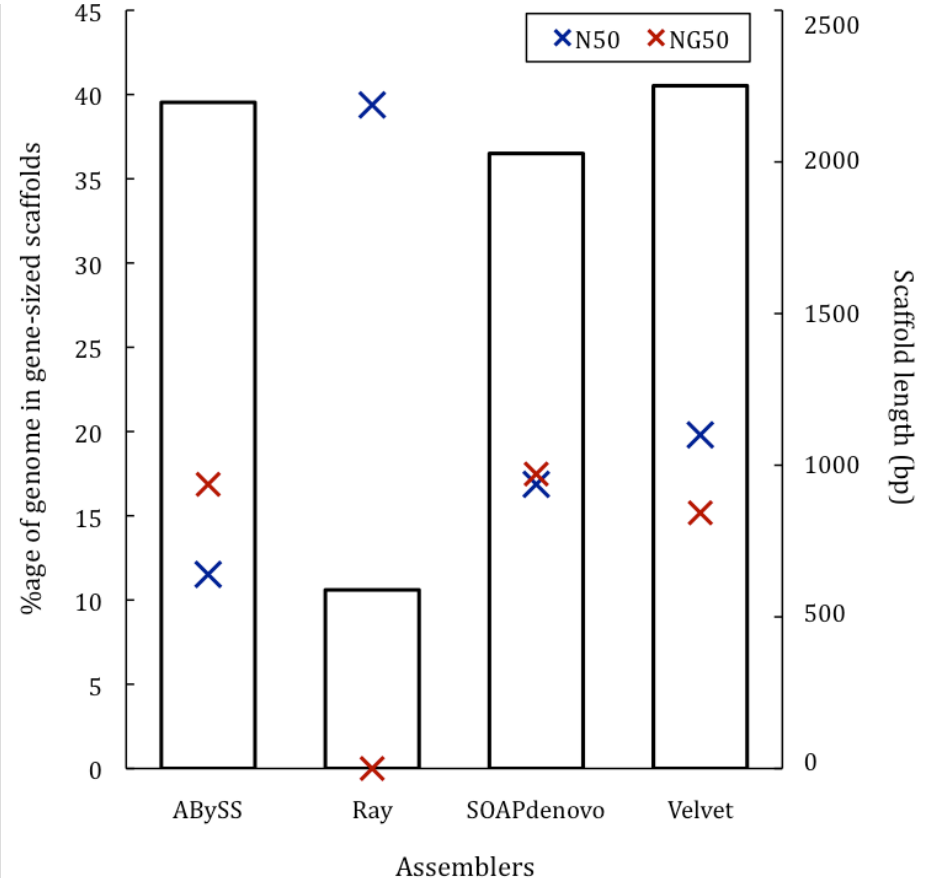


Figure 4.3.5. Comparison of assemblies' NG50 and N50 statistics, plotted with the proportions of the assembled genomes represented by gene sized scaffolds. A gene-sized scaffold is defined as one at least 1,280 bp in length. N50 and NG50 are plotted against the right axis; genome proportion is plotted against the left axis.

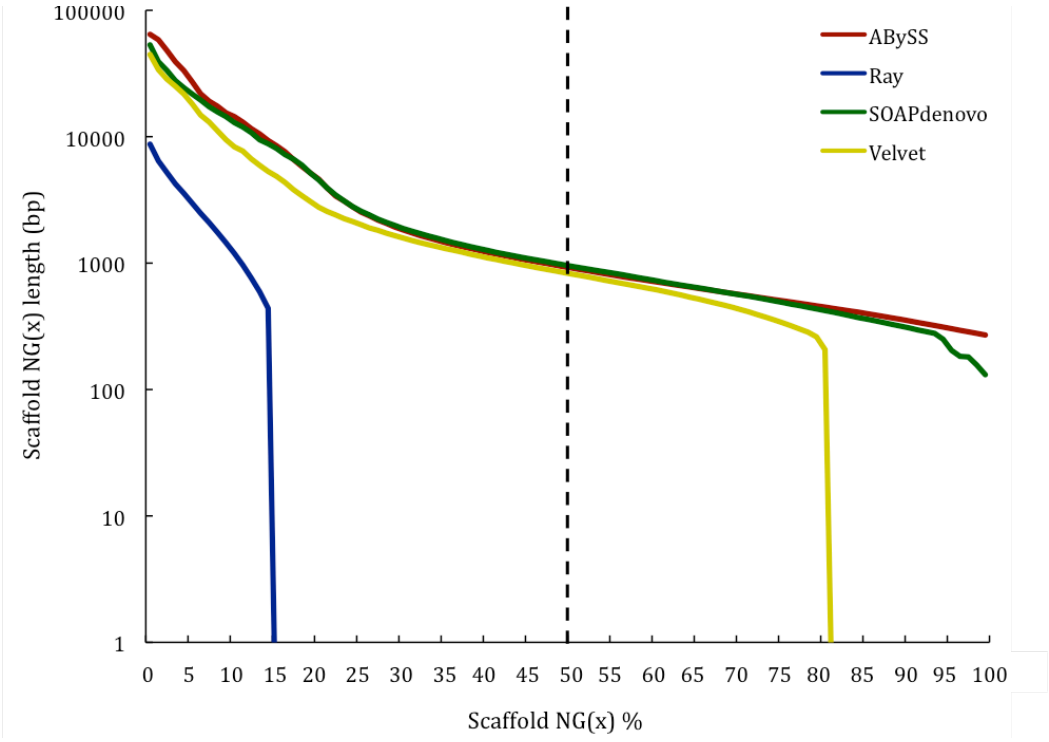


Figure 4.3.6. Comparison of the scaffold lengths at NG values of 1-100 in the four assemblies generated by ABySS, Ray, Velvet and SOAP. The dotted line represents the NG50 scaffold length.

From Figure 4.3.5, one can see that a considerably lower proportion of the assembly produced by Ray was comprised of gene-sized scaffolds than in the other three assemblies. The proportions seen in the ABySS, SOAP and Velvet assemblies, on the other hand, were comparable (between 36.5 and 40.5%). Initially, one might consider the low value in Ray to be indicative of a highly fragmented assembly of a similar length to the other three assemblies. However, the NG50 values give a better indication of the cause. Whilst the NG50 values in the ABySS, SOAP and Velvet assemblies varied, the Ray assembly's NG50 value could not be calculated. This is indicative of an assembly that is less than 50% of the length of the expected genome size. This is seen more clearly in Figure 4.3.6, above, which shows the NG values for 1-100% of the expected genome length. Were an assembly the exact same length as the genome, its scaffold length for values above NG100 would be 0. As such, the value of NG at which each assembly's line intercepts the x-axis reveals how long the assemblies were as proportions of the expected genome length. The assemblies of ABySS and SOAP were longer than the expected genome size (ABySS's was 130.9% of the expected length, whilst SOAP's was 102.5%). As such, it is likely that they either included reads that were not omitted from the assembly that should have been (i.e. reads that were

not actually from the *E. bangladeshi* genome) or they did not assemble overlapping scaffolds correctly. This could be due to sequencing errors preventing reads from being assembled together where they should be. A higher, more specific, k-mer length could potentially have reduced the assembly sizes, though it could not be guaranteed that this would have produced a more accurate and complete assembly. Velvet output a considerably shorter assembly, though it was, itself, a lot longer than that of Ray. Ray's assembly was just 15% of the length of the *E. histolytica* genome.

Whilst it appeared that Ray had produced a poor assembly, it was possible that this was a result of higher coverage depth across the scaffolds assembled by Ray than across the other three assemblies. As such, the coverage depths across each assembly's individual scaffolds were compared as functions of their lengths (Figure 4.3.7 a-d). The average depths for the assemblies clearly suggest that the assembly achieved using Ray was covered to a much greater depth than the others, with ABySS showing relatively poor coverage. The averages, however, were heavily skewed by the fact that the ABySS, SOAP and Velvet assemblies included scaffolds with low coverage values (<1x), as well as a large number with very high coverage depths (>100x), whilst the Ray assembly contained no coverage values below 4.5x. If one considers the numbers of scaffolds with coverage depths above Ray's average of 22.76x in each assembly, as well as those scaffolds sequenced at great depth (>=100x), it is clear that Ray's output was no more deeply sequenced than the other three assemblies (Figure 4.3.7.e).

In light of this, it was important to return to the comparison of the N50, NG50 and gene-sized scaffolds statistics (Figure 4.3.5) to choose which of the N50 and NG50 statistics would be used in comparing the assemblies. There was little to distinguish between the two statistics when comparing the ABySS, Velvet and SOAP assemblies – the NG50 values varied little whilst the N50 values increased slightly as assembly lengths decreased. If one were to base their judgement of assembly quality solely on the N50 value, it would appear that the Ray assembly was the best. It is clear from the NG50 statistic and the above comparisons that this was not the case and that the Ray assembly was simply the shortest, meaning any large scaffolds would have comprised a considerably greater proportion of the assembly than in the other three assemblies. Whilst the NG50 statistic is an imperfect measure of assembly completeness, it is clearly superior to the N50. As such, the assemblies were compared using the NG50 statistic and the proportion of each assembly that consisted of gene-sized scaffolds, as well as two other statistics discussed below.

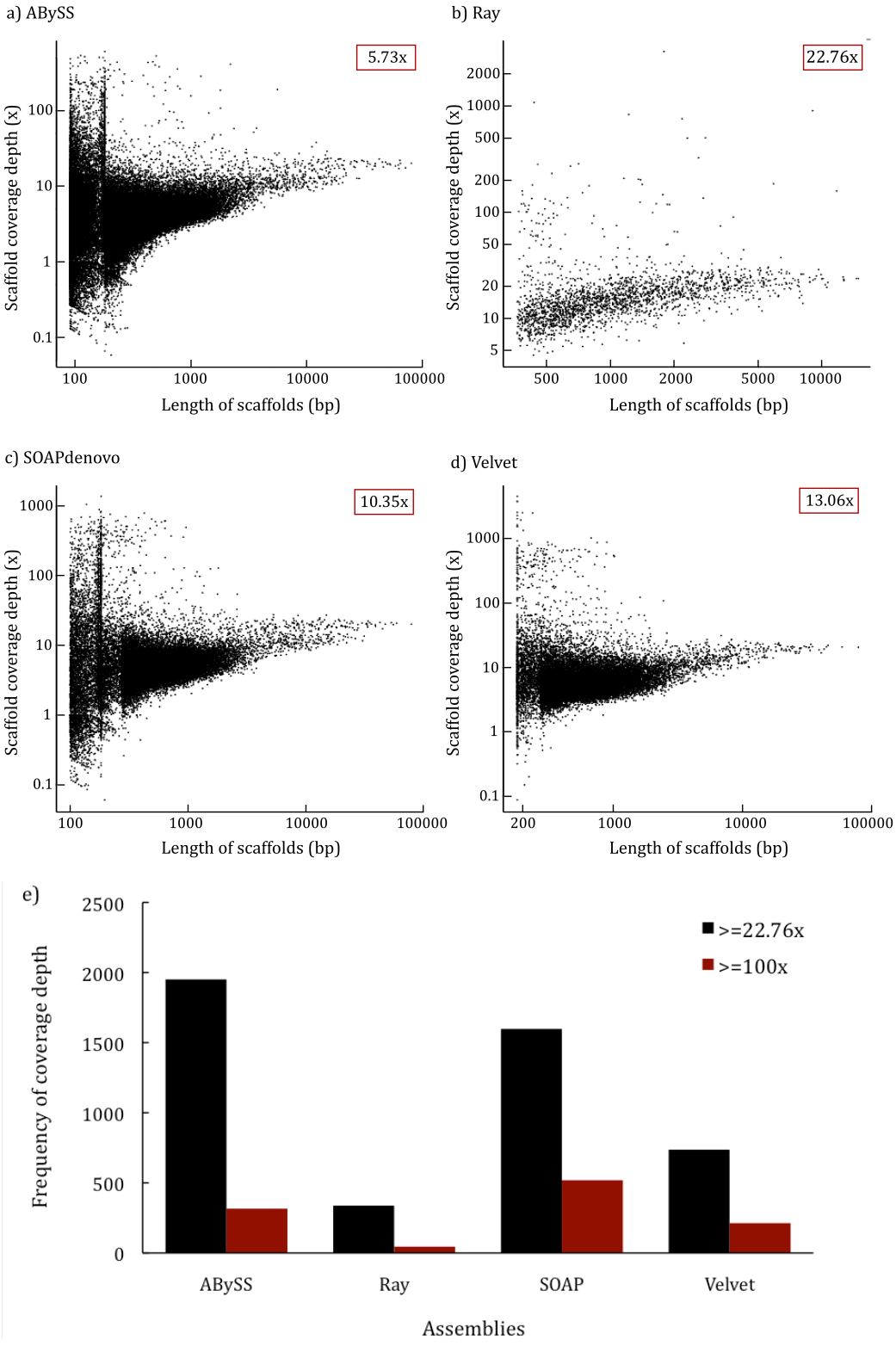


Figure 4.3.7. Comparison of scaffold coverage depths in the four assemblies.

a-d) Coverage depths across scaffolds in the assemblies as a function of their length. Values in red boxes are average coverage depths; **e)** Frequencies of coverage depths greater than the Ray assembly average and greater than 100x.

ii. Identifying presence of CEGs

Another method for measuring assembly completeness, as suggested by the Assemblathon 2 team [112], was to search the assemblies for the presence of the CEG set [114]. As described in Chapter Two, the CEGs are a set of 458 orthologous groups, consisting of six aligned proteins each (one from each of *Homo sapiens*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Caenorhabditis elegans*). Whilst it cannot be guaranteed, at this stage, that *E. bangladeshi* possesses orthologues of all 458 CEGs, it was reasonable to argue that the assembler whose output contained the most CEGs had produced the most complete assembly. Using the CEGMA program, sequences in the four assemblies orthologous to the CEGS were identified and any matching 70% or more of the CEG alignment were accepted as being present in the genome (Figure 4.3.8; Table 4.3.1). Many of the CEGs have paralogues, potentially making it difficult for assemblers to differentiate between them, however. As such, a subset of highly conserved, minimally paralogous CEGs, totalling 248 orthologue groups, was also searched for. Owing to the nature of the CEGMA output, in which the IDs of the 248 core CEGs found in each assembly are not given, it was not possible to see if any relationships seen between the assemblers in the 458 CEG set also applied for the subset. It was, however, possible to see how the numbers of core CEGs present in each assembly compared overall (Table 4.3.1). When detecting core CEGs, CEGMA reports both matches of 70% or greater and matches covering 20-70% of CEGs. The latter are reported as 'partial' matches and allow for a more inclusive analysis of the gene content of each assembly.

Of the 458 CEGs one might expect to exist in the *E. bangladeshi* genome, the assembly produced using ABySS contained the largest number. The SOAP and Velvet assemblies possessed slightly lower, but still comparable, numbers. The assembly generated using Ray, on the other hand, contained very few CEG orthologues, most likely as a result of its short length. For all four assemblers, fewer complete CEGs were identified from the core CEG set than in the full 458-sequence set. It is, therefore, possible that some of the matches in the larger set were caused by erroneous mapping to paralogous sequences. However, this could not be confirmed without a time-consuming study of the individual CEGs and, given that comparisons of both CEG sets ranked the assemblers in the same order, it seemed unnecessary to do this. It is interesting to note that, despite the poor number of CEG orthologues identified in the Ray assembly, all but two of them were only found in the Ray assembly (Figure 4.3.8).

Indeed, the majority of sequences identified in each assembly were unique to that assembly. This suggests that either each program was more successful at assembling particular regions of the genome than the others or, more likely, each assembled similar regions but with scaffolds that ended at different genomic positions. This would have resulted in different numbers of partial hits, which were not included in the initial comparison of 458 CEGs.

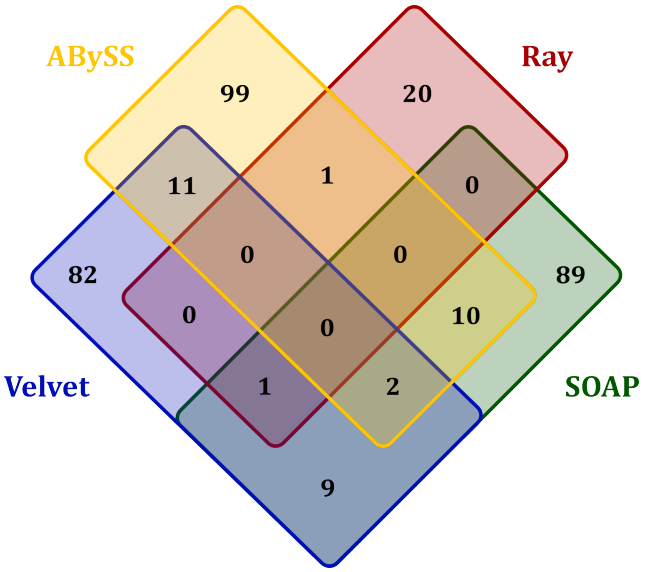


Figure 4.3.8. The number of the 458 CEGS to which orthologues were found in the assemblies output by ABySS, Ray, Velvet and SOAP. In total, 324 CEG orthologues were identified across the assemblies.

Table 4.3.1. Numbers of orthologues of CEGs and core CEGs identified in the ABySS, Ray, SOAP and Velvet *Entamoeba bangladeshi* assemblies. The full list of CEGs contains 458 alignments and the core list contains the 248 least paralogous CEGs.

Assembler	No. of total CEGs	%age of total CEGs	No. of complete core CEGs	No. of partial core CEGs	%age of core CEGs inc partials
ABySS	123	26.86	73	33	42.74
Ray	22	4.80	20	6	10.48
SOAPdenovo	111	24.24	67	30	39.11
Velvet	105	22.93	63	30	37.50

iii. Identifying presence of core *Entamoeba* gene clusters

As an additional measure of completeness, each assembly was tested for the presence of genes from the core *Entamoeba* gene set generated in Chapter Two. The gene set consisted of 4,704 clusters, containing at least one gene from each of *E. histolytica*, *E. dispar*, *E. invadens* and *E. moshkovskii*. The set contained 21,741 genes in total. The same principle regarding their usage applied as for the CEGs. These searches were performed using TBLASTN. The presence of an orthologue of a member of the core gene set in an assembly was accepted if the BLAST search results indicated that 70% or more of the core gene was present on one scaffold. Results were reported in terms of the number of clusters to which *E. bangladeshi* sequences belonged (Figure 4.3.9). Just one gene from a cluster needed to be matched in an assembly for that cluster to be considered present. The relative numbers of cluster orthologues identified in each assembly bore similarities to those seen in the CEG searches. The ABySS assembly contained the most orthologues, followed by SOAP's, then Velvet's, with the Ray assembly containing far fewer than the others. Once again, the different assemblies contained different gene sequences, for the same probable reason as outlined earlier. The consequences of this are explained in the next section.

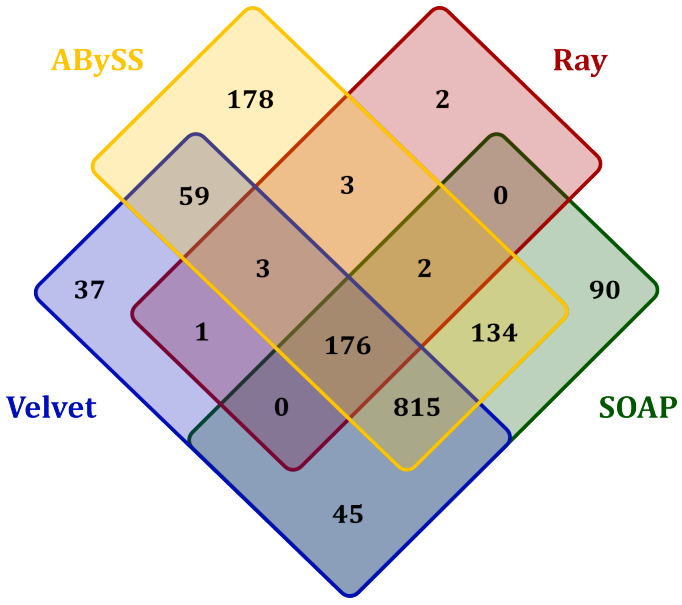


Figure 4.3.9. The number of the 4,704 core *Entamoeba* gene clusters to which orthologues were found in the assemblies output by ABySS, Ray, Velvet and SOAP. In total, 1,545 were identified across the assemblies.

iv. Identification of 'best' assembly

Four parameters have thus far been presented to define the 'best' assembly of the *E. bangladeshi* genome. The four used are not, of course, an exhaustive set. Indeed, arguably the most comprehensive assembler comparison study to date – Assemblathon 2 – identified ten key parameters that could be combined to rank assembly qualities [112]. In addition to the metrics compared in this study, the competition organisers compared statistics regarding fosmids and restriction fragment-based maps, as well as the results of REAPR analyses, which identify SNPs, indels and larger structural errors in assemblies [336].

Whilst it would have been ideal to perform a similarly inclusive and in-depth comparative study, time and resource limitations restricted the extent to which the assemblies could be analysed in this project. To combine the generated test results in order to more robustly identify the most effective assembler, the assemblers were ranked by their performance in each test, as in the Assemblathon 2 competition [112]. The best performing program for each parameter was given a rank of '1', with the poorest performer receiving a rank of '4'. The sum of each assembly's ranks across the four tests was then calculated, with the assembly with the lowest total rank being regarded as the best assembly (Table 4.3.2).

Table 4.3.2. The ranked performances of assemblers ABySS, Ray, SOAP and Velvet, according to multiple tests of assembly quality. '1' represents the best performance in a category, whilst '4' represents the worst. Final ranks calculated by summing each assembler's 4 ranks and ordering from lowest total '1' to highest '4'.

Assembler	%age of genome size made up by gene-sized scaffolds	NG50	CEGMA	Core *Entamoeba* genes	Final rank
ABySS	2	2	1	1	1
Ray	4	4	4	4	4
SOAP	3	1	2	2	2
Velvet	1	3	3	3	3

The ABySS assembly was found to be the best, overall, followed by the SOAP and Velvet assemblies. Ray was found to be the least effective assembler in all tests. As such, one can conclude that, if an *Entamoeba* species is to be assembled *de novo* using a single assembler, ABySS is likely to be highly effective at constructing a genome that can be used in gene prediction. However, it is unlikely that any research group should be limited to using just one assembler. The previous two sections revealed that each of the four assemblies compared here possessed gene sequences that the others did not. In light of this, it is more reasonable to conclude that read libraries passed through the Blobology protocol should be assembled using multiple programs in order to detect a greater number of gene models. As such, all core *Entamoeba* genes identified in Section 4.3.2 were entered into Section 4.3.3's comparison, not just those detected in the ABySS assembly.

It is important to reiterate that no set of parameters for measuring the quality and completeness of a genome assembly has been agreed upon. Indeed, multiple competitions to identify the best assembler have been held, with inconclusive results highlighting the difficult nature of this task [111, 112, 323]. In light of the different strengths and weaknesses of the myriad available assemblers, the practice of combining assemblies to produce one final superior genome has been introduced by a number of programs, employing a variety of approaches [337-340].

One such program is the Genome Assembler, Reconciliation and Merging (GARM) pipeline [337]. GARM makes use of Perl scripts and modules, as well as existing third-party software, to merge scaffolds and contigs produced by different assemblers or sequencing technologies [337]. It has been used to combine *de novo* assemblies of 454 reads and Illumina reads in the assembly of the genome of *Hymenolepis microstoma*, the mouse bile duct tapeworm [341], and to combine assemblies made from 454 reads and Ion Torrent reads when sequencing an avipoxvirus isolated from a Feral Pigeon (*Columba livia*) [342].

Also available are programs based upon performing local [338] and global [340] alignments between assemblies. For example, GAM-NGS (Genome Assemblies Merger for Next Generation Sequencing) [338] uses a weighted graph algorithm to perform local alignments between two or more assemblies, allowing regions of a genome that are difficult to sequence to be resolved. Another program, MAIA (Multiple Assembly IntegrAtor) [339], employs its own weighted graph protocol to merge two or

more assemblies. Taking a different approach, the GAA (Graph Accordance Assembly) program uses BLAT [343] to perform a global alignment between two assemblies only, generating a weighted graph that is subsequently used to merge the two sets of contigs [340].

Whilst these examples demonstrate the breadth of options available to those hoping to combine multiple assembies, another group has shown it is possible to perform a merger of assemblies without these dedicated programs or pipelines. The Rhesus Macaque genome was assembled through the merging of three individual assemblies, simply by mapping the assemblies consisting of shorter contigs to the most contiguous assembly [344]. Evidently, there are numerous options available to those who might choose to combine assemblies of an *Entamoeba* genome; however, one must take care when comparing these combined assemblies as the complexities of defining a superior assembly, as described above, remain.

4.3.3 Comparison of assembly techniques

The previous section considered one method for assembling the genome of an *Entamoeba* species extracted from a xenic culture, whereby contaminants were removed prior to assembly. That approach was useful for constructing the genome, but its efficacy with regards to gene prediction was thought to be improvable. This section will compare that protocol with another, less time-intensive, method in order to determine the most effective overall approach to gene prediction in a *de novo* genome assembly in *Entamoeba* species. For this second method, all reads generated from the sequencing libraries – amoebic and bacterial – were used. Forward and reverse reads from each of the three libraries were kept separate, treating them like singlet reads, and were used as databases against which all protein sequences from *E. histolytica*, *E. dispar*, *E. invadens* and *E. moshkovskii* were compared in a TBLASTN search. Reads were treated as singlets because it was unlikely, given the relatively small size of the gene sequences, that the members of many pairs of reads would both align to a gene sequence. Reads that showed a high quality hit to a gene sequence defined by the hit's E-value were then entered into a *de novo* assembly with only those other reads that hit that particular gene, generating a set of 'mini assemblies'.

As comparisons of assembly parameters with the assemblies from Section 4.3.2 would have been meaningless here, the most appropriate measure of effectiveness of

the two methods was to compare the numbers of genes they assembled (Figure 4.3.10). Each of the four sets of scaffolds based upon a species' gene set was used as input for the CEGMA program, as in Section 4.3.2, to identify complete and partial CEGs. Every scaffold constructed for the genes of each of the four species was also used as a database in a TBLASTN search featuring the core *Entamoeba* protein set as the query sequences. Only hits that matched at least 70% of the query sequence on one scaffold were accepted here. The complete *E. histolytica* protein set was also queried against the scaffolds based upon its genes (as a measure of completeness outside of core genes), using the same completeness filter.

