UNIVERSITY OF LIVERPOOL

Improved genomic assembly and genomic analyses of *Entamoeba histolytica*

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

Amber Leckenby

September 2018

Acknowledgements

There are many people without whom this thesis would not have been possible. The list is long and I am truly grateful to each and every one.

Firstly I have to thank my supervisors Gareth, Christiane, Neil and Steve for the continuous support throughout my PhD. Particularly, I am grateful to Gareth and Christiane, for their patience, motivation and immense knowledge that helped me through the entirety of the project from the initial research to the writing of this thesis. I cannot have imagined having better mentors and role models.

I also have to thank the staff at the CGR for their role in the sequencing aspects of this thesis. My further thanks extend to the CGR bioinformatics team, most notably Richard, Matthew, Sam and Luca, for not only tolerating the number of bioinformatics questions I have asked them, but also providing great friendship and warmth in the office.

I must also give a special mention to Graham Clark at the London School of Hygiene and Tropical Medicine for sending cultures of *Entamoeba* and providing general advice, especially around the tRNA arrays. I would also like to thank David Starns, for his efforts troubleshooting the Companion pipeline and to Laura Gardiner for providing advice around all things methylation.

My gratitude goes to the members of the many offices I have moved around during my PhD, many of which have become close friends who have got me through many bioinformatics conundrums, lab meltdowns and (some equally challenging) gym sessions. Stef, Jen, Charl and Dave – you guys kept me sane.

Further, I must also include Sophie and Grace who have also made it through the past seven years at the University of Liverpool. From the first day of being an undergraduate to the final day of being a PhD student. I'm so grateful I got to navigate the adventures of student life with you both – looks like we all made it out okay in the end, right?

On a personal note, I must give my warmest thanks to my family, particularly my parents for encouraging me in everything and making this endeavour possible. To my many sisters for being my personal cheerleaders as I made it to the end of this thesis. To my housemate, Adele, who was an excellent sounding board to probably a few too many rants. To John Newsham, the teacher who got me so interested in biology in the first place. And finally, to Troy for his unwavering support and belief that I could achieve such great things.

Abstract

Amoebiasis is the third most common cause of mortality worldwide from a parasitic infection. It affects up to 50 million people annually, of whom 100,000 will die from the disease each year. Amoebiasis is caused by the amoeba *Entamoeba histolytica,* an obligate parasite of humans. Our understanding of the biology of this pathogen has been greatly advanced by the sequencing of its genome. However, the unusual nature of the genome (an extreme nucleotide composition bias, abundant repetitive elements and unknown chromosome structures/ploidy) made it particularly challenging to sequence and the resulting reference genome assembly is highly fragmented and possibly incomplete, limiting its usefulness for some analyses. New sequencing technologies can overcome some of the problems of the previous genome assembly. Here, single molecule real time (SMRT) sequencing was applied to sequence long fragments of DNA and build an improved reference genome for *E. histolytica*.

This thesis describes the generation of sequence data and a comprehensive comparative analysis of genome assembly tools available for long-read SMRT sequencing data. This analysis showed that assembly using PacBio data only produced better quality genome assemblies than hybrid assembly approaches utilising both long- and short-read data together. The PacBio genome assembly is significantly better than the published reference genome assembly based on a range of quality metrics.

The new genome assembly was annotated, revealing an increase in gene number. The spatial organisation of key virulence gene families (AIG1, Ariel-1, BspA, cysteine proteases, Gal/GalNAc lectins and STIRP families) was analysed, revealing an association of virulence gene families with transposable elements.

The new assembly allowed analyses of two key, unusual features of the *E. histolytica* genome: the long arrays of multiple tRNA genes and the multi-copy, extra-chromosomal molecules containing the ribosomal DNA. Several lines of evidence were consistent with tRNA arrays capping chromosomes and acting as telomeres in *Entamoeba*. Variation among array units exists (relevant as they are used as population genetic markers), but the majority sequence was consistently retrieved when genotyping, suggesting they may be relatively robust markers. Analysis of the rDNA episomes present in the *E. histolytica* strain sequenced (the HM-1:IMSS strain used for previous whole genome sequencing) revealed that one of the two rDNA episome types described in this strain has apparently been lost during *in vitro* culture.

Genome-wide 5-methylcytosine methylation profiles for trophozoite stage parasites in culture were determined using bisulphite sequencing for the new *E. histolytica* genome assembly and two additional species (*Entamoeba moshkovskii* and *Entamoeba invadens*). The analyses confirmed previous reports of sparse methylation of the genome as a whole but highlighted interesting patterns of methylation. While there was virtually no methylation of genes, there was extensive methylation of transposable elements and tRNA arrays. These patterns suggest methylation functions to suppress active transposition and may play a role in the structural control of tRNA arrays, again consistent with telomeric role.

The work presented here improves our understanding of the structure, content and regulation of the *E. histolytica* genome and provides a platform for improved future analyses for the *Entamoeba* research community.

Table of Contents

Table of figures

Table of tables

Abbreviations

5-MeC	5-Methylcytosine
ACT	Artemis Comparison Tool
ALA	Amoebic Liver Abscesses
Ariel	Asparagine-rich *Entamoeba histolytica* antigens
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BUSCO	Benchmarking Universal Single Copy Orthologues
BWA	Burrows Wheeler Alignment
CCS	Circular Consensus Sequencing
CDC	Center for Disease Control
CGR	Centre for Genomic Research
CLR	Continuous Long Read
DNA	Deoxyribonucleic Acid
Dnmt	DNA Methyltransferase
E-value	Exponent Value
EhSTIRP	*Entamoeba histolytica* serine, threonine and isoleucine rich protein
FDA	Food and Drug Administration
GO	Gene Ontology
HGAP	Hierarchical Genome Assembly Process
HPLC	High-Pressure Liquid Chromatography
IGS	Intergenic Spacers
Indels	Insertions/Deletions
MTase	Methyltransferase
MUSCLE	MUltiple Sequence Comparison by Log- Expectation
NGS	Next Generation Sequencing
OLC	Overlap Layout Consensus
ORF	Open Reading Frame
P-value	Probability Value
PacBio	Pacific Biosciences
PBS	Phosphate-buffered Saline
PFGE	Pulse-Field Gel Electrophoresis
qPCR	Quantitative PCR
RBC	Red Blood Cell
RNA	Ribonucleic Acid
RNAi	RNA interference
rRNA	Ribosomal Ribonucleic Acid
SBS	Sequencing By Synthesis
SMRT	Single Molecule Real Time
SNP	Single Nucleotide Polymorphism
SREHP	Serine rich *Entamoeba histolytica* protein
sRNAs	Small RNAs
SSU	Small Subunit
STR	Short Tandem Repeat

TE	Transposable Element
TGS	Transcriptional Gene Silencing
tRNA	Transfer Ribonucleic Acid
VSG	Variant Surface Glycoprotein
WBC	White Blood Cell
WHO	World Health Organisation
YAC	Yeast Artificial Chromosome

Chapter 1 – Introduction

1.1. *Entamoeba* phylogeny

The genus *Entamoeba* is part of the phylum *Amoebazoa*, which forms a sister group to the Opisthokonts (animals and fungi), diverging from it after the divergence of plants [1]. The *Amoebazoa* are separated into two lineages: the *Mycetozoa*, which are free-living and include the slime-mold *Dictyostelium discoideum*; and the mitochondria-lacking *Archamoebae* which in turn can be divided into *Mastigamoeba* and *Entamoeba* lineages [1]. Species within the *Amoebazoa* phylum are poorly sampled and represent a largely unknown part of the tree of life. They are highly diverse; for instance, the divergence between the *Archamoebae* and the *Mycetozoa* may be as great as that observed between animals and fungi [2].

The genus *Entamoeba* is diverse and many of its species parasitize a broad range of hosts, from reptiles to mammals [3–5] though zoonotic transmission of *Entamoeba* species between different host restrictions is thought to be extremely rare. Recently however, there has been a reported cases of *Entamoeba nuttalli*, a parasite of wild macaques, having jumped host restriction into humans [6]; Figure 1.1.1 highlights the host restriction for each species. Different species are also able to parasitize different niches within the same host type; for example *E. gingivalis* parasitizes the oral cavity whereas *E. histolytica* colonises the gut [7–9]. Not all species are pathogenic and some, such as *E. dispar and E. moshkovskii*, which both infect the human gut, are generally thought to be harmless to humans, unlike the pathogenic human parasite *E. histolytica* [10].

Defining species within the *Entamoeba* genus is difficult; as with many unicellular organisms, it is hampered by the fact that many species are morphologically indistinguishable. In some cases, morphological markers can be used to identify different species, such as the number of visible nuclei present at certain life cycle stages. Briefly, the *Entamoeba* species undergo a two stage life cycle composed of the active trophozoite stage and the dormant (infectious) cyst stage [11]. The number of nuclei per cyst, commonly one, four or eight, can help to distinguish species from one another. However, this method is limited by the observation that multiple species fit into the different nucleus groups and some species, like the oral parasite *Entamoeba gingivalis*, do not form cysts and therefore, are indistinguishable [12]. In these instances, species must be defined by genetic divergence [10]. Species can be identified based on their ribosomal lineage whereby an individual can be identified by its 18S small subunit ribosomal RNA (18S SSU rRNA) sequence [13]. Utilisation of this method on *Entamoeba* species has shown that some previously defined *Entamoeba* species show diversity within this 18S SSU rRNA sequence suggesting that some species, such as *Entamoeba moshkovskii* [14] and *Entamoeba coli* [13], are in fact species complexes.

Figure 1.1.1. also illustrates the phylogenetic relationship among a small sample of *Entamoeba* species and indicates species for which genome sequence data are available. The phylogeny of the genus often shows large evolutionary distances between different *Entamoeba* species, even in those that occupy the same host. Few *Entamoeba* species have been extensively studied and many species, including *E. bangladeshi* and *E. nuttalli*, are only very recently identified and barely investigated. The difficulty of culturing *Entamoeba* species suggests that the true diversity of this genus is likely vast however, sequencing of the few known organisms can help to understand the evolution of the genus, especially those with the potential to cause disease.

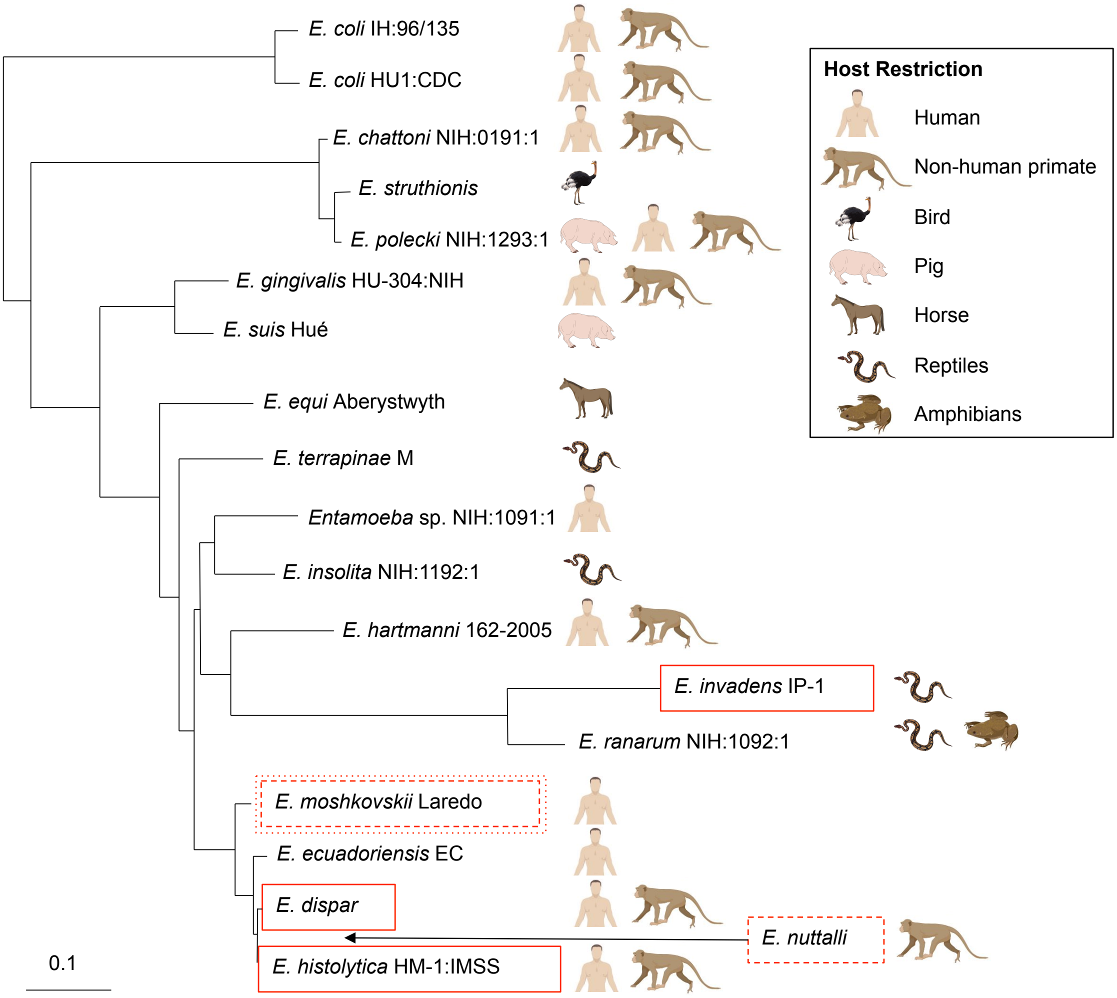


Figure 1.1.1. *Entamoeba* phylogeny and host restriction. Phylogeny is based on the small subunit rRNA sequence. At the time of its original publication, species surrounded by dashed boxes were due to be sequenced, low coverage shotgun sequencing data existed for those in dotted boxes and fully sequenced species are in solid boxes. Modified from Weedall and Hall, 2006 [15].

1.2. *Entamoeba histolytica* and other important *Entamoeba* species

The species most relevant to human health, *Entamoeba histolytica,* accounts for a substantial portion of *Entamoeba* research. It is an invasive, enteric protozoan pathogen and the causative agent of amoebiasis, an important cause of diarrhoea and diarrhoeal death in developing countries. *Entamoeba histolytica* affects approximately 500 million people worldwide of whom approximately 4-10% will develop clinical symptoms within one year. 100,000 cases are estimated to be fatal per year, making *E. histolytica* the third leading cause of death from a parasitic disease worldwide after malaria and schistosomiasis [16,17]. What triggers pathogenesis in a small subset of individuals is still unknown. However, we are becoming increasingly aware of the complexity of host-parasite interactions as drivers of virulence in *Entamoeba* infections.

1.2.1. The *Entamoeba histolytica* life cycle

The *E. histolytica* life cycle is completed within one host, usually humans, and does not require an intermediate host. Its life cycle (Figure 1.2.1) consists of two stages, infective trophozoites and dormant cysts [11]. In the environment, the parasite exists as a quadrinucleate cyst that can be ingested by the human host. Once in the small intestine, the parasite undergoes excystation and develops into (potentially) pathogenic trophozoites that colonise the large intestine. Trophozoites replicate via binary fission and (under unknown stimuli) produce cysts that are passed in the faeces. Unlike the cysts, the anaerobic trophozoites (sometimes passed in diarrhoea) rapidly die in the aerobic external environment. Nor would they survive the gastric environment if they were re-ingested. Cysts can exist in the external environment for days to weeks and are responsible for further infections.