As was established in Section 4.3.2, the ABySS assembler alone was inferior to a combination of all four whole genome assemblies when assembling, and allowing detection of the presence of, CDS sets (Figure 4.3.10). The single assembly did, however, possess a greater number of *E. histolytica* genes than were detected in the mini assemblies approach. It included a greater number of CEGs too, suggesting that many of the CEG sequence orthologues in the mini assemblies were incomplete on single scaffolds. Whilst the smaller sequence fragments that result from the mini assemblies approach appeared to affect the number of sequences that could be detected in those two categories, the same is not true of the core *Entamoeba* clusters. In this case, the mini assemblies method outperformed even the combination of full genome assemblies. As such, it would appear that, once again, neither of the methods being compared can be resolutely recommended over the other. To illustrate the point that both can be considered effective approaches, it was useful to identify how many genes/clusters each method assembled that the other could not (Figure 4.3.11).

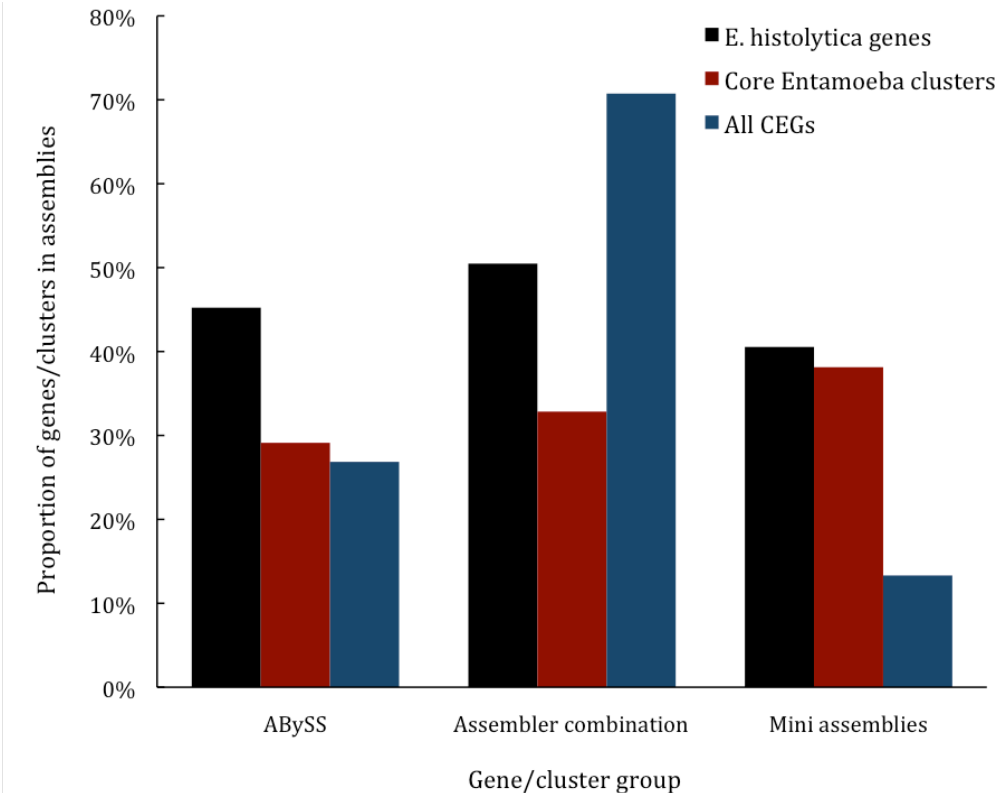
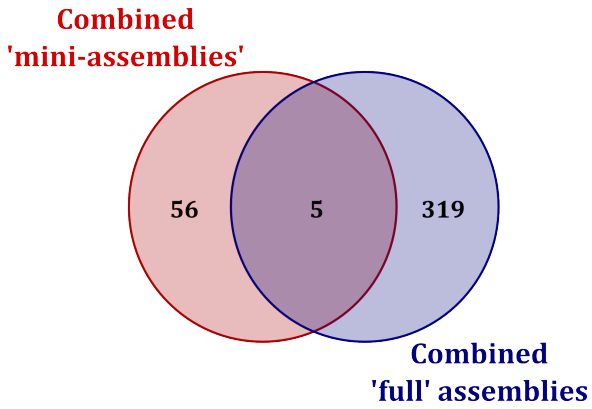
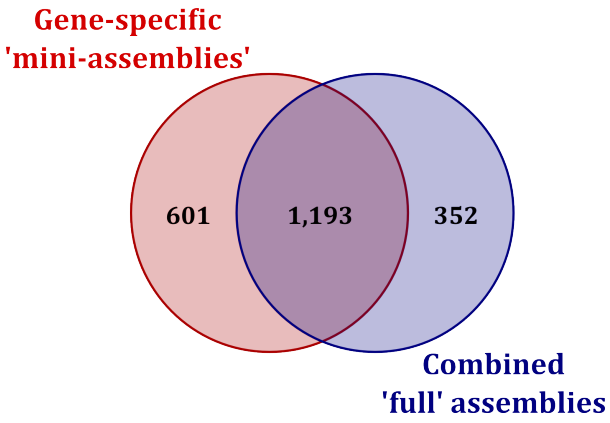


Figure 4.3.10. Comparison of assembly methods based upon proportions of several gene groups detected within assemblies of the *Entamoeba bangladeshi* genome. Genes detected by the ABySS *de novo* assembler's genomic assembly are presented alone, and with other genes similarly detected by the Ray, Velvet and SOAP assemblers. 'Mini assemblies' refers to assemblies based upon individual gene sequences from *E. histolytica*, *E. dispar*, *E. invadens* and *E. moshkovskii*, except in the case of the *E. histolytica* gene set for which only assemblies based upon *E. histolytica* gene sequences were used. The *E. histolytica* gene set totals 8,306; there are 4,704 core *Entamoeba* clusters; and there are 458 CEG orthologues.

a)



b)



c)

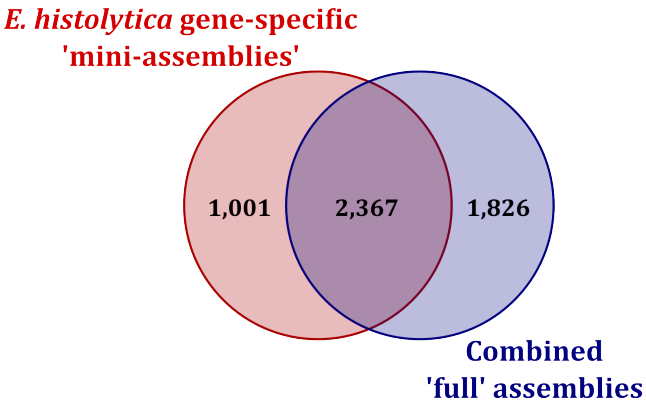


Figure 4.3.11. Comparison of genes and clusters identified in all full genome assemblies combined and all 'mini-assemblies' combined. Gene and cluster sets are **a)** 458 CEG orthologue groups; **b)** 4,704 core *Entamoeba* gene clusters; **c)** all 8,306 *Entamoeba histolytica* genes. Note that, for 'c', only mini-assemblies based upon *E. histolytica* coding sequences were compared with the combined full genome assemblies.

Table 4.3.3. Proportions of the 20 most prevalent gene functions from the core *Entamoeba* gene set that were detected in *Entamoeba bangladeshi* using every whole genome and mini assembly

Function	Count	%age of total
Serine/Threonine/Tyrosine Protein Kinases	85	73.28
Ras family GTPases	83	96.51
RasGAP family proteins	29	50.00
Ras family GEFs	18	40.00
Serine/threonine/tyrosine protein phosphatases	35	79.55
Zinc finger domains	31	73.81
F-box domain-/WD domain-/Leucine-rich repeat- or combination thereof containing protein	23	56.10
Cell membrane transporters and pumps	35	87.50
Cytoskeleton-related proteins	21	63.64
Vesicle/vacuole transport- and membrane-related proteins	25	80.65
BspA	27	100.00
RNA-binding proteins	16	66.67
DNA repair proteins	16	69.57
Ubiquitination proteins	12	54.55
Heat Shock Proteins and transcription factors	18	85.71
Nuclear receptors and transporters	2	10.53
DNA packaging, assembly and maintenance	11	61.11
Ubiquitin carboxyl-terminal hydrolase	5	33.33
Calcium-binding proteins	13	100.00
Helicases	8	66.67

The numbers in Figure 4.3.11 demonstrate that, as was seen in the various *de novo* assemblers, the different methods effectively sequenced some regions that the other could not or else one could not concatenate scaffolds where the other could. The outcome of this is that each could assemble coding sequences that the other could not. As such, the logical conclusion would be to perform both assembly methods when detecting gene sequences in novel *Entamoeba* genomes. None of the programs or methods described in this chapter are particularly memory- or time-intensive, meaning very few research groups would be limited in what they could achieve. The

benefit of this is the ability to better identify gene sequences in incomplete genomes. In the case of *E. bangladeshi*, combining all genes detected in the myriad assemblies generated allows one to begin to identify which members of the core *Entamoeba* gene set can still be regarded as 'core' genes when considering this novel species. In total, genes belonging to 2,146 of the 4,704 clusters were identified. Within this set, genes with the most prevalent functions from the core gene set were detected (Table 4.3.3). With greater sequencing depth and a further refinement of the assembly processes, it is likely that a greater proportion of the genes in *E. bangladeshi* could be detected and, with that, a better understanding of *E. bangladeshi*'s virulence capabilities.

4.3.4 Phylogenetic relationship of *Entamoeba bangladeshi* with other *Entamoeba* species

To date, the position of *E. bangladeshi* within the *Entamoeba* genus has only been determined through alignment of a small ribosomal subunit, for which an incomplete sequence was known in *E. bangladeshi* [26, 46]. Here, I analysed the phylogenetic arrangement within the *Entamoeba* genus using a highly conserved gene sequence completely sequenced in all five species involved (*E. histolytica*, *E. dispar*, *E. invadens, E. moshkovskii* and *E. bangladeshi*). This comparison is the first to define the relationship between *E. invadens* and *E. bangladeshi*. The orthologous gene sequences used to construct the phylogeny form a cluster within the core *Entamoeba* gene set. They encode the 60S acidic ribosomal protein P2. Within the ABySS assembly of *E. bangladeshi*, the coding sequence existed on the forward strand of scaffold 1175, positions 334 – 657, as indicated by the result of the TBLASTN search described in Section 4.3.3.

The phylogram generated shows *E. bangladeshi* to be most closely related to *E. histolytica* and *E. dispar* (Figure 4.3.12). It is more distant from *E. moshkovskii*, which is shown to have diverged at an earlier date. Given that we have accepted that *E. moshkovskii* is a human-infective species, this suggests that *E. bangladeshi* is, indeed, human-infective too, as opposed to a free-living, transient species. The tree is completed by the distantly related *E. invadens*. The tree previously generated using an incomplete *E. bangladeshi* sequence [26] corroborates the phylogeny presented here.

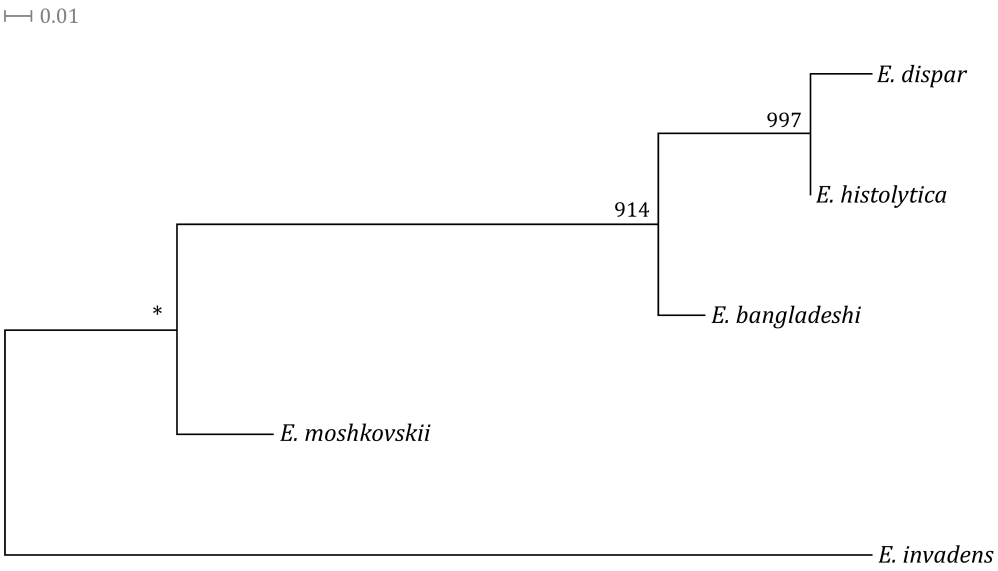


Figure 4.3.12. Phylogeny of orthologous genes encoding 60S acidic ribosomal protein P2. The midpoint-rooted tree was created using the additive tree model. Bootstrapping was done 1,000 times. The asterisk represents a bootstrap value of 1,000.

4.4 Concluding remarks

In this chapter I have compared methods for performing a *de novo* assembly of the recently discovered species *Entamoeba bangladeshi*, which was sequenced as part of an undefined xenic culture. Initially, reads found to contain sequences from genomes other than that of *E. bangladeshi* were filtered out of the read libraries through an iterative use of the publicly available Blobplot protocol. Whilst it is unlikely that this process could be used to remove all contaminants from a metagenomic sample, it was effective in improving the performance of the ABySS assembler, resulting in a greatly improved genomic assembly of *E. bangladeshi*. Other researchers aiming to extract *Entamoeba* reads from a metagenomic sample would be encouraged to consider using the method described here prior to performing a final assembly.

Once reads confirmed as, or suspected of, containing sequence data from the *E. bangladeshi* genome had been extracted, the performances of four *de novo* assemblers based upon the de Bruijn algorithm – AbySS, Ray, Velvet and SOAPdenovo – in assembling the *E. bangladeshi* genome were compared. No program can be deemed the best for assembling all genomes, nor can any one parameter be considered a universally good measure of genotype quality. As such, it was necessary to investigate which parameters provided the most practical and informative assessment of assemblies of this member of the *Entamoeba* genus, and judge the assemblies using these statistics. The NG50 statistic, combined with a measure of the proportion of gene-sized scaffolds in an assembly and a count of expected core genes, produced rational results. The oft-used N50 statistic, on the other hand, was found to be easily skewed by a short, inferior assembly. It would seem that this statistic, certainly on its own, is insufficient as a measure of assembly quality.

ABySS produced the best assembly overall, with Ray assembling a relatively short set of scaffolds. If one were to perform a single assembly for an *Entamoeba* genome, it would be advisable not to utilise the Ray platform. The optimisation of parameters used in any of the other three programs tested here would, however, be encouraged. It is important to note, however, that the assembly produced using Ray did include gene sequences not seen in the other three assemblies. So, too, did a final method that involved performing individual assemblies for each orthologue of a known *Entamoeba* gene found in the *E. bangladeshi* reads. As such, neither of these general methods can be recommended more highly than the other. Indeed, one can conclude

that researchers investigating the gene content of an *Entamoeba* species for which a *de novo* assembly is necessary would be well advised to perform multiple full genome assemblies using different platforms, as well as individual assemblies for the core gene sets described in this chapter. These assemblies are far from costly to generate and, together, can provide a much clearer picture of the coding sequences contained within newly assembled genomes. It would be very interesting to see which genes are found to exist in the *E. bangladeshi* genome as a result of such methods when it is more completely assembled than was possible to achieve here.

Finally, the phylogenetic relationship of *E. bangladeshi* to its confirmed human-infective relatives and the more distant *E. invadens* were presented here for the first time. This is also the first time a complete gene sequence has been used to test the phylogeny of the novel species. *E. bangladeshi* was found to be most closely related to *E. histolytica*. It diverged from a common ancestor after *E. moshkovskii* but before *E. histolytica* and *E. dispar*, implying that it is a human-infective species. I am hopeful that further research will investigate this species in order to gain a greater understanding of its genome structure and gene content. With such information, it may be possible to determine whether *E. bangladeshi* is a non-invasive parasite or another species capable of causing symptomatic amoebiasis for which preventative measures will need to be investigated.

Chapter Five – Conclusions and Future work

Entamoeba histolytica is an obligate parasite of humans and is the aetiological agent of the disease amoebiasis [1]. The genome of *E. histolytica* has been studied in depth over the past decade with much being learned about the genes and proteins it employs in parasitising and invading hosts [49-51]. However, comparatively few studies have looked into the genetic content of other members of the genus. Recent years have seen an increase in evidence suggesting that species other than *E. histolytica* are pathogenic [4, 33], as well as the discovery of a novel human-infective species in Bangladesh [26]. This project sought to utilise modern comparative genomic techniques to improve understanding of these species, their capacity for causing disease and their potential impact upon the epidemiology of amoebiasis.

5.1 *Entamoeba moshkovskii* – assembly, annotation and diversity studies

Entamoeba moshkovskii was originally thought to be a free-living species [36, 38]. However, recent studies have shown it to be human-infective, with the suggestion that it can cause disease [3, 4, 44]. A draft assembly and annotation of the 25 Mb genome of *E. moshkovskii* strain Laredo was presented here. The genome contained 12,449 gene models. It is likely that, as has happened in similar draft genomic annotations [51], this number of genes will be an over-estimate of the true count. Re-assembly and re-annotation of the Laredo genome would aid in rectifying any errors made here. Indeed, Chapter Three of this thesis demonstrated the improvements that can be achieved through alignment of additional sequencing reads, identifying incorrectly called bases and heterozygous bases, to an existing genome.

Three additional strains of *E. moshkovskii* were subsequently sequenced and mapped to the Laredo genome. The overall diversity demonstrated within the species (based upon homozygous SNP rates) was compared with that of *E. histolytica*. Significantly greater diversity was seen across all sequence classes of the *E. moshkovskii* genome, with coding regions proving to be approximately 200 times more variable in *E. moshkovskii* than in *E. histolytica*. This prompted the use of the four-haplotype test, which failed to detect evidence of genetic recombination between the strains. This suggests that *E. moshkovskii* should no longer be considered a single species, where all strains exchange genetic material. *E. moshkovskii* should henceforth be referred to as a species complex. It would be encouraging to see further studies

expand upon this by investigating which members of the *E. moshkovskii* species complex are strains of the same species. Determining which particular sequence types are human-infective, and possibly pathogenic, would also greatly improve our understanding of the epidemiology of amoebiasis.

5.2 Comparative analyses of *Entamoeba* species' genetic content

The generation of a reference genome for *E. moshkovskii* was exploited to allow a comparative genomic analysis with three other members of the *Entamoeba* genus – *E. histolytica*, *Entamoeba dispar* and *Entamoeba invadens*. Orthologous gene families present in all four species were identified. These were compared with gene families present in other genera in the Unikonts clade to generate a set of orthologous gene families unique to the *Entamoeba* genus. The functions most prevalent amongst the *Entamoeba*-exclusive gene families included those whose presence allows for survival in the environmental niche into which *Entamoeba* species have adapted to fit. For example, all *Entamoeba* possess cell membrane transporters and pumps not seen in other genera, as well as cytoskeleton-related and vesicle-related proteins. These may be required to acquire nutrients from hosts and to phagocytose bacteria. It should be recognised that this core gene set is incomplete and further efforts would be encouraged to improve upon it, reducing it where necessary as more species and strains of *Entamoeba* are sequenced. It would also be helpful to increase the numbers of species against which the *Entamoeba* are compared, as this will further refine the list of gene families supposedly unique to *Entamoeba*. A comprehensive list of the gene families required by all parasitic members of this genus would aid studies investigating key families required for parasitism of *Entamoeba* hosts, as well as those factors involved in causing invasive disease.

Not all infections by *Entamoeba* species are marked by symptomatic disease. Whilst some *E. histolytica* strains are capable of invading the host's intestinal mucosa, *E. dispar* is considered avirulent [2, 32]. The presence of orthologous virulence factor families in the genus was investigated in order to ascertain the virulence potential of *E. moshkovskii* and to identify those virulence factors predominantly seen in organisms able to cause amoebic colitis. The results presented here support the literature in showing *E. dispar* to be a non-pathogenic species [2, 32]. Conversely, *E. moshkovskii* has a genome consistent with that of a human-infective parasite, but also possesses similar families to those seen in the pathogenic reference strain of *E. histolytica*, including

members of the cysteine protease super-family. It would be very interesting to see further evidence of *E. moshkovskii*'s pathogenicity, as the existence of another human-infective virulent *Entamoeba* species would have a profound impact upon our definition of the aetiological agent of amoebiasis, as well as, potentially, the morbidity and mortality rates of *E. histolytica* infections.

Of particular importance in survival of a parasitic, and potentially invasive, life cycle were surface-bound virulence factors including the heavy Gal/GalNAc lectins and members of the vast BspA family. These proteins, the latter of which are not seen in the avirulent *E. dispar*, are suspected of involvement in adherence of trophozoites to host cells [200], implying this stage is a key determinant of the course of an infection. As relatively little is known regarding the BspA family, research into its sequence diversity, as well as functional analysis of its precise role, would be highly recommended. The Gal/GalNAc lectins have previously been mooted as potential vaccine targets, which the findings presented here support, as the CRD domain is present in all strains studied.

Interesting results were also obtained regarding *E. invadens*. This parasite can infect a range of reptilian hosts [345], leading to speculation as to how it might achieve this. Presented in this thesis was evidence that *E. invadens* possesses expanded sets of several virulence factors, notably cysteine proteases and surface proteins such as the Gal/GalNAc lectin complex [81, 314]. It is probable that these proteins, and the genes that encode them, are responsible for allowing *E. invadens* to attack such a wide host range. This may explain the large genome and gene set observed in *E. invadens* relative to those of human-infective species [345].

5.3 Comparisons of *de novo* genome assemblers used to construct *Entamoeba bangladeshi* genome

Many *Entamoeba* cultures cannot be easily axenised, making DNA extraction and sequencing considerably more difficult. Multiple methods of assembling DNA sequenced from such a mixed culture were compared in Chapter Four of this thesis using a xenic culture of the recently discovered species *Entamoeba bangladeshi* [26]. The publicly available Blobology protocol [325] was utilised to separate *Entamoeba* reads from bacterial reads, before four *de novo* assembly programs – AbySS, Ray, Velvet and SOAPdenovo [141-144] - were compared using the amoebic reads. The

assemblers were tested using comparable parameters under the assumption that the majority of genome assemblies are carried out with the intention of identifying coding sequences. Whilst ABySS was found to produce the best individual assembly in this respect, it was found that groups performing similar assemblies in future would be advised to combine the gene prediction abilities of multiple assemblers. Each program is capable of detecting coding sequences that the others cannot. Attempting to assemble individual genes, rather than identifying genes within an assembled genome, produced fewer gene models yet still identified genes that the larger assemblies omitted. As such, a combination of these techniques is advised for future studies attempting to assemble novel *Entamoeba* species.

These findings do come with the caveat that different parameters and different measures of assembly quality may produce significantly different results. As such, further work to investigate and refine the optimal parameters required by individual genome assemblers would be strongly advised. Statistics for determining the best assembly, meanwhile, are more complex; however, the results of this thesis strongly suggest that the N50 is an inadequate measure of assembly quality. A combination of the NG50 statistic, proportion of gene-sized scaffolds and a count of expected core genes produced far more logical results. I would urge groups undertaking similar studies to investigate this matter further, however, and to use additional parameters, such as those listed by the Assemblathon 2 team [112], wherever possible, so as to improve the scientific community's understanding of the best set of statistics by which to measure genome assemblies.

With regards to *E. bangladeshi* itself, genetic information was limited in its use. However, the phylogenetic relationship of *E. bangladeshi*, compared with *E. histolytica*, *E. dispar*, *E. moshkovskii* and *E. invadens* was elucidated, suggesting that *E. bangladeshi* is a human-infective species. As such, future work would ideally further investigate the gene set of this human parasite in order to determine whether it possesses the gene families identified here as important to the survival of a pathogenic lifestyle. As is the case with *E. moshkovskii*, the more information that can be accumulated regarding *E. bangladeshi*, the greater our understanding of this parasitic and, in some cases, pathogenic genus of protozoa will be. The work described here on sequencing from xenic culture opens up new avenues for sampling genomes from more *Entamoeba* strains and species in the future.

References

1) Walsh JA: **Problems in recognition and diagnosis of amebiasis: estimation of the global magnitude of morbidity and mortality.** *Rev Infect Dis* 1986, **8**:228–238.

2) Diamond LS, Clark CG: **A redescription of *Entamoeba histolytica* Schaudinn, 1903 (Emended Walker, 1911) separating it from *Entamoeba dispar* Brumpt, 1925.** *J Eukaryot Microbiol* 1993, **40**:340–344.

3) Ali IKM, Hossain MB, Roy S, Ayeh-Kumi PF, Petri Jr. WA, Haque R, Clark CG: ***Entamoeba moshkovskii* infections in children, Bangladesh.** *Emerging Infect Dis* 2003, **9**:580–584.

4) Shimokawa C, Kabir M, Taniuchi M, Mondal D, Kobayashi S, Ali IKM, Sobuz SU, Senba M, Houpt E, Haque R, Petri Jr. WA, Hamano S: ***Entamoeba moshkovskii* is associated with diarrhea in infants and causes diarrhea and colitis in mice.** *J Infect Dis* 2012, **206**:744–751.

5) Stark D, Fotedar R, van Hal S, Beebe N, Marriott D, Ellis JT, Harkness J: **Prevalence of enteric protozoa in human immunodeficiency virus (HIV)-positive and HIV-negative men who have sex with men from Sydney, Australia.** *Am J Trop Med Hyg* 2007, **76**:549–552.

6) Stark D, van Hal SJ, Matthews G, Harkness J, Marriott D: **Invasive Amebiasis in Men Who Have Sex with Men, Australia**. *Emerging Infect Dis* 2008, **14**:1141–1143.

7) Rivera WL, Santos SR, Kanbara H: **Prevalence and genetic diversity of *Entamoeba histolytica* in an institution for the mentally retarded in the Philippines**. *Parasitol Res* 2006, **98**:106–110.

8) Nishise S, Fujishima T, Kobayashi S, Otani K, Nishise Y, Takeda H, Kawata S: **Mass infection with *Entamoeba histolytica* in a Japanese institution for individuals with mental retardation: epidemiology and control measures**. *Ann Trop Med Parasitol* 2010, **104**:383–390.

9) Barwick RS, Uzicanin A, Lareau S, Malakmadze N, Imnadze P, Iosava M, Ninashvili N, Wilson M, Hightower AW, Johnston S, Bishop H, Petri Jr. WA, Juranek DD: **Outbreak of amebiasis in Tbilisi, Republic of Georgia, 1998.** *Am J Trop Med Hyg* 2002, **67**:623–631.

10) Stanley SL Jr: **Amoebiasis**. *The Lancet* 2003, **361**:1025–1034.

11) Savage DC: **Microbial ecology of the gastrointestinal tract.** *Annu Rev Microbiol* 1977, **31**:107–133.