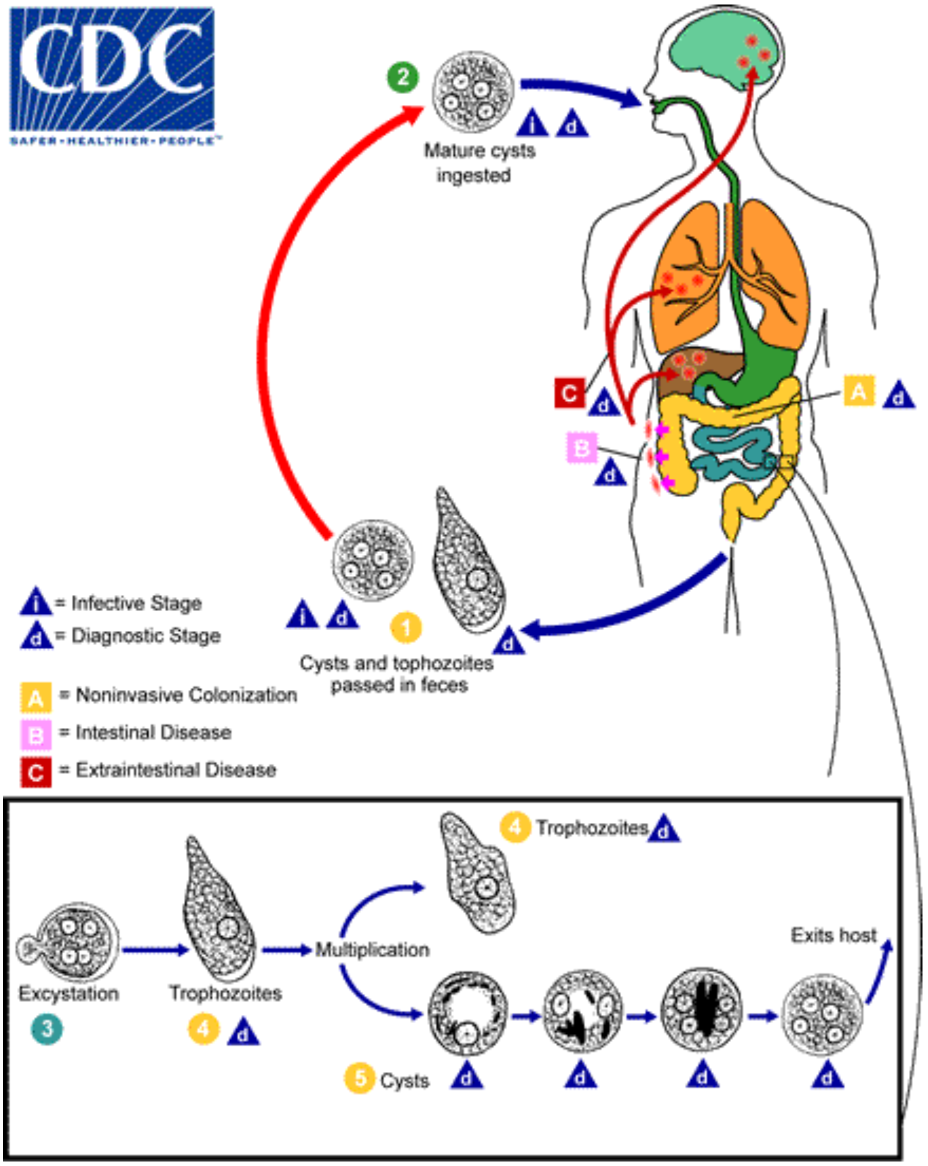


Figure 1.2.1. The life cycle of *Entamoeba histolytica.* The different infection outcomes within the human host are labelled A, B and C. Stages of the life cycle are number numerically (1-5) in chronological order. Image taken by the US Centers for Disease Control and Prevention (CDC).

1.2.2. Pathogenicity and treatment of amoebiasis

E. histolytica is the causative agent of amoebiasis, which can present across a range of severities from asymptomatic to invasive and extra-intestinal disease. Asymptomatic infections account for the majority of infections and are defined as the presence of *E. histolytica* cysts in the stool in the absence of colitis or extra-intestinal disease. These individuals do not exhibit clinical manifestations and present with no history of blood in the stools. Cysts and trophozoites lacking ingested red blood cells (RBCs) may be visible in the stool under a microscope [18] indicating that asymptomatic patients can still be infective carriers of *E. histolytica*. Most individuals will also produce serum antibody responses to the parasite even in the absence of invasive disease [19]. It is important that asymptomatic patients are also treated to prevent spread of amoebiasis from these carriers. Untreated asymptomatic patients usually self-resolve the disease over time until full clearance of the parasite is reached though, some individuals may develop colitis after a period of months [20].

In a small subset of individuals, the disease can persist and progress into invasive amoebiasis, also called amoebic dysentery, where the infective *E. histolytica* stage (trophozoites) can penetrate the intestinal mucosa [21]. In invasive disease, trophozoites damage and kill epithelial cells and invade the epithelium that lines the colon leading to abdominal pain and tenderness. Other common symptoms of invasive amoebiasis are watery, bloody or mucous stools with up to 10 bowel movements per day, appetite loss followed by weight loss, and fever (in one third of patients) [22]. The presence of Charcot-Leyden crystals, the lack of faecal leukocytes, and blood are the most common stool findings in the acute stage of amoebic dysentery. Detection of the parasite can be poor from a single stool sample and the best diagnostic method is detection of the *E. histolytica* antigen or *E. histolytica* DNA in the stool [23,24]. Left untreated, the disease leads to amoebic colitis resulting in flask shaped ulcers within the colon in which trophozoites replicate at a high rate [25]. The incubation time for invasive amoebiasis is variable though usually symptoms appear one to four weeks after ingestion of cysts; however, the range may be from a few days to years [18,22,26].

Occasionally, *E. histolytica* can spread to other organs leading to extra-intestinal amoebiasis. This organ is most commonly the liver, with the parasite gaining access through the hepatic portal venous system and causing amoebic liver abscesses (ALA) [27]. The incubation period between invasive infection and extra-intestinal infection is not well documented however, ALA occurs more commonly in adults than children. The disease presents with similar symptoms to amoebic dysentery with most patients experiencing abdominal pain followed by fever and more diffuse abdominal pain in the sub-acute phase [28]. In addition, many patients have elevated peripheral white blood cell counts and alkaline phosphate levels [29–31]. Definitive diagnosis of ALA is confirmed by serological testing for antibodies against *E. histolytica* and through the demonstration of lesions to the liver through imaging techniques such as computed tomography ultrasonography or magnetic resonance imaging [22]. Complications of these liver abscesses such as abscess rupture and secondary bacterial infections are usually fatal [32]. Spread of trophozoites to other sites does occur and has been reported in the lungs, brain and skin however, these cases are much rarer than colonisation of the liver [33–35].

Vaccines using native and recombinant forms of an *E. histolytica* lectin (Gal-lectin) can protect animals against intestinal amoebiasis and ALA [36]. However as these have not yet been developed for testing on humans, treatment of amoebiasis still heavily relies on drug therapy. Treatment options for all stages of infection are limited and there is a great need to identify novel drug targets to aid new drug developments. Currently, asymptomatic infections are treated with paromomycin (25-35 mg/kg, 3x daily, 7 days) followed by a luminal agent, diloxanide furoate (500mg, 3x daily, 10 days) [37]. When used in combination, the drugs are effective at clearing *E. histolytica* from asymptomatic individuals, however, both drugs can cause gastrointestinal upset, nausea, vomiting and diarrhoea, which can make it hard to distinguish any emergence of invasive disease during the treatment period as symptoms and side effects are similar.

For invasive and extra-intestinal infection, metronidazole (750 mg, 3x daily, 7-10 days) is the most effective treatment. How metronidazole works to resolve *E. histolytica* infections is unclear. However, it is thought to work by disrupting redox regulation mechanisms, specifically the thioredoxin system. Thioredoxin is able to reduce metronidazole, producing highly reactive molecules that generate increasing levels of oxidative stress leading to cell death. When used in conjunction with a luminal amoebacide such as diloxanide furoate as before, complete clearance of infection is likely within two weeks. The side effects of metronidazole are more severe than those following treatment of asymptomatic disease. Metronidazole side effects are primarily gastrointestinal and include anorexia, nausea, vomiting, diarrhoea and abdominal pain as well as a metallic mouth taste and an intolerance reaction with alcohol [37]. As before, the side effects and the disease symptoms are very similar and hence, it can be difficult to observe whether a patient is responding well to the metronidazole treatment and truly getting better.

There is no major second-line drug at present; this becomes particularly important, as metronidazole is both expensive and not easily available in some countries. Many areas where amoebiasis is endemic occur in the tropics and the rate of infection is directly linked to socio-economic factors such as income and access to adequate hygiene infrastructure. This means that there are situations where patients simply do no have access or cannot afford effective treatment of the disease and as a result, the disease continues to spread within these poorer areas. In addition, metronidazole-resistant *E. histolytica* strains have been reported *in vitro* [38,39], highlighting that metronidazole resistance could emerge *in vivo* as has been in seen in other parasites such as *Trichomonas vaginalis* [40–43].

Where standard metronidazole treatment has been ineffective, toxic levels of metronidazole or other drugs might be prescribed with dangerous side effects. Paramomycin, an orally delivered aminoglycoside amoebacide, can be used as a second-line treatment to treat invasive amoebiasis and patients in comas resulting from liver damage however, it is ineffective against extra-intestinal disease and has many serious side effects [44,45]. The final alternative is dehydroemetine. It is highly toxic and an irritant when taken orally meaning that it is delivered by injection directly into muscle tissue. The drug inhibits protein synthesis and can cause fatal myocardial toxicity in sufferers with cardiac problems. Chloroquine and needle aspiration of abscesses are also recommended in extreme cases [45,46]. The severity of the side effects associated with these last resort treatments have led to active research to identify other drugs effective against both *E. histolytica* trophozoites and cysts. Recently, drug screens have been performed with some success. Auranofin, a US Food and Drug Administration (FDA) approved drug used in patients suffering from rheumatoid arthritis, was discovered to be as active against *E. histolytica* trophozoites in culture as metronidazole [47]. It is yet to be determined whether auranofin will be active against the more resistant cysts and clinical trials are being performed to assess the effectiveness of using auranofin to treat amoebiasis patients [48].

However, it could be argued that it is not beneficial to the *E. histolytica* species to be pathogenic as by killing the host, the parasite is unable to continue its life cycle. Recent investigations into the prevalence of *Entamoeba histolytica* in the rural African gut microbiome have suggested that *E. histolytica* may be better termed a pathobiont (i.e. a potentially pathogenic organism, which under normal circumstances, lives as a symbiont). The study observed that the faecal microbiota of Pygmy hunter-gathers as well as Bantu individuals from bother faming and fishing populations in Southwest Cameroon, presence of *E. histolytica* in the gut was significantly correlated with microbiome composition and increased diversity suggesting *E. histolytica* may usually act as a symbiont in normal circumstances [49]. Further, the study noted that colonisation of *Entamoeba* in the gut could be predicted with 79% accuracy based on the composition of an individuals gut microbiome and that several of the taxa most important for distinguishing *Entamoeba* absence from the microbiome were signature taxa for autoimmune disorders [49].

Though, as described in previously in Section 1.2.1, invasion of the colon by *E. histolytica* trophozoites during amoebic colitis, results in flask shaped ulcers within the colon in which trophozoites replicate at a high rate [25]. It is likely that by forming these ulcers, the trophozoites create an environment that is partly shielded from the harsh environment of the gut and under these conditions the parasite can proliferate at a high rate [25]. Therefore, the generation of these flask shaped ulcers could perhaps be beneficial to the survival and reproduction of the *E. histolytica* trophozoites effectively selecting for pathogenicity during the evolution of the parasite. The origins of these virulence genes involved in pathogenicity of *E. histolytica* are unknown. It was speculated that horizontal gene transfer may be responsible for the acquisition of some genes important in virulence as is seen the evolution of many other eukaryotic pathogens. However, despite 22 transferences of HGT being predicted in *E. histolytica* between 31.45 Mya and 253.59 Mya, no virulence factors have been identified as being transferred [50].

1.2.3. Epidemiology of amoebiasis

Amoebiasis is a major cause of morbidity and mortality worldwide and the disease can be both endemic and epidemic. Endemic disease is most prevalent in tropical and sub-tropical countries. These settings are characterised by poor sanitation infrastructure, such as the lack of access to clean water and toilet systems, and inadequate health care infrastructure [51]. For instance, amoebiasis is endemic in regions of Mexico and the disease consistently ranks fifth or sixth in the list of the 20 major causes of disease in Mexico [52,53]. Similarly, high instances of endemic amoebiasis are seen in the Hué region in Vietnam, where approximately 11% of the population are estimated to be infected [54] and in Bangladesh, where 15.6% of tested children aged two to five had an *E. histolytica* infection, confirmed using an antigen detection kit [55]. *E. histolytica,* is transmitted between hosts via the faecal-oral route and as such, the majority of amoebiasis deaths are characterised by intestinal or extra-intestinal disease.

In addition to individuals who live in areas where amoebiasis is endemic, there are other groups of individuals who are at a higher risk of becoming infected with *E. histolytica.* Outbreaks have been observed in individuals who engage in oral and anal sex, most commonly homosexual men [56,57] and also in institutionalised individuals in Japan and the Philippines [58,59]. More recently, a study of 346 individuals suffering from amoebic liver abscesses (ALAs), where *E. histolytica* trophozoites metastasize from the intestine to the liver causing invasive liver disease, was performed in Sri Lanka. The study found that among the cohort of ALA-sufferers, almost all (98.6%) were male and all (100%) reported a history of heavy alcohol consumption, especially a local drink that consists of the fermented sap of the Palmyra palm (called toddy) [60].

Amoebiasis epidemics sometimes occur in more affluent countries. Accidental sewage contamination of public water supplies can lead to *E. histolytica* outbreaks as was witnessed in Sweden in 1986 [61], Taiwan in 1993 [62] and in the Republic of Georgia in 1998 [63].

It was estimated in 1986, that between 10% and 20% of the global population were infected with the *E. histolytica* parasite of whom, 1% would develop the invasive form of the disease and 100,000 would die annually [7]. 99% of cases were reported as asymptomatic and it is thought that some infections may be due to non-pathogenic amoebas such as *E. dispar* and *E. moshkovskii*, which are morphologically identical to *E. histolytica* and were not recognised as separate species until 1993 [10] and 1991 [64], respectively. As a result, it is likely that the infection rate was over-estimated however, as these other *Entamoeba* species are non-pathogenic, it is likely that the global mortality remains at 100,000 people annually [65]. More recently the World Health Organisation (WHO) has revised these figures and it is thought *E. histolytica* infects approximately 500 million people *per annum*. Of these, it is estimated that 50 million will develop symptomatic disease and 100,000 cases will result in death [66].

1.2.4. Other *Entamoeba* species relevant to *E. histolytica* research

Research into the most closely related *Entamoeba* species to *E. histolytica* can be useful in determining the evolution of pathogenicity and the narrowing of host range within this part of the *Entamoeba* phylum. These species are *E. nuttalli, E. dispar* and *E. moshkovskii* (Figure 1.1.1 for reference). Comparisons of the gene content of these species (described in more detail in section 1.3.4) provides insights into the gene differences (present/absent/polymorphisms) between pathogenic and non-pathogenic *Entamoeba* spp. as well as between human-infecting and non-human primate-infecting species.

Entamoeba nuttalli is the closest related species to *E. histolytica* that has currently been identified; it is pathogenic and its main host species so far has been identified as captive and wild macaques including *Macaca mulatta, M. fascicualis, M. fuscata, M. thibetana* and *M. sinica* [67–74]. Most macaques with *E. nuttalli* infections are asymptomatic, perhaps indicating that the host-parasite relationship in macaques may be commensal in natural infection [71]. More recently, zoonotic concerns surround the *E. nuttalli* parasite have been raised after cysts of *E. nuttalli* were detected in a care-taker of non-human primates in a zoo [6].

Entamoeba dispar is the closest related species to *E. histolytica* that also infects humans however, *E. dispar* has a wider host range and can also infect non-human primates [70]. *E. dispar* is unlike *E. histolytica* in regards to its pathogenicity; *E. dispar* is non-pathogenic in humans although this has been questioned [75] as a Brazilian strain of *E. dispar* has been seen to cause amoebic liver abscesses (ALAs) in hamsters that were occasionally indistinguishable from ALAs produced by *E. histolytica*. This and other findings, such as the detection of DNA sequences from *E. dispar* in ALA-sufferers, has revived the possibility that this species can produce human lesions [76].

Entamoeba moshkovskii was long thought to be a free-living amoeba [14,77,78] however, the observation of its ability to infect humans has made it clinically relevant over the last decade. It has frequently been found in areas where amoebiasis is prevalent [79]. It has been suggested that *E. moshkovskii* is associated with gastrointestinal symptoms and studies of human-derived clinical isolates and cases of diarrhoea have been directly associated with *E. moshkovskii* [55,80]. It has the potential to infect large numbers of people and one study in Bangladesh revealed that 21.1% of sampled children (two to five years old) were infected with *E. moshkovskii* [55], highlighting the importance of research into this *Entamoeba* species.

Though distantly related to *E. histolytica, E. dispar and E. moshkovskii* [79], *Entamoeba invadens* is one of the more researched *Entamoeba* species because it is the only *Entamoeba* species that can be induced to encyst and excyst in axenic culture [81]. This unique characteristic, alongside the observation that the life cycle, symptoms and infection caused by *E. invadens* are the same as those in *E. histolytica* [82,83], has made it a model for *Entamoeba* encystation*;* gene expression data are available for the entirety of its life cycle [84]. The model has helped to elucidate the genes involved in the regulation of the *Entamoeba* life cycle and identify genes that are differentially regulated during the encystation and excystation transition stages [84]. The ability to encyst this species *in vitro* has also been utilised within high-throughput screens to identify drugs effective against both the trophozoite and cyst life forms of metronidazole-resistant *E. invadens* strains [85]. Compounds with good activity against *E. invadens* were tested in *E. histolytica* trophozoites resulting in identification of five compounds with good activity (EC50 <25 μM, >50% inhibition of cysts) against both *E. histolytica* trophozoites and *E. invadens* cysts [85].

1.3. The power of genomics and comparative analyses amongst *Entamoeba* species

Whole genome sequencing (WGS) has played a significant role in tackling human pathogens. Candidate targets for drugs or vaccines have been discovered from the vast amount of functional data that can be derived from the annotation of these pathogen genomes. Reference genome assemblies that have been developed for each pathogen provide a vital resource for post-genomic data analyses such as analyses of gene expression, through microarrays or whole transcriptome sequencing, and the identification of epigenetic modifications to DNA through whole genome bisulphite sequencing. Underlying all of this, the genome provides an open resource for understanding the organism's biology. EuPathDB is a good example of this, and acts as a resource for the collective storage of multi-omics data of eukaryotic protist pathogens, including *Entamoeba*, and provides integrated tools that aid exploration and analyses of these numerous datasets [86]. The growing number of organisms in this database includes many species of *Plasmodium* [87]*, Trypanosoma* [88,89]*, Leishmania* [90]*, Trichomonas* [91]*, Giardia* [92,93]*, Cryptosporidium* [94,95] *and Neospora* [96]*.* These genomes have provided an invaluable resource for understanding the biology of these organisms, including mechanisms of virulence, and many of the omics datasets have been utilised to understand the organism's biology and combat the diseases they cause. The draft *Entamoeba histolytica* genome [97,98] is also hosted on EuPathDB, alongside many other omics datasets, including those for other members of the *Entamoeba* genus. These genomes have facilitated genomic and comparative genomic analyses of the *Entamoeba* species, though limitations remain. These limitations will be discussed in this section; it is hoped that new genome sequencing (Chapter 2), annotation (Chapter 3) and post-genomic analyses (Chapter 5) will help to solve some of these problems.

1.3.1. The *Entamoeba histolytica* genome

The *E. histolytica* strain sequenced, HM-1:IMSS, is the most widely studied culture-adapted strain and was originally isolated from the rectal ulcer of an adult human male suffering from amoebic dysentery in Mexico City in 1967 [99,100]. The *Entamoeba histolytica* HM-1:IMSS genome was sequenced prior to the arrival of second generation (massively parallel) and third-generation (single molecule) sequencing technologies and as a result the genome is largely composed of whole-genome shotgun (WGS) Sanger reads. The *E. invadens* and *E. dispar* genomes were sequenced with the same technology. At the time of sequencing, assembling WGS Sanger sequencing data was difficult for organisms with complex, repetitive genomes as the reads, around 750 bp in length [101], rarely spanned repetitive regions of the genome leading to breaks in the assembly where a contig could no longer be extended. As a result, strategies were developed to sequence chromosomes or long stretches of DNA individually. Bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs) can aid assembly processes by allowing the isolated sequencing of long stretches of contiguous DNA whilst retaining more spatial information than whole shotgun sequencing alone. BACs can be created with inserts up to 300 Kbp and YACs are able to incorporate inserts of 100-2000 Kbp [102]. Markers or restriction digest patterns can be identified on the assembled long inserts that allow them to be physically mapped to the chromosomes of an organism. However, the method was not suitable for *E. histolytica* genome sequencing due to the high AT content that makes the DNA unstable in BACs. In addition, the process relies on physical mapping, which is not possible in *E. histolytica* as the chromosomes in *E. histolytica* do not condense.

Entamoeba histolytica's closest fully sequenced relative, *Dictyostelium discoideum,* also has a very high AT content (77.6 %) that meant the genome was unsuitable for large insert BAC library generation. To overcome this problem, researchers used pulse-field gel electrophoresis (PFGE) to separate the *D. discoideum* chromosomes so they could be individually isolated, sequenced and assembled [103]. During PFGE, genomic DNA can be separated to reveal distinct DNA bands that represent whole chromosomes. These chromosomes can be then isolated, sequenced and assembled separately to improve the likelihood of assembling whole chromosomes. This sequence data was combined with HAPPY mapping to help guide the sequence assembly of each chromosome. This hybrid method of assembly produced six chromosomes of the *D. discoideum* genome consisting of 34 Mbp [103].

Unfortunately, a similar approach could not be adopted for the *E. histolytica* genome. The karyotype of *E. histolytica* appears to be very complex and cannot clearly be separated using PFGE, unlike many other organisms whose chromosomes can be well separated using this method. The best method of performing PGFE on *E. histolytica* trophozoites includes embedding whole cells into agarose plugs prior to digestion with proteinase K and separation of DNA using rotating field gel electrophoresis (ROFE). The whole cells are required to be embedded as *Entamoeba* contain large numbers of endogenous nucleases which rapidly degrade its genomic DNA upon lysis and result in a smear of low molecular weight DNA when used in PFGE [104]. Unfortunately, as whole cells are required to be embedded, only 1-3x10^6 cells are able to be loaded into a single agarose plug resulting in weak staining of the chromosomes. From this, numerous faint, but visible, bands were observed ranging from 0.3 Mbp and 2.2 Mbp (Figure 1.3.1).

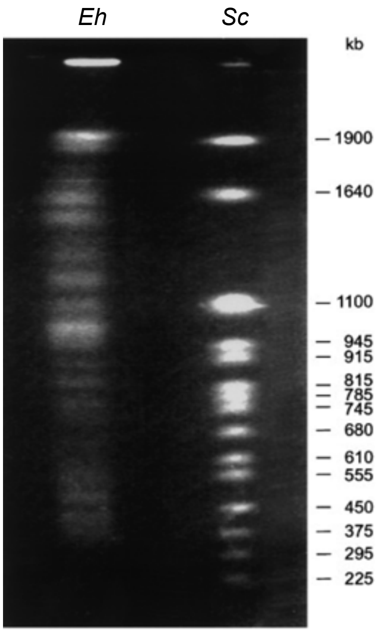


Figure 1.3.1. The complex karyotype of *Entamoeba histolytica.* PFGE-separation of *E. histolytica* DNA (*Eh*) and *Saccharomyces cerevisiae* (*Sc*) shows a complex pattern of separation of *E. histolytica* DNA and a distinct separation of *S. cerevisiae* DNA. Figure adapted from Willhoeft and Tannich, 1999 [105].

Staining intensity is not constant between the different bands suggesting that brighter bands may consist of two or more distinct DNA structures (chromosomes or plasmid molecules) or represent cases where entire DNA molecules exhibit increased copy number. Adding to this complexity, it is suggested that the *E. histolytica* genome is composed of a mixture of both linear and circular molecules [106–109]. Circular molecules contain the rRNA genes of *E. histolytica*; Estimates suggest approximately 200 copies of these molecules may occur per cell [110].

HAPPY mapping could not be used to complement the assembly of the *E. histolytica* genome as was done for the assembly of *D. discoideum.* In genomics, HAPPY mapping can be used to assess the orientation of various DNA sequences across a particular genome and generate a genomic map that can be used to guide the assembly of other omics sequencing datasets. The process defines linkage groups based on the frequency or co-occurrence of the markers in samples generated by fragmentation [111]. The rationale being that the closer two markers are to one another spatially, the less likely they are to be separated by the fragmentation stage. These markers can be matched to those in sequence data and those reads/contigs containing markers not separated by fragmentation can be assumed to occur close together in the genome [111]. HAPPY mapping has been performed on *E. histolytica* DNA however the results were not very successful due to the repetitive nature of the genome [112].

Optical mapping is another technique to guide the assembly of sequence data based on spatial information. In optical mapping, ordered restriction maps of very long, single DNA molecules are created [113]. *In silico* restriction mapping of assembled contigs or scaffolds is performed and the restriction patterns mapped to those created during the optical mapping. The placement of contigs within the optical map can bring together contigs close enough to produce spatially accurate scaffolds (containing variable sized gaps). Optical mapping has also been performed on *E. histolytica* with limited success (Dr. Elisabet Caler, personal communications). A small number of putative linkage groups [14] were produced, totaling a small amount of the genome size. Subsequent mapping of sequence scaffolds to these maps proved difficult owing to the short scaffold lengths that meant unique restriction mapping patterns of these scaffolds were difficult to produce. Attempts to reproduce a newer optical map utilizing newer technologies are outlined in Chapter 2.

As a result of these complexities, the *E. histolytica* genome was assembled entirely from WGS Sanger reads in 2005 without any spatial information or physical mapping [97]. Sequencing produced around 580,000 reads of which around 170,000 were removed before assembly due to them being episomal-derived or tRNA-containing. The remaining ~416,000 reads had an average length of 645 bp and were assembled using the assembler, phusion [114]. Scaffolds smaller than 2 Kbp were removed and scaffolds smaller than 5 Kbp that shared 98% or more nucleotide sequence identify over >95% of their lengths were also removed. The remaining 888 scaffolds had a total length of 23,751,783 bp. The genome was annotated using the Combiner algorithm using two gene finder programs, Phat [115] and GlimmerHMM [116], trained using a set of published *E. histolytica* gene sequences. Functional annotations for the predicted proteins were automatically generated by searching protein sequences against a non-redundant protein database and the Pfam database [117].

The genome was re-assembled and re-annotated with additional sequence data in 2010 [98]. Assembly was performed in a similar way as was done previously; reads containing episomal DNA or tRNA models were removed before assembly and remaining reads were assembled using UMD Overlapper [118] and Celera Assembler [119]. 300 known genes and 60 full length cDNAs were used to create a training set for the gene finder programs Genezilla [120] and GlimmerHMM [116]. EVidenceModeler [121] was used to generate the new gene data set, as a weighted consensus of all available evidence such as proteins and protein-domains alignments, cDNAs and gene finder output predictions [98]. This final assembly consists of 20,800,560 bp assembled across 1,496 scaffolds. Sequencing of *Entamoeba* DNA was challenging owing to its high AT content (75-80%) and as a result, the assembly produced remained highly fragmented [97,98]. This meant that a lot of information about the genome structure was lost and subsequently, very few structural features are defined for *E. histolytica.* There is no evidence as to how many chromosomes the genome contains nor is there any information as to what the telomeric or centromeric structures of the genome look like. However, it was revealed that the genome is highly repetitive and transposable elements accounted for approximately 20% of the assembly [122]. This observation, alongside the observation that many scaffolds in the assembly ended with transposable elements (TEs), suggests that the repetitive nature of the genome may have been the major limiting factor in assembling the *E. histolytica* genome.

1.3.2. Genome structure and gene content of the *Entamoeba histolytica* genome

Transfer RNA genes show a unique organization within the *Entamoeba* genus and form a considerable portion of the repetitive DNA in the *E. histolytica* genome. The tRNA genes occur in mixed structures, separated by spaces of DNA which may or may not contain short tandem repeats (STRs) depended on the species. These units are then tandemly duplicated to form large stretches of repetitive DNA [4,5]. One hypothesis is that these structures could cap the chromosomes and act as telomeres; no telomeric sequences have so far been identified in previous *E. histolytica* sequencing attempts and the presence of tRNA-based telomeres would represent an analogous mechanism to that seen in *D. discoideum* where ribosomal DNA repeats act as telomeres [123]. However, there is no evidence of linkage of these tRNA arrays to non-repetitive DNA and hence, little evidence exists surrounding the genomic location of these structures.

The rRNA genes within the *E. histolytica* genome exist in extra-chromosomal molecules comprised of the rDNA genes and multiple short tandem repeat families [110]. Two described rDNA episomes, EhR1 and EhR2, differ in the number of rDNA genes they contain. EhR1 contains two copies of the rDNA genes in an inverted orientation to one another whereas, EhR2 contains only one set of rDNA genes and likely results from genetic recombination of EhR1 facilitated by the short tandem repeat families [124,125]. In *D. discoideum,* a chromosomal region containing the rRNA genes appears to act as a master copy from which linear extra-chromosomal copies of the rRNA genes are generated [103,126]. No chromosomal copy of the rRNA genes has been identified in *E.*

histolytica although until now, the assembly has been highly fragmented and assumed to be incomplete, so the possibility of a chromosomal master copy of rRNA genes could not be excluded. Further characterisation of this molecule with regard to the organization of the rRNA genes across the chromosomal and extra-chromosomal portions of the *E. histolytica* genome is performed in Chapter Four.

The presence of rDNA episomes and tRNA arrays causes problems when performing genome assembly as these regions constitute a large proportion of the sequence library compared to the non-repetitive regions that are more useful in genome assembly. However, *Entamoeba histolytica*'s repetitive nature does not end here; its genome contains a large fraction of transposable and repetitive elements [98]. Up to 20% of the genome was predicted as repetitive in the original sequencing attempts [97,98] and it is these repetitive elements that can lead to chromosomal instability, resulting in chromosomal breaks points where breakage and rejoining can occur [98]. These repetitive regions and associated break points have proven a huge problem for previous *E. histolytica* genome assembly attempts as, unless reads (or read pairs) span the entire length of the repetitive region, regions either side of it cannot be unambiguously linked. Some repetitive elements are several kilobases long and Chapter Four highlights that some of these tRNA arrays can reach tens of kilobases long. These lengths are much longer than those producible by Sanger and NGS sequencing methods that have previously been used to study *E. histolytica* and as a result, the current *E. histolytica* reference genome remains very fragmented. Supporting the theory that it is the repetitive elements causing this fragmentation is the observation that many of the current *E. histolytica* genomic scaffolds end in repetitive DNA, often transposable elements, further suggesting that it is the presence of these that has caused the fragmented nature of the current reference genome.

The genome was reported to be gene rich with around half of the assembled sequence corresponding to predicted coding sequence. In total 8,333 protein-coding genes have been detected in the *E. histolytica* genome. However, functional gene annotation across the *E. histolytica* genome remains largely incomplete owing to the low throughput nature of functional genetic analysis and the lack of researchers dedicated to the annotation of *Entamoeba* genomes. In addition, Protist genomes are inherently difficult to annotate, as they are highly divergent from the well-described model organisms that most gene annotation programs are optimized for. Figure 1.3.2 highlights the lack of gene annotation across the *Entamoeba* genus. This data was generated from the AmoebaDB website, a database that forms part of the EuPathDB database that hosts the most current annotations for many protist genomes [127]. The majority of these GO terms were assigned automatically by the annotation software used to annotate the *E. histolytica* genome and as a result, very little human curation has been performed on the genome; only a few hundred genes on AmoebaDB have user comments associated with them and the majority of annotated genes still encode a "hypothetical protein". Ultimately, annotation of the *E. histolytica* genome will require manual curation to ensure their accuracy.

Many genes are members of multi-gene families (Figure 1.3.3) [98,128], though their organization is unknown. The fragmented nature of the *E. histolytica* genome means that many members of the same gene family are spread across multiple scaffolds with no spatial information available that could suggest their relative genomic positions to one another. 897 protein families were identified containing 4,564 proteins (56% of the proteome). The average gene family in *E. histolytica* is five members however some gene families in the *E. histolytica* genome are very large in size and contain more than 50 members [98]. 82 families are specific to *E. histolytica* though almost all of these families are entirely composed of hypothetical or functionally unannotated proteins; six families have function annotated and these families include the BspA, Gal/GalNAc lectin and Ariel families which have all been implicated in parasite virulence [105,129,130].