12) Voigt H, Olivo JC, Sansonetti P, Guillén N: **Myosin IB from *Entamoeba histolytica* is involved in phagocytosis of human erythrocytes.** *Journal of Cell Science* 1999, **112 (Pt 8)**:1191–1201.

13) Santi-Rocca J, Rigothier MC, Guillén N: **Host-Microbe Interactions and Defense Mechanisms in the Development of Amoebic Liver Abscesses**. *Clin Microbiol Rev* 2009, **22**:65–75.

14) Davis PH, Zhang X, Guo J, Townsend RR, Stanley SL: **Comparative proteomic analysis of two *Entamoeba histolytica* strains with different virulence phenotypes identifies peroxiredoxin as an important component of amoebic virulence**. *Mol Microbiol* 2006, **61**:1523–1532.

15) Biller L, Davis PH, Tillack M, Matthiesen J, Lotter H, Stanley SL, Tannich E, Bruchhaus I: **Differences in the transcriptome signatures of two genetically related *Entamoeba histolytica* cell lines derived from the same isolate with different pathogenic properties.** *BMC Genomics* 2010, **11**:63.

16) Ayeh-Kumi PF, Ali IM, Lockhart LA, Gilchrist CA, Petri Jr. WA, Haque R: ***Entamoeba histolytica*: Genetic Diversity of Clinical Isolates from Bangladesh as Demonstrated by Polymorphisms in the Serine-Rich Gene**. *Experimental Parasitology* 2001, **99**:80–88.

17) Ali IKM, Mondal U, Roy S, Haque R, Petri Jr. WA, Clark CG: **Evidence for a Link between Parasite Genotype and Outcome of Infection with *Entamoeba histolytica***. *Journal of Clinical Microbiology* 2007, **45**:285–289.

18) Venable S, Peterson AM: **Unit VI - Pharmacotherapy for Gastrointestinal Tract Disorders: Parasitic Infections**. In *Pharmacotherapeutics for Advanced Practice: A Practical Approach*. *Volume 536*. 2nd edition. Edited by Peterson AM. Lippincott Williams and Wilkins; 2006:430–452.

19) McAuley JB, Herwaldt BL, Stokes SL, Becher JA, Roberts JM, Michelson MK, Juranek DD: **Diloxanide furoate for treating asymptomatic *Entamoeba histolytica* cyst passers: 14 years' experience in the United States.** *Clin Infect Dis* 1992, **15**:464–468.

20) World Health Organization: *WHO Model Prescribing Information*. 2nd edition. Geneva; 1995:3–12.

21) Salles JM, Moraes LA, Salles MC: **Hepatic Amoebiasis**. *The Brazilian Journal of Infectious Diseases* 2003, **7**:96–110.

22) McAuley JB, Juranek DD: **Luminal agents in the treatment of amebiasis.** *Clin Infect Dis* 1992, **14**:1161–1162.

23) Haque R, Ali IM, Sack RB, Farr BM, Ramakrishnan G, Petri Jr. WA: **Amebiasis and mucosal IgA antibody against the *Entamoeba histolytica* adherence lectin in Bangladeshi children.** *J Infect Dis* 2001, **183**:1787–1793.

24) Haque R, Mondal D, Duggal P, Kabir M, Roy S, Farr BM, Sack RB, Petri Jr. WA: ***Entamoeba histolytica* infection in children and protection from subsequent amebiasis.** *Infect Immun* 2006, **74**:904–909.

25) Haque R, Mondal D, Karim A, Molla IH, Rahim A, Faruque ASG, Ahmad N, Kirkpatrick BD, Houpt E, Snider C, Petri Jr. WA: **Prospective case-control study of the association between common enteric protozoal parasites and diarrhea in Bangladesh.** *Clin Infect Dis* 2009, **48**:1191–1197.

26) Royer TL, Gilchrist C, Kabir M, Arju T, Ralston KS, Haque R, Clark CG, Petri Jr. WA: ***Entamoeba bangladeshi*** **nov. sp., Bangladesh.** *Emerging Infect Dis* 2012, **18**:1543–1545.

27) Beck DL, Tanyuksel M, Mackey AJ, Haque R, Trapaidze N, Pearson WR, Loftus B, Petri Jr. WA: ***Entamoeba histolytica*****: sequence conservation of the Gal/GalNAc lectin from clinical isolates.** *Experimental Parasitology* 2002, **101**:157–163.

28) Blessmann J, Van Linh P, Nu PAT, Thi HD, Muller-Myhsok B, Buss H, Tannich E: **Epidemiology of amebiasis in a region of high incidence of amebic liver abscess in central Vietnam.** *Am J Trop Med Hyg* 2002, **66**:578–583.

29) Blessmann J, Le Van A, Tannich E: **Epidemiology and Treatment of Amebiasis in Hué, Vietnam**. *Archives of Medical Research* 2006, **37**:269–271.

30) Acuna-Soto R, Maguire JH, Wirth DF: **Gender distribution in asymptomatic and invasive amebiasis.** *Am J Gastroenterol* 2000, **95**:1277–1283.

31) Hamano S, Becker S, Asgharpour A, Ocasio YPR, Stroup SE, McDuffie M, Houpt E: **Gender and genetic control of resistance to intestinal amebiasis in inbred mice**. *Genes Immun* 2008, **9**:452–461.

32) Bansal D, Ave P, Kerneis S, Frileux P, Boché O, Baglin AC, Dubost G, Leguern A-S, Prevost M-C, Bracha R, Mirelman D, Guillén N, Labruyère E: **An *ex-vivo* Human Intestinal Model to Study *Entamoeba histolytica* Pathogenesis**. *PLoS Neglected Tropical Diseases* 2009, **3**:e551.

33) Ximénez C, Cerritos R, Rojas L, Dolabella S, Morán P, Shibayama M, González E, Valadez A, Hernández E, Valenzuela O, Limón A, Partida O, Silva EF: **Human Amebiasis: Breaking the Paradigm?** *Int J Environ Res Public Health* 2010, **7**:1105–1120.

34) Davis PH, Chen M, Zhang X, Clark CG, Townsend RR, Stanley SL: **Proteomic Comparison of *Entamoeba histolytica* and *Entamoeba dispar* and the Role of *E. histolytica* Alcohol Dehydrogenase 3 in Virulence**. *PLoS Neglected Tropical Diseases* 2009, **3**:e415.

35) Leitsch D, Wilson IB, Paschinger K, Duchene M: **Comparison of the proteome profiles of *Entamoeba histolytica* and its close but non-pathogenic relative *Entamoeba dispar***. *Wien Klin Wochenschr* 2006, **118**:37–41.

36) Tshalaia LE: **On a species of *Entamoeba* detected in sewage effluents [in Russian]**. *Med Parasit* 1941, **10**:244.

37) Neal RA: **Studies on the morphology and biology of *Entamoeba moshkovskii* Tshalaia, 1941.** *Parasitology* 1953, **43**:253–268.

38) Clark CG, Diamond LS: **Intraspecific variation and phylogenetic relationships in the genus *Entamoeba* as revealed by riboprinting.** *J Eukaryot Microbiol* 1997, **44**:142–154.

39) Fotedar R, Stark D, Beebe N, Marriott D, Ellis J, Harkness J: **Laboratory diagnostic techniques for *Entamoeba* species.** *Clin Microbiol Rev* 2007, **20**:511–532.

40) Fotedar R, Stark D, Beebe N, Marriott D, Ellis J, Harkness J: **PCR detection of *Entamoeba histolytica*, *Entamoeba dispar*, and *Entamoeba moshkovskii* in stool samples from Sydney, Australia.** *Journal of Clinical Microbiology* 2007, **45**:1035–1037.

41) Hamzah Z, Petmitr S, Mungthin M, Leelayoova S, Chavalitshewinkoon-Petmitr P: **Differential detection of *Entamoeba histolytica*, *Entamoeba dispar*, and *Entamoeba moshkovskii* by a single-round PCR assay.** *Journal of Clinical Microbiology* 2006, **44**:3196–3200.

42) El-Bakri A, Samie A, Ezzedine S, Odeh RA: **Differential detection of *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* in fecal samples by nested PCR in the United Arab Emirates (UAE).** *Acta Parasitol* 2013, **58**:185–190.

43) Nazemalhosseini Mojarad E, Nochi Z, Sahebekhtiari N, Rostami Nejad M, Dabiri H, Zali MR, Kazemi B, Haghighi A: **Discrimination of *Entamoeba moshkovskii* in patients with gastrointestinal disorders by single-round PCR.** *Jpn J Infect Dis* 2010, **63**:136–138.

44) Fotedar R, Stark D, Marriott D, Ellis J, Harkness J: ***Entamoeba moshkovskii* infections in Sydney, Australia**. *Eur J Clin Microbiol Infect Dis* 2007, **27**:133–137.

45) Weedall GD, Hall N: **Evolutionary genomics of *Entamoeba***. *Research in Microbiologoy* 2011, **162**:637–645.

46) Stensvold CR, Lebbad M, Victory EL, Verweij JJ, Tannich E, Alfellani M, Legarraga P, Clark CG: **Increased Sampling Reveals Novel Lineages of *Entamoeba*: Consequences of Genetic Diversity and Host Specificity for Taxonomy and Molecular Detection**. *Annals of Anatomy* 2011, **162**:525–541.

47) Balamuth W: **Effects of some environmental factors upon growth and encystation of *Entamoeba invadens*.** *J Parasitol* 1962, **48**:101–109.

48) Wang Z, Samuelson J, Clark CG, Eichinger D, Paul J, Van Dellen K, Hall N, Anderson I, Loftus B: **Gene discovery in the *Entamoeba invadens* genome**. *Molecular and Biochemical Parasitology* 2003, **129**:23–31.

49) Loftus B, Anderson I, Davies R, Alsmark UCM, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, Nozaki T, Suh B, Pop M, Duchene M, Ackers J, Tannich E, Leippe M, Hofer M, Bruchhaus I, Willhoeft U, Bhattacharya A, Chillingworth T, Churcher C, Hance Z, Harris B, Harris D, Jagels K, Moule S, Mungall K, Ormond D, et al.: **The genome of the protist parasite *Entamoeba histolytica*.** *Nature* 2005, **433**:865–868.

50) Clark CG, Alsmark UCM, Tazreiter M, Saito Nakano Y, Ali V, Marion S, Weber C, Mukherjee C, Bruchhaus I, Tannich E, Leippe M, Sicheritz Ponten T, Foster PG, Samuelson J, Noël CJ, Hirt RP, Embley TM, Gilchrist CA, Mann BJ, Singh U, Ackers JP, Bhattacharya S, Bhattacharya A, Lohia A, Guillén N, Duchêne M, Nozaki T, Hall N: **Structure and Content of the *Entamoeba histolytica* Genome**. *Advances in Parasitology* 2007 **65**:51-190.

51) Lorenzi HA, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, Caler EV: **New Assembly, Reannotation and Analysis of the *Entamoeba histolytica* Genome Reveal New Genomic Features and Protein Content Information**. *PLoS Neglected Tropical Diseases* 2010, **4**:e716.

52) Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C: **Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly.** *PLoS ONE* 2013, **8**:e62856.

53) Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36**:e105.

54) Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ: **Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.** *Nat Methods* 2009, **6**:291–295.

55) Lorenzi H, Thiagarajan M, Haas B, Wortman J, Hall N, Caler E: **Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species**. *BMC Genomics* 2008, **9**:595.

56) Ostertag EM, Kazazian HH: **Biology of mammalian L1 retrotransposons.** *Annu Rev Genet* 2001, **35**:501–538.

57) Bakre AA, Rawal K, Ramaswamy R, Bhattacharya A, Bhattacharya S: **The LINEs and SINEs of *Entamoeba histolytica*: comparative analysis and genomic distribution.** *Experimental Parasitology* 2005, **110**:207–213.

58) Mandal PK, Bagchi A, Bhattacharya A, Bhattacharya S: **An *Entamoeba histolytica* LINE/SINE pair inserts at common target sites cleaved by the restriction enzyme-like LINE-encoded endonuclease.** *Eukaryotic Cell* 2004, **3**:170–179.

59) Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song X-Z, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452**:872–876.

60) Alkan C, Sajjadian S, Eichler EE: **Limitations of next-generation genome sequence assembly.** *Nat Methods* 2011, **8**:61–65.

61) Salzberg SL, Yorke JA: **Beware of mis-assembled genomes.** *Bioinformatics* 2005, **21**:4320–4321.

62) Clark CG, Ali IKM, Zaki M, Loftus BJ, Hall N: **Unique organisation of tRNA genes in *Entamoeba histolytica***. *Molecular and Biochemical Parasitology* 2006, **146**:24–29.

63) Tawari B, Ali IKM, Scott C, Quail MA, Berriman M, Hall N, Clark CG: **Patterns of Evolution in the Unique tRNA Gene Arrays of the Genus *Entamoeba***. *Molecular Biology and Evolution* 2007, **25**:187–198.

64) Willhoeft U, Tannich E: **The electrophoretic karyotype of *Entamoeba histolytica*.** *Molecular and Biochemical Parasitology* 1999, **99**:41–53.

65) Ghosh S, Frisardi M, Ramirez-Avila L, Descoteaux S, Sturm-Ramirez K, Newton-Sanchez OA, Santos-Preciado JI, Ganguly C, Lohia A, Reed S, Samuelson J: **Molecular epidemiology of *Entamoeba* spp.: evidence of a bottleneck (Demographic sweep) and transcontinental spread of diploid parasites.** *Journal of Clinical Microbiology* 2000, **38**:3815–3821.

66) Mukherjee C, Clark CG, Lohia A: ***Entamoeba* shows reversible variation in ploidy under different growth conditions and between life cycle phases.** *PLoS Neglected Tropical Diseases* 2008, **2**:e281.

67) Mackenstedt U, Schmidt M, Raether W, Mehlhorn H, Uphoff M: **Increase of DNA content in tissue stages of *Entamoeba histolytica* strain SFL 3.** *Parasitol Res* 1990, **76**:373–378.

68) Weedall GD, Clark CG, Koldkjaer P, Kay S, Bruchhaus I, Tannich E, Paterson S, Hall N: **Genomic diversity of the human intestinal parasite *Entamoeba histolytica***. *Genome Biology* 2012, **13**:R38.

69) Bhattacharya A, Satish S, Bagchi A, Bhattacharya S: **The genome of *Entamoeba histolytica*.** *International Journal For Parasitology* 2000, **30**:401–410.

70) Lejeune M, Rybicka JM, Chadee K: **Recent discoveries in the pathogenesis and immune response toward *Entamoeba histolytica***. *Future Microbiology* 2009, **4**:105–118.

71) Mortimer L, Chadee K: **The immunopathogenesis of *Entamoeba histolytica***. *Experimental Parasitology* 2010, **126**:366–380.

72) Wilson IW, Weedall GD, Hall N: **Host-Parasite interactions in *Entamoeba histolytica* and *Entamoeba dispar*: what have we learned from their genomes?** *Parasite Immunology* 2012, **34**:90–99.

73) Casados-Vázquez LE, Lara-González S, Brieba LG: **Crystal structure of the cysteine protease inhibitor 2 from *Entamoeba histolytica*: functional convergence of a common protein fold.** *Gene* 2011, **471**:45–52.

74) Bruchhaus I, Jacobs T, Leippe M, Tannich E: ***Entamoeba histolytica* and *Entamoeba dispar*: differences in numbers and expression of cysteine proteinase genes.** *Mol Microbiol* 1996, **22**:255–263.

75) Ankri S, Stolarsky T, Bracha R, Padilla-Vaca F, Mirelman D: **Antisense inhibition of expression of cysteine proteinases affects *Entamoeba histolytica*-induced formation of liver abscess in hamsters.** *Infect Immun* 1999, **67**:421–422.

76) Stanley SL, Zhang T, Rubin D, Li E: **Role of the *Entamoeba histolytica* cysteine proteinase in amebic liver abscess formation in severe combined immunodeficient mice.** *Infect Immun* 1995, **63**:1587–1590.

77) Meléndez-López SG, Herdman S, Hirata K, Choi M-H, Choe Y, Craik C, Caffrey CR, Hansell E, Chávez-Munguía B, Chen YT, Roush WR, McKerrow J, Eckmann L, Guo J, Stanley SL, Reed SL: **Use of recombinant *Entamoeba histolytica* cysteine proteinase 1 to identify a potent inhibitor of amebic invasion in a human colonic model.** *Eukaryotic Cell* 2007, **6**:1130–1136.

78) Jacobs T, Bruchhaus I, Dandekar T, Tannich E, Leippe M: **Isolation and molecular characterization of a surface-bound proteinase of *Entamoeba histolytica*.** *Mol Microbiol* 1998, **27**:269–276.

79) Moncada D, Keller K, Chadee K: ***Entamoeba histolytica* cysteine proteinases disrupt the polymeric structure of colonic mucin and alter its protective function.** *Infect Immun* 2003, **71**:838–844.

80) Moncada D, Keller K, Chadee K: ***Entamoeba histolytica*-secreted products degrade colonic mucin oligosaccharides.** *Infect Immun* 2005, **73**:3790–3793.

81) Petri Jr. WA, Haque R, Mann BJ: **The Bittersweet Interface of Parasite and Host: Lectin-Carbohydrate Interactions During Human Invasion by the Parasite *Entamoeba histolytica***. *Annu Rev Microbiol* 2002, **56**:39–64.

82) Li E, Becker A, Stanley SL: **Use of Chinese hamster ovary cells with altered glycosylation patterns to define the carbohydrate specificity of *Entamoeba histolytica* adhesion.** *J Exp Med* 1988, **167**:1725–1730.

83) Li E, Becker A, Stanley SL: **Chinese hamster ovary cells deficient in N-acetylglucosaminyltransferase I activity are resistant to *Entamoeba histolytica*-mediated cytotoxicity.** *Infect Immun* 1989, **57**:8–12.

84) Ravdin JI, Croft BY, Guerrant RL: **Cytopathogenic mechanisms of *Entamoeba histolytica*.** *J Exp Med* 1980, **152**:377–390.

85) Ravdin JI, Stanley P, Murphy CF, Petri Jr. WA: **Characterization of cell surface carbohydrate receptors for *Entamoeba histolytica* adherence lectin.** *Infect Immun* 1989, **57**:2179–2186.

86) MacFarlane RC, Singh U: **Identification of an *Entamoeba histolytica* Serine-, Threonine-, and Isoleucine-Rich Protein with Roles in Adhesion and Cytotoxicity**. *Eukaryotic Cell* 2007, **6**:2139–2146.

87) Maldonado C, Trejo W, Ramírez A, Carrera M, Sánchez J, López-Macías C, Isibasi A: **Lipophosphopeptidoglycan of *Entamoeba histolytica* induces an antiinflammatory innate immune response and downregulation of toll-like receptor 2 (TLR-2) gene expression in human monocytes.** *Archives of Medical Research* 2000, **31**:S71–3.

88) Maldonado-Bernal C, Kirschning CJ, Rosenstein Y, Rocha LM, Rios-Sarabia N, Espinosa-Cantellano M, Becker I, Estrada I, Salazar-González RM, López-Macías C, Wagner H, Sánchez J, Isibasi A: **The innate immune response to *Entamoeba histolytica* lipopeptidophosphoglycan is mediated by toll-like receptors 2 and 4.** *Parasite Immunology* 2005, **27**:127–137.

89) Guo X, Houpt E, Petri Jr. WA: **Crosstalk at the initial encounter: interplay between host defense and ameba survival strategies**. *Current Opinion in Immunology* 2007, **19**:376–384.

90) Bruchhaus I, Tannich E: **Induction of the iron-containing superoxide dismutase in *Entamoeba histolytica* by a superoxide anion-generating system or by iron chelation.** *Molecular and Biochemical Parasitology* 1994, **67**:281–288.

91) Choi M-H, Sajed D, Poole L, Hirata K, Herdman S, Torian BE, Reed SL: **An unusual surface peroxiredoxin protects invasive *Entamoeba histolytica* from oxidant attack**. *Molecular and Biochemical Parasitology* 2005, **143**:80–89.

92) Lo H, Reeves RE: **Purification and properties of NADPH:flavin oxidoreductase from *Entamoeba histolytica*.** *Molecular and Biochemical Parasitology* 1980, **2**:23–30.

93) Fournier M, Dermoun Z, Durand MC, Dolla A: **A New Function of the *Desulfovibrio vulgaris* Hildenborough [Fe] Hydrogenase in the Protection against Oxidative Stress**. *Journal of Biological Chemistry* 2004, **279**:1787–1793.

94) MacFarlane RC, Singh U: **Identification of differentially expressed genes in virulent and nonvirulent *Entamoeba* species: potential implications for amebic pathogenesis.** *Infect Immun* 2006, **74**:340–351.

95) Kelsall BL, Ravdin JI: **Degradation of human IgA by *Entamoeba histolytica*.** *J Infect Dis* 1993, **168**:1319–1322.

96) Garcia-Nieto RM, Rico-Mata R, Arias-Negrete S, Avila EE: **Degradation of human secretory IgA1 and IgA2 by *Entamoeba histolytica* surface-associated proteolytic activity**. *Parasitology International* 2008, **57**:417–423.

97) Carrero JC, Díaz MY, Viveros M, Espinoza B, Acosta E, Ortiz-Ortiz L: **Human secretory immunoglobulin A anti-*Entamoeba histolytica* antibodies inhibit adherence of amebae to MDCK cells.** *Infect Immun* 1994, **62**:764–767.

98) Baxt LA, Baker RP, Singh U, Urban S: **An *Entamoeba histolytica* rhomboid protease with atypical specificity cleaves a surface lectin involved in phagocytosis and immune evasion**. *Genes & Development* 2008, **22**:1636–1646.

99) Espinosa-Cantellano M, Martínez-Palomo A: ***Entamoeba histolytica*: mechanism of surface receptor capping.** *Experimental Parasitology* 1994, **79**:424–435.

100) Arhets P, Gounon P, Sansonetti P, Guillén N: **Myosin II is involved in capping and uroid formation in the human pathogen *Entamoeba histolytica*.** *Infect Immun* 1995, **63**:4358–4367.

101) Bennett S: **Solexa Ltd.** *Pharmacogenomics* 2004, **5**:433–438.

102) Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53–59.

103) Metzker ML: **Sequencing technologies - the next generation.** *Nat Rev Genet* 2010, **11**:31–46.

104) Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, **24**:133–141.

105) Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M: **Comparison of next-generation sequencing systems.** *Journal of Biomedicine and Biotechnology* 2012, **2012**:251364.

106) Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M: **Performance comparison of whole-genome sequencing platforms.** *Nat Biotechnol* 2012, **30**:78–82.

107) Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms.** *Nat Biotechnol* 2012, **30**:434–439.

108) Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB: **The real cost of sequencing: higher than you think!** *Genome Biology* 2011, **12**:125.

109) Schatz MC, Delcher AL, Salzberg SL: **Assembly of large genomes using second-generation sequencing.** *Genome Research* 2010, **20**:1165–1173.

110) Hall N: **After the gold rush**. *Genome Biology* 2013, **14**:115.

111) Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung W-K, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, *et al.*: **Assemblathon 1: a competitive assessment of *de novo* short read assembly methods.** *Genome Research* 2011, **21**:2224–2241.

112) Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou W-C, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, *et al.*: **Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species.** *Gigascience* 2013, **2**:10.

113) Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y: **A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.** *BMC Genomics* 2012, **13**:341.

114) Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23**:1061–1067.

115) Birney E, Clamp M, Durbin R: **GeneWise and Genomewise.** *Genome Research* 2004, **14**:988–995.

116) Parra G, Blanco E, Guigó R: **GeneID in *Drosophila.*** *Genome Research* 2000, **10**:511–515.

117) Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268**:78–94.

118) Haubold B, Wiehe T: **Gene Prediction**. In *Introduction to Computational Biology: An Evolutionary Approach*. Birkhäuser, Basel; 2006:117–140.

119) Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Research* 2004, **14**:942–950.

120) Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19**:ii215–ii225.

121) Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW: **Extensive error in the number of genes inferred from draft genome assemblies.** *PLoS Comput Biol* 2014, **10**:e1003998.

122) Yandell M, Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nat Rev Genet* 2012, **13**:329–342.

123) Denoeud F, Aury J-M, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F: **Annotating genomes with massive-scale RNA sequencing.** *Genome Biology* 2008, **9**:R175.

124) Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**:470–476.

125) Baker M: ***De novo* genome assembly: what every biologist should know**. *Nat Methods* 2012, **9**:333–337.

126) Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621–628.

127) Reid AJ, Vermont SJ, Cotton JA, Harris D, Hill-Cawthorne GA, Könen-Waisman S, Latham SM, Mourier T, Norton R, Quail MA, Sanders M, Shanmugam D, Sohal A, Wasmuth JD, Brunk B, Grigg ME, Howard JC, Parkinson J, Roos DS, Trees AJ, Berriman M, Pain A, Wastling JM: **Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy.** *PLoS Pathog* 2012, **8**:e1002567.

128) El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, Ghedin E, Peacock C, Bartholomeu DC, Haas BJ, Tran A-N, Wortman JR, Alsmark UCM, Angiuoli S, Anupama A, Badger J, Bringaud F, Cadag E, Carlton JM, Cerqueira GC, Creasy T, Delcher AL, Djikeng A, Embley TM, Hauser C, Ivens AC, *et al.*: **Comparative genomics of trypanosomatid parasitic protozoa.** *Science* 2005, **309**:404–409.

129) Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, Crabtree J, Angiuoli SV, Merino EF, Amedeo P, Cheng Q, Coulson RMR, Crabb BS, Del Portillo HA, Essien K, Feldblyum TV, Fernandez-Becerra C, Gilson PR, Gueye AH, Guo X, Kang'a S, Kooij TWA, Korsinczky M, Meyer EV-S, Nene V, Paulsen I, White O, Ralph SA, Ren Q, Sargeant TJ, *et al.*: **Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*.** *Nature* 2008, **455**:757–763.

130) Tenter AM, Heckeroth AR, Weiss LM: ***Toxoplasma gondii*: from animals to humans.** *International Journal For Parasitology* 2000, **30**:1217–1258.

131) McAllister MM, Dubey JP, Lindsay DS, Jolley WR, Wills RA, McGuire AM: **Dogs are definitive hosts of *Neospora caninum*.** *International Journal For Parasitology* 1998, **28**:1473–1478.

132) King JS, Slapeta J, Jenkins DJ, Al-Qassab SE, Ellis JT, Windsor PA: **Australian dingoes are definitive hosts of *Neospora caninum*.** *International Journal For Parasitology* 2010, **40**:945–950.

133) Gondim LFP, McAllister MM, Pitt WC, Zemlicka DE: **Coyotes (*Canis latrans*) are definitive hosts of *Neospora caninum*.** *International Journal For Parasitology* 2004, **34**:159–161.

134) Pays E, Vanhamme L, Pérez-Morga D: **Antigenic variation in *Trypanosoma brucei*: facts, challenges and mysteries.** *Curr Opin Microbiol* 2004, **7**:369–374.