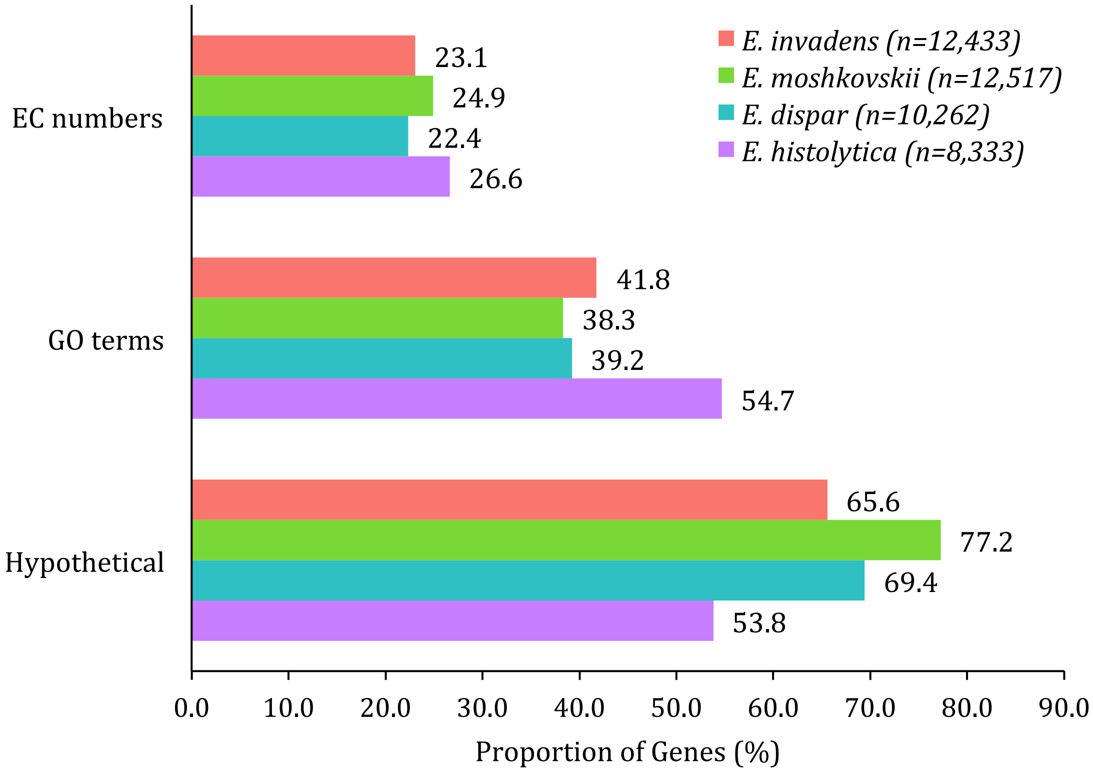


Figure 1.3.2. Functional gene annotation of *Entamoeba histolytica, Entamoeba moshkovskii, Entamoeba dispar* and *Entamoeba invadens.* The plot shows proportions of genes whose products are annotated as "hypothetical protein" or "unspecified product" however this does not include genes annotated with less informative name descriptions such as "X-domain-containing protein. The plot also shows proportions of genes associated with at least one gene ontology (GO) term of any class (cellular component, molecular function of biological process); and genes associated with an enzyme commission (EC) number. All statistics correct at time of writing (August 2018). Data collected from AmoebaDB and plot based on a previous publication [131].

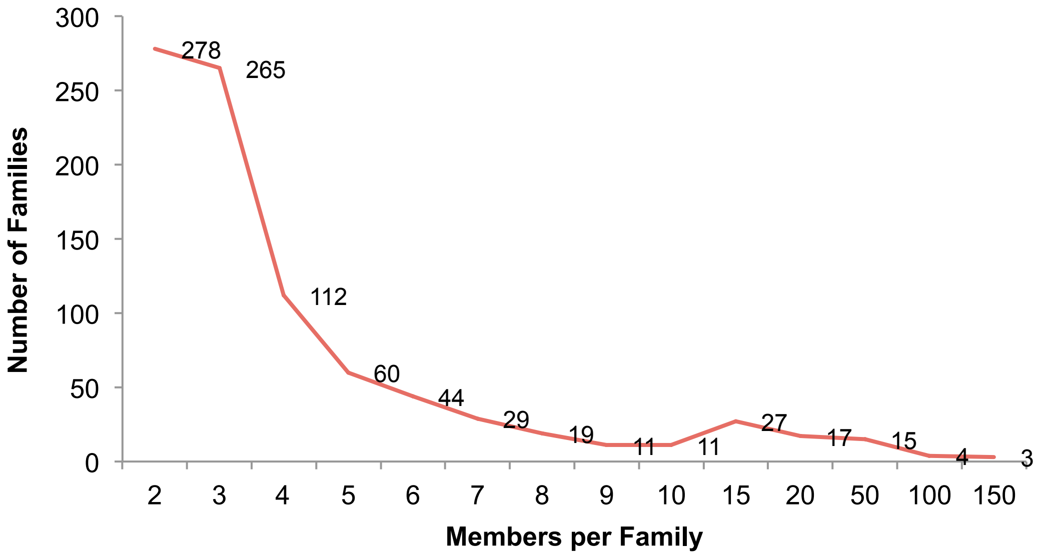


Figure 1.3.3. Size distribution of protein families in *Entamoeba histolytica*. Recreated with data published by Lorenzi *et al*, 2010 [98].

The majority of gene families in *E. histolytica* are unstudied, though some have been implicated in virulence in *E. histolytica.* The genome contains multiple multi-gene surface protein families that have been associated with virulence however, the mechanisms by which these families are regulated are not yet known. At the beginning of *E. histolytica* infection, trophozoites must first degrade and cross the mucosal layer that covers and protects the gut lining. To achieve this a group of enzymes known a cysteine proteases (CPs) are secreted. The CPs are a family of at least 50 endopeptidases, 36 of which form three major clades (A/B/C) [128,132]. Although the family is collectively regarded as virulent, evidence suggests that around 90% of the CP-derived proteolytic activity is provided by three proteins, EhCPSA1, EhCPSA2 and EhCPSA5 [133,134]. EhCPSA5 is particularly interesting, as no orthologue exists in the non-pathogenic *E. dispar* [135]. In concert with amoebic glycosidases, an unknown number of cysteine proteases degrade the MUC2 polymers that constitute much of the mucosal layer [136,137].

The *E. histolytica* trophozoites utilise surface-bound proteins to bind to host mucins and, once the mucosal layer has been degraded, the epithelial cells that line the gut. Two major gene families involved in this are the Gal/GalNAc lectins and the serine, threonine and isoleucine rich proteins (EhSTIRPs). The Gal/GalNAc lectin is a heterodimer, comprised of a 170 kDa heavy subunit and a 35 kDa light subunit, associated with a 150 kDa intermediate subunit [130]. The lectin binds to galactose and N-acetyl-galactosamine on host cell membranes and without it, the ability for *E. histolytica* to adhere to host cells is significantly reduced. *E. histolytica*'s cytotoxic impact upon the host cells is also reduced without the Gal/GalNAc lectin leading to the understanding that the cytokine cascade by which *E. histolytica* degrades cells is contact-dependent [11,138–140]. Down-regulation of EhSTIRP, 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 [141], implying that both proteins play a key role in amoebiasis.

Other gene families implicated in *E. histolytica* virulence include the BspA, Ariel-1 and AIG-1 gene families. The BspA-like proteins [98], which number more than 100 members and form the BspA family, are thought to localise to the cell surface. [129]. One member of this BspA-like family has been proven to be expressed at the parasite surface [129] and BspA proteins are known to play roles in adhesion to extracellular membranes in *Bacteroides forsythus* and *Trichomonas vaginalis* [142–144]. As such, the family has a clear potential role in amoebiasis, which could explain why the virulent *E. histolytica* has such a uniquely expanded set of BspA genes.

The smaller *E. histolytica* specific surface protein family, Ariel-1, contains 15 genes [105]. The family encodes asparagine-rich *Entamoeba histolytica* antigens (Ariel-1), which are constitutively expressed by trophozoites. The family also belongs to the same large family as the serine-rich *Entamoeba histolytica* protein (SREHP), which has been shown to have some use in immunising against amoebic infection [145]. It is interesting to note that the avirulent *E. dispar* does not have any unique copies of the surface-bound BspA and Ariel-1 families, which supports the theory that proteins involved in adhesion play essential roles in causing the invasive infections that distinguish *E. histolytica* from *E. dispar*.

Finally, the large AIG-1 family in *E. histolytica* contains 49 AIG1-like GTPases. This gene family is not specific to *E. histolytica* like the BspA family, though 12 members in the family are *E. histolytica* specific [98,146,147]. AIG genes are small GTPases, originally identified in *Arabidopsis thaliana* where they are thought to be associated with resistance to bacterial infections [148]. Their function in *Entamoeba* is less well understood however, studies have suggested that they may be involved in virulence or the adaption to the intestinal environment of the host [146,147] perhaps through the formation of protrusions on the plasma membrane that help with adherence to host cells [149]. Further supporting this is the observation that the AIG1 proteins are more highly expressed in more virulent *E. histolytica* cell lines compared to less virulent lines [146].

1.3.3. Gene regulation mechanisms in *Entamoeba histolytica*

Annotation of the *E. histolytica* genome revealed that introns and upstream open reading frames in the 3'-untranslated regions (UTRs) of genes were rare and hence, alternative splicing was an unlikely facilitator of gene regulation in *E. histolytica* [128]. Further supporting this, UTRs of *E. histolytica* genes are very short (<20 bp) [150,151] and subsequent whole transcriptome sequencing of *E. histolytica* confirmed the lack of alternative splicing [152].

RNA interference (RNAi) is an important biological process involved in gene regulation and genome stability, it has also been used by researchers as a robust tool for manipulation of gene expression [153–155]. Many pathways have been identified as being involved with the biogenesis and function of small RNAs though ultimately, all mature small RNAs associate with Argonaute (Ago) protein to produce an RNA-induced silencing complex, which mediates gene silencing [156–158]. Silencing occurs through transcriptional gene silencing (TGS), repression of translation or RNA cleavage [159]. During TGS, RNAi components recruit histone modification enzymes to target loci to induce silencing. Post-translational modifications to amino terminal tails of histones promote a change in the condensation state of the chromatin, regulating the accessibility of DNA-binding sites where transcriptional machinery can bind [160].

E. histolytica regulates gene expression through its non-canonical endogenous RNAi pathway [161,162]. *Entamoeba* have an abundant population of 27nt small RNAs that have 5'-polyphosphate (PolyP) termini, indicating that they are not Dicer products and mimics an observation only seen in the other amoebae, *Caenorhabditis elegans* and parasitic nematodes [161,163,164]. More recently, the repertoire of non-canonical RNAi proteins in *E. histolytica* was expanded with the characterization of EhRNaseIII, a minimal and non-canonical Dicer-like protein. Having a single RNaseIII domain and devoid of all domains typically associated with Dicer enzymes in other systems, EhRNaseIII is capable of processing double-stranded RNA into smaller RNA fragments that productively contribute to gene silencing [165]. It has been shown that *E. histolytica* genes are targeted by these small RNAs and are effectively silenced, although the mechanism by which TGS is initiated and maintained in *Entamoeba* species are unclear, though it may include chromatin remodeling to regulate gene expression [166]. Investigations into endogenous RNAi in *Entamoeba* discovered that genes to which abundant small RNAs map can induce silencing of genes fused to it [167,168]. The biological significance of RNAi in *E. histolytica* also remains unclear; studies have found that small RNA populations and the genes they target did not change in abundance or expression under various stress conditions (heat shock and oxidative stress) nor do they change between life cycle stages suggesting that the genes regulated by RNAi are not associated with stage conversion [169]. The RNAi pathway has been reported to silence genes relevant to virulence and this contributes to strain-specific virulence profiles [170].

The *Entamoeba* RNAi pathway is not responsible for the regulation of many genes, hence other regulation mechanisms must be involved. Epigenetic regulation of protein expression has been long recognized to be a key component in cellular development, adaptability and physiology of all living things, ranging from simple prokaryotes and Archaea to plants and mammals.

Epigenetics refers to chemical or structural modifications of DNA that ultimately result in altered RNA transcription and protein expression. The most studied epigenetic modification is DNA methylation and covalent modifications of histone proteins such as acetylation and phosphorylation. These modifications result in changes to the chromatin structure and accessibility of the DNA to transcription factors [171] and other nuclear proteins including methyl-binding domain proteins [172]. Recently, high-pressure liquid chromatography (HPLC) coupled to mass spectrometry revealed low amounts of 5-Methylcytosines (5-Me) in *E. histolytica* corresponding to around 0.05% of the genome [173], though the distribution of this methylation is unknown. Some evidence has suggested that DNA methylation in *Entamoeba histolytica* occurs in the repetitive DNA elements that litter the genome as a method of silencing these regions [174].

1.3.4. Comparative genomics among *Entamoeba* species

The arrival of next generation sequencing (NGS) revolutionised many areas of biology and enabled rapid sequencing of entire genomes, transcriptomes and epigenomes. Few comparative genomic studies of *Entamoeba* species exist. Large structural comparisons of *Entamoeba* species have been limited by the fragmented nature of the *E. histolytica* reference genome. For example, sequencing of *Entamoeba nuttalli,* the most closely related known species to *E. histolytica*, revealed that the *E. nuttalli* genome is smaller however it's assembly is more fragmented than *E. histolytica* and hence, it cannot be confirmed that this observation is due to real genomic differences between the two species or whether the sequencing technologies used to produce these genomes (Sanger for *E. histolytica* and Illumina 100 bp paired-end reads for *E. nuttalli*) are unsuitable for the assembly of such repetitive genomes.

As structural comparisons between *Entamoeba* species' genomes are difficult, most comparative genomics studies have focused on the comparison of gene content and single nucleotide differences among different *Entamoeba* species. A few studies of this kind exist however, most of these have investigated single nucleotide polymorphisms (SNPs) present in small numbers of loci in two or more species [112,131]. Genome-wide comparisons have still not been comprehensively performed. Intraspecific genomic diversity of the *Entamoeba histolytica* strains has been performed as multiple other *Entamoeba histolytica* strains and *Entamoeba* species have been sequenced using NGS [175,176]. SNPs have been identified between strains, as well as polymorphisms in gene copy numbers. The evidence of differential gene duplications [175] and the duplications of large chromosomal regions [98] among *E. histolytica* strains suggests the *E. histolytica* genome is very dynamic.

RNA-seq data exist for several *Entamoeba* species and comparisons of gene expression have also been performed [131]. It is possible that epigenetic differences among *Entamoeba* strains may be important in driving differential gene expression between species. However, no genome-wide methylation profiles exist for any *Entamoeba* species. The development and sequencing of bisulphite libraries of *Entamoeba* species (Chapter 5) would enable detection of any differential methylation between *Entamoeba* species and strains that may drive infection outcome.

An improved *E. histolytica* reference genome would significantly aid such comparative studies by providing a high quality genome that is rich in spatial information on genes and other sequences, on to which other *Entamoeba* species/strain sequence data can be mapped to reveal structural variations between different species and strains.

1.4. Improvements to the assembly of repetitive, complex genomes

When assembling repetitive and complex genomes, fragmented assemblies are often produced. This usually results from the technical difficulties in reconstructing the genome sequence and is a feature observed in the highly fragmented *E. histolytica* genome assembly. Single molecule real time (SMRT) sequencing, which produces significantly longer reads than current platforms may provide a solution to assembly problems by producing reads that are longer than the repetitive regions of the *E. histolytica* genome, producing reads whose ends contains unique sequences to which other reads can be confidently joined. Incorporating these reads with complementary techniques such as optical mapping and chromosome conformation capture techniques such as Hi-C to guide assembly of the raw reads further promises to produce a *E. histolytica* reference genome that is more contiguous and information-rich than the existing reference genome. Alongside this, bisulphite sequencing of *Entamoeba* species allows methylation in *Entamoeba* genomes to be quantified and compared.

1.4.1. Single Molecule Real Time (SMRT) sequencing

Accurate assemblies of the genomes of organisms are crucial to understanding organism diversity, speciation, evolution of species and the impact of genomic diversity on health and disease. Prior to the recent development of single molecule sequencing, the most advanced methods of DNA sequencing involved amplification of template DNA. These methods often referred to as second generation sequencing, generally use massively parallel methods for amplifying and then sequencing by DNA synthesis. This new third generation method of sequencing allows longer sequenced read lengths of up to 10,000 bases long and produces these reads more quickly than second generation sequencing as the reaction is observed in real time [177,178].

SMRT sequencing has been used to improve the genome assemblies of many organisms whose genomes have already been sequenced. High quality, highly contiguous assemblies have proven invaluable for population genomic studies, most notably in humans. SMRT sequencing of the first Chinese [179] and Korean [180] human reference genomes have led to the discovery of population-specific sequences in these populations compared to the human reference genome, which is largely derived from European individuals. Further, SMRT sequencing of human genomes has revealed structural variation that was undetected by NGS data suggesting that long-read sequencing data and associated assemblies significantly increase the sensitivity when detecting structural variations between two individuals or populations [181].

SMRT sequencing has also improved the research into other protists. SMRT sequencing was used to improve the genome assembly of the monkey malaria parasite, *Plasmodium cynomolgi.* SMRT sequencing produced a genome assembly of significantly higher quality than the existing reference, comprising 56 contigs, no gaps and an improved average gene length. 1,000 more genes were annotated and the new assembly improved understanding of the sub-telomeric sequence of *P. cynomolgi,* which constitutes nearly 40% of the genome sequence. The new assembly revealed a novel expansion of 36 methyltransferase pseudogenes in the sub-telomeric regions from what was thought to be a single copy gene in the previous reference assembly [182].

The ability to accurately reconstruct more complete genomes and unlock the comparative genomics capabilities that comes with more complete genomes makes the Pacific Biosciences RSII platform (PacBio) desirable for the assembly of genomes, especially those which are highly fragmented and require long reads to span arrays of repetitive elements. This is why PacBio sequencing is suitable for sequencing the highly fragmented and repetitive *E. histolytica* genome. The process by which PacBio was used and the subsequent genome reconstructed in described in Chapter 2.

1.4.2. BioNano optical mapping

Often when constructing *de novo* genome assemblies, short-read or long-read sequence alone is not enough to produce a chromosome level assembly. As a result, sequence data is often combined with complementary technologies such as optical mapping to produce a physical map that can be used to guide the assembly of sequencing reads produced by NGS and/or SMRT sequencing.

BioNano maps have been used to improve the assembly of many repetitive genomes, including plant genomes that are notoriously repetitive and complex to assemble. BioNano genome maps have been used in a hybrid approach with long-read read data to improve the genome of *Zea mays* (maize) [183] and *Trifolium subterraneum L.* (clover) [184]. In both cases, the genome assembly aided by BioNano mapping resolved previously uncharacterized regions of the genome and novel genes, transposable elements and structural variations.

1.5. Gaps still remaining in *Entamoeba* knowledge

Previous sequencing attempts have been invaluable in contributing to current knowledge of *Entamoeba histolytica* biology. However, a large amount of knowledge surrounding the genome of *E. histolytica*, and other *Entamoeba* species, remains elusive. Some of this is a result of the current quality of the *E. histolytica* reference genome, which remains largely fragmented owning to the repetitive nature of the genome.

1.5.1. Organisation of genes and gene families

The repetitive nature of the genome has meant that despite continuous sequencing efforts, the assembly remains fragmented and a chromosome level assembly is yet to be reached. The utilisation of single molecule sequencing, as previously described, promises improved contiguity of the *E. histolytica* assembly. Currently, however, very little is known about the organisation of genes within the *E. histolytica* genome. No information on gene family organisation is available, as often members of the same gene family occur over multiple scaffolds (AmoebaDB data). In addition, the lack of information regarding the wide-scale structure of *E. histolytica* has made it impossible to understand the evolution of large gene families and it still remains unclear how such families have expanded throughout the genome. Understanding the organisation of gene families has been important in understanding the biology of other parasites. For example, single-copy expression of gene family members has been shown to regulate variation of surface proteins in a process called antigen switching. *P. falciparum* differentially expresses genes from the *var* gene family [185,186], which are involved in evasion of the host immune system and cytoadhesion of infection erythrocytes [87,187–189]. The majority of the *var* genes are located in gene clusters in the sub-telomeric regions where recombination between different *var* paralogues produces novel *var* genes [87,190]. In addition, the close proximity of the genes to one another is thought to facilitate antigen switching which is mediated epigenetically [191]. Similarly, antigen switching is observed in *Trypanosoma brucei* in the Variant Surface Glycoproteins (VSGs), which mediate immune evasion. The VSGs are mono-allelically expressed from expression sites (ES) in the sub-telomere [192–194], and VSG switching exploits subtelomere plasticity [195].

1.5.2. Structural features of the *E. histolytica* genome and associated genes

Many structural features of the *E. histolytica* genome remain unresolved; these include information about the chromosome number, ploidy and the structure and sequence of the telomeres of this species. Related to these structures, it is unknown what genes occur in close proximity to the telomeres. Research into other protists has shown that the sub-telomeric regions are often enriched for virulence genes (see above). As previously mentioned, the *E. histolytica* genome contains a variety of multi-gene families, some of which encode surface proteins. Many of these have been shown to be expressed on the cell surface but the regulation of individual gene family members is not understood and hence, it is unknown whether any surface protein-based virulence mechanisms, analogous to those seen in *Trypanosoma* and *Plasmodium*, exist in *Entamoeba*.

Complicating this analysis further is the poor annotation of the *Entamoeba* genomes. Even if the telomeric and sub-telomeric regions had been resolved by previous sequencing projects, the majority of annotated genes encode proteins of unknown function (53.8%) making functional studies difficult. Methods for *in vitro* down-regulation of genes, for use in functional studies in *E. histolytica,* have been limited by the unknown ploidy of the parasite and the lack of robust homologous recombination*.* RNAi approaches that include ‘feeding’ the parasite bacteria expressing double-stranded RNA to a gene of interest or soaking parasites in small RNAs (sRNAs) have shown some success [196,197]. Though these techniques vary widely in their efficiency of down-regulation and long-term silencing stability, as loss of silencing has been observed [198]. Additionally, achieving silencing using shRNA is labour-intensive [199]. Owing to the lack of a method for stable down-regulation of genes of interest, very few experimental functional studies have been performed on *E. histolytica* and many of the genes remain functionally uncharacterised.

1.5.3. Distribution of DNA methylation across the *E. histolytica* genome

As previously mentioned, DNA methylation is present in *E. histolytica* in small amounts. Its distribution is largely unknown. Epigenetics is an important factor in the virulence, differentiation and lifecycle control of a range of protists including *Toxoplasma gondii, Plasmodium falciparum and Trypanosoma brucei* [200–203]. Evidence of alternate transcriptomes has also been obtained for *Entamoeba histolytica* HM-1:IMSS and the avirulent *Entamoeba histolytica* Rahman, with differential expression profiles for key virulence genes including the cysteine proteases (CPs) and Gal/GalNAc lectins [204]. Explanations behind these differential expression profiles have been suggested as being mediated by differential DNA-methylation between the two species [205]. Although many of the fundamental principles of epigenetic gene regulation in protists are similar to those observed in mammals, the protist parasites demonstrate unique and diverse mechanisms of epigenetic gene regulation. DNA methylation is also an essential virulence regulation mechanism in several pathogenic bacteria [206,207]. For example in *Salmonella enterica,* the lack of *Dam* (DNA adenine methyltransferase) methylation causes reduced mobility and an impaired ability to invade the host intestinal epithelium [208].

The observation of differential expression profiles between *Entamoeba* species suggest there may be an important role of DNA methylation in the regulation of genes that distinguish the virulent *E. histolytica* from the avirulent *E. dispar.* This information cannot be obtained from the current HPLC data and a whole-genome approach, such as whole-genome bisulphite sequencing, will be required before the identification of genes that are methylated can be performed.

1.6. Aims of thesis

This thesis describes the re-sequencing and assembly of the *E. histolytica* HM-1:IMSS genome using third generation sequencing technology in a bid to improve understanding of genome structure and gene-family organisation. In addition, bisulphite sequencing has been utilised to study the pattern of methylation across the *E. histolytica* genome (and two other *Entamoeba* species) with the aim of understanding the role that low-level epigenetic modification plays in *Entamoeba* genomes. It is hoped these analyses will help explain the mechanisms of expansion of gene families, elucidate unknown structures such as telomeres and their associated sub-telomeric regions, and reveal the pattern and role of DNA methylation across the *E. histolytica* genome.

Chapter Two describes the production the improved reference genome and comprehensively compares the assembly methods currently available for long-read data. *E. histolytica* HM-1:IMSS is sequenced and assembled using a variety of the available long-read assemblers. At the time of sequencing, comparisons of long-read genome assemblers were not available so this chapter compares four publicly available assembly programs: Canu, HGAP, Falcon and Miniasm. The results of this technical evaluation offer guidance for future similar assemblies whilst also producing a better quality *E. histolytica* reference genome. Chapter 2 also highlights how, even with more advanced sequencing technologies, assembly of the *Entamoeba histolytica* genome remains challenging due to the complex nature of the genome.

Chapter Three describes the genome-wide annotation of genes and other features within the new reference assembly. This is compared to the previous reference genome to identify novel genes and investigate novel gene family expansions (whether a result of biological or technical differences). The chapter describes the expansion of a previously reported single copy gene. Further analysis is performed into the identification of gene families and investigations into their structure and organisation. The hypothesis that expansion of virulence gene families has been facilitated by the propagation of transposable elements throughout the *E. histolytica* HM-1:IMSS is investigated and evidence presented. Investigation of gene enrichment of sub-telomeric regions of the genome is also performed with the aim of determining whether any analogous virulence mechanisms to those observed in the sub-telomeric regions of other eukaryotic parasites are present in *E. histolytica*. Overall, Chapter Three aims to identify novel genes, gene family expansions or unique organisations of gene families that may be related to the outcome of *E. histolytica* infections and/or can explain virulence within the parasite.

Chapter Four explores the repetitive features of the *E. histolytica* HM-1:IMSS genome. First this chapter explores the tRNA array structures that are unique to *Entamoeba* species in an attempt to determine their genomic location and function. This chapter presents evidence to suggest that the tRNA arrays are the telomeres in *Entamoeba* species, analogous to the rRNA-based telomeres seen in *Dictyostelium discoideum* [103]. The identification of the telomeres enables the analyses of gene enrichment in sub-telomeric regions performed in Chapter Three. The tRNA arrays are currently used for genotyping *Entamoeba* parasites, as the genus does not contain generic microsatellites. Further analyses of these tRNA arrays are performed to identify the efficiency and accuracy of using these sequences as genotyping markers. The analysis reveals some tRNA arrays contain significant sequence variation while others appear extremely stable and the chapter provides guidance as to which tRNA arrays should be used for detection of the *E. histolytica* infections. Chapter Four also investigates the extra-chromosomal rDNA episomes that exist in hundreds of copies per cell. The full sequence of the EhR2 episome is assembled in the SMRT sequencing assembly, a feat that could not be performed using short read sequencing due to the repetitive nature of the rDNA episomes. The chapter reports the loss of EhR1, thought to be the main rDNA-carrying episome in *E. histolytica* HM-1:IMSS. The Chapter also reports the absence of a chromosomal copy of the rDNA genes.

Chapter Five explores the epigenetic landscape of three *Entamoeba* genomes. Firstly, the DNA methylation pattern of *E. histolytica* HM-1:IMSS is explored with the aim of defining 5-Methyl cytosine methylation (5-MeC) throughout the genome and where this methylation occurs. The chapter further develops the analyses of methylated genes through examining correlations between methylated genes and gene expression utilizing existing RNA-seq data. A tiny proportion of genes are methylated in the *E. histolytica* genome and the function of these genes is investigated to assess if any obvious importance exists for the methylation of these specific genes (e.g. are any of the methylated genes involved in virulence as observed in other protists?). Most methylation is detected in repetitive regions of the genome such as in transposable elements and in the tRNA arrays and Chapter Five aims to determine the function of the methylation of these regions and hypothesises that DNA methylation is a protective mechanism in *Entamoeba* species, protecting the genome from deleterious expansion of transposable elements and stabilising the telomeric tRNA arrays. Finally, the chapter investigates the extent of methylation in two other *Entamoeba* species, *E. moshkovskii* and *E. invadens*, to determine whether it is a species-specific phenomenon or whether methylation of particular genes and/or genome features is conserved among *Entamoeba* species.

Chapter 2 – SMRT sequencing and assembly of the *Entamoeba histolytica* HM-1:IMSS genome

2.1 Introduction

As outlined in Chapter 1, single molecule real time (SMRT) sequencing using the Pacific Biosciences RS II (PacBio sequencing) generates long reads capable of spanning repetitive regions that cannot be spanned by short read technologies and so are capable of producing more contiguous assemblies. This made it very desirable as a method to re-sequence the *Entamoeba histolytica* HM-1:IMSS genome as it was reported previously as being approximately 20% repetitive [97,98]. Consequently, the previous sequencing attempts using Sanger sequencing, and some further attempts using 454 sequencing, have produced highly fragmented assemblies arranged in relatively small scaffolds and therefore, there is currently little knowledge of the genome structure and organisation of *E. histolytica* HM-1:IMSS (technologies producing approximate read lengths up to 1Kbp, but generally shorter; Assembly unpublished; Data available from EuPathDB).

2.1.1. Assembling long single molecule real time (SMRT) sequence reads

PacBio sequencing and other long-read sequencing platforms are known to produce reads with a high rate of error (around 11%-15%) in a single continuous long read (CLR) [178,209]. Producing circular consensus sequences (CCS) can mitigate these errors with sufficient passing of a CLR. However, the length of a CLR is limited by the lifetime of a polymerase molecule therefore the number of sequencing passes and the CCS read lengths are negatively correlated. That is, shorter sequences yield more passes in a CLR, and hence a higher accuracy and *vice versa* [210]. Therefore, errors can remain in the assembled reads if coverage is low and CCS read lengths are long.

Previous programs developed for genome assembly have been designed to cope with short next generation sequencing (NGS) reads with a high level of accuracy. PacBio sequencing produces the opposite type of reads and hence, new programs needed to be developed to cope with the long, error-prone reads. At the time of assembly, few of these had been developed and even fewer had been compared to one another to determine which of these new assemblers performed the best. To carry out such a comparison, four assemblers were selected for analysis: HGAP [211], Canu [212,213], Falcon [214] and Miniasm [215].

The SMRT portal is an open-source, browser-based program provided by Pacific Biosciences that can directly interpret the raw output data produced by the Pacific Biosciences RS II instrument. It offers a suite of analysis applications optimised for single-molecule sequencing data, including *de novo* assembly, variant detection and epigenetic motif detection [211]. The suite provides two main SMRT analysis applications for *de novo* assembly. These are RS_HGAP Assembly.2 (HGAP2) and RS_HGAP Assembly.3 (HGAP3). Both use the Hierarchical Genome Assembly Process (HGAP) SMRT Analysis algorithm to generate *de novo* genome assemblies using a single library type. HGAP follows a pipeline comprising: (i) pre-assembly (mapping single pass reads to seed reads, which represent the longest portion of the read length distribution); (ii) *de novo* assembly; (iii) assembly polishing (correcting miscalled bases and erroneous indels using read coverage) [211]. The main difference between HGAP2 and HGAP3 is the assembler used in the *de novo* assembly step. HGAP3 is optimised for speed and therefore carries out *de novo* assembly with PacBio's AssembleUnitig whereas HGAP2 performs *de novo* assembly using the Celera Assembler [216]. The AssembleUnitig assembler replaces the most time-consuming step in the genome assembly process, which is the Celera Assembler step.

Three other assemblers tested were open-source, long read assemblers designed to cope with the noisy, error-prone raw reads produced by PacBio and

Oxford Nanopore technologies. Falcon is based on the HGAP assembly process. However, it is able to split haplotypes in a way that is more reflective of a diploid genome [214]. The Canu assembler is different to HGAP and Falcon in that it incorporates a novel overlapping and assembly algorithm based on it's predecessor, Celera [212]. Miniasm is another long-read *de novo* assembler but differs from the previous assemblers in that it does not have a consensus step. The Miniasm program produces final contigs by concatenating pieces of read sequences. As a result, the per-base error rate in the final contigs is similar to that seen in the raw input reads [215].

2.1.2. Metrics used to compare assemblies produced by different long read assemblers

A perfect genome assembly would be an assembly with fully contiguated, telomere-to-telomere, chromosomes, to allow for accurate gene model annotation and analysis of genome organisation. Virtually no eukaryotic genome assemblies meet this ideal, so it is important to quantify what a good quality (and biologically useful) assembly looks like. A good quality assembly should have the following features; a total assembly length close to the estimated genome size assembled into a reasonable number of contigs or scaffolds that is manageable for downstream analyses. These contigs should be larger than the average gene size of the organism in question to allow downstream gene-annotation of all possible genes.

A number of papers have discussed the problems associated with trying to determine metrics for assessing the quality of an assembly [217–219]. No single parameter is universally agreed upon as an accurate predictor of a 'good' genome assembly. Assembly quality and the performance of an assembler is dependent on the genome in question, as some are inherently much more difficult to assemble due to their highly repetitive nature or low complexity (genomes with high GC% or AT%). However, a range of metrics can be applied to get an overall idea of the quality the genomes produces by each assembler. Key metrics are described below.

2.1.2.1. N Statistics

One of the parameters often used as a measure of a good assembly is the N50 The N50 is a weighted median statistic such that 50% of the entire assembly can be contained in contigs or scaffolds equal to or larger than this value (Figure 2.1.1 A-B). Comparing N50s across a range of different assemblies can be used to give relative merit to one assembly over another. A similar metric, and often argued as a more useful metric, is the NG50 length. The NG50 is calculated in the same way as the N50 except the total assembly size is replaced with the estimated genome size when making the calculation, meaning that comparison of assemblies is standardised (Figure 2.1.1 A/C).