135) Eichinger L, Pachebat JA, Glöckner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, HALL N, Anjard C, *et al.*: **The genome of the social amoeba *Dictyostelium discoideum*.** *Nature* 2005, **435**:43–57.

136) Clarke M, Lohan AJ, Liu B, Lagkouvardos I, Roy S, Zafar N, Bertelli C, Schilde C, Kianianmomeni A, Bürglin TR, Frech C, Turcotte B, Kopec KO, Synnott JM, Choo C, Paponov I, Finkler A, Heng Tan CS, Hutchins AP, Weinmeier T, Rattei T, Chu JS, Gimenez G, Irimia M, Rigden DJ, Fitzpatrick DA, Lorenzo-Morales J, Bateman A, Chiu C-H, Tang P, *et al.*: **Genome of *Acanthamoeba castellanii* highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling.** *Genome Biology* 2013, **14**:R11.

137) Anderson IJ, Watkins RF, Samuelson J, Spencer DF, Majoros WH, Gray MW, Loftus BJ: **Gene discovery in the *Acanthamoeba castellanii* genome.** *Annals of Anatomy* 2005, **156**:203–214.

138) Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG: **Life with 6000 genes.** *Science* 1996, **274**:563–567.

139) van Valen L: **A new evolutionary law**. *Evolutionary Theory* 1973, **1**:1–30.

140) Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, Thomson NR, Quail M, Smith F, Walker D, Libberton B, Fenton A, Hall N, Brockhurst MA: **Antagonistic coevolution accelerates molecular evolution**. *Nature* 2010, **464**:275–278.

141) Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs.** *Genome Research* 2008, **18**:821–829.

142) Boisvert S, Laviolette F, Corbeil J: **Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies.** *J Comput Biol* 2010, **17**:1519–1533.

143) Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J: **SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler.** *Gigascience* 2012, **1**:18.

144) Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Research* 2009, **19**:1117–1123.

145) Entner N, Most H: **Genetics of *Entamoeba*: Characterization of Two New Parasitic Strains Which Grow at Room Temperature (and at 37 Degrees C).** *J Protozool* 1965, **12**:10–13.

146) Clark CG, Diamond LS: **The Laredo strain and other *"Entamoeba histolytica*-like" amoebae are *Entamoeba moshkovskii***. *Molecular and Biochemical Parasitology* 1991, **46**:11–18.

147) Dreyer DA: **Growth of a strain of *Entamoeba histolytica* at room temperature.** *Tex Rep Biol Med* 1961, **19**:393–396.

148) Goldman M: ***Entamoeba histolytica*-like amoebae occurring in man.** *Bull World Health Organ* 1969, **40**:355–364.

149) Khairnar K, Parija SC: **A novel nested multiplex polymerase chain reaction (PCR) assay for differential detection of *Entamoeba histolytica*, *E. moshkovskii* and *E. dispar* DNA in stool samples.** *BMC Microbiol* 2007, **7**:47.

150) Ayed SB, Aoun K, Maamouri N, Abdallah RB, Bouratbine A: **First molecular identification of *Entamoeba moshkovskii* in human stool samples in Tunisia.** *Am J Trop Med Hyg* 2008, **79**:706–707.

151) Lau YL, Anthony C, Fakhrurrazi SA, Ibrahim J, Ithoi I, Mahmud R: **Real-time PCR assay in differentiating *Entamoeba histolytica*, *Entamoeba dispar*, and *Entamoeba moshkovskii* infections in Orang Asli settlements in Malaysia.** *Parasit Vectors* 2013, **6**:250.

152) Heredia RD, Fonseca JA, López MC: ***Entamoeba moshkovskii* perspectives of a new agent to be considered in the diagnosis of amebiasis**. *Acta Tropica* 2012, **123**:139–145.

153) Fitch WM: **Distinguishing homologous from analogous proteins.** *Syst Zool* 1970, **19**:99–113.

154) Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM, Farmer AD, Sheridan J, Iwata A, Tuteja R, Penmetsa RV, Wu W, Upadhyaya HD, Yang S-P, Shah T, Saxena KB, Michael T, McCombie WR, Yang B, Zhang G, Yang H, Wang J, Spillane C, Cook DR, May GD, Xu X, *et al.*: **Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers.** *Nat Biotechnol* 2012, **30**:83–89.

155) Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B, Millan T, Zhang X, Ramsay LD, Iwata A, Wang Y, Nelson W, Farmer AD, Gaur PM, Soderlund C, Penmetsa RV, Xu C, Bharti AK, He W, Winter P, Zhao S, Hane JK, Carrasquilla-Garcia N, Condie JA, Upadhyaya HD, Luo M-C, *et al.*: **Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement.** *Nat Biotechnol* 2013, **31**:240–246.

156) Diguistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, Birol I, Holt RA, Hirst M, Mardis E, Marra MA, Hamelin RC, Bohlmann J, Breuil C, Jones SJ: ***De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data.** *Genome Biology* 2009, **10**:R94.

157) Zhan S, Merlin C, Boore JL, Reppert SM: **The monarch butterfly genome yields insights into long-distance migration.** *Cell* 2011, **147**:1171–1185.

158) Clark CG, Diamond LS: **Methods for cultivation of luminal parasitic protists of clinical importance.** *Clin Microbiol Rev* 2002, **15**:329–341.

159) Diamond LS, Cunnick CC: **A serum-free, partly defined medium, PDM-805, for axenic cultivation of *Entamoeba histolytica* Schaudinn, 1903 and other *Entamoeba*.** *J Protozool* 1991, **38**:211–216.

160) Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437**:376–380.

161) Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.

162) Hon C-C, Weber C, Sismeiro O, Proux C, Koutero M, Deloger M, Das S, Agrahari M, Dillies M-A, Jagla B, Coppée J-Y, Bhattacharya A, Guillén N: **Quantification of stochastic noise of splicing and polyadenylation in *Entamoeba histolytica*.** *Nucleic Acids Res* 2013, **41**:1936–1952.

163) Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* 2009, **10**:R25.

164) Dimon MT, Sorber K, DeRisi JL: **HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data.** *PLoS ONE* 2010, **5**:e13875.

165) Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA: **Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data.** *Bioinformatics* 2012, **28**:464–469.

166) Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16**:944–945.

167) Parra G, Bradnam K, Ning Z, Keane T, Korf I: **Assessing the gene space in draft genomes.** *Nucleic Acids Res* 2009, **37**:289–297.

168) Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.

169) Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC, Kraemer ET, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Srinivasamoorthy G, Stoeckert CJ, Thibodeau R, Treatman C, Wang H: **EuPathDB: a portal to eukaryotic pathogen databases.** *Nucleic Acids Res* 2010, **38**(Database issue):D415–D419.

170) Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler EV, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Iodice J, Kissinger JC, Kraemer ET, Li W, Nayak V, Pennington C, Pinney DF, Pitts B, Roos DS, Srinivasamoorthy G, Stoeckert CJ, Treatman C, Wang H: **AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species.** *Nucleic Acids Res* 2011, **39**(Database issue):D612–D619.

171) Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups**. *Nucleic Acids Res* 2006, **34**:D363–D368.

172) Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–29.

173) Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**:3674–3676.

174) Zhi-Liang H, Bao J, Reecy JM: **CateGOrizer: A Web-Based Program to Batch Analyze Gene Ontology Classification Categories**. *Online J Bioinformatics* 2008, **9**(2):108-112.

175) Kreppel L, Fey P, Gaudet P, Just E, Kibbe WA, Chisholm RL, Kimmel AR: **dictyBase: a new *Dictyostelium discoideum* genome database.** *Nucleic Acids Res* 2004, **32**(Database issue):D332–333.

176) Basu S, Fey P, Pandit Y, Dodson R, Kibbe WA, Chisholm RL: **DictyBase 2013: integrating multiple Dictyostelid species.** *Nucleic Acids Res* 2013, **41**(Database issue):D676–683.

177) Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED: ***Saccharomyces*** **Genome Database: the genomics resource of budding yeast.** *Nucleic Acids Res* 2012, **40**(Database issue):D700–705.

178) Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, Weng S, Wong ED, Lloyd P, Skrzypek MS, Miyasato SR, Simison M, Cherry JM: **The reference genome sequence of *Saccharomyces cerevisiae*: then and now.** *G3 (Bethesda)* 2014, **4**:389–398.

179) Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792–1797.

180) Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2)**. *Cladistics* 1989, **5**:164–166.

181) Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8**:275–282.

182) Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Molecular Biology and Evolution* 2011, **28**:2731–2739.

183) Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7**:562–578.

184) Sipos B, Massingham T, Stütz AM, Goldman N: **An improved protocol for sequencing of repetitive genomic regions and structural variations using mutagenesis and next generation sequencing.** *PLoS ONE* 2012, **7**:e43359.

185) Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nat Rev Genet* 2012, **13**:36–46.

186) Kumari V, Sharma R, Yadav VP, Gupta AK, Bhattacharya A, Bhattacharya S: **Differential distribution of a SINE element in the *Entamoeba histolytica* and *Entamoeba dispar* genomes: role of the LINE-encoded endonuclease.** *BMC Genomics* 2011, **12**:267.

187) Shire AM, Ackers JP: **SINE elements of *Entamoeba dispar*.** *Molecular and Biochemical Parasitology* 2007, **152**:47–52.

188) Willhoeft U, Buss H, Tannich E: **The abundant polyadenylated transcript 2 DNA sequence of the pathogenic protozoan parasite *Entamoeba histolytica* represents a nonautonomous non-long-terminal-repeat retrotransposon-like element which is absent in the closely related nonpathogenic species *Entamoeba dispar*.** *Infect Immun* 2002, **70**:6798–6804.

189) Clark CG: **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* 2006, **56**:2235–2239.

190) Kelley DR, Salzberg SL: **Detection and correction of false segmental duplications caused by genome mis-assembly.** *Genome Biology* 2010, **11**:R28.

191) Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, Davis RW, Scherer S: **The diploid genome sequence of *Candida albicans*.** *Proc Natl Acad Sci USA* 2004, **101**:7329–7334.

192) Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JMC, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, *et al.*: **The genome sequence of the malaria mosquito *Anopheles gambiae*.** *Science* 2002, **298**:129–149.

193) Cheung J, Wilson MD, Zhang J, Khaja R, MacDonald JR, Heng HHQ, Koop BF, Scherer SW: **Recent segmental and gene duplications in the mouse genome.** *Genome Biology* 2003, **4**:R47.

194) Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE: **Analysis of segmental duplications and genome assembly in the mouse.** *Genome Research* 2004, **14**:789–801.

195) Li Z, Zhang Z, Yan P, Huang S, Fei Z, Lin K: **RNA-Seq improves annotation of protein-coding genes in the cucumber genome.** *BMC Genomics* 2011, **12**:540.

196) Li L, Chen E, Yang C, Zhu J, Jayaraman P, De Pons J, Kaczorowski CC, Jacob HJ, Greene AS, Hodges MR, Cowley AW, Liang M, Xu H, Liu P, Lu Y: **Improved rat genome gene prediction by integration of ESTs with RNA-Seq information.** *Bioinformatics* 2015, **31**:25–32.

197) Devisetty UK, Covington MF, Tat AV, Lekkala S, Maloof JN: **Polymorphism identification and improved genome annotation of *Brassica rapa* through Deep RNA sequencing.** *G3 (Bethesda)* 2014, **4**:2065–2078.

198) Andersson JO, Andersson SG: **Insights Into the Evolutionary Process of Genome Degradation**. *Current Opinion in Genetics & Development* 1999, **9**:664–671.

199) Davis PH, Zhang Z, Chen M, Zhang X, Chakraborty S, Stanley SL: **Identification of a family of BspA like surface proteins of *Entamoeba histolytica* with novel leucine rich repeats.** *Molecular and Biochemical Parasitology* 2006, **145**:111–116.

200) Sharma A, Sojar HT, Glurich I, Honma K, Kuramitsu HK, Genco RJ: **Cloning, expression, and sequencing of a cell surface antigen containing a leucine-rich repeat motif from *Bacteroides forsythus* ATCC 43037.** *Infect Immun* 1998, **66**:5703–5710.

201) Hirt RP, Harriman N, Kajava AV, Embley TM: **A novel potential surface protein in *Trichomonas vaginalis* contains a leucine-rich repeat shared by micro-organisms from all three domains of life.** *Molecular and Biochemical Parasitology* 2002, **125**:195–199.

202) Noël CJ, Diaz N, Sicheritz-Ponten T, Safarikova L, Tachezy J, Tang P, Fiori P-L, Hirt RP: ***Trichomonas vaginalis* vast BspA-like gene family: evidence for functional diversity from structural organisation and transcriptomics.** *BMC Genomics* 2010, **11**:99.

203) Willhoeft U, Buβ H, Tannich E: **DNA sequences corresponding to the ariel gene family of *Entamoeba histolytica* are not present in *E. dispar***. *Parasitol Res* 1999, **85**:787–789.

204) Mai Z, Samuelson J: **A new gene family (ariel) encodes asparagine-rich *Entamoeba histolytica* antigens, which resemble the amebic vaccine candidate serine-rich *E. histolytica* protein.** *Infect Immun* 1998, **66**:353–355.

205) Zhang T, Cieslak PR, Stanley SL: **Protection of gerbils from amebic liver abscess by immunization with a recombinant *Entamoeba histolytica* antigen.** *Infect Immun* 1994, **62**:1166–1170.

206) Reuber TL, Ausubel FM: **Isolation of *Arabidopsis* genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes.** *Plant Cell* 1996, **8**:241–249.

207) Gilchrist CA, Houpt E, Trapaidze N, Fei Z, Crasta O, Asgharpour A, Evans C, Martino-Catt S, Baba DJ, Stroup S, Hamano S, Ehrenkaufer G, Okada M, Singh U, Nozaki T, Mann BJ, Petri Jr. WA: **Impact of intestinal colonization and invasion on the *Entamoeba histolytica* transcriptome**. *Molecular and Biochemical Parasitology* 2006, **147**:163–176.

208) Lidell ME, Moncada DM, Chadee K, Hansson GC: ***Entamoeba histolytica* cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel.** *Proc Natl Acad Sci USA* 2006, **103**:9298–9303.

209) Poole LB, Chae HZ, Flores BM, Reed SL, Rhee SG, Torian BE: **Peroxidase activity of a TSA-like antioxidant protein from a pathogenic amoeba.** *Free Radic Biol Med* 1997, **23**:955–959.

210) Singh LN, Hannenhalli S: **Functional diversification of paralogous transcription factors via divergence in DNA binding site motif and in expression.** *PLoS ONE* 2008, **3**:e2345.

211) Zhang J: **Evolution by gene duplication: an update**. *TRENDS in Ecology and Evolution* 2003, **18**:292–298.

212) Nadeau JH, Sankoff D: **Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution.** *Genetics* 1997, **147**:1259–1266.

213) Goeman JJ, Mansmann U: **Multiple testing on the directed acyclic graph of gene ontology.** *Bioinformatics* 2008, **24**:537–544.

214) Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, Alaoui El H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivarès CP: **Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*.** *Nature* 2001, **414**:450–453.

215) Sakharkar KR, Dhar PK, Chow VTK: **Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis.** *International Journal of Systematic and Evolutionary Microbiology* 2004, **54**:1937–1941.

216) Lahr DJG, Grant J, Nguyen T, Lin JH, Katz LA: **Comprehensive phylogenetic reconstruction of amoebozoa based on concatenated analyses of SSU-rDNA and actin genes.** *PLoS ONE* 2011, **6**:e22780.

217) Glöckner G, Noegel AA: **Comparative genomics in the Amoebozoa clade.** *Biol Rev Camb Philos Soc* 2013, **88**:215–225.

218) Wennerberg K, Rossman KL, Der CJ: **The Ras superfamily at a glance.** *Journal of Cell Science* 2005, **118**:843–846.

219) Orozco E, Guarneros G, Martínez-Palomo A, Sánchez T: ***Entamoeba histolytica*. Phagocytosis as a virulence factor.** *J Exp Med* 1983, **158**:1511–1521.

220) Tannich E, Scholze H, Nickel R, Horstmann RD: **Homologous cysteine proteinases of pathogenic and nonpathogenic *Entamoeba histolytica*. Differences in structure and expression.** *J Biol Chem* 1991, **266**:4798–4803.

221) Braga LL, Ninomiya H, McCoy JJ, Eacker S, Wiedmer T, Pham C, Wood S, Sims PJ, Petri Jr. WA: **Inhibition of the complement membrane attack complex by the galactose-specific adhesion of *Entamoeba histolytica*.** *J Clin Invest* 1992, **90**:1131–1137.

222) Ravdin JI, Guerrant RL: **Role of adherence in cytopathogenic mechanisms of *Entamoeba histolytica*. Study with mammalian tissue culture cells and human erythrocytes.** *J Clin Invest* 1981, **68**:1305–1313.

223) Arias DG, Gutierrez CE, Iglesias AA, Guerrero SA: **Thioredoxin-linked metabolism in *Entamoeba histolytica*.** *Free Radic Biol Med* 2007, **42**:1496–1505.

224) Arnér ES, Holmgren A: **Physiological functions of thioredoxin and thioredoxin reductase.** *Eur J Biochem* 2000, **267**:6102–6109.

225) Beutner EH, Jordon RE, Chorzelski TP: **The immunopathology of pemphigus and bullous pemphigoid.** *J Invest Dermatol* 1968, **51**:63–80.

226) Cai SCS, Allen JC, Lim YL, Chua SH, Tan SH, Tang MBY: **Mortality of bullous pemphigoid in Singapore: risk factors and causes of death in 359 patients seen at the National Skin Centre.** *Br J Dermatol* 2014, **170**:1319–1326.

227) Chanter N, Talbot NC, Newton JR, Hewson D, Verheyen K: ***Streptococcus equi* with truncated M-proteins isolated from outwardly healthy horses.** *Microbiology* 2000, **146 (Pt 6)**:1361–1369.

228) Gustafsson MCU, Lannergård J, Nilsson OR, Kristensen BM, Olsen JE, Harris CL, Ufret-Vincenty RL, Stålhammar-Carlemalm M, Lindahl G: **Factor H Binds to the Hypervariable Region of Many *Streptococcus pyogenes* M Proteins but Does Not Promote Phagocytosis Resistance or Acute Virulence**. *PLoS Pathog* 2013, **9**:e1003323.

229) Reyes-Lopez M, Bermudez-Cruz RM, Avila EE, de la Garza M: **Acetaldehyde/alcohol dehydrogenase-2 (EhADH2) and clathrin are involved in internalization of human transferrin by *Entamoeba histolytica***. *Microbiology* 2011, **157**:209–219.

230) Leippe M: **Amoebapores**. *Parasitol Today* 1997, **13**:178–183.

231) Edman U, Meraz MA, Rausser S, Agabian N, Meza I: **Characterization of an immuno-dominant variable surface antigen from pathogenic and nonpathogenic *Entamoeba histolytica*.** *J Exp Med* 1990, **172**:879–888.

232) Vicente JB, Ehrenkaufer GM, Saraiva LM, Teixeira M, Singh U: ***Entamoeba histolytica* modulates a complex repertoire of novel genes in response to oxidative and nitrosative stresses: implications for amebic pathogenesis**. *Cell Microbiol* 2009, **11**:51–69.

233) Baxt LA, Rastew E, Bracha R, Mirelman D, Singh U: **Downregulation of an *Entamoeba histolytica* Rhomboid Protease Reveals Roles in Regulating Parasite Adhesion and Phagocytosis**. *Eukaryotic Cell* 2010, **9**:1283–1293.

234) Coulter ED, Kurtz DM: **A role for rubredoxin in oxidative stress protection in *Desulfovibrio vulgaris*: catalytic electron transfer to rubrerythrin and two-iron superoxide reductase.** *Arch Biochem Biophys* 2001, **394**:76–86.

235) Das A, Coulter ED, Kurtz DM, Ljungdahl LG: **Five-gene cluster in *Clostridium thermoaceticum* consisting of two divergent operons encoding rubredoxin oxidoreductase- rubredoxin and rubrerythrin-type A flavoprotein- high-molecular-weight rubredoxin.** *J Bacteriol* 2001, **183**:1560–1567.

236) Somlata, Bhattacharya S, Bhattacharya A: **A C2 domain protein kinase initiates phagocytosis in the protozoan parasite *Entamoeba histolytica*.** *Nat Commun* 2011, **2**:230.

237) Santi-Rocca J, Weber C, Guigon G, Sismeiro O, Coppée J-Y, Guillén N: **The lysine- and glutamic acid-rich protein KERP1 plays a role in *Entamoeba histolytica* liver abscess pathogenesis.** *Cell Microbiol* 2007, **10**:202–217.

238) Riahi Y, Siman-Tov R, Ankri S: **Molecular cloning, expression and characterization of a serine proteinase inhibitor gene from *Entamoeba histolytica*.** *Molecular and Biochemical Parasitology* 2004, **133**:153–162.

239) González-Zorn B, Domínguez-Bernal G, Suárez M, Ripio MT, Vega Y, Novella S, Rodríguez A, Chico I, Tierrez A, Vázquez-Boland JA: **SmcL, a novel membrane-damaging virulence factor in *Listeria*.** *Int J Med Microbiol* 2000, **290**:369–374.

240) Bruchhaus I, Richter S, Tannich E: **Recombinant expression and biochemical characterization of an NADPH:flavin oxidoreductase from *Entamoeba histolytica*.** *Biochem J* 1998, **330 (Pt 3)**:1217–1221.

241) Arias DG, Regner EL, Iglesias AA, Guerrero SA: ***Entamoeba histolytica*** **thioredoxin reductase: molecular and functional characterization of its atypical properties.** *Biochim Biophys Acta* 2012, **1820**:1859–1866.

242) Nakada-Tsukui K, Tsuboi K, Furukawa A, Yamada Y, Nozaki T: **A novel class of cysteine protease receptors that mediate lysosomal transport.** *Cell Microbiol* 2012, **14**:1299–1317.

243) Nickel R, Jacobs T, Leippe M: **Molecular characterization of an exceptionally acidic lysozyme-like protein from the protozoon *Entamoeba histolytica*.** *FEBS Lett* 1998, **437**:153–157.

244) Labruyère E, Zimmer C, Galy V, Olivo-Marin J-C, Guillén N: **EhPAK, a member of the p21-activated kinase family, is involved in the control of *Entamoeba histolytica* migration and phagocytosis.** *Journal of Cell Science* 2003, **116**:61–71.

245) Elnekave K, Siman-Tov R, Ankri S: **Consumption of L-arginine mediated by *Entamoeba histolytica* L-arginase (EhArg) inhibits amoebicidal activity and nitric oxide production by activated macrophages.** *Parasite Immunology* 2003, **25**:597–608.

246) López-Vancell R, Montfort I, Pérez-Tamayo R: **Galactose-specific adhesin and cytotoxicity of *Entamoeba histolytica*.** *Parasitol Res* 2000, **86**:226–231.

247) Nickel R, Jacobs T, Urban B, Scholze H, Bruhn H, Leippe M: **Two novel calcium-binding proteins from cytoplasmic granules of the protozoan parasite *Entamoeba histolytica*.** *FEBS Lett* 2000, **486**:112–116.

248) Ravdin JI, Murphy CF, Guerrant RL, Long-Krug SA: **Effect of antagonists of calcium and phospholipase A on the cytopathogenicity of *Entamoeba histolytica*.** *J Infect Dis* 1985, **152**:542–549.

249) Dodson JM, Lenkowski PW, Eubanks AC, Jackson TF, Napodano J, Lyerly DM, Lockhart LA, Mann BJ, Petri Jr. WA: **Infection and immunity mediated by the carbohydrate recognition domain of the *Entamoeba histolytica* Gal/GalNAc lectin.** *J Infect Dis* 1999, **179**:460–466.

250) Gaucher D, Chadee K: **Prospect for an *Entamoeba histolytica* Gal-lectin-based vaccine.** *Parasite Immunology* 2003, **25**:55–58.

251) Soong CJ, Kain KC, Abd-Alla M, Jackson TF, Ravdin JI: **A recombinant cysteine-rich section of the *Entamoeba histolytica* galactose-inhibitable lectin is efficacious as a subunit vaccine in the gerbil model of amebic liver abscess.** *J Infect Dis* 1995, **171**:645–651.

252) Lotter H, Zhang T, Seydel KB, Stanley SL, Tannich E: **Identification of an epitope on the *Entamoeba histolytica* 170-kD lectin conferring antibody-mediated protection against invasive amebiasis.** *J Exp Med* 1997, **185**:1793–1801.

253) Arif IAI, Khan HAH, Shobrak MM, AA Homaidan Al AA, MA Sadoon Al M, AH Farhan Al AH: **Measuring the genetic diversity of Arabian Oryx using microsatellite markers: implication for captive breeding.** *Genes Genet Syst* 2010, **85**:141–145.

254) Glenn TC, Lance SL, McKee AM, Webster BL, Emery AM, Zerlotini A, Oliveira G, Rollinson D, Faircloth BC: **Significant variance in genetic diversity among populations of *Schistosoma haematobium* detected using microsatellite DNA loci from a genome-wide database.** *Parasit Vectors* 2013, **6**:300–300.

255) Henrichs DW, Renshaw MA, Gold JR, Campbell L: **Genetic diversity among clonal isolates of *Karenia brevis* as measured with microsatellite markers**. *Harmful Algae* 2013, **21-22**:30–35.

256) Serrano M, Calvo JH, Martínez M, Marcos-Carcavilla A, Cuevas J, González C, Jurado JJ, de Tejada PD: **Microsatellite based genetic diversity and population structure of the endangered Spanish Guadarrama goat breed.** *BMC Genet* 2009, **10**:61–61.

257) Viguera EE, Canceill DD, Ehrlich SDS: **Replication slippage involves DNA polymerase pausing and dissociation.** *EMBO J* 2001, **20**:2587–2595.

258) Ball AD, Stapley J, Dawson DA, Birkhead TR, Burke T, Slate J: **A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*).** *BMC Genomics* 2010, **11**:218.

259) Evans DM, Cardon LR: **Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps.** *Am J Hum Genet* 2004, **75**:687–692.

260) Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES: **Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.** *Science* 1998, **280**:1077–1082.

261) Gray IC, Campbell DA, Spurr NK: **Single nucleotide polymorphisms as tools in human genetics.** *Hum Mol Genet* 2000, **9**:2403–2408.

262) Hoffman JI, Amos W: **Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion.** *Mol Ecol* 2005, **14**:599–612.

263) Pompanon F, Bonin A, Bellemain E, Taberlet P: **Genotyping errors: causes, consequences and solutions.** *Nat Rev Genet* 2005, **6**:847–859.