The idea of using N50/NG50 as a measure of assembly quality has come under scrutiny. If the majority of contigs/scaffolds in an assembly are short, ranking assemblies by their N50/NG50 size can be very inaccurate and misleading. This is because an assembly that contains a few very large contigs/scaffolds, despite having a large majority of smaller contigs/scaffolds, can still produce the largest N50/NG50 when compared to better quality assemblies (Figure 2.1.1 D). Therefore, it has been recommended that the N50/NG50 metric is much more informative when used in comparison with other metrics [220]. Alongside analysing N50 values that has been proposed incorporates other NG(X) values (e.g. N25 and N75 values) to display a more representative contig length distribution rather then relying on a single N50/NG50 value to compare genome assemblies [218].

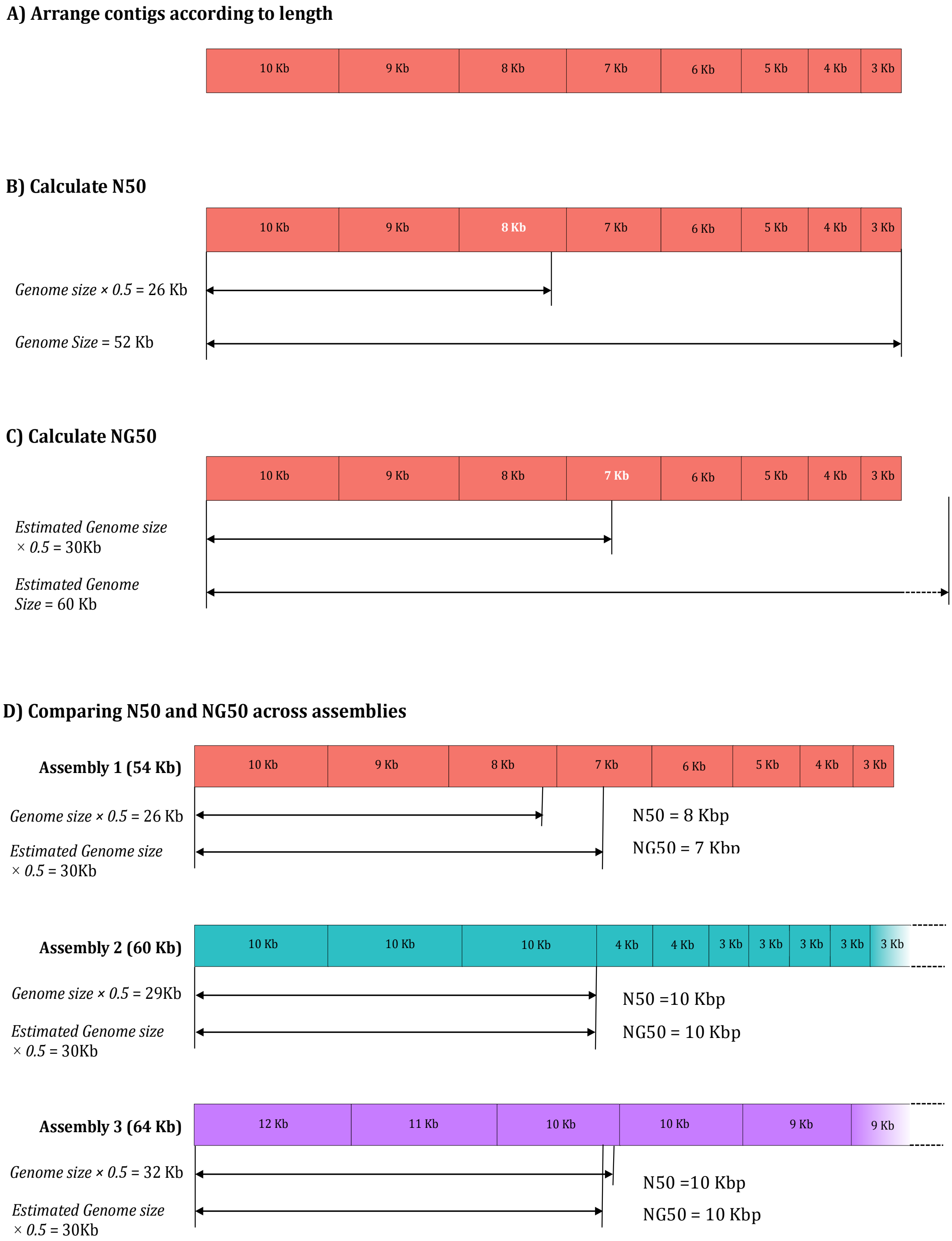


Figure 2.1.1. Calculating and comparing N50 and NG50 values. A) To calculate N50 and NG50, contigs/scaffolds are first ordered by length (high to low). **B)** To calculate N50, contig lengths are consecutively added together from longest to shortest until 50% of the *total assembly size* is reached. The N50 is the length of the last contig to be added. **C)** To calculate NG50, contig lengths are consecutively added together from longest to shortest until 50% of the *estimated genome size* is reached. The NG50 is the length of the last contig to be

added. **D) Comparing N50 and NG50 across assemblies.** N50 values can be poor indicators of demonstrating the distribution of sizes of contigs in an assembly. Assemblies with a few long contigs and many small contigs (Assembly 2) can produce the similar N50/NG50 values as an assembly composed of a more even distribution of contig lengths (Assemblies 1 and 3).

2.1.2.2. The proportion of 'gene-sized' scaffolds

The normal progression upon completion of a genome assembly is genome annotation using *ab initio* or *de novo* methods of gene prediction [221,222]. A metric of assembly quality directly relevant to this is the proportion of contigs/scaffolds that are longer than the average gene size of the organism. It has been proposed that an assembly with a large proportion of scaffolds longer than the average gene size may be an indicator of a genome assembly of sufficient quality to carry out gene annotation [223].

2.1.2.3. Gene content-based assembly completeness

To measure the completeness of a genome assembly, Benchmarking Universal Single-Copy Orthologues (BUSCOs) have been generated from large collections of sequenced genomes. These are sets of single-copy orthologous genes conserved throughout a phylogenetic clade (BUSCO sets have been determined for Bacteria, Eukaryotes, Protists, Metazoa, Plants and Fungi) [224].

The BUSCO program for assessing genome assembly completeness uses tBLASTn [225], Augustus [226] and HMMER 3 [227] to determine whether a conserved gene set is present in a genome. The identified orthologues are defined as single-copy, fragmented or duplicated. The total number of orthologues can give a measure of how complete a genome is and also can be used between genome assemblies to identify which assembly has the most complete gene set. A large number of duplicated BUSCOs can indicate 'over-assembly': the representation of haplotypes that should have been collapsed multiple times in the assembly. This is because the real duplication of BUSCOs is expected to be rare as they tend to occur as single copy genes [228].

2.1.3. Aims of chapter

Previous sequencing attempts have been unable to assemble a chromosome-level, gold standard genome for *Entamoeba histolytica*; new long-read sequencing technology may help to overcome this problem and cope with the repetitive nature of the genome. At the time of sequencing the *E. histolytica* genome using PacBio sequencing, few long-read assemblers had been rigorously tested and good practice procedures were still yet to be defined for long-read data. This chapter aims to assemble a new *E. histolytica* genome that can used as a tool for the remaining chapters whilst also bench-marking some of the available long-read assemblers available at the time of assembly. Specifically the chapter aims to:

- Produce a range of PacBio assemblies to benchmark assembler programs for long-read data
- Conclude whether utilising old assembly data with PacBio data generates a better assembly
- Improve PacBio sequencing attempts with additional technologies (Hi-C, Optical mapping).

2.2. Materials and methods

2.2.1. Origins and growth of *Entamoeba histolytica* HM1:IMSS trophozoites

Entamoeba histolytica HM-1:IMSS is a long-established laboratory strain, originally isolated in 1967 from a patient suffering from amoebic dysentery in Mexico City [100]. HM-1:IMSS trophozoites were cultured in LYI-S-2 (liver extract, yeast extract, iron, serum growth medium) with 15% adult bovine serum (PAN-Biotech, Aidenbach, Germany) as described [229].

Trophozoites were grown in 15 mL borosilicate tubes (Fisher Scientific, Hampton, NH, USA) with screw caps, filled with 13 mL of LYI-S-2 to minimise the amount of oxygen in the culture. Cultures were kept upright at 36°C and sub-cultured into new tubes every 3-4 days with an inoculum of 150 µL - 250 µL taken and added to fresh media. The inoculum was increased to 1.5 mL – 2.5 mL when adapting cells to a new batch of serum and progressively decreased to the normal inoculum volume over several weeks. High cell density was usually reached between 72-96 hours growth at 36°C and cells were harvested at this point.

2.2.2. Purifying high molecular weight genomic DNA for whole genome sequencing

Entamoeba spp can be difficult organisms extract large quantities of high quality genomic DNA from; this is for a number of reasons. There is little DNA per cell and one 13 mL culture tube can support only up to around 100,000 cells (Clark, G., 2015, Pers. Comms.); many lytic enzymes are released when cells are lysed, degrading the DNA; and cells are rich in polysaccharides that co-precipitate with DNA, meaning several rounds of DNA purification are required, with loss of DNA and degradation with each round. Therefore, three DNA extraction methods were tested before large-scale DNA extraction for sequencing. Cell cultures were centrifuged (1,000 rpm, 10 mins) and medium removed, then cells were washed twice in phosphate-buffered saline (PBS) and resuspended in a 50 µL of PBS in DNA LoBind 1.5 mL microcentrifuge tubes (Eppendorf, Hamburg, Germany). Samples were subjected to one of three DNA extraction methods.

2.2.2.1. QIAGEN DNeasy Blood and Tissue Kit

DNA was extracted using the spin-column-based QIAgen DNeasy Blood and Tissue kit (QIAgen, Crawley, UK) as per the manufacturer's protocol for cell culture samples. Mixing by vortexing was avoided at all steps except the initial lysis stage; instead mixing was done by pipetting. Purified DNA was re-suspended in nuclease-free water.

2.2.2.2. QIAGEN Gentra Puregene Cell Kit

DNA was extracted using the QIAgen Gentra Puregene Cell Kit (QIAgen) as per the manufacturer's protocol for cell culture samples. This method uses high salt buffers to precipitate proteins, separating them from the DNA. Purified DNA was re-suspended in nuclease-free water.

2.2.2.3. CTAB:Phenol:Chloroform DNA purification

Washed cell cultures were lysed in 300 µL of QIAgen cell lysis buffer (QIAgen). A modified version of a phenol:chloroform extraction using cetyl trimethylammonium bromide (CTAB) to remove polysaccharides, described elsewhere (Ali et al., 2005; Clark & Diamond, 1991; Clark, 2015a), was used to isolate genomic DNA. Specifically, 250 µL of lysis buffer (0.25% SDS in 0.1 M EDTA, pH 8.0) and Proteinase K (QIAgen) to 100 µg/mL was added to washed cell cultures. Samples were vortexed and incubated at 55°C for 20 minutes. 75 µL of 3.5 M NaCl and 42 µL of CTAB solution (10% w/v CTAB in 0.7 M NaCl, preheated to 65°C) was added to samples and incubated for 10 minutes at 65°C. At room temperature, 500 µL of Chloroform (Sigma-Aldrich, MI, USA) was added; samples were mixed by inversion and centrifuged at 13,000 rpm for 5 minutes. Supernatant was transferred to a new DNA LoBind 1.5 mL tube, 500 µL of Phenol:Chloroform:Isoamyl Alcohol 25:24:1 (Sigma-Aldrich) was added, mixed by inversion and centrifuged as before. Supernatant was transferred to a new DNA LoBind 1.5 mL tube, 2 volumes of ethanol (100% EtOH, room temperature) were added and samples incubated overnight at -20°C before being centrifuged as before. The supernatant was discarded and the DNA pellet washed with 200 µL of room temperature 70% ethanol then centrifuged as before (this was repeated twice). The DNA pellet was air-dried before being resuspended in 50 µL of sterile water. Vortexing was avoided at all stages, except initial lysis, to avoid shearing the DNA.

2.2.3. Further DNA purification and DNA quality assessment

Extracted DNA was subjected to RNase treatment using QIAgen RNase A solution (QIAgen). 3% (v/v) RNase A solution was added to samples and incubated at 37°C for 30 minutes before RNase was removed using QIAGEN protein precipitation solution and subsequently washed with rounds of 100% and 70% ethanol. Samples were stored in nuclease-free water, ready for DNA clean up.

A Solid Phase Reversible Immobilization (SPRI) based method, which utilises paramagnetic carboxyl coated beads, was used for DNA clean up opposed to a column based clean up, to minimise shearing of DNA. The bead mix (known as MagNA) was made following a protocol described elsewhere [232].

1.8 volumes of MagNA were added to RNase treated DNA, mixed and, incubated at room temperature for 5 minutes, to allow DNA to bind to the beads. Separation was carried out using a magnetic Eppendorf rack for approximately two minutes, until the solution had cleared. Supernatant was removed and 500 µL of freshly made 70% ethanol added. The solution was incubated for 1 minute before ethanol was removed and another ethanol-wash carried out. Beads were then air-dried on the magnet at room temperature for 5-10 minutes. Elution of DNA was carried out by the addition of nuclease-free water. This MagNA-water solution was mixed and incubated (off the magnet) at room temperature for 5 minutes to elute DNA. The solution was returned to the magnet for approximately two minutes, until the solution had cleared. The supernatant, containing purified DNA, was then removed and stored for DNA library preparation, unless further clean up was required as indicated by poor 260:230 ratios.

DNA concentration was determined using a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA) as per the manufacturer's instructions. DNA purity was assessed by analysis of 260:280 and 260:230 ratios as determined by a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE, USA). 2 µL of sample were loaded onto the NanoDrop instrument and the absorption ratios measured as per the manufacturer's instructions. Pure, clean DNA was determined as having a 260:280 ratio of 1.8-2.0 and a 260:230 ratio of >2.0. Agarose gel electrophoresis was used to assess the size (i.e. integrity) of the DNA. Samples were run on a 1.0% w/v agarose gel for 16-18 hours at 30V with a high molecular weight ladder for size reference. Ethidium bromide was used to visualise the DNA fragments using a transilluminator.

2.2.4. Library preparation and SMRT sequencing of *Entamoeba histolytica* HM-1:IMSS DNA

CTAB:phenol:chloroform-based DNA purification was found to be the only method to produce high quantity, high quality, high molecular weight DNA and these samples were used for PacBio sequencing. 10 µg of DNA was submitted to The Centre for Genomic Research (CGR, Liverpool, UK) for PacBio library generation and sequencing using the following protocol.

DNA was purified once more using the AMPure XP purification system (Agencourt, Brea, CA, USA), using a 1:1 volume ratio of sample to AMPure beads as per the manufacturer's instructions. DNA quantity and quality were again assessed using NanoDrop and Qubit assays, as well as an Agilent Bioanalyser (Agilent Technologies, Santa Clara, CA, USA), using a high sensitivity kit, to determine the average size of the DNA. DNA was sheared using a Covaris G tube with the S2 focused-ultrasonicator (Covaris, Woburn, MA, USA) to generate 10 Kbp fragments, and the sample cleaned again using the AMPure XP purification system.

DNA was treated with Exonuclease V11 at 37 °C for 15 minutes. The ends of the DNA were repaired; the sample was incubated for 20 minutes at 37°C with the damage repair mix supplied in the SMRTbell library kit (Pacific Biosciences, CA, USA). This was followed by a 5 minute incubation at 25 °C with end repair mix. DNA was cleaned using 1:1 volume ratio of AMPure beads and 70% ethanol washes.

DNA was ligated to adapters overnight at 25 °C. Ligation was terminated by incubation at 65°C for 10 minutes followed by exonuclease treatment for 1 hour at 37°C. The SMRTbell library was purified with a 1:1 volume ratio of AMPure beads. The quantity of library, and therefore the recovery, was determined by Qubit assay and the average fragment size determined by the Agilent Bioanalyser. Size selection was performed on a Sage Blue Pippin Prep (Sage Science Inc., Beverly, MA, USA) using a 0.75% agarose cassette and S1 marker. The final SMRT bell was recovered as before and quantified.

The SMRTbell library was annealed to the sequencing primer at values predetermined by the Binding Calculator (Pacific Biosciences, CA, USA) and a complex made with the DNA Polymerase (P6/C4 chemistry). The complex was bound to Magbeads and this was used to set up the SMRT cells. Sequencing was done using 360 minute movie times.