264) Samie A, Obi CL, Bessong PO, Houpt E, Stroup S, Njayou M, Sabeta C, Mduluza T, Guerrant RL: ***Entamoeba histolytica*: genetic diversity of African strains based on the polymorphism of the serine-rich protein gene.** *Experimental Parasitology* 2008, **118**:354–361.

265) Haghighi A, Kobayashi S, Takeuchi T, Masuda G, Nozaki T: **Remarkable Genetic Polymorphism among *Entamoeba histolytica* Isolates from a Limited Geographic Area**. *Journal of Clinical Microbiology* 2002, **40**:4081–4090.

266) Escueta-de Cadiz A, Kobayashi S, Takeuchi T, Tachibana H, Nozaki T: **Identification of an avirulent *Entamoeba histolytica* strain with unique tRNA-linked short tandem repeat markers**. *Parasitology International* 2010, **59**:75–81.

267) Zaki M, Reddy SG, Jackson TFHG, Ravdin JI, Clark CG: **Genotyping of *Entamoeba* species in South Africa: diversity, stability, and transmission patterns within families.** *J Infect Dis* 2003, **187**:1860–1869.

268) Bhattacharya D, Haque R, Singh U: **Coding and noncoding genomic regions of *Entamoeba histolytica* have significantly different rates of sequence polymorphisms: implications for epidemiological studies.** *Journal of Clinical Microbiology* 2005, **43**:4815–4819.

269) Haghighi A, Rasti S, Mojarad EN, Kazemi B, Bandehpour M, Nochi Z, Hooshyar H, Rezaian M: ***Entamoeba dispar*: genetic diversity of Iranian isolates based on serine-rich *Entamoeba dispar* protein gene.** *Pak J Biol Sci* 2008, **11**:2613–2618.

270) Mojarad EN, Haghighi A, Kazemi B, Nejad MR, Abadi A, Zali MR: **High genetic diversity among Iranian *Entamoeba dispar* isolates based on the noncoding short tandem repeat locus D-A.** *Acta Tropica* 2009, **111**:133–136.

271) Kumar V, Westra H-J, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, Almeida R, Zhernakova A, Reinmaa E, Võsa U, Hofker MH, Fehrmann RSN, Fu J, Withoff S, Metspalu A, Franke L, Wijmenga C: **Human disease-associated genetic variation impacts large intergenic non-coding RNA expression.** *PLoS Genet* 2013, **9**:e1003201.

272) Gilchrist CA, Ali IKM, Kabir M, Alam F, Scherbakova S, Ferlanti E, Weedall GD, Hall N, Haque R, Petri Jr. WA, Caler E: **A Multilocus Sequence Typing System (MLST) reveals a high level of diversity and a genetic component to *Entamoeba histolytica* virulence**. *BMC Microbiol* 2012, **12**:151.

273) Nei M, Li WH: **Mathematical model for studying genetic variation in terms of restriction endonucleases.** *Proc Natl Acad Sci USA* 1979, **76**:5269–5273.

274) Blanc G, Ngwamidiba M, Ogata H, Fournier P-E, Claverie J-M, Raoult D: **Molecular Evolution of *Rickettsia* Surface Antigens: Evidence of Positive Selection**. *Molecular Biology and Evolution* 2005, **22**:2073–2083.

275) Buckling A, Rainey PB: **Antagonistic coevolution between a bacterium and a bacteriophage**. *Proceedings of the Royal Society B: Biological Sciences* 2002, **269**:931–936.

276) Mu J, Awadalla P, Duan J, McGee KM, Keebler J, Seydel K, McVean GAT, Su X-Z: **Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome.** *Nat Genet* 2007, **39**:126–130.

277) Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**:1754–1760.

278) Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**:2078–2079.

279) Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13**:555–556.

280) Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Molecular Biology and Evolution* 2007, **24**:1586–1591.

281) Jukes TH, Cantor CR: **Evolution of protein molecules**. In *Mammalian protein metabolism*. *Volume III*. Edited by Munro MN. Academic Press; 1969:21–132.

282) Kumari V, Iyer LR, Roy R, Bhargava V, Panda S, Paul J, Verweij JJ, Clark CG, Bhattacharya A, Bhattacharya S: **Genomic distribution of SINEs in *Entamoeba histolytica* strains: implication for genotyping.** *BMC Genomics* 2013, **14**:432.

283) Kim D, Kim W-Y, Lee S-Y, Lee S-Y, Yun H, Shin S-Y, Lee J, Hong Y, Won Y, Kim S-J, Lee YS, Ahn S-M: **Revising a personal genome by comparing and combining data from two different sequencing platforms.** *PLoS ONE* 2013, **8**:e60585.

284) Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, Schuster SC: **Comparison of sequencing platforms for single nucleotide variant calls in a human sample.** *PLoS ONE* 2013, **8**:e55089.

285) Biller L, Schmidt H, Krause E, Gelhaus C, Matthiesen J, Handal G, Lotter H, Janssen O, Tannich E, Bruchhaus I: **Comparison of two genetically related *Entamoeba histolytica* cell lines derived from the same isolate with different pathogenic properties**. *Proteomics* 2009, **9**:4107–4120.

286) Ungar BL, Yolken RH, Quinn TC: **Use of a monoclonal antibody in an enzyme immunoassay for the detection of *Entamoeba histolytica* in fecal specimens.** *Am J Trop Med Hyg* 1985, **34**:465–472.

287) Ajay SS, Parker SCJ, Abaan HO, Fajardo KVF, Margulies EH: **Accurate and comprehensive sequencing of personal genomes.** *Genome Research* 2011, **21**:1498–1505.

288) Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP: **Sequencing depth and coverage: key considerations in genomic analyses.** *Nat Rev Genet* 2014, **15**:121–132.

289) Anbar M, Bracha R, Nuchamowitz Y, Li Y, Florentin A, Mirelman D: **Involvement of a short interspersed element in epigenetic transcriptional silencing of the amoebapore gene in *Entamoeba histolytica*.** *Eukaryotic Cell* 2005, **4**:1775–1784.

290) Bracha R, Nuchamowitz Y, Anbar M, Mirelman D: **Transcriptional Silencing of Multiple Genes in Trophozoites of *Entamoeba histolytica***. *PLoS Pathog* 2006, **2**:e48.

291) Bracha R, Nuchamowitz Y, Mirelman D: **Transcriptional silencing of an amoebapore gene in *Entamoeba histolytica*: molecular analysis and effect on pathogenicity.** *Eukaryotic Cell* 2003, **2**:295–305.

292) Mar-Aguilar F, Trevino V, Salinas-Hernández JE, Taméz-Guerrero MM, Barrón-González MP, Morales-Rubio E, Treviño-Neávez J, Verduzco-Martínez JA, Morales-Vallarta MR, Reséndez-Pérez D: **Identification and characterization of microRNAs from *Entamoeba histolytica* HM1-IMSS.** *PLoS ONE* 2013, **8**:e68202.

293) Ludwig MZ: **Functional evolution of noncoding DNA.** *Current Opinion in Genetics & Development* 2002, **12**:634–639.

294) Nelson CE, Hersh BM, Carroll SB: **The regulatory content of intergenic DNA shapes genome architecture.** *Genome Biology* 2004, **5**:R25.

295) Wilusz JE, Sunwoo H, Spector DL: **Long noncoding RNAs: functional surprises from the RNA world.** *Genes & Development* 2009, **23**:1494–1504.

296) Riethoven J-JM: **Regulatory regions in DNA: promoters, enhancers, silencers, and insulators.** *Methods Mol Biol* 2010, **674**:33–42.

297) Hayashi S, Watanabe J, Kawajiri K: **Genetic polymorphisms in the 5'-flanking region change transcriptional regulation of the human cytochrome P450IIE1 gene.** *J Biochem* 1991, **110**:559–565.

298) Marcos-Carcavilla A, Mutikainen M, González C, Calvo JH, Kantanen J, Sanz A, Marzanov NS, Pérez-Guzmán MD, Serrano M: **A SNP in the HSP90AA1 gene 5' flanking region is associated with the adaptation to differential thermal conditions in the ovine species.** *Cell Stress Chaperones* 2010, **15**:67–81.

299) Peñaloza C, Hamilton A, Guy DR, Bishop SC, Houston RD: **A SNP in the 5' flanking region of the myostatin-1b gene is associated with harvest traits in Atlantic salmon (*Salmo salar*).** *BMC Genet* 2013, **14**:112.

300) Sun T, Gao Y, Tan W, Ma S, Shi Y, Yao J, Guo Y, Yang M, Zhang X, Zhang Q, Zeng C, Lin D: **A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers.** *Nat Genet* 2007, **39**:605–613.

301) Kimura M: **Evolutionary Rate at the Molecular Level**. *Nature* 1968, **217**:624–626.

302) King JL, Jukes TH: **Non-Darwinian evolution.** *Science* 1969, **164**:788–798.

303) Kumar S, Subramanian S: **Mutation rates in mammalian genomes.** *Proc Natl Acad Sci USA* 2002, **99**:803–808.

304) Neafsey DE, Galinsky K, Jiang RHY, Young L, Sykes SM, Saif S, Gujja S, Goldberg JM, Young S, Zeng Q, Chapman SB, Dash AP, Anvikar AR, Sutton PL, Birren BW, Escalante AA, Barnwell JW, Carlton JM: **The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum***. *Nat Genet* 2012, **44**:1046–1050.

305) Lawrie DS, Messer PW, Hershberg R, Petrov DA: **Strong purifying selection at synonymous sites in *D. melanogaster*.** *PLoS Genet* 2013, **9**:e1003527.

306) Künstner A, Nabholz B, Ellegren H: **Significant selective constraint at 4-fold degenerate sites in the avian genome and its consequence for detection of positive selection.** *Genome Biol Evol* 2011, **3**:1381–1389.

307) Haghighi A, Kobayashi S, Takeuchi T, Thammapalerd N, Nozaki T: **Geographic Diversity among Genotypes of *Entamoeba histolytica* Field Isolates**. *Journal of Clinical Microbiology* 2003, **41**:3748–3756.

308) Hazes B, Read RJ: **A mosquitocidal toxin with a ricin-like cell-binding domain.** *Nat Struct Biol* 1995, **2**:358–359.

309) Hazes B: **The $(QxW)_3$ domain: a flexible lectin scaffold.** *Protein Sci* 1996, **5**:1490–1501.

310) Hirabayashi J, Dutta SK, Kasai K: **Novel galactose-binding proteins in Annelida. Characterization of 29-kDa tandem repeat-type lectins from the earthworm *Lumbricus terrestris*.** *J Biol Chem* 1998, **273**:14450–14460.

311) Laufs J, Traut W, Heyraud F, Matzeit V, Rogers SG, Schell J, Gronenborn B: ***In vitro* cleavage and joining at the viral origin of replication by the replication initiator protein of tomato yellow leaf curl virus.** *Proc Natl Acad Sci USA* 1995, **92**:3879–3883.

312) Ilyina TV, Koonin EV: **Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria.** *Nucleic Acids Res* 1992, **20**:3279–3285.

313) Beg QK, Kapoor M, Mahajan L, Hoondal GS: **Microbial xylanases and their industrial applications: a review.** *Appl Microbiol Biotechnol* 2001, **56**:326–338.

314) Ankri S, Padilla-Vaca F, Stolarsky T, Koole L, Katz U, Mirelman D: **Antisense inhibition of expression of the light subunit (35 kDa) of the Gal/GalNac lectin complex inhibits *Entamoeba histolytica* virulence.** *Mol Microbiol* 1999, **33**:327–337.

315) Padilla-Vaca F, Ankri S, Bracha R, Koole LA, Mirelman D: **Down regulation of *Entamoeba histolytica* virulence by monoxenic cultivation with *Escherichia coli* O55 is related to a decrease in expression of the light (35-kilodalton) subunit of the Gal/GalNAc lectin.** *Infect Immun* 1999, **67**:2096–2102.

316) Katz U, Ankri S, Stolarsky T, Nuchamowitz Y, Mirelman D: ***Entamoeba histolytica* expressing a dominant negative N-truncated light subunit of its Gal-lectin are less virulent.** *Mol Biol Cell* 2002, **13**:4256–4265.

317) Kobe B, Deisenhofer J: **A structural basis of the interactions between leucine-rich repeats and protein ligands.** *Nature* 1995, **374**:183–186.

318) Kobe B, Kajava AV: **The leucine-rich repeat as a protein recognition motif.** *Curr Opin Struct Biol* 2001, **11**:725–732.

319) Meskiene I, Baudouin E, Schweighofer A, Liwosz A, Jonak C, Rodriguez PL, Jelinek H, Hirt H: **Stress-induced protein phosphatase 2C is a negative regulator of a mitogen-activated protein kinase.** *J Biol Chem* 2003, **278**:18945–18952.

320) Diamond LS: **Techniques of axenic cultivation of *Entamoeba histolytica* Schaudinn, 1903 and *E. histolytica*-like amebae.** *J Parasitol* 1968, **54**:1047–1056.

321) Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proc Natl Acad Sci USA* 2001, **98**:9748–9753.

322) Compeau PEC, Pevzner PA, Tesler G: **How to apply de Bruijn graphs to genome assembly.** *Nat Biotechnol* 2011, **29**:987–991.

323) Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marçais G, Pop M, Yorke JA: **GAGE: A critical evaluation of genome assemblies and assembly algorithms.** *Genome Research* 2012, **22**:557–567.

324) Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data.** *Genomics* 2010, **95**:315–327.

325) Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M: **Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots.** *Front Genet* 2013, **4**:237.

326) Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9**:357–359.

327) Wooley JC, Ye Y: **Metagenomics: Facts and Artifacts, and Computational Challenges**. *J Comput Sci Technol* 2009, **25**:71–81.

328) Charuvaka A, Rangwala H: **Evaluation of short read metagenomic assembly.** *BMC Genomics* 2011, **12 (Suppl 2)**:S8.

329) Nederbragt AJ, Rounge TB, Kausrud KL, Jakobsen KS: **Identification and Quantification of Genomic Repeats and Sample Contamination in Assemblies of 454 Pyrosequencing Reads**. *Sequencing* 2010, **2010**:1–12.

330) Godel C, Kumar S, Koutsovoulos G, Ludin P, Nilsson D, Comandatore F, Wrobel N, Thompson M, Schmid CD, Goto S, Bringaud F, Wolstenholme A, Bandi C, Epe C, Kaminsky R, Blaxter M, Mäser P: **The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets.** *FASEB J* 2012, **26**:4650–4661.

331) Kumar S, Blaxter ML: **Simultaneous genome sequencing of symbionts and their hosts.** *Symbiosis* 2011, **55**:119–126.

332) D'haeseleer P, Gladden JM, Allgaier M, Chain PSG, Tringe SG, Malfatti SA, Aldrich JT, Nicora CD, Robinson EW, Paša-Tolić L, Hugenholtz P, Simmons BA, Singer SW: **Proteogenomic analysis of a thermophilic bacterial consortium adapted to deconstruct switchgrass.** *PLoS ONE* 2013, **8**:e68465.

333) Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M: **The 160-kilobase genome of the bacterial endosymbiont *Carsonella*.** *Science* 2006, **314**:267.

334) Hildebrand F, Meyer A, Eyre-Walker A: **Evidence of selection upon genomic GC-content in bacteria.** *PLoS Genet* 2010, **6**:e1001107.

335) Meader S, Hillier LW, Locke D, Ponting CP, Lunter G: **Genome assembly quality: assessment and improvement using the neutral indel model.** *Genome Research* 2010, **20**:675–684.

336) Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD: **REAPR: a universal tool for genome assembly evaluation.** *Genome Biology* 2013, **14**:R47.

337) Soto-Jimenez LM, Estrada K, Sanchez-Flores A: **GARM: genome assembly, reconciliation and merging pipeline.** *Curr Top Med Chem* 2014, **14**:418–424.

338) Vicedomini R, Vezzi F, Scalabrin S, Arvestad L, Policriti A: **GAM-NGS: genomic assemblies merger for next generation sequencing.** *BMC Bioinformatics* 2013, **14 (Suppl 7)**:S6.

339) Nijkamp J, Winterbach W, van den Broek M, Daran J-M, Reinders M, de Ridder D: **Integrating genome assemblies with MAIA.** *Bioinformatics* 2010, **26**:i433–i439.

340) Yao G, Ye L, Gao H, Minx P, Warren WC, Weinstock GM: **Graph accordance of next-generation sequence assemblies.** *Bioinformatics* 2012, **28**:13–16.

341) Olson PD, Zarowiecki M, Kiss F, Brehm K: **Cestode genomics - progress and prospects for advancing basic and applied aspects of flatworm biology.** *Parasite Immunology* 2012, **34**:130–150.

342) Offerman K, Carulei O, van der Walt AP, Douglass N, Williamson A-L: **The complete genome sequences of poxviruses isolated from a penguin and a pigeon in South Africa and comparison to other sequenced avipoxviruses.** *BMC Genomics* 2014, **15**:463.

343) Kent WJ: **BLAT - The BLAST-like alignment tool.** *Genome Research* 2002, **12**:656–664.

344) Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, *et al.*: **Evolutionary and biomedical insights from the Rhesus Macaque genome.** *Science* 2007, **316**:222–234.

345) Ehrenkaufer GM, Weedall GD, Williams D, Lorenzi HA, Caler E, Hall N, Singh U: **The genome and transcriptome of the enteric parasite *Entamoeba invadens*, a model for encystation.** *Genome Biology* 2013, **14**:R77.

Appendices

Appendix C – Chapter Two additional digital materials

File C.1. AUGUSTUS training set of 197 predicted coding sequences

File C.2. Core *Entamoeba* gene set - orthologous gene clusters

File C.3. *Entamoeba*-specific gene set – orthologous gene clusters

Appendix D– Chapter Three additional digital materials

File D.1. Positions at which high quality SNPs identified in reference genomes

File D.2. *Entamoeba histolytica* strains: SNPs per gene, dN/dS ratios and dN-dS differences

File D.3. *Entamoeba dispar* strains: SNPs per gene, dN/dS ratios and dN-dS differences

File D.4. *Entamoeba moshkovskii* strains: SNPs per gene, dN/dS ratios and dN-dS differences

Appendix E– Chapter Four additional digital materials

File E.1. CEG IDs, Core *Entamoeba* cluster IDs and *E. histolytica* gene IDs of sequences detected in assemblies of *Entamoeba bangladeshi* reads

Appendix F – Supplementary material

Table A.1. Virulence factors search terms. Search terms enclosed in quotation marks were similarly enclosed when entered into AmoebaDB or NCBI's Gene DB. C2PK = C2-domain-containing protein kinase; KERP = lysine and glutamic acid-rich protein; rhomboid is also known as 'peptidase S54'; NADPH:flavin oxidoreductase is also known as 'p34 thioredoxin reductase'. Rows are shaded for ease of reading.

Gene Function Group	AmoebaDB Search Terms	Other Searches
Adhesin 112	EhADH112	-
Arginase	Arginase	-
C2PK	C2PK	-
Cysteine proteases	Cysteine protease "Cysteine proteinase" Cysteine peptidase	Referred to [50]
Cysteine protease binding proteins	"Cysteine protease" cpbf11	Referred to [242]
Fe-Hydrogenase	Fe-hydrogenase "Iron hydrogenase"	-
Gal/GalNAc lectin subunits	Galactose lectin	-
Grainins	Grainin	-
KERPs	-	Searched NCBI Gene DB for accession numbers from [237]
Lysozyme	Lysozyme	-

Gene Function Group	AmoebaDB Search Terms	Other Searches
NADPH:flavin oxidoreductase	NADPH "Thioredoxin reductase"	-
P21-activated kinase	P21 PAK	-
Rhomboid	Rhomboid	-
Peroxiredoxin	Peroxiredoxin	-
Phospholipases	Phospholipase Phospho	Searched NCBI Gene DB using: "Phospholipase [A/B/C/D]" "*Entamoeba* histolytica"
Poreformers	-	Searched NCBI Gene DB for accession numbers from [50]
Rubredoxin	Rubredoxin	-
Rubrerythrin	Rubrerythrin	-
Serine protease inhibitor (Serpin)	"Serine protease inhibitor"	-
STIRP	"Serine-threonine-isoleucine rich protein"	-
Sphingomyelinase C	"Sphingomyelinase C"	-
SOD	SOD	-
Thioredoxin/Thioredoxin reductase	Thioredoxin	-
Type A flavoprotein	"Type A flavoprotein"	-

Table A.2. Coefficient of Variation and proportion of invariant sites determined for each virulence factor family. Values in unshaded cells were calculated using protein sequences; those in shaded cells were calculated using DNA sequences, including pseudogenes. CoV = $1/\sqrt{\alpha}$.

Virulence Factor Family	Coefficient of Variation		Proportion of invariant sites (%)	
Adhesin	0.5290		0.7975	
Arginase	0.8164		0.0000	
C2 protein kinase	0.5942		16.7723	
Cysteine protease Family A	0.8753	0.8289	1.9743	2.3815
Cysteine protease Family B	0.7196		0.2580	
Cysteine protease Family C	0.4226		0.4519	
CPBP	0.4071		1.0174	
Fe Hydrogenase	0.7520		0.0000	
Heavy Gal/GalNAc lectin	0.9174		0.0000	
Intermediate Gal/GalNAc lectin	0.5284		5.9373	
Light Gal/GalNAc lectin	0.9946		8.1210	
Grainins	0.8198	0.9219	5.3432	7.8916
KERP	0.7875		0.0000	
Lysozyme	0.6571	0.7642	0.0000	1.9041
NADPH:flavin oxidoreductase	0.3168	0.5491	0.0000	0.7409
P21-activated kinase	0.8740		0.8127	
Peroxiredoxin	0.7239	0.6887	1.3446	0.7023
Phospholipases	0.5981	0.6187	0.0000	0.0000
Poreformers	0.2615	0.5911	0.0000	0.4851
Rhomboid	0.0034		0.0000	
Rubredoxin	0.9280		0.0000	
Rubrerythrin	1.2622		0.0000	
Serpin	0.7568		4.2376	
SOD	0.8084		32.1017	
Sphingomyelinase C	0.5051	0.8657	4.1837	0.0000
STIRP	0.6560		0.0835	
Thioredoxin	0.6756	0.7931	4.5217	12.1551
Thioredoxin reductase	1.1206		7.2837	
Type A Flavoprotein	1.0556	0.9441	0.0000	16.7244

Table A.3. Original and updated coordinates of genes manually curated, having been incorrectly annotated by AUGUSTUS. For gene g10524, each line of the row represents an exon.

Gene ID	Scaffold/ Contig ID	+/-	Original prediction			Corrected prediction		
			Start	End	Frame	Start	End	Frame
g10524	Scaffold 00956	+	312	925	2	239	925	0
			1,013	1,519	0	1,013	1,519	0
g3629	Contig 00295	+	309	2,773	2	347	2,773	0
g12434	Contig 00399	+	35	851	1	39	851	0
g12665	Contig 01277	+	37	1,211	2	63	1,211	0
g13750	Contig 06118	+	50	1,226	1	261	1,226	0
g14162	Contig 07626	+	309	649	2	242	649	0

Table A.4. Gene clusters unique to *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba invadens*. Asterisks (*) indicate families that contain pseudogenes.

Gene count	Cluster count	Cluster function
Entamoeba histolytica		
22	6	BspA family
18	4	Surface antigen ariel1
12	2	AIG1 family
12	2	Mucins
7	2	Cylicin-2
7	2	Cysteine protease*
6	1	Acetyltransferase
3	1	Peroxiredoxin*
2	1	60S ribosomal subunit biogenesis protein NIP7*
2	1	Dentin sialophosphoprotein
2	1	DNA repair protein
2	1	Gal/GalNAc light subunit
2	1	Dextranase
2	1	Chaperone ClpB
2	1	Midasin
2	1	Nucleosome binding protein
2	1	Alcohol dehydrogenase
2	1	Proteoglycan-4
2	1	Serine-rich 25 kDa antigen protein
2	1	Immediate-early protein
Entamoeba dispar		
13	1	AIG1 family
5	2	Heat shock protein
3	1	Serine/threonine/tyrosine protein kinase
2	1	Casein kinase I
2	2	Chaperone ClpB
2	1	dab2-interacting protein
2	1	Ras family GTPase

Gene count	Cluster count	Cluster function
Entamoeba invadens		
214	46	Serine/threonine/tyrosine kinase
34	9	Ras family GTPase
32	2	Ribonuclease
27	8	Heat shock protein
21	1	Cylicin
21	2	Myosin
19	2	Glutamine/asparagine-rich protein pqn-25
16	5	Actin
15	1	Thioredoxin
12	1	Profilin
11	1	Capsular polysaccharide phosphotransferase
11	2	DNA double-strand break repair Rad50 ATPase
9	1	Embryonic protein DC-8
8	3	Serine/threonine protein phosphatase
8	1	Tropomyosin alpha-1 chain
7	2	ADP ribosylation factor
7	2	Cysteine protease
7	1	Elongation factor 1-alpha
7	1	Furin
6	2	Actophorin
6	1	Gal/GalNAc lectin light subunit
6	1	Nitrogen fixation protein nifU
5	1	Calcium-binding protein/Caltractin/Centrin-1
5	2	Chaperone ClpB
5	1	DNA repair and recombination protein rad52
5	1	GRIP domain-containing protein RUD3
5	2	Serpin (serine protease inhibitor)
5	1	Vacuolar protein sorting-associated protein
4	1	Acyltransferase
4	1	Ephrin type-A receptor 4A
3	1	Calmodulin-4
3	1	Capsular polysaccharide posphotransferase fcs1
3	1	Centromeric protein E

Gene count	Cluster count	Cluster function
3	1	F-box and WD domain protein
3	1	Lysozyme
3	1	Ornithine decarboxylase
3	1	R2r3-MYB transcription factor
3	1	Ribonucleoprotein
2	1	40S ribosomal protein
2	1	60S ribosomal protein
2	1	Abnormal long morphology protein
2	1	Coactosin
2	1	Developmentally-regulated GTP-binding protein
2	1	Histone
2	1	Leucine-rich repeat-containing protein 33 precursor
2	1	Nucleoporin nup45
2	1	Paramyosin
2	1	Ras family guanine nucleotide exchange factors
2	1	Replication factor-A protein
2	1	S-adenosylmethionine synthetase
2	1	Trichoplein keratin filament-binding protein
2	1	Type A flavoprotein
2	1	UDP-N-acetylglucosamine transporter
2	1	Zinc finger protein DHHC domain containing protein
2	1	Zinc homeostasis factor

Table A.5. GOA Slim categories to which GO terms were ascribed in significantly different proportions, based upon pairwise χ^2 comparisons of *Entamoeba histolytica* HM-1:IMSS, *Entamoeba dispar* SAW760, *Entamoeba invadens* IP-1 and *Entamoeba moshkovskii* Laredo. Each section of the table contains GOA Slim categories sharing a common trait regarding differences in proportions. White rows are GOA Slim categories from the 'Components' subgroup; light grey rows are 'Functions'; and dark grey rows are 'Processes'. Asterisked (*) comparisons required the Monte Carlo simulation.

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics for significantly different pairings
GOA Slims in which *E. invadens* proportions differ from all other species					
GO:0005737: Cytoplasm	20.81	21.37	14.62	17.90	EHI v EIN: χ^2 = 11.5173, p-value < 0.001 EDI v EIN: χ^2 = 11.2962, p-value < 0.001 EIN v EMO: χ^2 = 4.5452, p-value = 0.033
GO:0005622: Intracellular	44.49	43.89	51.25	42.61	EHI v EIN: χ^2 = 7.6069, p-value = 0.006 EDI v EIN: χ^2 = 7.5311, p-value = 0.006 EIN v EMO: χ^2 = 17.7831, p-value < 0.001
GO:0016209: Antioxidant activity	1.55	1.26	0.05	1.58	EHI v EIN: χ^2 = 42.0497, p-value < 0.001* EDI v EIN: χ^2 = 55.0524, p-value < 0.001* EIN v EMO: χ^2 = 58.5489, p-value < 0.001*
GO:0015075: Ion transporter activity	2.41	2.52	0.62	2.31	EHI v EIN: χ^2 = 32.414, p-value < 0.001 EDI v EIN: χ^2 = 33.695, p-value < 0.001 EIN v EMO: χ^2 = 40.9933, p-value < 0.001

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics for significantly different pairings
GO:0004872: Receptor activity	0.85	0.56	0.15	0.73	EHI v EIN: χ^2 = 5.3917, p-value = 0.020 EDI v EIN: χ^2 = 15.4052, p-value < 0.001 EIN v EMO: χ^2 = 15.0217, p-value < 0.001
GO:0009056: Catabolism	2.79	3.05	4.35	2.40	EHI v EIN: χ^2 = 5.5669, p-value = 0.018 EDI v EIN: χ^2 = 11.5706, p-value = 0.001 EIN v EMO: χ^2 = 34.8623, p-value < 0.001
GO:0006810: Transport	3.11	3.86	6.61	3.54	EHI v EIN: χ^2 = 18.0073, p-value < 0.001 EDI v EIN: χ^2 = 42.8352, p-value < 0.001 EIN v EMO: χ^2 = 58.7267, p-value < 0.001
GOA Slims in which only *E. histolytica* and *E. dispar* proportions are similar					
GO:0016301: Kinase activity	7.17	6.23	12.01	14.97	EHI v EIN: χ^2 = 36.8345, p-value < 0.001 EHI v EMO: χ^2 = 75.3238, p-value < 0.001 EDI v EIN: χ^2 = 33.0213, p-value < 0.001 EDI v EMO: χ^2 = 79.2615, p-value < 0.001 EIN v EMO: χ^2 = 17.1608, p-value < 0.001
GO:0016491: Oxidoreductase activity	9.12	7.91	2.77	5.26	EHI v EIN: χ^2 = 69.0693, p-value < 0.001 EHI v EMO: χ^2 = 14.3481, p-value < 0.001 EDI v EIN: χ^2 = 114.2331, p-value < 0.001

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics for significantly different pairings
					EDI v EMO: χ^2 = 37.1669, p-value < 0.001 EIN v EMO: χ^2 = 35.3053, p-value < 0.001
GO:0004871: Signal transducer activity	3.06	2.38	0.50	4.35	EHI v EIN: χ^2 = 35.9702, p-value = < 0.001 EHI v EMO: χ^2 = 11.1945, p-value < 0.001 EDI v EIN: χ^2 = 63.5128, p-value < 0.001 EDI v EMO: χ^2 = 6.1215, p-value = 0.013 EIN v EMO: χ^2 = 128.7255, p-value < 0.001
GO:0005198: Structural molecule activity	1.85	2.45	0.27	0.62	EHI v EIN: χ^2 = 56.8079, p-value < 0.001 EHI v EMO: χ^2 = 37.222, p-value < 0.001 EDI v EIN: χ^2 = 39.9077, p-value < 0.001 EDI v EMO: χ^2 = 22.7671, p-value < 0.001 EIN v EMO: χ^2 = 5.3278, p-value = 0.021
GO:0016740: Transferase activity	11.23	11.48	15.36	17.43	EHI v EIN: χ^2 = 12.5531, p-value < 0.001 EHI v EMO: χ^2 = 29.316, p-value < 0.001 EDI v EIN: χ^2 = 18.5526, p-value < 0.001 EDI v EMO: χ^2 = 42.3429, p-value < 0.001 EIN v EMO: χ^2 = 7.1924, p-value = 0.007
GO:0005215: Transporter activity	4.41	4.90	1.45	2.86	EHI v EIN: χ^2 = 53.0682, p-value < 0.001 EHI v EMO: χ^2 = 14.4305, p-value < 0.001

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics for significantly different pairings
					EDI v EIN: χ^2 = 47.9749, p-value < 0.001 EDI v EMO: χ^2 = 10.7883, p-value = 0.001 EIN v EMO: χ^2 = 20.537, p-value < 0.001
GO:0005488: Binding	30.98	30.39	37.91	27.42	EHI v EIN: χ^2 = 25.5298, p-value < 0.001 EHI v EMO: χ^2 = 4.8536, p-value = 0.028 EDI v EIN: χ^2 = 27.6263, p-value < 0.001 EDI v EMO: χ^2 = 9.0409, p-value = 0.003 EIN v EMO: χ^2 = 119.5158, p-value < 0.001
GO:0007154: Cell communication	5.89	5.20	8.51	9.70	EHI v EIN: χ^2 = 20.6316, p-value < 0.001 EHI v EMO: χ^2 = 36.9197, p-value < 0.001 EDI v EIN: χ^2 = 17.1897, p-value < 0.001 EDI v EMO: χ^2 = 36.7935, p-value < 0.001 EIN v EMO: χ^2 = 4.6999, p-value = 0.030
GO:0007275: Development	2.05	1.55	4.14	2.99	EHI v EIN: χ^2 = 26.5254, p-value < 0.001 EHI v EMO: χ^2 = 11.0861, p-value < 0.001 EDI v EIN: χ^2 = 23.4589, p-value < 0.001 EDI v EMO: χ^2 = 6.4413, p-value = 0.011 EIN v EMO: χ^2 = 11.1099, p-value < 0.001

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics for significantly different pairings
GO:0043170: Macromolecule metabolism	19.13	19.08	12.70	14.88	EHI v EIN: χ^2 = 43.7116, p-value < 0.001 EHI v EMO: χ^2 = 19.3222, p-value < 0.001 EDI v EIN: χ^2 = 57.1822, p-value < 0.001 EDI v EMO: χ^2 = 26.8109, p-value < 0.001 EIN v EMO: χ^2 = 11.1445, p-value = 0.001
GO:0050789: Regulation of biological process	12.00	11.09	16.18	17.73	EHI v EIN: χ^2 = 27.2697, p-value < 0.001 EHI v EMO: χ^2 = 47.0566, p-value < 0.001 EDI v EIN: χ^2 = 24.6159, p-value < 0.001 EDI v EMO: χ^2 = 48.7176, p-value < 0.001 EIN v EMO: χ^2 = 4.7392, p-value = 0.029
GOA Slims in which proportions are similar between *E. histolytica* and *E. dispar,* and between *E. invadens* & *E. moshkovskii*					
GO:0016020: Membrane	22.02	20.99	26.66	30.13	EHI v EIN: χ^2 = 5.9679, p-value = 0.015 EHI v EMO: χ^2 = 14.933, p-value < 0.001 EDI v EIN: χ^2 = 4.6634, p-value = 0.031 EDI v EMO: χ^2 = 13.8309, p-value < 0.001
GO:0009058: Biosynthesis	5.68	5.36	2.34	2.10	EHI v EIN: χ^2 = 38.8107, p-value < 0.001 EHI v EMO: χ^2 = 56.8122, p-value < 0.001 EDI v EIN: χ^2 = 56.8463, p-value < 0.001

GOA Slim ID	EDI	EHI	EIN	EMO	Statistics for significantly different pairings
					EDI v EMO: χ^2 = 84.6367, p-value < 0.001
GO:0030154: Cell differentiation	0.42	0.32	2.67	2.19	EHI v EIN: χ^2 = 36.454, p-value < 0.001 EHI v EMO: χ^2 = 28.0414, p-value < 0.001 EDI v EIN: χ^2 = 48.7331, p-value < 0.001 EDI v EMO: χ^2 = 36.9405, p-value < 0.001
GO:0006928: Cell motility	0.14	0.43	1.28	1.42	EHI v EIN: χ^2 = 8.7129, p-value = 0.003 EHI v EMO: χ^2 = 11.3672, p-value = 0.001 EDI v EIN: χ^2 = 27.3161, p-value < 0.001* EDI v EMO: χ^2 = 31.8374, p-value < 0.001*

Table A.6. GOA Slim categories to which GO terms were ascribed in significantly different proportions based on pairwise χ^2 comparisons between *Entamoeba histolytica* HM-1:IMSS, and *Acanthamoeba castellanii* Neff, *Dictyostelium discoideum* AX4 and *Saccharomyces cerevisiae* S822c. White rows are GOA Slim categories from the 'Components' subgroup; light grey rows are 'Functions'; and dark grey rows are 'Processes'. Asterisked (*) comparisons required the Monte Carlo simulation.

GOA Slim ID	EHI	ACA	DDI	SCA	Statistics for significantly different pairings
GO:0005622: Intracellular	51.21	49.44	33.66	41.31	EHI v DDI: χ^2 = 220.6421, p-value < 0.001 EHI v SCE: χ^2 = 80.8169, p-value < 0.001
GO:0005737: Cytoplasm	17.67	20.60	18.21	23.24	EHI v ACA: χ^2 = 8.5665, p-value = 0.003 EHI v SCE: χ^2 = 35.1247, p-value < 0.001
GO:0016020: Membrane	21.47	16.74	30.73	18.02	EHI v DDI: χ^2 = 70.0117, p-value < 0.001 EHI v ACA: χ^2 = 24.3738, p-value < 0.001 EHI v SCE: χ^2 = 15.9325, p-value < 0.001
GO:0005634: Nucleus	8.21	10.50	8.78	14.00	EHI v ACA: χ^2 = 9.2585, p-value = 0.002 EHI v SCE: χ^2 = 57.0086, p-value < 0.001
GO:0005694: Chromosome	1.07	1.58	0.65	2.26	EHI v SCE: χ^2 = 12.9585, p-value < 0.001
GO:0005576: Extracellular region	0.19	0.60	5.25	0.36	EHI v DDI: χ^2 = 108.877, p-value < 0.001* EHI v ACA: χ^2 = 5.5889, p-value = 0.022*
GO:0030312: External encapsulating structure	0.09	0.32	0.16	0.81	EHI v SCE: χ^2 = 13.7197, p-value = 0.001*
GO:0009986: Cell surface	0.09	0.06	0.13	0.00	EHI v SCE: χ^2 = 18.281, p-value = 0.014*

GOA Slim ID	EHI	ACA	DDI	SCA	Statistics for significantly different pairings
GO:0005615: Extracellular space	0.00	0.08	2.41	0.00	EHI v DDI: $\chi^2 = 53.0119$, p-value < 0.001*
GO:0005488: Binding	36.07	42.37	38.27	31.41	EHI v DDI: $\chi^2 = 7.8568$, p-value = 0.005 EHI v ACA: $\chi^2 = 71.8218$, p-value < 0.001 EHI v SCE: $\chi^2 = 42.1398$, p-value < 0.001
GO:0016787: Hydrolase activity	13.57	8.36	9.50	12.28	EHI v DDI: $\chi^2 = 64.5563$, p-value < 0.001 EHI v ACA: $\chi^2 = 131.2465$, p-value < 0.001 EHI v SCE: $\chi^2 = 6.2946$, p-value = 0.012
GO:0016740: Transferase activity	11.78	9.37	17.15	13.24	EHI v DDI: $\chi^2 = 87.4312$, p-value < 0.001 EHI v ACA: $\chi^2 = 27.4659$, p-value < 0.001 EHI v SCE: $\chi^2 = 8.1691$, p-value = 0.004
GO:0016301: Kinase activity	7.31	5.46	8.03	3.64	EHI v ACA: $\chi^2 = 26.089$, p-value < 0.001 EHI v SCE: $\chi^2 = 124.0981$, p-value < 0.001
GO:0003676: Nucleic acid binding	4.11	5.27	7.73	14.36	EHI v DDI: $\chi^2 = 85.6063$, p-value < 0.001 EHI v ACA: $\chi^2 = 12.4374$, p-value < 0.001 EHI v SCE: $\chi^2 = 460.7881$, p-value < 0.001
GO:0016491: Oxidoreductase activity	4.13	2.48	3.57	3.40	EHI v ACA: $\chi^2 = 39.9904$, p-value < 0.001 EHI v SCE: $\chi^2 = 6.1997$, p-value = 0.013

GOA Slim ID	EHI	ACA	DDI	SCA	Statistics for significantly different pairings
GO:0005515: Protein binding	11.51	17.71	4.41	5.03	EHI v DDI: χ^2 = 291.0861, p-value < 0.001 EHI v ACA: χ^2 = 126.6165, p-value < 0.001 EHI v SCE: χ^2 = 269.8634, p-value < 0.001
GO:0005215: Transporter activity	1.92	1.16	1.52	4.50	EHI v ACA: χ^2 = 17.6407, p-value < 0.001 EHI v SCE: χ^2 = 80.7731, p-value < 0.001
GO:0005198: Structural molecule activity	0.65	0.51	0.50	1.94	EHI v SCE: χ^2 = 46.5943, p-value < 0.001
GO:0004871: Signal transducer activity	2.24	1.15	2.37	0.51	EHI v ACA: χ^2 = 34.6093, p-value < 0.001 EHI v SCE: χ^2 = 117.9236, p-value < 0.001
GO:0030234: Enzyme regulator activity	1.76	1.12	1.09	2.07	EHI v DDI: χ^2 = 12.4531, p-value < 0.001 EHI v ACA: χ^2 = 13.1219, p-value < 0.001
GO:0016874: Ligase activity	0.83	1.11	1.37	1.38	EHI v DDI: χ^2 = 9.5219, p-value = 0.002 EHI v SCE: χ^2 = 10.5668, p-value = 0.001
GO:0015075: Ion transporter activity	1.07	0.75	0.68	2.55	EHI v DDI: χ^2 = 6.6731, p-value = 0.010 EHI v ACA: χ^2 = 5.2263, p-value = 0.022 EHI v SCE: χ^2 = 45.6585, p-value < 0.001
GO:0004386: Helicase activity	0.47	0.48	0.60	0.90	EHI v SCE: χ^2 = 10.1129, p-value = 0.001
GO:0004872: Receptor activity	0.90	0.50	1.17	0.24	EHI v ACA: χ^2 = 10.7967, p-value = 0.001 EHI v SCE: χ^2 = 40.276, p-value < 0.001

GOA Slim ID	EHI	ACA	DDI	SCA	Statistics for significantly different pairings
GO:0016829: Lyase activity	0.42	1.26	0.78	0.80	EHI v DDI: χ^2 = 7.1927, p-value = 0.007 EHI v ACA: χ^2 = 30.575, p-value < 0.001 EHI v SCE: χ^2 = 8.6506, p-value = 0.003
GO:0016209: Antioxidant activity	0.31	0.11	0.15	0.22	EHI v ACA: χ^2 = 9.2956, p-value = 0.002
GO:0008565: Protein transporter activity	0.25	0.04	0.09	0.35	EHI v DDI: χ^2 = 5.2625, p-value = 0.022 EHI v ACA: χ^2 = 15.2297, p-value < 0.001
GO:0015267: Channel or pore class transporter activity	0.06	0.22	0.19	0.35	EHI v DDI: χ^2 = 4.7422, p-value = 0.022* EHI v ACA: χ^2 = 6.6795, p-value = 0.008* EHI v SCE: χ^2 = 13.7396, p-value = 0.001*
GO:0045182: Translation regulator activity	0.00	0.01	0.00	0.12	EHI v SCE: χ^2 = 7.4066, p-value = 0.011*
GO:0008152: Metabolism	24.99	23.43	28.29	27.85	EHI v DDI: χ^2 = 30.9301, p-value < 0.001 EHI v ACA: χ^2 = 8.7349, p-value = 0.003 EHI v SCE: χ^2 = 30.6059, p-value < 0.001
GO:0043170: Macromolecule metabolism	15.01	13.86	15.12	22.27	EHI v ACA: χ^2 = 7.1372, p-value = 0.008 EHI v SCE: χ^2 = 238.4408, p-value < 0.001
GO:0050789: Regulation of biological process	16.80	16.27	10.71	7.99	EHI v DDI: χ^2 = 183.3241, p-value < 0.001 EHI v SCE: χ^2 = 667.394, p-value < 0.001

GOA Slim ID	EHI	ACA	DDI	SCA	Statistics for significantly different pairings
GO:0050896: Response to stimulus	11.51	11.78	8.63	4.14	EHI v DDI: χ^2 = 52.8262, p-value < 0.001 EHI v SCE: χ^2 = 779.4157, p-value < 0.001
GO:0006139: Nucleobase, nucleoside, nucleotide and nucleic acid metabolism	5.72	6.63	9.39	14.33	EHI v DDI: χ^2 = 102.8679, p-value < 0.001 EHI v ACA: χ^2 = 9.0406, p-value = 0.003 EHI v SCE: χ^2 = 499.6046, p-value < 0.001
GO:0009058: Biosynthesis	3.53	5.18	5.89	12.32	EHI v DDI: χ^2 = 65.5539, p-value < 0.001 EHI v ACA: χ^2 = 39.3025, p-value < 0.001 EHI v SCE: χ^2 = 606.2695, p-value < 0.001
GO:0007154: Cell communication	8.20	7.18	4.77	1.26	EHI v DDI:χ^2 = 115.0102, p-value < 0.001 EHI v ACA: χ^2 = 9.9052, p-value = 0.002 EHI v SCE: χ^2 = 1442.618, p-value < 0.001
GO:0006810: Transport	4.62	4.17	10.45	6.40	EHI v DDI: χ^2 =254.2104, p-value < 0.001 EHI v SCE: χ^2 = 41.5705, p-value < 0.001
GO:0007275: Development	3.10	4.33	1.95	0.00	EHI v DDI: χ^2 = 31.449, p-value < 0.001 EHI v ACA: χ^2 = 25.9575, p-value < 0.001 EHI v SCE: χ^2 = 1247.861, p-value < 0.001*
GO:0009056: Catabolism	3.72	2.78	3.16	2.56	EHI v DDI: χ^2 = 5.2516, p-value = 0.022 EHI v ACA: χ^2 = 19.5234, p-value < 0.001 EHI v SCE: χ^2 = 36.5236, p-value < 0.001

GOA Slim ID	EHI	ACA	DDI	SCA	Statistics for significantly different pairings
GO:0030154: Cell differentiation	1.39	2.37	0.77	0.58	EHI v DDI: χ^2 = 20.5514, p-value < 0.001 EHI v ACA: χ^2 = 30.6476, p-value < 0.001 EHI v SCE: χ^2 = 65.1084, p-value < 0.001
GO:0006928: Cell motility	0.70	0.91	0.40	0.04	EHI v DDI: χ^2 = 9.4919, p-value = 0.002 EHI v SCE: χ^2 = 189.944, p-value < 0.001
GO:0008219: Cell death	0.43	0.69	0.12	0.07	EHI v DDI: χ^2 = 20.8153, p-value < 0.001 EHI v ACA: χ^2 = 6.8955, p-value = 0.009 EHI v SCE: χ^2 = 71.2897, p-value < 0.001
GO:0007610: Behavior	0.10	0.16	0.07	0.00	EHI v SCE: χ^2 = 39.0475, p-value < 0.001*
GO:0009405: Pathogenesis	0.02	0.02	0.10	0.00	EHI v DDI: χ^2 = 4.7041, p-value = 0.042* EHI v SCE: χ^2 = 8.676, p-value = 0.037*

Table B.1. Concentrations of DNA samples from *Entamoeba moshkovskii* strains FIC, 15114 and Snake, *Entamoeba dispar* strains SAW760 and AS16IR, and *Entamoeba bangladeshi* strain 8237, at different stages in creation of pooled libraries. Concentrations and volumes of purified DNA extracts were recorded after dilution to within range required by TruSeq protocol. Concentrations and volumes for libraries of individual strains were recorded after dilution to within range of Agilent 2100 Bioanalyzer. These diluted samples constitute the final pooled libraries.

Library	Strains	Extracted DNA		Strain-specific libraries		Pooled library conc. (ng/μL)
		Conc. (ng/μL)	Vol. (μL)	Conc. (ng/μL)	Vol. (μL)	
1	FIC	41.5	29	-	-	-
	15114	13.0	64	-	-	
	Snake	34.4	49	-	-	
2	Laredo	21.6	100	1.95	10.84	0.25
	SAW760	10.0	75	2.59	3.99	
	AS16IR	9.50	55	2.62	10.90	
	8237	24.8	110	2.36	4.43	
3	SAW760	10.0	75	7.45	12.42	2.18
	8237	24.8	110	6.77	17.58	

Table B.2. SNP rates in 4D synonymous sites common to ten strains of *Entamoeba histolytica*. Rates are given as the number of SNPs per 4D site across the reference genome. The greatest SNP rate between a pair of strains is in bold.

	HM-1:IMSS	**DS4**	**2952100**	**HK-9**	**MS27**	**MS84**	**MS96**	**PVB**	**PVF**
DS4	0.000268								
2592100	0.000189	0.000127							
HK-9	0.000262	0.000324	0.000251						
MS27	0.000248	0.000032	0.000112	0.000316					
MS84	0.000236	0.000174	0.000153	0.000310	0.000159				
MS96	0.000313	**0.000363**	0.000318	0.000351	0.000360	0.000313			
PVB	0.000206	0.000292	0.000224	0.000310	0.000283	0.000277	0.000330		
PVF	0.000203	0.000295	0.000221	0.000313	0.000286	0.000274	0.000333	0.000021	
Rahman	0.000212	0.000180	0.000159	0.000239	0.000171	0.000189	0.000318	0.000254	0.000262

Parasite Immunology

Parasite Immunology, 2012, **34,** 90–99 DOI: 10.1111/j.1365-3024.2011.01325.x

Review Article

Host–Parasite interactions in *Entamoeba histolytica* and *Entamoeba dispar*: what have we learned from their genomes?

I. W. WILSON, G. D. WEEDALL & N. HALL

Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool, UK

SUMMARY

Invasive amoebiasis caused by Entamoeba histolytica *is a major global health problem. Virulence is a rare outcome of infection, occurring in fewer than 1 in 10 infections. Not all strains of the parasite are equally virulent, and understanding the mechanisms and causes of virulence is an important goal of* Entamoeba *research. The sequencing of the genome of* E. histolytica *and the related avirulent species* Entamoeba dispar *has allowed whole-genome-scale analyses of genetic divergence and differential gene expression to be undertaken. These studies have helped elucidate mechanisms of virulence and identified genes differentially expressed in virulent and avirulent parasites. Here, we review the current status of the* E. histolytica *and* E. dispar *genomes and the findings of a number of genome-scale studies comparing parasites of different virulence.*

Keywords *differential gene expression,* Entamoeba*, genome, virulence*

INTRODUCTION

Amoebiasis is a disease of global importance, caused by the eukaryotic parasite *Entamoeba histolytica*. It is the most common worldwide cause of mortality from a protozoon after malaria, killing an estimated 40 000–110 000 people annually, and causing 34–50 million cases of severe disease. However, fewer than 10% of those infected develop invasive amoebiasis (1). Those most at risk are people living in areas of poor sanitation, as the parasite is transmitted via a faecal–oral route. In such environments, exposure may be very high. For example, acquisition of *Entamoeba*-specific antibodies indicated an annual incidence of infection of 40% in children living in a slum in Bangladesh (2). In Hué, Vietnam prevalence of Amoebic Liver Abscesses (ALA) was higher in a more densely populated area than in the city as a whole (3,4). In more affluent countries, where poor living conditions are less common, amoebiasis tends to be seen in certain groups, such as travellers returning from endemic areas (5), men who have sex with men and institutionalised individuals (6–9). Heterosexual and female homosexual activity can also transmit amoebiasis (10). Overall, men are more susceptible to invasive amoebiasis than women, despite similar infection rates (11). It is hypothesised that, in pathogenic *E. histolytica* infections, resistance to invasion is determined by a relatively small number of host genes (12).

The molecular biology of *Entamoeba* is complex, and much remains unknown, including chromosome number, ploidy and whether they undergo sexual reproduction. In an effort to better understand the biology of *E. histolytica*, its genome was sequenced along with that of the related species *Entamoeba dispar*. Since the first assembly and annotation of the *E. histolytica* genome in 2005 (13,14), significant advances have been made in understanding host–parasite interactions and virulence in *Entamoeba*. In this review, we describe the current status of genome annotation in virulent and nonvirulent *Entamoeba* species and review some of the important genes identified by genomic, proteomic and transcriptomic studies in the context of the pathogenic *E. histolytica* life cycle.

Entamoeba histolytica's pathogenic life cycle

Entamoeba histolytica has a two-stage life cycle, existing as resistant infective cysts in the environment and potentially

Correspondence: Gareth D. Weedall, Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK (e-mail: gweedall@liv.ac.uk).
Disclosure: None.