2.2.5. Library preparation and Illumina sequencing of *Entamoeba histolytica* HM1:IMSS

DNA was harvested as described in section 2.2.2.3. 400 ng of high quality DNA was submitted to The Centre for Genomic Research (CGR, Liverpool, UK) for Illumina library generation. Libraries were generated, assessed for quality and sequenced using the following protocols.

200 ng of this DNA was sheared with the S2 Focused-Ultrasonicator (Covaris, Woburn, MA, USA) to generate fragments approximately 350 bp in length, following the manufacturer's guidelines. The sample was cleaned using the AMPure XP purification system (Agencourt, Brea, CA, USA) using a 1:1 volume ratio of sample to AMPure beads as per the manufacturer's instructions. The sample was end-repaired and size selected to retrieve ~350 bp fragments. The sample was A-tailed and adapter ligated before being amplified with 8 cycles of PCR. The library was cleaned with a 1:1 volume ratio of AMPure beads.

The quantity of library, and therefore the recovery, was determined by a Qubit assay (Invitrogen) and the average fragment size determined by Agilent Bioanalyser (Agilent Technologies).

A quantitative real-time PCR (qPCR) assay, designed to specifically detect adapter sequences flanking the Illumina libraries, was performed using an Illumina KAPA Library Quantification Kit (Kapa Biosystems, Wilmington, USA). Quantification of cDNA templates containing adaptor sequences on both ends of the template was determined using the qPCR output.

Following calculation of the molarity using qPCR data, template DNA was diluted to 3 nM concentration using the resuspension buffer. 5 μL of the 3nM stock DNA was denatured for 8 minutes at room temperature using 5 μL of freshly diluted 0.1 N sodium hydroxide (NaOH) and the reaction was subsequently terminated by the addition of the 5 μL TrisCl (pH 8.0, 0.5M). 35 μL enzyme mix was added to the denatured DNA library so that the final loading concentration was 300 pM and cBot clustering started immediately. During clustering, the templates were immobilized onto a proprietary flow cell surface, designed to present the DNA in such a manner to facilitate access to enzymes whilst maintaining high stability of the surface-bound template and low non-specific binding of fluorescently labelled nucleotides.

The pooled libraries were sequenced on an Illumina HiSeq 4000 platform using sequencing by synthesis (SBS) technology to generate 2 x 150 bp paired-end reads.

2.2.6. *De novo* assembly of the *Entamoeba histolytica* HM-1:IMSS genome

Raw PacBio reads were processed into filtered sub-reads by the CGR (Liverpool, UK) by breaking the polymerase reads into single passes and removing adapter sequences.

These PacBio sub-reads were used as an input to a range of assemblers, assemblers without their own polishing step were polished using Pilon [233] and the paired-end reads produced from the Illumina TruSeq library. Analysis of assembly metrics, using custom perl scripts (assemblyStats.pl, Appendix 2), followed to compare assembly quality. Assemblies that passed this test (HGAP2 and Canu) were analysed to determine differences in the assembly. tRNA arrays and rDNA episomes represented a large number of contigs therefore these were masked from the assembly. The remaining assemblies showed little difference once these regions were removed and it was decided to merge the best-assembled regions from the HGAP and Canu assemblies. The final assembly was polished using Pilon and the paired-end reads produced from the Illumina TruSeq library. This process is outlined in Figure 2.2.1.

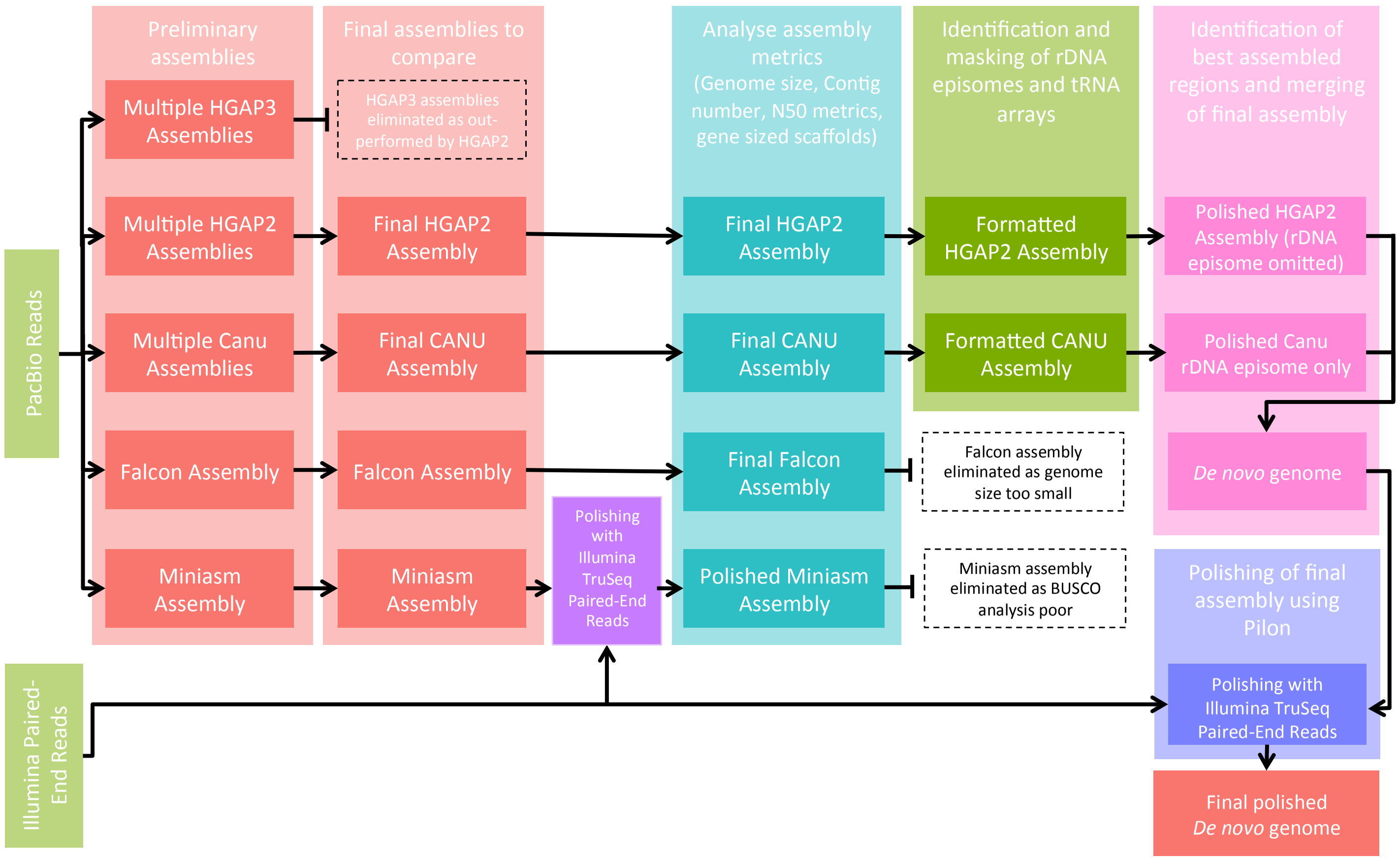

Preliminary assemblies
Multiple HGAP3 Assemblies
Multiple HGAP2 Assemblies
Multiple Canu Assemblies
Falcon Assembly
Miniasm Assembly
PacBio Reads
Illumina Paired-End Reads
Final assemblies to compare
HGAP3 assemblies eliminated as out-performed by HGAP2
Final HGAP2 Assembly
Final CANU Assembly
Falcon Assembly
Miniasm Assembly
Polishing with Illumina TruSeq Paired-End Reads
Analyse assembly metrics (Genome size, Contig number, N50 metrics, gene sized scaffolds)
Final HGAP2 Assembly
Final CANU Assembly
Final Falcon Assembly
Polished Miniasm Assembly
Identification and masking of rDNA episomes and tRNA arrays
Formatted HGAP2 Assembly
Formatted CANU Assembly
Falcon assembly eliminated as genome size too small
Miniasm assembly eliminated as BUSCO analysis poor
Identification of best assembled regions and merging of final assembly
Polished HGAP2 Assembly (rDNA episome omitted)
Polished Canu rDNA episome only
De novo genome
Polishing of final assembly using Pilon
Polishing with Illumina TruSeq Paired-End Reads
Final polished *De novo* genome

Figure 2.2.1. *De novo* assembly of the *Entamoeba histolytica* HM-1:IMSS genome assembly. PacBio reads were used as input for a range of assemblers. Multiple assemblies were produced for the HGAP3, HGAP2 and Canu assemblers. The assemblies with the best metrics were taken forward for comparison. The Miniasm assembly was polished using Pilon as the assembler does not have a polishing step. Final assemblies were analysed using a range of metrics before tRNA arrays and rDNA episomes were masked from the remaining assemblies. The best-assembled regions were merged and polished using Pilon to produce the final polished *de novo* assembly.

2.2.6.1. Pacific Biosciences SMRT portal HGAP2 and HGAP3 assemblers

The HGAP assemblers work using three steps [211]. First, preassembly is performed to generate long, highly accurate sequences. This step maps single pass reads to seed reads, which represent the longest portion of the read length distribution. After mapping, a consensus sequence is generated for each seed read. Second, assembly is performed using an overlap layout consensus (OLC). OLC generally works in three stages: initial overlaps (O) among all reads are identified; a layout (L) of all the reads and overlap information is represented as a graph; finally, the consensus (C) sequence is inferred [234]. At this stage, HGAP2 uses the Celera assembler whereas HGAP3 uses the Pacific Biosciences AssembleUnitig assembler. Finally, the assembly is polished to reduce the remaining indel and base substitution errors in the draft assembly.

Filtered PacBio sub-reads were assembled using the Pacific Biosciences SMRT portal HGAP3 assembler [211] using standard options with an expected genome size of 24 Mbp. The filtered PacBio sub-reads were also assembled using the Pacific Biosciences SMRT Portal HGAP2 assembler [211] using the following non-standard assembly options to create a range of preliminary assemblies. A range of expected genome sizes were set, ranging from 20 to 30 Mbp in 2 Mbp intervals. Assemblies were also tested using an increased minimum seed read length from 9-14 Kbp (default: auto-calculated).

2.2.6.2. Canu assembler

The Canu assembler consists of three steps: correct, trim and assemble [212,213]. At all stages, the first step constructs an indexed store of input sequences, generates a k-mer histogram and constructs an indexed store of all-vs-all overlaps. The correction stage selects the best overlaps to use for corrections, estimates corrected read lengths and generates corrected reads. The trimming stage identifies unsupported regions in the input and trims or splits reads to their longest supported range. The assembly stage makes a final pass to identify sequencing errors, constructs the best overlap graph and outputs contigs.

Filtered PacBio sub-reads were assembled using Canu [212], using default parameters with an estimated genome size of 26, 28.5 and 30 Mbp based on the general consensus on genome size produced from the HGAP2 outputs.

2.2.6.3. Miniasm assembler

Miniasm is another OLC-based assembler that uses a four-stage process: crude read selection, fine read selection, string graph generation and merging [215]. During crude read selection each read is analysed to find the longest contiguous region covered by three good mappings. This step also predicts read coverage. Fine read selection builds on the first stage and uses the coverage information to find the good regions again but with more stringent thresholds. A string graph is then generated to remove any weak overlaps and collapse short bubbles in the assembly. Finally merging of unambiguous overlaps is used to produce unitig sequences. The Miniasm assembler uses no consensus step and therefore, the final unitig sequences have a similar per-base error rate as the raw reads.

Filtered PacBio sub-reads were assembled using Miniasm [215], using default parameters. Unitigs were error-corrected as follows:

The Burrows-Wheeler Aligner (BWA) [235] was used to map the short reads generated from the *Entamoeba histolytica* 350bp insert paired end library to the Miniasm unitigs. Mapping was performed using default parameters to produce BAM alignment files. Determination of mapping statistics was carried out using the SAMTools view function [236].

The BAM alignment file and the final genome FASTA file were used as input for the error-correcting program, Pilon [233] following the pipeline outlined in the

Pilon manual. The consensus sequence generated by Pilon was used as the final Miniasm assembly.

2.2.6.4. Falcon assembler

Like HGAP2 and HGAP3, Falcon uses a hierarchical genome assembly process. However, it can identify haplotypes and cope with assemblies of organisms that are diploid [214]. An initial assembly is computed by Falcon, which error-corrects the raw reads and assembles them using a string graph of the read overlaps. The assembled contigs are further refined into a final set of contigs and haplotigs. Phasing of heterozygous SNPs is performed and reads grouped by haplotype. The phased reads are used to open up the haplotype-fused regions of the genome and generate as output a set of primary contigs and associated haplotigs.

Filtered PacBio sub-reads were assembled using Falcon [214], using default parameters and an estimated genome size of 28 Mbp.

2.2.6.5. Estimating N statistics and gene-sized scaffolds

N50 values were calculated using custom written perl scripts that calculated a range of assembly metrics (S2.1, Appendix 2). NG10-NG100 values (in intervals of 10) were calculated manually for each assembly and plotted in R using the ggplot2 package [237,238]. For example, to calculate a NG50 value, contigs were consecutively added together from longest to shortest until 50% of the estimated genome size was assembled. The smallest contig required to reconstitute 50% of the genome size was categorised as the NG50. The estimated genome size used when calculating the NG50 for all assemblies was 28 Mbp.

2.2.6.6. Estimating gene content-based assembly completeness

In order to gain an insight into how the well the different assemblers assembled the core genome, the assemblies were processed using BUSCO v3.0.1 (Benchmark Universal Single-Copy Orthologues) [224]. The program assigns a score to an assembled genome based on its BUSCO content. The Eukaryota dataset of conserved orthologues, available in the BUSCO package, was used as the reference set of core genes to be searched by the program. The content for each assembler was manually inspected and results plotted for comparison using R [237].

2.2.6.7. Creating an optimal assembly that best represents chromosomal, episomal and repetitive DNA

The Canu-produced rDNA episome sequence was merged with the modified HGAP2 genome contigs to produce a final assembly. This was necessary as the HGAP2 assembly failed to assemble the rDNA episomes and instead split the episomal sequence across many contigs. A FASTA file was created containing the finalised PacBio genome assembly. The fully assembled rDNA episome sequence was isolated from the Canu assembly output. This rDNA episome sequence was used as a query in a BLASTn search against the contigs assembled using the HGAP2 assembler. An exponent value (E-value) threshold of 0.01 was applied to the BLAST query. BLAST [239] is available from the National Center for Biotechnology Information (NCBI). Contigs containing rDNA episome sequence were removed from the HGAP2 assembly.

2.2.6.8. Error correction of the final genome assembly

The Burrows-Wheeler Alignment Tool (BWA) [235] was used to map the short reads generated from the *Entamoeba histolytica* 350bp insert paired end library to the final assembly. Mapping was performed using default parameters to produce BAM alignment files. Determination of mapping statistics was carried out using the SAMTools flagstats function [236].

The BAM alignment file and the final genome FASTA file were used as input for error correction using Pilon [233] following the pipeline outlined in the Pilon manual. The consensus sequence output from the Pilon pipeline was determined as the frozen assembly and was used for all further analyses.

2.2.7. Identification of arrays of tRNA genes and ribosomal DNA episomes in the assembly

2.2.7.1. tRNA array identification

Transfer RNAs (tRNAs) were detected using tRNAscan-SE [240], using default parameters. The main default parameters used the eukaryotic tRNA model for tRNA analysis and allowing for pseudogene checking.

2.2.7.2. rDNA episome identification

Entamoeba histolytica ribosomal RNA (rRNA) genes have previously been published [241] and are accessible from NCBI [Acc: X65163]. This (rRNA) gene sequence was used a query in a BLASTN search against the new genome assembles generated by the single molecule sequencing data. An E-value threshold of 0.01 was applied to the BLAST query.

2.2.8. Additional approaches to scaffold the PacBio genome assembly

2.2.8.1. Scaffolding of published *Entamoeba histolytica* HM-1:IMSS assembly with long PacBio reads

A crude hybrid assembly was also performed using SSPACE-Long Read, a program designed to improve inaccurate draft assemblies produced by next generation sequencing [242]. Scaffolding of the published reference assembly [98] was performed in an iterative manner using the PacBio RS II long read information as a backbone.