Received: 21 March 2011
Accepted for publication: 7 July 2011

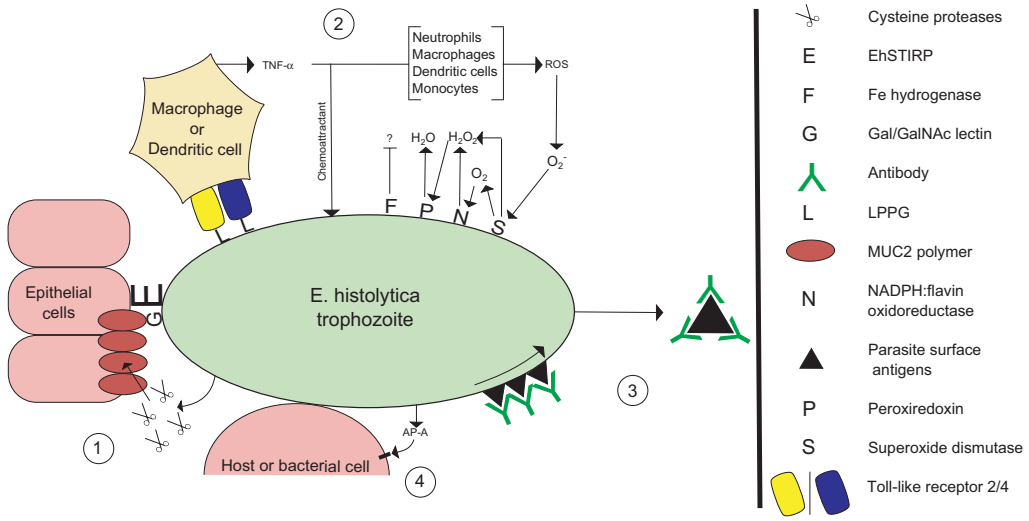


Figure 1 Key virulence factors of *Entamoeba histolytica* involved in pathogenic infections that have been identified by genome-scale investigations. 1 = Binding to epithelial extracellular matrix via Gal/GalNac lectin and EhSTIRP; and degradation of MUC2 polymers via secreted cysteine proteases. 2 = Subversion of host immune response, following binding of LPPG to host Toll-Like receptors 2 and 4, via degradation of reactive oxygen species by superoxide dismutase, NADPH:flavin oxidoreductase and peroxiredoxin. Fe-hydrogenase inhibits immune response by unknown mechanism. 3 = 'Capping and Shedding' of trophozoite surface antigens by host antibodies and lectins, involving cytoskeletal rearrangement to translocate antigen–antibody complexes to 'uroid' of cell for shedding. Putative function for EhROM1 in translocation. 4 = Direct contact between trophozoite and host or bacterial cell, leading to secretion of amoebapore-A, which forms pores in target cell membrane without need for receptor.

pathogenic trophozoites in the human colon. Upon excystation, trophozoites follow one of two paths. The more common path is commensal colonisation, where trophozoites inhabit the gut lumen and feed on enteric bacteria by phagocytosis, a process involving rearrangement of the amoebic cytoskeleton to internalise bacteria in lytic phagosomes (15). The less common path leads to invasive amoebiasis. Virulence factors allow the parasite to cause pathogenic amoebiasis via a variety of mechanisms, crucially including those that allow it to resist and subvert the host's innate and adaptive immune responses (Figure 1). Upon activation, previously commensal trophozoites degrade the colonic mucosal layer then bind to host epithelial cells (16,17). As reviewed by Lejeune *et al.* (18), the bound trophozoites trigger pathology in the host tissues, promoting penetration and infection. Apoptosis is induced in the trophozoite-bound epithelial cells as a result of cascading secretory proinflammatory cytokines. This cellular damage and the subsequent lateral invasion through the submucosa result in tissue inflammation and characteristic flask-shaped ulcers (19). The importance of apoptosis in amoebic virulence (20) is highlighted by studies on the leptin signalling pathway. Leptin signalling has multiple roles in the human body including regulation of the immune response to infection (towards a Th1 inflammatory response) and preventing apoptosis; however, experiments in mice show that it is leptin's anti-apoptotic role in gut epithelia, rather than its role in immune effector cells, which mediates susceptibility (21). An amino acid substitution (glutamine to arginine) in the leptin receptor is associated with increased susceptibility to, and severity of, infection in both mice and humans (22).

In many respects, the immune response to *E. histolytica* infection resembles that raised against the intestinal parasites *Cryptosporidium* and *Giardia* (23,24), with important roles for reactive oxygen species (ROS), nitric oxide (NO) and secreted IgA (25,26). Host immunity and pathology are closely linked. Human immune cells are recruited to the site of trophozoite invasion and, whilst attacking trophozoites, enhance the pathology caused by the invasion. NO and ROS released by immune effector cells damage *E. histolytica* trophozoites; however, the parasites have evolved means to minimise damage caused by these oxygen species, including the expression of various surface molecules (27–31) and internalisation and destruction of host immune cells (as well as other host cells) by phagocytosis (15).

Entamoeba histolytica also faces challenges from adaptive immunity. Adaptive immunity appears to protect against symptomatic disease, although not reinfection (32,33). The occurrence of subsequent infections indicates that immunity is either incomplete, ineffective against heterologous parasite strains or that the parasite utilises effective immune evasion strategies. For example, immunoglobulins binding to surface proteins may block adhesion

and activate the complement pathway. Trophozoites appear to be able to evade this arm of immunity by a process of 'capping and shedding' where bound antibodies are moved to the rear of the trophozoite, forming an 'uroid', and are shed. The host immune system is temporarily 'blind' to the parasite until different surface receptors are bound, at which point the process begins again (34,35).

Trophozoites that penetrate and cross the intestinal epithelium can be disseminated to other organs, most commonly the liver, where they form abscesses. Entering the relatively oxygen-rich environment of the bloodstream exposes the trophozoites to greater oxidative stress. In addition, greater exposure to humoral immunity and the complement system places the trophozoites at greater risk of inhibition and degradation. Consequently, it is likely that trophozoites require different molecular pathways to cause ALA, rather than remain as intestinal infections (36,37).

In support of this theory, virulent *E. histolytica* trophozoites exposed to conditions inducing heat shock demonstrate differential gene expression. According to a microarray analysis of 1131 transcripts, 471 genes were downregulated and 40 upregulated when cells grown at 37°C were incubated at 42°C for 4 h. It has been hypothesised that the large number of downregulated genes is indicative of a general molecular reaction to a heat shock-induced homeostatic imbalance (38).

After entering the hepatic sinusoids, pathogenic trophozoites invade the parenchyma. The hepatocytes and trophozoites are physically separated by a barrier of polymorphonuclear leukocyte (PMNs) and mononuclear host cells. The trophozoites make direct contact with the PMNs, resulting in lysis of the immune cells and the release of their own lytic enzymes, which damage surrounding hepatocytes. As surviving trophozoites reproduce and spread, the necrotic regions coalesce into abscesses. Immune epithelioid cells segregate these regions from healthy tissue, forming granulomas, in which the trophozoites are trapped with the expanding necrotic zones (37,39).

Differential virulence between *Entamoeba* species and strains

Evidently, not all *E. histolytica* strains are equally virulent. The genomic reference strain, *E. histolytica* HM-1:IMSS, is the best-studied virulent strain of *E. histolytica*, derived from a colonic biopsy taken from a man with dysentery in Mexico in 1967 (13,40). The *E. histolytica* Rahman strain was isolated from the stool of an Indian sailor in the UK in 1964 (41) and is considered to be avirulent. It has reduced ability to phagocytose erythrocytes, damage colonic epithelia and cause ALA, relative to HM-1:IMSS (29,42). A close human-infective relative of *E. histolytica* is *E. dispar*, which is morphologically indistinguishable from *E. histolytica* by microscopic analysis. Only in 1993 was it described as a distinct species, under the name '*dispar*' originally used by Brumpt in 1925 (43). *E. dispar* is avirulent. Tracking *E. dispar* (strain SAW1734) cells on human colonic explants shows that they do not break down the mucus barrier or cause epithelial cell damage, unlike *E. histolytica* HM-1:IMSS (Figure 2) (44). Recently, however, *E. dispar* has been associated both with cases of amoebic colitis and ALA, and its avirulence status has been questioned (45).

The genomes of *Entamoeba histolytica* and *Entamoeba dispar*

The draft assembly and annotation of the *E. histolytica* HM-1:IMSS genome was published in 2005 (13,14). A reassembly of the genome, including more sequence data and new annotation, was published in 2010 (46). The genome assembly and annotation was held on the Pathema website (http://pathema.jcvi.org/Pathema/) (47). More recently, the data have been made available on AmoebaDB (http://www.amoebadb.org), part of the EuPathDB web resource (48–50), along with the as-yet unpublished genome sequence of *E. dispar*. The *E. histolytica* genome assembly represents approximately 20 Mb of sequence, covered to >12.5 × depth (13,46). It remains fragmentary, comprising 1496 scaffolds, most likely due to the high number of repetitive elements in the genome (51). The *E. dispar* assembly is slightly larger than that of *E. histolytica* (∼22 Mb), but is sequenced to lower coverage depth (∼4.5×) and is more fragmentary (3312 scaffolds). A total of 8745 genes are predicted, slightly more than the ∼8300 for *E. histolytica*. Average divergence between orthologous genes of the two species is approximately 38% at synonymous sites (Weedall G., unpublished observations).

The reassembly and reannotation of the *E. histolytica* genome reduced the estimated number of genes from ∼10 000 to 8333, largely because of the removal of apparently artefactual paralogues, very short gene models and truncated genes (46). The majority of genes (∼55%) encode unknown proteins (Figure 3). This can be compared to other gut parasites, the apicomplexan *Cryptosporidium parvum* (40% of 4367 genes are annotated as hypothetical) and the diplomonad *Giardia lamblia* (75% of 9747 genes are annotated as hypothetical in isolate WB from assemblage A) (data from EuPathDB). The predominance of uncharacterised genes presents a problem for genome-wide analyses because the majority of genes of interest are often of unknown function. The facility to

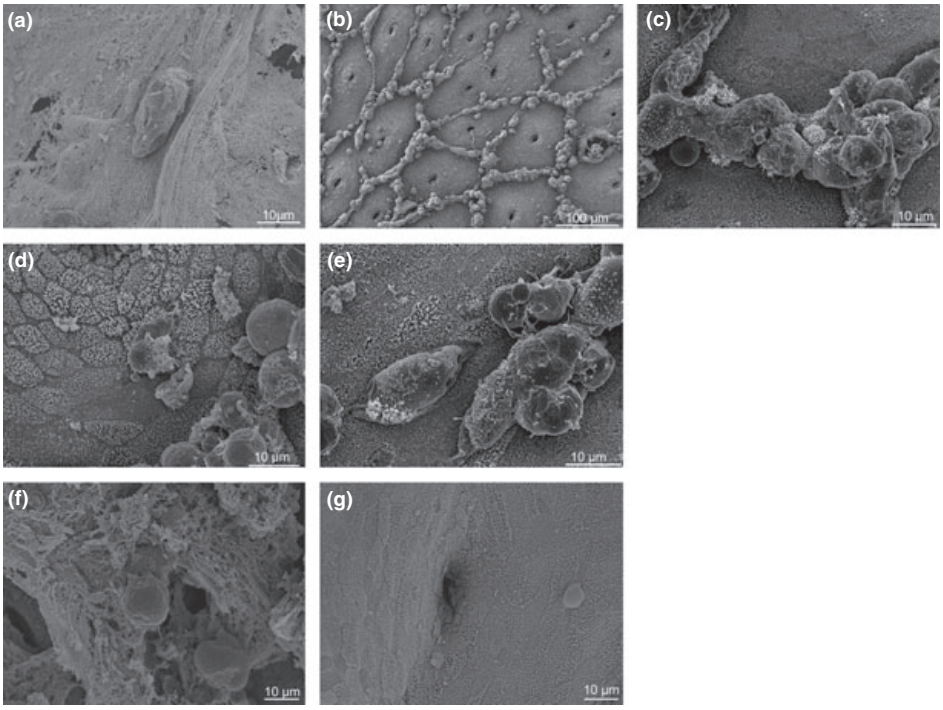


Figure 2 Comparison of colonisation of the colonic surface by *Entamoeba histolytica* and *Entamoeba dispar*. Panels show breakdown of mucus by *E. histolytica* after 0 h (a) and 2 h (b). Enlargement of region shows aggregates of trophozoites and recruited human cells (c). After 4 h, trophozoites begin to damage (d) and to penetrate epithelia (e). Conversely, after 4 h, *E. dispar* binds to, but does not degrade, the mucus barrier (f) and, as shown by manually removing the mucus layer, does not recruit immune cells to the epithelial surface (g). [Reprinted, with permission, from (44)].

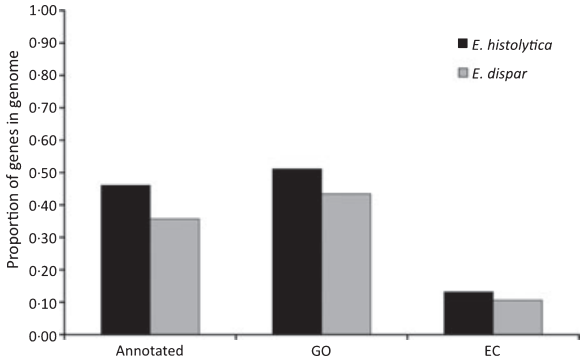


Figure 3 Comparison of the current status of the *Entamoeba histolytica* and *Entamoeba dispar* genome annotations, indicating the relative proportions of genes with putative functions. 'Annotated' = Percentage of non-hypothetical genes in the annotation; 'GO' = Percentage of genes associated with a 'Gene Ontology' term, i.e. those with either a cellular component, molecular function or biological process; 'EC' = Percentage of genes with 'Enzyme Commission' numbers, i.e. enzymes identified as being involved in known chemical reactions. Based upon figures from AmoebaDB (48–50). Based on most recent genome annotation (46).

upload corrected annotations to the genome is available (47,49), and such 'community annotation' has been encouraged (52). Researchers can post corrections to gene models, links to validating data and functional annotations that can be incorporated into future annotations. Annotation of hypothetical proteins in other species, such as *Plasmodium falciparum*, has been improved by using annotated genes with similar transcriptional profiles, annotated orthologues and automated literature mining (53). Similar methods may aid the annotation of *E. histolytica*.

Genome-scale analyses of *Entamoeba* virulence

Entamoeba genome sequences are used either as a means to identify sequences generated by processes such as mass spectrometry of peptides (29,54,55) or sequencing of cDNA from differential display PCR (36,56), or to design microarrays for hybridisation-based analyses (57–61). Many genes involved in amoebic pathogenesis have been identified by genome-wide analyses. Investigations have compared gene expression in the same strain in different environments, identifying genes that may be important for

survival in these environments (36,59), and compared cell lines that show different virulence characteristics (29,57,58,60,62).

A DNA microarray created from a clone library representing 2110 unique genes has been used to compare diversity of genomic DNA among *E. histolytica* and *E. dispar* strains (63) and transcriptional differences between *E. histolytica* HM-1:IMSS, *E. histolytica* Rahman and *E. dispar* SAW760 (60). A 70-bp oligonucleotide DNA microarray representing 6242 unique *E. histolytica* HM-1:IMSS genes has been used to compare transcriptional differences between HM-1:IMSS and Rahman and to compare syngenic cell lines of differential virulence derived from HM-1:IMSS (57,58). A different microarray using 25-bp oligonucleotide probes representing 9435 *E. histolytica* HM-1:IMSS open reading frames has been used to compare *E. histolytica* trophozoites from murine intestinal infections and from *in vitro* culture (59) and to compare the transcriptional responses of HM-1:IMSS and Rahman to oxidative and nitrosative stress (61).

The numbers of putative differentially expressed genes among strains vary with the different methods and criteria used to define differential expression. However, broad trends are apparent. A greater proportion of *E. dispar* genes than Rahman genes appear to be downregulated relative to HM-1:IMSS (58,60). A number of genes are downregulated in both avirulent cell lines. For instance, of 32 genes with lower mRNA expression in Rahman, 29 also showed lower expression in *E. dispar* (60). The following sections describe some of the genes identified by these studies as potentially important virulence factors.

Genes involved in survival and virulence in the intestine

Experimental infections of mouse intestines induced differential expression (twofold or greater) of 523 genes: 326 on day 1 post-infection, 109 on day 29 post-infection and 88 at both time points (59). The authors speculated that an initial stress response associated with adaptation to the new environment might partly explain the large number of genes differentially regulated early in infection. Among putative virulence factors showing differential expression were cysteine proteases and members of the galactose- and *N*-acetyl-$_D$-galactosamine-binding lectin (Gal/GalNAc–lectin) complex on the parasite surface (16).

An important process in amoebic virulence is the degradation of the mucus layer, which enables the trophozoites to reach the gut epithelial layer (Figure 2). Trophozoites release cysteine proteases to degrade the main component of the mucus barrier, MUC2. Different members of the cysteine protease gene family are expressed in culture and in mouse intestine, suggesting that different gene family members may play unique roles important in different environments (59). Cysteine protease expression is lower overall in *E. dispar* than in *E. histolytica* (64), indicating their role in virulence, and CP-A5 (gene ID, EHI_168240), a key protease for the degradation of the MUC2 polymer (17,65,66), is a pseudogene in *E. dispar* (67). However, CP-A5 showed no statistically significant differential expression between *E. histolytica* HM-1:IMSS and Rahman. CP-A4 (EHI_050570), CP-A6 (EHI_151440) and CP-B1 (EHI_117650) were expressed to a greater degree in HM-1:IMSS, whereas CP-A3 (EHI_159610), CP-A7 (EHI_039610) and CP-B9 (EHI_181230) were greater in Rahman (58). Numerous cysteine protease genes (e.g. EHI_127470, EHI_019390, EHI_144040 and EHI_132640) are pseudogenes in the *E. histolytica* genome, and this, along with their divergence from *E. dispar* orthologues (64), suggests a degree of evolutionary plasticity in this gene family.

Trophozoites bind to the mucus layer and to epithelial cells via the Gal/GalNAc–lectin complex (16). Two genes encoding light subunits in the Gal/GalNAc lectin – *lgl*2 (EHI_049690) and *lgl3* (EHI_027800) – were downregulated to different degrees during the course of intestinal infection (59). The importance of the downregulation of *lgl3* in invasive infection was supported by transcriptional analysis showing 22-fold higher expression in the nonvirulent Rahman strain, compared with HM-1:IMSS (58). Other molecules involved in binding to host cells include the *E. histolytica* serine-, threonine- and isoleucine-rich proteins (EhSTIRP) (68). These proteins are encoded by a small gene family in *E. histolytica* (EHI_012330, EHI_004340 and EHI_025700).

Surviving host responses to invasive amoebiasis

Nitric oxide and ROS released by neutrophils, macrophages, monocytes and dendritic cells constitute a major threat to the trophozoites, which they can counteract by the actions of a number of molecules expressed on their surfaces: peroxiredoxin, superoxide dismutase (SOD) and NADPH:flavin oxidoreductase (27–29,31). SOD generates H_2O_2 in the presence of O_2^-, NADPH:flavin oxidoreductase catalyses the reduction of O_2 to H_2O_2, and peroxiredoxin reduces the H_2O_2 from both pathways to H_2O (69). Fe-hydrogenase, which, in bacteria, is involved in survival of oxidative stress (30), is also expressed by the trophozoites (58). Large numbers of genes show differential regulation in response to oxidative and nitrosative stress, and there is a substantial overlap in the genes involved in these responses. HM-1:IMSS shows a more robust response to stress than Rahman, with more genes differentially regulated overall and to a greater degree (61).

Peroxiredoxin is more highly expressed in *E. histolytica* HM-1:IMSS than in Rahman according to analyses of

protein (29) and mRNA (60) expression. Furthermore, it is downregulated in *E. dispar* relative to *E. histolytica* HM-1:IMSS (28,60). In *E. histolytica* HM-1:IMSS, peroxiredoxin is expressed on the surface where it is co-localised with the Gal/GalNAc lectin, possibly to degrade ROS released from bound immune cells (28). In contrast, expression in *E. dispar* is restricted to the cytoplasm, suggesting an inability of *E. dispar* to survive the oxidative burst that would be inflicted upon it following host invasion. *E. histolytica* peroxiredoxin sequences are highly divergent from their *E. dispar* orthologues (63). The current *E. histolytica* genome annotation contains a number of putative peroxiredoxin genes (EHI_001420, EHI_061980, EHI_114010, EHI_122310, EHI_123390, EHI_201250, EHI_145840, EHI_018740, EHI_183180 and EHI_084260) and pseudogenes (EHI_121620, EHI_139570 and EHI_172720). Whether all of these genes are real, functional and expressed remains to be determined. If so, it is possible that gene copy number variations between strains and species of *Entamoeba* affect overall gene expression levels.

Involvement of other putative oxidative stress response genes in virulence is less clear. Fe-hydrogenase (EHI_073390) is more highly expressed in *E. histolytica* HM-1:IMSS than in *E. dispar*, suggesting a role in virulence, yet within *E. histolytica*, it is more highly expressed in Rahman than HM-1:IMSS (58,60). SOD (EHI_159160) is also more highly expressed in *E. histolytica* Rahman than in HM-1:IMSS (29). SOD does appear to play a role in oxidative stress resistance: increased expression of SOD and peroxiredoxin is associated with metronidazole resistance, implying an involvement in detoxification of nitrogen-based free radicals generated by metronidazole activation (70,71).

Immunoglobulins binding to amoebic surface proteins can disrupt trophozoite cell functions, block adhesion to host receptors and activate the complement pathway. The parasite can avoid these outcomes by cysteine protease-mediated clipping of bound antibodies and complement (72,73) and by shedding the bound antibodies from its surface. Binding of host antibodies to amoebic surface antigens induces actin- and myosin-mediated redistribution to a membranous posterior appendage of the cell, the 'uroid', where this 'cap' is shed mostly as membrane-bound vesicles (34,35,74). A rhomboid protease involved in shedding surface proteins, EhROM1 (EHI_197460), was identified by searching the *E. histolytica* genome sequence for motifs conserved across known rhomboid proteases (75). EhROM1 specifically cleaves the heavy chain subunit of the Gal/GalNAc lectin and localises to the uroid (75). However, EhROM1 knock-down mutants showed no significant change in cap formation or complement resistance, but did show reduced ability to adhere to host cells and reduced phagocytic ability (76), suggesting a novel role for this protein (75,76). Proteomic analysis of uroid-extruded vesicles identified several surface-linked proteins, in addition to the Gal/GalNAc lectin, that are apparently capped and discarded, implying that they are involved in host–amoeba interactions. These included calreticulin, a multifunctional antigen with a notable involvement in calcium signalling, and the variable surface antigen M17 (77).

A number of proteins with uncertain functions show differential expression between *E. histolytica* HM-1:IMSS and the noninvasive Rahman and *E. dispar* (29,59). Grainin-1 was upregulated in *E. histolytica* Rahman, and grainin-2 was upregulated in both nonvirulent cell lines. The sequences of both grainins contain at least one metal-ligating EH-hand motif, commonly seen in proteins that bind calcium. Both genes are upregulated in culture in response to inducers of programmed cell death (PCD), and a stress-response role diminishing intracellular Ca^{2+} was suggested (78), as was a role in calcium-dependent endocytosis and granular exocytosis, aiding pathology (79). Lower levels of expression of both genes in mouse intestines, and of grainin 1 in ALA samples, were seen relative to *in vitro* cultures, possibly owing to higher stress levels in *in vitro* conditions (36,59). In the current genome assembly, seven putative grainin genes are annotated: grainin 1 (EHI_167300) is relatively divergent from its nearest paralogues (EHI_120360, 71% amino acid identity; EHI_060380, 57% identity); grainin 2 (EHI_167310) has a shorter, near-identical, paralogue (EHI_111720); both are relatively divergent from their nearest paralogues (EHI_164430, EHI_164440, both 56% amino acid identity).

A LIM domain-containing protein (EHI_096420) was more highly expressed in *E. histolytica* HM-1:IMSS than in Rahman or *E. dispar* (29,54). Its function is not known, but it has been shown to localise to the plasma membrane and to bind to the actin cytoskeleton via its LIM domain (80). Alcohol dehydrogenase 3 (ADH3) was more highly expressed in HM-1:IMSS relative to Rahman and *E. dispar* (29,54). ADH3 (EHI_125950 and EHI_198760) is expressed at greater levels on the cell surface of HM-1:IMSS than *E. dispar* and, when overexpressed in HM-1:IMSS and Rahman cells, increased host inflammatory response, although no definite role in virulence was determined (54). ADH2 (EHI_024240, EHI_150490 and EHI_160940) was more highly expressed in HM-1:IMSS than in *E. dispar* (54). ADH2 is associated with the cell membrane and is involved in iron scavenging from the host's transferrin (81).

Trophozoites can phagocytose host epithelial cells, erythrocytes and immune cells. Phagocytosis is modulated by

the motor protein myosin IB, which cross-links actin filaments, to restructure the amoebic cytoskeleton as necessary (82). Proteomic analysis, by liquid chromatography and tandem mass spectroscopy (LC-MS/MS), of phagosome proteins allows identification of proteins differentially expressed over time and in different conditions. In wild-type *E. histolytica* HM-1:IMSS and a strain overexpressing myosin IB (MyoIB+), approximately 1000 proteins were identified overall. Of these, about 150 proteins present in the early phagosome were associated with the cytoskeleton (including actin, coactosin and talin), were signalling molecules (including PI3-K and Ras GAP) or were involved in intracellular trafficking (including calreticulin). Of those associated with the cytoskeleton, seven proteins were functionally linked to myosin IB, demonstrable by their expression in detectable levels in MyoIB+ only (83). Also in HM-1:IMSS, of 159 phagosomal proteins detected, 51 were constitutively expressed, whilst the remaining 108 showed differential expression across the monitored 2-h period. Those constitutively expressed included CP-A5, actin and the Gal/GalNAc lectin. The more numerous transient proteins included many Rab GTPases and several of the Rac cytoskeletal proteins, reflecting the necessary fluidity of the cytoskeleton in the phagocytic process. The same study reported inter-strain variation in expressed *E. histolytica* phagosome proteins, suggesting a role in differential virulence (84).

Virulence factors involved in amoebic liver abscess

Death from amoebiasis results mainly from the formation of abscesses on the liver after trophozoites escape the gut, so understanding the molecular basis of abscess formation is of considerable interest. Comparisons of the transcriptomes of *E. histolytica* trophozoites axenically cultured *in vitro* with those isolated from liver abscesses using differential display PCR (DD-PCR) identified small numbers of genes differentially expressed between the two (36,56). Among these were genes encoding grainin-1, a flavoprotein, a GTP-binding protein and ribosomal proteins (36,56).

A cell line derived from HM-1:IMSS ('HM1A'), which has lost the ability to cause ALA, has been compared to virulent HM-1:IMSS ('HM1B') at both proteomic and transcriptomic levels (57,62). Eighty-seven genes showed twofold or greater differential (mRNA) expression between HM1A (47 genes upregulated) and HM1B (40 genes upregulated) (57). Thirty-one proteins showed 2.3-fold or greater differential protein expression between HM1A (21 upregulated) and HM1B (10 upregulated) (62). Only two genes, Fe-hydrogenase-2 (EHI_005060) and a C2-domain-containing protein (EHI_069320), were found differentially expressed (upregulated in HM1A) at both the proteomic and the transcriptomic levels. Despite using the same microarray, little overlap was seen in the transcripts downregulated in HM-1:IMSS clone A and in Rahman, relative to HM-1:IMSS clone B (57,58). Only 1 gene was significantly downregulated in both Rahman and avirulent HM-1:IMSS, and of the 152 transcripts upregulated in *E. histolytica* HM-1:IMSS relative to Rahman, only five were also significantly upregulated in the pathogenic HM-1:IMSS clone B relative to clone A. Two of these five genes encoded AIG1-like proteins. AIG1 proteins are small GTPases originally identified in *Arabidopsis thaliana* (85) where they confer resistance to bacterial infections. AIG1-like proteins are encoded by a large gene family in *E. histolytica* (57,59) and may be involved in bacterial interactions. This lack of overlap suggests that the nature of avirulence in Rahman and HM-1:IMSS clone A may be quite different.

Another investigation comparing virulent and avirulent lines derived from the *E. histolytica* HM-1:IMSS strain compared mRNA expression in 1130 genes and showed downregulation (>twofold, $P<0.05$) of 21 genes and upregulation of 29 genes in the virulent line (86). Among the upregulated genes in the virulent line were the surface antigen ariel-1, which has been shown to be absent from *E. dispar* (87), and several lysine-rich proteins ('KRiPs') and lysine- and glutamic acid-rich proteins ('KERPs'). Gene knock-down of KERP1 using antisense RNA reduced the formation of liver abscesses (86).

None of these studies identified the virulence factor amoebapore-A (AP-A; EHI_159480). The amoebapore's role in pathogenesis has been demonstrated in hamster and severe combined immunodeficient (SCID) mouse livers (88,89). AP-A appeared to be essential for ALA formation in hamsters, but suppression in the mouse model did not completely prevent ALA, suggesting that other processes are important in ALA formation. AP-A is inserted into host plasma membranes, without the need to bind to a host receptor, upon direct contact between a trophozoite and a host cell (90), forming pores and lysing the host cell (91). Amoebapores also have a bacteriolytic function, being able to lyse gram-positive bacteria (88,90).

Characterising candidate virulence factors

Characterisation of gene function has proven difficult in *Entamoeba* as gene knock-outs have not been achieved. There has, however, been some success with transcriptional gene silencing and, more recently, with RNAi-mediated gene knock-down (92–94). The gene encoding AP-A has been silenced in some, although not all, cell lines, using what was originally designed to be a putative overexpression vector. The mechanism of silencing is not known for certain, although involvement of a short interspersed

element (SINE) and of tRNA repeat arrays in the vector have been proposed (95,96). The 'G3' *E. histolytica* cells this silencing mutation gave rise to are virulence attenuated, being impaired in their ability to digest phagocytosed cells (although not impaired in their ability to phagocytose them in the first instance) and unable to cause ALA (88,97). Cell lines that have been silenced for AP-A expression continued to show AP-A silencing even when selection for the vector was removed, although in other cases, silencing has not been integrated permanently into the cell lineage and future generations have reverted to their wild type (98). Moreover, additional silencing of genes could be achieved in this line using a vector with an additional gene in it. By this method, CP-A5 and *Ehlgl* were silenced (99,100). Gene silencing affected multiple members of the gene families containing the target gene (99), and, interestingly, downregulation of several *lgl* genes led to upregulation of others, a possible compensatory mechanism. Silencing of EhLIM-A – the gene encoding a LIM-like protein – has been achieved in a similar fashion (80). RNAi-mediated gene knock-down has been achieved using different methods of administering the siRNA: bacterial expression of double-stranded RNA followed by either adding the bacteria to *Entamoeba* culture or extracting the dsRNA and soaking *Entamoeba* trophozoites in them (94), or addition of vectors expressing short hairpin RNA (sense and antisense linked by a loop) to the trophozoites (92). Beta-tubulin, KERP1, URE3-BP, IGl and EhC2A have been 'knocked down' by these methods. Continued improvement of molecular tools for targeted gene silencing will help to characterise the roles of specific genes and gene families in host–parasite interactions.

Concluding Remarks

The genome-scale studies made possible by sequencing of the *E. histolytica* genome have greatly improved our knowledge of the pathogenesis of *E. histolytica* and identified many genes that may play important roles in host–parasite interactions. Comparisons of different strains of *E. histolytica* and of the related species *E. dispar* show differences in sequence and in expression that may account for different virulence profiles. In order for further genome-scale studies into genetic and gene expression differences to be successful in the future improved gene annotation is vital. It is hoped that a model of 'community annotation' may help rapidly improve and disseminate information characterising *Entamoeba* genes. Much work has yet to be done before we understand the complexities of *Entamoeba* virulence. Continual improvement to the assembly and annotation of *Entamoeba* genomes is central to this effort.

ACKNOWLEDGEMENTS

We are grateful to Dr Elisabeth Labruyère (Institut Pasteur, Paris), Dr Vikas Sharma (Imperial College, London) and Dr Kevin Tetteh (London School of Hygiene and Tropical Medicine), as well as two anonymous reviewers, for helpful comments on the manuscript.

REFERENCES

1 Walsh JA. Problems in recognition and diagnosis of amebiasis: estimation of the global magnitude of morbidity and mortality. *Rev Infect Dis* 1986; **8**: 228–238.
2 Haque R, Ali IM, Sack RB, Farr BM, Ramakrishnan G & Petri WA. Amebiasis and mucosal IgA antibody against the *Entamoeba histolytica* adherence lectin in Bangladeshi children. *J Infect Dis* 2001; **183**: 1787–1793.
3 Blessmann J, Van LP, Nu PAT, *et al.* Epidemiology of amebiasis in a region of high incidence of amebic liver abscess in central Vietnam. *Am J Trop Med Hyg* 2002; **66**: 578–583.
4 Blessmann J, Le Van A & Tannich E. Epidemiology and treatment of amebiasis in Hué, Vietnam. *Arch Med Res* 2006; **37**: 270–272.
5 Weinke T, Friedrich-Jänicke B, Hopp P & Janitschke K. Prevalence and clinical importance of *Entamoeba histolytica* in two high-risk groups: travelers returning from the tropics and male homosexuals. *J Infect Dis* 1990; **161**: 1029–1031.
6 Stark D, Fotedar R, van Hal S, *et al.* Prevalence of enteric protozoa in human immunodeficiency virus (HIV)-positive and HIV-negative men who have sex with men from Sydney, Australia. *Am J Trop Med Hyg* 2007; **76**: 549–552.
7 Stark D, van Hal SJ, Matthews G, Harkness J & Marriott D. Invasive amebiasis in men who have sex with men, Australia. *Emerg Infect Dis* 2008; **14**: 1141–1143.
8 Rivera WL, Santos SR & Kanbara H. Prevalence and genetic diversity of *Entamoeba histolytica* in an institution for the mentally retarded in the Philippines. *Parasitol Res* 2006; **98**: 106–110.
9 Nishise S, Fujishima T, Kobayashi S, *et al.* Mass infection with *Entamoeba histolytica* in a Japanese institution for individuals with mental retardation: epidemiology and control measures. *Ann Trop Med Parasitol* 2010; **104**: 383–390.
10 Salit IE, Khairnar K, Gough K & Pillai DR. A possible cluster of sexually transmitted *Entamoeba histolytica*: genetic analysis of a highly virulent strain. *Clin Infect Dis* 2009; **49**: 346–353.
11 Acuna-Soto R, Maguire JH & Wirth DF. Gender distribution in asymptomatic and invasive amebiasis. *Am J Gastroenterol* 2000; **95**: 1277–1283.
12 Hamano S, Becker S, Asgharpour A, *et al.* Gender and genetic control of resistance to intestinal amebiasis in inbred mice. *Genes Immun* 2008; **9**: 452–461.
13 Loftus B, Anderson I, Davies R, *et al.* The genome of the protist parasite *Entamoeba histolytica*. *Nature* 2005; **433**: 865–868.
14 Clark CG, Alsmark UCM, Tazreiter M, *et al.* Structure and content of the *Entamoeba histolytica* genome. *Adv Parasitol* 2007; **65**: 51–190.
15 Voigt H, Olivo J-C, Sansonetti P & Guillén N. Myosin IB from *Entamoeba histolytica* is involved in phagocytosis of human erythrocytes. *J Cell Sci* 1999; **112**: 1191–1201.
16 Petri WA, Haque R & Mann BJ. The bittersweet interface of parasite and host: lectin-carbohydrate interactions during human

invasion by the parasite *Entamoeba histolytica*. *Annu Rev Microbiol* 2002; **56**: 39–64.
17 Lidell ME, Moncada DM, Chadee K & Hansson GC. *Entamoeba histolytica* cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel. *Proc Natl Acad Sci USA* 2006; **103**: 9298–9303.
18 Lejeune M, Rybicka JM & Chadee K. Recent discoveries in the pathogenesis and immune response toward *Entamoeba histolytica*. *Future Microbiol* 2009; **4**: 105–118.
19 Stanley SJ. Amoebiasis. *Lancet* 2003; **361**: 1025–1034.
20 Becker SM, Cho K-N, Guo X, *et al.* Epithelial cell apoptosis facilitates *Entamoeba histolytica* infection in the gut. *Am J Pathol* 2010; **176**: 1316–1322.
21 Guo X, Roberts MR, Becker SM, *et al.* Leptin signaling in intestinal epithelium mediates resistance to enteric infection by *Entamoeba histolytica*. *Mucosal Immunol* 2011; **4**: 294–303.
22 Duggal P, Guo X, Haque R, *et al.* A mutation in the leptin receptor is associated with *Entamoeba histolytica* infection in children. *J Clin Investig* 2011; **121**: 1191–1198.
23 Petry F, Jakobi V & Tessema TS. Host immune response to *Cryptosporidium parvum* infection. *Exp Parasitol* 2010; **126**: 304–309.
24 Solaymani-Mohammadi S & Singer SM. *Giardia duodenalis*: the double-edged sword of immune responses in giardiasis. *Exp Parasitol* 2010; **126**: 292–297.
25 Guo X, Houpt E & Petri WA. Crosstalk at the initial encounter: interplay between host defense and ameba survival strategies. *Curr Opin Immunol* 2007; **19**: 376–384.
26 Carrero JC, Cervantes-Rebolledo C, Aguilar-Díaz H, Díaz-Gallardo MY, Laclette JP & Morales-Montor J. The role of the secretory immune response in the infection by *Entamoeba histolytica*. *Parasite Immunol* 2007; **29**: 331–338.
27 Bruchhaus I & Tannich E. Induction of the iron-containing superoxide dismutase in *Entamoeba histolytica* by a superoxide anion-generating system or by iron chelation. *Mol Biochem Parasitol* 1994; **67**: 281–288.
28 Choi M-H, Sajed D, Poole L, *et al.* An unusual surface peroxiredoxin protects invasive *Entamoeba histolytica* from oxidant attack. *Mol Biochem Parasitol* 2005; **143**: 80–89.
29 Davis PH, Zhang X, Guo J, Townsend RR & Stanley SL. Comparative proteomic analysis of two *Entamoeba histolytica* strains with different virulence phenotypes identifies peroxiredoxin as an important component of amoebic virulence. *Mol Microbiol* 2006; **61**: 1523–1532.
30 Fournier M, Dermoun Z, Durand M-C & Dolla A. A new function of the *Desulfovibrio vulgaris* Hildenborough [Fe] hydrogenase in the protection against oxidative stress. *J Biol Chem* 2004; **279**: 1787–1793.
31 Lo HS & Reeves RE. Purification and properties of NADPH: flavin oxidoreductase from *Entamoeba histolytica*. *Mol Biochem Parasitol* 1980; **2**: 23–30.
32 Blessmann J, Ali IKM, Nu PAT, *et al.* Longitudinal study of intestinal *Entamoeba histolytica* infections in asymptomatic adult carriers. *J Clin Microbiol* 2003; **41**: 4745–4750.
33 Haque R, Mondal D, Duggal P, *et al.* *Entamoeba histolytica* infection in children and protection from subsequent amebiasis. *Infect Immun* 2006; **74**: 904–909.
34 Calderón J & Avila EE. Antibody-induced caps in *Entamoeba histolytica*: isolation and electrophoretic analysis. *J Infect Dis* 1986; **153**: 927–932.
35 Espinosa-Cantellano M & Martínez-Palomo A. *Entamoeba histolytica*: mechanism of surface receptor capping. *Exp Parasitol* 1994; **79**: 424–435.
36 Bruchhaus I, Roeder T, Lotter H, Schwerdtfeger M & Tannich E. Differential gene expression in *Entamoeba histolytica* isolated from amoebic liver abscess. *Mol Microbiol* 2002; **44**: 1063–1072.
37 Santi-Rocca J, Rigothier M-C & Guillén N. Host-microbe interactions and defense mechanisms in the development of amoebic liver abscesses. *Clin Microbiol Rev* 2009; **22**: 65–75.
38 Weber C, Guigon G, Bouchier C, *et al.* Stress by heat shock induces massive down regulation of genes and allows differential allelic expression of the Gal/GalNAc lectin in *Entamoeba histolytica*. *Eukaryot Cell* 2006; **5**: 871–875.
39 Tsutsumi V, Mena-Lopez R, Anaya-Velazquez F & Martinez-Palomo A. Cellular bases of experimental amebic liver abscess formation. *Am J Pathol* 1984; **117**: 81–91.
40 Diamond LS, Mattern CFT & Bartgis IL. Viruses of *Entamoeba histolytica*. I. Identification of transmissible virus-like agents. *J Virol* 1972; **9**: 326–341.
41 Sargeaunt PG, Williams JE & Neal RA. A comparative study of *Entamoeba histolytica* (NIH:200, HK9, etc.), "*E. histolytica*-like" and other morphologically identical amoebae using isoenzyme electrophoresis. *Trans R Soc Trop Med Hyg* 1980; **74**: 469–474.
42 Burchard GD & Mirelman D. *Entamoeba histolytica*: virulence potential and sensitivity to Metronidazole and Emetine of four isolates possessing nonpathogenic zymodemes. *Exp Parasitol* 1988; **66**: 231–242.
43 Diamond LS & Clark CG. A redescription of *Entamoeba histolytica* Schaudinn, 1903 (Emended Walker, 1911) separating it from *Entamoeba dispar* Brumpt, 1925. *J Eukaryot Microbiol* 1993; **40**: 340–344.
44 Bansal D, Ave P, Kerneis S, *et al.* An *ex-vivo* human intestinal model to study *Entamoeba histolytica* pathogenesis. *PLoS Negl Trop Dis* 2009; **3**: e551.
45 Ximénez C, Cerritos R, Rojas L, *et al.* Human amebiasis: breaking the paradigm? *Int J Environ Res Public Health* 2010; **7**: 1105–1120.
46 Lorenzi HA, Puiu D, Miller JR, *et al.* New assembly, reannotation and analysis of the *Entamoeba histolytica* genome reveal new genomic features and protein content information. *PLoS Negl Trop Dis* 2010; **4**: e716.
47 Brinkac LM, Davidsen T, Beck E, *et al.* Pathema: a clade-specific bioinformatics resource center for pathogen research. *Nucleic Acids Res* 2010; **38**: D408–D414.
48 Aurrecoechea C, Barreto A, Brestelli J, *et al.* AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Res* 2011; **39**: D612–D619.
49 Aurrecoechea C, Brestelli J, Brunk BP, *et al.* EuPathDB: a portal to eukaryotic pathogen databases. *Nucleic Acids Res* 2010; **38**: D415–D419.
50 Aurrecoechea C, Heiges M, Wang H, *et al.* ApiDB: integrated resources for the apicomplexan bioinformatics resource center. *Nucleic Acids Res* 2007; **35**: D427–D430.
51 Lorenzi H, Thiagarajan M, Haas B, Wortman J, Hall N & Caler E. Genome wide survey, discovery and evolution of repetitive elements in three *Entamoeba* species. *BMC Genomics* 2008; **9**: 595.
52 Gilchrist CA & Petri WA. Using differential gene expression to study *Entamoeba histolytica* pathogenesis. *Trends Parasitol* 2009; **25**: 124–131.
53 Zhou Y, Ramachandran V, Kumar KA, *et al.* Evidence-based annotation of the malaria parasite's genome using comparative expression profiling. *PLoS ONE* 2008; **3**: e1570.
54 Davis PH, Chen M, Zhang X, Clark CG, Townsend RR & Stanley SL. Proteomic comparison of *Entamoeba histolytica* and *Entamoeba dispar* and the role of *E. histolytica* alcohol dehydrogenase 3 in virulence. *PLoS Negl Trop Dis* 2009; **3**: e415.
55 Leitsch D, Wilson IB, Paschinger K & Duchêne M. Comparison of the proteome profiles of *Entamoeba histolytica* and its close but non-pathogenic relative *Entamoeba dispar*. *Middle Eur J Med* 2006; **118** (Suppl. 3): 37–41.
56 Balderas-Renteria I, García-Lázaro JF, Carranza-Rosales P, Morales-Ramos LH, Galan-Wong LJ & Muñoz-Espinosa LE. Transcriptional upregulation of genes related to virulence activation in *Entamoeba histolytica*. *Arch Med Res* 2007; **38**: 372–379.
57 Biller L, Davis PH, Tillack M, *et al.* Differences in the transcriptome signatures of two genetically related *Entamoeba histolytica* cell lines derived from the same isolate with different pathogenic properties. *BMC Genomics* 2010; **11**: 63.
58 Davis PH, Schulze J & Stanley SL. 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. *Mol Biochem Parasitol* 2007; **151**: 118–128.
59 Gilchrist CA, Houpt E, Trapaidze N, *et al.* Impact of intestinal colonization and invasion on the *Entamoeba histolytica* transcriptome. *Mol Biochem Parasitol* 2006; **147**: 163–176.

60 MacFarlane RC & Singh U. Identification of differentially expressed genes in virulent and nonvirulent *Entamoeba* species: potential implications for amebic pathogenesis. *Infect Immun* 2006; **74**: 340–351.

61 Vicente JB, Ehrenkaufer GM, Saraiva LM, Teixeira M & Singh U. *Entamoeba histolytica* modulates a complex repertoire of novel genes in response to oxidative and nitrosative stresses: implications for amebic pathogenesis. *Cell Microbiol* 2009; **11**: 51–69.

62 Biller L, Schmidt H, Krause E, *et al.* Comparison of two genetically related *Entamoeba histolytica* cell lines derived from the same isolate with different pathogenic properties. *Proteomics* 2009; **9**: 4107–4120.

63 Shah PH, MacFarlane RC, Bhattacharya D, *et al.* Comparative genomic hybridizations of *Entamoeba* strains reveal unique genetic fingerprints that correlate with virulence. *Eukaryot Cell* 2005; **4**: 504–515.

64 Tannich E, Scholze H, Nickel R & Horstmann RD. Homologous cysteine proteinases of pathogenic and nonpathogenic *Entamoeba histolytica.* Differences in structure and expression. *J Biol Chem* 1991; **266**: 4798–4803.

65 Bruchhaus I, Jacobs T, Leippe M & Tannich E. *Entamoeba histolytica* and *Entamoeba dispar*: differences in numbers and expression of cysteine proteinase genes. *Mol Microbiol* 1996; **22**: 255–263.

66 Mortimer L & Chadee K. The immunopathogenesis of *Entamoeba histolytic*a. *Exp Parasitol* 2010; **126**: 366–380.

67 Willhoeft U, Hamann L & Tannich E. A DNA sequence corresponding to the gene encoding Cysteine Proteinase 5 in *Entamoeba histolytica* is present and positionally conserved but highly degenerated in *Entamoeba dispar*. *Infect Immun* 1999; **67**: 5925–5929.

68 MacFarlane RC & Singh U. Identification of an *Entamoeba histolytica* Serine-, Threonine-, and isoleucine-rich protein with roles in adhesion and cytotoxicity. *Eukaryot Cell* 2007; **6**: 2139–2146.

69 Nandi N, Sen A, Banerjee R, *et al.* Hydrogen peroxide induces apoptosis-like death in *Entamoeba histolytica* trophozoites. *Microbiology* 2010; **156**: 1926–1941.

70 Samarawickrema NA, Brown DM, Upcroft JA, Thammapalerd N & Upcroft P. Involvement of superoxide dismutase and pyruvate:ferredoxin oxidoreductase in mechanisms of metronidazole resistance in *Entamoeba histolytica*. *J Antimicrob Chemother* 1997; **40**: 833–840.

71 Wassmann C, Hellberg A, Tannich E & Bruchhaus I. Metronidazole resistance in the protozoan parasite *Entamoeba histolytica* is associated with increased expression of iron-containing superoxide dismutase and peroxiredoxin and decreased expression of ferredoxin 1 and flavin reductase. *J Biol Chem* 1999; **274**: 26051–26056.

72 Tran VQ, Herdman DS, Torian BE & Reed SL. The neutral cysteine proteinase of *Entamoeba histolytica* degrades IgG and prevents its binding. *J Infect Dis* 1998; **177**: 508–511.

73 Garcia-Nieto RM, Rico-Mata R, Arias-Negrete S & Avila EE. Degradation of human secretory IgA1 and IgA2 by *Entamoeba histolytica* surface-associated proteolytic activity. *Parasitol Int* 2008; **57**: 417–423.

74 Arhets P, Olivo JC, Gounon P, Sansonetti P & Guillén N. Virulence and functions of myosin II are inhibited by overexpression of light meromyosin in *Entamoeba histolytica*. *Mol Biol Cell* 1998; **8**: 1537–1547.

75 Baxt LA, Baker RP, Singh U & Urban S. An *Entamoeba histolytica* rhomboid protease with atypical specificity cleaves a surface lectin involved in phagocytosis and immune evasion. *Genes Dev* 2008; **22**: 1636–1646.

76 Baxt LA, Rastew E, Bracha R, Mirelman D & Singh U. Downregulation of an *Entamoeba histolytica* rhomboid protease reveals roles in regulating parasite adhesion and phagocytosis. *Eukaryot Cell* 2010; **9**: 1283–1293.

77 Marquay Markiewicz J, Syan S, Hon C-C, Weber C, Faust D & Guillén N. A proteomic and cellular analysis of uropods in the pathogen *Entamoeba histolytica*. *PLoS Negl Trop Dis* 2011; **5**: e1002.

78 Monroy VS, Flores MOM, Villalba-Magdaleno JD, Garcia CG & Ishiwara DGP. *Entamoeba histolytica*: differential gene expression during programmed cell death and identification of early pro- and anti-apoptotic signals. *Exp Parasitol* 2010; **126**: 497–505.

79 Nickel R, Jacobs T, Urban B, Scholze H, Bruhn H & Leippe M. Two novel calcium-binding proteins from cytoplasmic granules of the protozoan parasite *Entamoeba histolytica*. *FEBS Lett* 2000; **486**: 112–116.

80 Wender N, Villalobo E & Mirelman D. EhLimA, a novel LIM protein, localizes to the plasma membrane in *Entamoeba histolytica*. *Eukaryot Cell* 2007; **6**: 1646–1655.

81 Reyes-López M, Bermúdez-Cruz RM, Avila EE & de la Garza M. Acetaldehyde/alcohol dehydrogenase-2 (*Eh*ADH2) and clathrin are involved in internalization of human transferrin by *Entamoeba histolytica*. *Microbiology* 2011; **157**: 209–219.

82 Marion S, Wilhelm C, Voigt H, Bacri J-C & Guillén N. Overexpression of myosin IB in living *Entamoeba histolytica* enhances cytoplasm viscosity and reduces phagocytosis. *J Cell Sci* 2004; **117**: 3271–3279.

83 Marion S, Laurent C & Guillén N. Signalization and cytoskeleton activity through myosin IB during the early steps of phagocytosis in *Entamoeba histolytica*: a proteomic approach. *Cell Microbiol* 2005; **7**: 1504–1518.

84 Okada M, Huston CD, Oue M, *et al.* Kinetics and strain variation of phagosome proteins of *Entamoeba histolytica* by proteomic analysis. *Mol Biochem Parasitol* 2006; **145**: 171–183.

85 Reuber TL & Ausubel FM. Isolation of *Arabidopsis* genes that differentiate between resistance responses mediated by the *RPS2* and *RPM1* disease resistance genes. *Plant Cell* 1996; **8**: 241–249.

86 Santi-Rocca J, Weber C, Guigon G, Sismeiro O, Coppée J-Y & Guillén N. The lysine- and glutamic acid-rich protein KERP1 plays a role in *Entamoeba histolytica* liver abscess pathogenesis. *Cell Microbiol* 2008; **10**: 202–217.

87 Willhoeft U, Buss H & Tannich E. DNA sequences corresponding to the ariel gene family of *Entamoeba histolytica* are not present in *E. dispar*. *Parasitol Res* 1999; **85**: 787–789.

88 Bracha R, Nuchamowitz Y & Mirelman D. Transcriptional silencing of an amoebapore gene in *Entamoeba histolytica*: molecular analysis and effect on pathogenicity. *Eukaryot Cell* 2003; **2**: 295–305.

89 Zhang X, Zhang Z, Alexander D, Bracha R, Mirelman D & Stanley SL. Expression of amoebapores is required for full expression of *Entamoeba histolytica* virulence in amebic liver abscess but is not necessary for the induction of inflammation or tissue damage in amebic colitis. *Infect Immun* 2004; **72**: 678–683.

90 Leippe M. Amoebapores. *Parasitol Today* 1997; **13**: 178–183.

91 Lynch EC, Rosenberg IM & Gitler C. An ion-channel forming protein produced by *Entamoeba histolytica*. *EMBO J* 1982; **1**: 801–804.

92 Linford AS, Moreno H, Good KR, Zhang H, Singh U & Petri WA. Short hairpin RNA-mediated knockdown of protein expression in *Entamoeba histolytica*. *BMC Microbiol* 2009; **9**: 38.

93 Mirelman D, Anbar M & Bracha R. Epigenetic transcriptional gene silencing in *Entamoeba histolytica*. *IUBMB Life* 2008; **60**: 598–604.

94 Solis CF, Santi-Rocca J, Perdomo D, Weber C & Guillén N. Use of bacterially expressed dsRNA to downregulate *Entamoeba histolytica* gene expression. *PLoS ONE* 2009; **4**: e8424.

95 Anbar M, Bracha R, Nuchamowitz Y, Li Y, Florentin A & Mirelman D. Involvement of a short interspersed element in epigenetic transcriptional silencing of the amoebapore gene in *Entamoeba histolytica*. *Eukaryot Cell* 2005; **4**: 1775–1784.

96 Irmer H, Hennings I, Bruchhaus I & Tannich E. tRNA gene sequences are required for transcriptional silencing in *Entamoeba histolytica*. *Eukaryot Cell* 2010; **9**: 306–314.

97 Bujanover S, Katz U, Bracha R & Mirelman D. A virulence attenuated amoebapore-less mutant of *Entamoeba histolytica* and its interaction with host cells. *Int J Parasitol* 2003; **33**: 1655–1663.

98 MacFarlane RC & Singh U. Loss of dsRNA-based gene silencing in *Entamoeba histolytica*: implications for approaches to genetic analysis. *Exp Parasitol* 2008; **119**: 296–300.

99 Bracha R, Nuchamowitz Y, Anbar M & Mirelman D. Transcriptional silencing of multiple genes in trophozoites of *Entamoeba histolytica*. *PLoS Pathog* 2006; **2**: e48.

100 Bracha R, Nuchamowitz Y, Wender N & Mirelman D. Transcriptional gene silencing reveals two distinct groups of *Entamoeba histolytica* Gal/GalNAc-lectin light subunits. *Eukaryot Cell* 2007; **6**: 1758–1765.