2.2.8.2. Alignment of existing *Entamoeba histolytica* HM1:IMSS optical map data

Previous unpublished optical mapping data were obtained from Dr Elisabet Caler (Caler, E., 2015, Pers. Comms.). This data had being generated as follows - HindIII was used to digest high molecular weight *E. histolytica* DNA using the OpGen Argus optical mapping platform. 21 linkage groups were generated from this data and strings of lengths (kbp) between digestion sites were provided for these groups ranging from 0.47-2.08 Mbp.

The final *E. histolytica* PacBio assembly was digested *in silico* using HindIII. These restriction digestion patterns were then reformatted and aligned to the original optical map using Soma, an optical map aligning software which aligns sequence contigs and/or scaffolds from *de novo* genome assembly against a restriction map [243]. All original optical map restriction maps were concatenated before HGAP contigs or scaffolds were aligned to the restriction map. An error rate of 1% was assumed for the optical map data to allow leniency in the alignment of *de novo* contigs to the restriction map, resulting in more contigs to be mapped.

2.2.8.3. Attempts to generate a BioNano optical map

An attempt was made to produce a new optical map using the BioNano optical mapping pipeline [244–246]. *E. histolytica* trophozoites were cultured axenically as described previously (2.2.1) and DNA extracted in agarose plugs using the BioNano animal cell culture DNA extraction protocol [64,230,231,247]. Agarose plugs were processed and loaded onto the BioNano Irys instrument in line with the BioNano Irys Processing and User Guide [248]. Nicking of the DNA was done using the single-strand restriction enzyme, BspQ1. Before loading of DNA, DNA quality was assessed using an OpGen Argus Q-card, produced using the standard usage guide (OpGen, Maryland, USA).

2.2.8.4. Attempts to scaffold assembled PacBio contigs using Hi-C data

A Hi-C dataset generated at the Institut Pasteur, Paris by the Gullién Lab using the Illumina HiSeq 2000 platform (Gullién, N. & Koszul, R., 2017, Pers. Comms.) was used in an attempt to scaffold the final PacBio assembly using the long-range sequence information provided by the Hi-C sequencing. The scaffolding attempt was performed externally at the Pasteur Institut (Paris, France) by Dr Romain Koszul using GRAAL [249], a Hi-C data based reassembler.

2.3. Results

2.3.1. Growth of *Entamoeba histolytica* HM1:IMSS, DNA extraction and purification of genomic DNA for whole genome sequencing

Entamoeba histolytica HM-1:IMSS can be grown in axenic culture (in the absence of bacteria). After an initial inoculum of trophozoite cells is added to a culture tube containing growth medium, growth follows a sigmoidal pattern with a long lag phase followed by logarithmic growth until a finite population is reached in the stationary phase at 3-4 days at which point cells are sub-cultured into a new culture tube containing growth medium. DNA was harvested at mid-log stage, to ensure minimal degraded DNA from dead cells was collected, which is an important consideration for SMRT sequencing.

The only time this pattern was not observed was during adaptation of the cells to LYI-S-2 media completed with a new batch of adult bovine serum. During this period of adaptation, the initial inoculating volume needed to enable survival of the subsequent culture had to be significantly increased up to 10x the usual volume and cultures showed an increased lag phase meaning the stationary phase was reached at 120-168 hours. This initial increase in inoculum volume decreases throughout subsequent cultures and showed full adaptation to the new serum within approximately 60 days and around 16 subcultures. By day 60, cultures present cells with normal morphology and are motile, with no rounding of the cells, which further supports the evidence of complete adaptation to new sera. They also generally grow to a high density within 72 hours and require an initial inoculum of 150-300 μL.

DNA extraction trials showed that the QIAGEN Gentra Puregene kit produced the lowest DNA yield per 13 mL culture with an average of 34.24 ng. This was followed by the QIAGEN DNeasy Blood and Tissue Kit method and then the CTAB:Phenol:Chloroform method with average yields of 157 ng and 282.4 ng per culture respectively. Pairwise comparisons of the purified DNA produced by the three different extraction methods (Section 2.2.2-2.2.3) were performed using a paired t-test (t-test of difference of means). A significant difference (p-value < 0.001) was seen between all pairwise comparisons with the CTAB:Phenol:Chloroform method yielding a significantly larger amount of DNA per 13 mL culture, followed by the QIAGEN DNeasy Blood and Tissue Kit and lastly, the QIAGEN Gentra Puregene Kit. No significant differences in DNA quality (as determined by 260:230 absorption ratios) were detected between the different extraction methods.

2.3.2. Analysis of sequence data generated from the *Entamoeba histolytica* HM-1:IMSS sequencing using the PacBio and Illumina platforms

Entamoeba histolytica HM-1:IMSS was re-sequenced to provide a better quality reference genome for the species, from which comparative analysis with other *Entamoeba* species and *E. histolytica* strains could be carried out. A PacBio sequencing library was produced from 10 µg of clean, high molecular weight genomic DNA and sequenced across eight SMRT cells on a Pacific Biosciences RS II. 2,613,934 filtered sub-reads were produced from the eight SMRT cells. These ranged from 35 to 63,710 bp, with an average length of 3,891 bp and an N50 length of 5,761 bp (Figure 2.3.1). A more comprehensive breakdown of the N25 – N95 distribution can be seen in Table 2.3.1.

To error-correct the *de novo* PacBio genome assemblies of *Entamoeba histolytica* HM-1:IMSS, a paired-end Illumina library was produced. 250 ng of clean genomic DNA was used to create a TruSeq library. The DNA required for Illumina sequencing is much less than was required for the PacBio sequencing due to the DNA being sheared in to smaller read lengths and then clonally amplified before sequencing. Therefore, the method of error correcting using shorter Illumina reads was more desirable than further PacBio sequencing to produce a deeper coverage and hence, a more confident sequence consensus. Further, the PacBio reads are error-prone (~10% error rate) whereas the Illumina reads are much less error prone and therefore, more suitable for error-correction. At the same time as generating the 350 bp-insert paired-end library, a subset was isolated and bisulphite treated to generate a MethylSeq library that could be used to detect any methylation within the *E. histolytica* HM-1:IMSS genome (described and analysed in Chapter 5). The Illumina HiSeq 4000 generated both of these *E. histolytica* HM-1:IMSS Illumina data sets on a single lane of an Illumina FlowCell. 3.81 x 10^8 pairs of reads were produced from the Truseq library, producing a sequencing depth of ~3600x (Table 2.3.2). After trimming, the median length of R1 reads was 147.4 bp (close to the maximum 150 bp), but it was shorter for R2 reads (~110 bp for the TruSeq and ~120 bp for the MethylSeq library).

Table 2.3.1. Sequence length distribution and nucleotide composition (% GC) of reads from PacBio and Illumina sequencing of *Entamoeba histolytica* HM-1:IMSS gDNA.

Feature	PacBio RS II 10Kb insert library	Illumina Truseq paired-end library
Total read sequences	2,613,934	774,434,170
Number of R1/R2 pairs	N/A	380,931,829
Number of R0 reads	N/A	12,570,512
Total bases (Gbp)	10.1	102.5
Min. read length (bp)	35	19
Max. read length (bp)	63,714	150 (R0/R1/R2)
Mean read length (bp)	3,864	-
Mean R0 length (bp)	N/A	139.19
Mean R1 length (bp)	N/A	147.40
Mean R2 length (bp)	N/A	117.21
N50 read length (bp)	5,761	-
N50 R0 length (bp)	N/A	150
N50 R1 length (bp)	N/A	150
N50 R2 length (bp)	N/A	139
N90 length (bp)	1,554	-
N90 R0 length (bp)	N/A	123
N90 R1 length (bp)	N/A	147
N90 R2 length (bp)	N/A	85
GC Content (%)	26.31	24.0

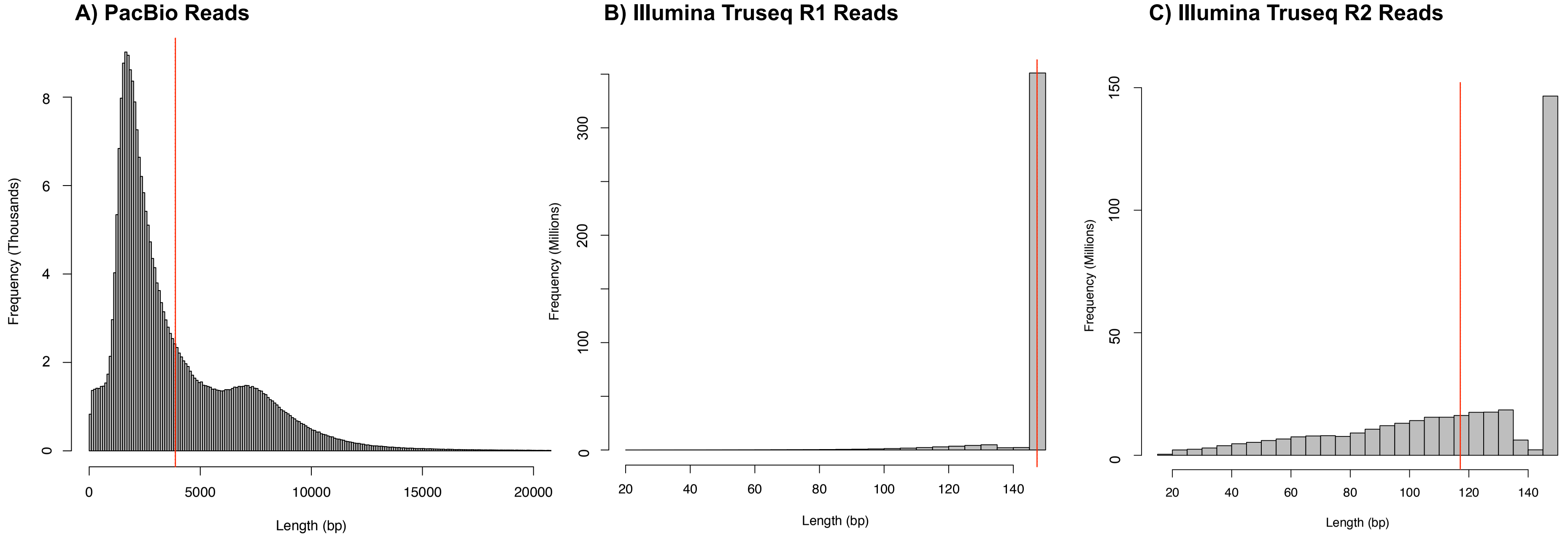


Figure 2.3.1. Read length distributions for PacBio and Illumina sequencing of *Entamoeba histolytica* HM-1:IMSS gDNA. Mean read lengths are indicated by a solid red line. **A) PacBio sub-reads.** The mean sub-read length was 3,891 bp and the N50 length was 5,761 bp. Sub-reads with lengths ranging from 20 to 63.71 Kbp were not shown in order to improve the visualisation of the majority of data that was <20 Kbp. **B) Illumina 350 bp-insert TruSeq R1 reads.** The mean read length was 147.4 bp. **C) Illumina 350 bp-insert TruSeq R2 reads.** The mean read length was 117.21 bp

Table 2.3.2. Comparison of data from PacBio and Illumina sequencing of *Entamoeba histolytica* HM-1:IMSS gDNA.

*Estimated genome coverage = Total bases/estimated genome size (28Mbp)

Sequence library type	Read Type	Number of reads	Mean read length (bp)	Equivalent genome coverage*
Pacific Bioscience RS II	Sub-reads	2.61 x 10^6	3,864	360.2x
Illumina HiSeq paired-end	R0 reads	1.26 x 10^7	139.19	62.6x
	R1 reads	3.81 x 10^8	147.40	2005.7x
	R2 reads	3.81 x 10^8	117.21	1594.9x

2.3.3. *De novo* assembly of the *Entamoeba histolytica* HM-1:IMSS genome using PacBio Single Molecule Real Time (SMRT) sequence reads

Having determined that the PacBio data contained a large number of long (multi-kilobase) reads, these were used to assemble the genome. Determining the 'best' assembler to use is not a trivial task; at the time of analysis, only a handful of genome assemblers existed which could produce *de novo* genome assemblies from sequence data produced by third-generation technologies. Of these, few had been comprehensively compared or rigorously tested across a range of taxa. These were HGAP, Canu, Falcon and Miniasm [211,213–215]. They were all run using the same set of reads and an estimated genome size (required by HGAP, Canu and Falcon) of 28 Mbp. In addition, initial tests were performed using HGAP2 and HGAP3 and a range of estimated genome sizes. Then the best of these was tested against the other assemblers.

2.3.3.1. SMRT Portal HGAP optimisation

Initial assemblies were performed using HGAP2 and HGAP3 to determine which produced the highest quality reference. Both HGAP2 and HGAP3 were run using default parameters and an estimated genome size of 24 Mb, the estimated genome size reported previously [97,98]. However, previous genome assembly attempts omitted reads matching tRNA arrays and multi-copy rDNA episomes, so were likely to under-estimate the true genome size.

HGAP3 assembled the sub-reads into 924 contigs with an N50 of 97,055 bp and a maximum contig length of 976,537 bp. HGAP2 produced an assembly of 712 contigs with an N50 of 105,410 bp and a maximum contig length of 1,014,864 bp. From these preliminary assemblies it was decided that, although HGAP3 was the faster of the two, more and longer contigs were produced by HGAP2. Therefore, HGAP2 was chosen for further analyses.

As the true genome size is probably larger than previous reported estimates (see above), the effect of different estimated genome sizes on the final assembly was tested. A range of expected genome sizes from 20 Mbp to 30 Mbp in 2 Mbp intervals were specified. The new HGAP2 genome assemblies produced range from 26.8 to 29.8 Mbp arranged in 611 to 804 contigs (Table 2.3.3.).

Table 2.3.3. Effect of predicted genome size on HGAP2 assembly

Data were assembled using HGAP2 with default assembly parameters except predicted genome size, which was varied from 20 to 30 Mbp in 2 Mbp intervals.

Predicted Genome Size (Mbp)	Genome Size (Mbp)	Number of Contigs	Max Contig (Kbp)	N50 (Kbp)
20	26.8	611	489	91.0
22	27.8	653	514	94.4
24	28.7	728	654	89.8
26	28.5	682	578	108.7
28	29.0	712	1015	105.4
30	29.0	804	1091	102.0

The assembly created specifying a predicted genome size of 28 Mbp was used in further analyses. It was chosen as it produced a genome most similar to the estimate the program was given, arranged in a reasonable number of contigs.

These contigs also had the second largest N50 and maximum contig length of all of the assemblies.

By default, HGAP2 automatically calculates the minimum seed read length to ensure 30X target genome coverage by the longest sub-reads. By lowering both the minimum seed read length and target genome coverage, it was hoped that an improved PreAssembly, onto which the remaining reads are assembled, could be generated. Conversely, increasing the minimum seed read length and target genome coverage makes PreAssembly more stringent creating a more confident final genome, at the cost of reduced genome size and assembly N50 values.

This automatically-calculated minimum seed length calculated for the final HGAP2 assembly was 11,504 bp. To test for the optimum minimum seed read length, input seed lengths from 10 to 14 Kbp were tested using a target genome coverage of 15X (Table 2.3.4).

The resulting assemblies ranged from 19.1 to 30.8 Mbp, assembled across 662 to 910 contigs; N50 ranged from 106.9 to 40.7 Kbp with the N50 read lengths of assemblies produced with a seed of 10 to 12 Kbp showing large N50 values around the 100 Kbp mark. Assemblies with a minimum seed higher than this (13 – 14 Kbp) showed smaller N50 lengths, suggesting that these assemblies did not have a sufficient number of reads over the seed length threshold to produce a contiguous, high-quality PreAssembly. This trend was also observed within the genome sizes of the assemblies. The genome size of the assemblies generated with minimum seed lengths of 13 to 14 Kbp were smaller than the assemblies generated with a minimum seed of 10 to 12 Kbp, further supporting predictions that these larger-seed assemblies did not have sufficient numbers of reads to generate a high-quality, contiguous PreAssembly and hence, the downstream processes in the HGAP2 assembly pipeline performed poorly, producing assemblies with low N50s and an underestimated genome size. Therefore, these assemblies were rejected from further analyses.

Table 2.3.4. Effect of seed size on the HGAP2 assembly Data were assembled using HGAP2 with default assembly parameters and an estimated genome size of 28 Mbp. Non-default parameters were a coverage target of 15X and a minimum seed read length from 10 to 14 Kbp, in 1 Kbp intervals. The automatically calculated data from the 28 Mbp HGAP2 assembly is shown for comparison.

Seed Input (Kbp)	Genome Size (Mbp)	Number of Contigs	Max Contig (Kbp)	N50 (Kbp)
10	30.8	910	503.512	106.9
11	28.6	715	527.9	93.8
11.5 (automatically calculated)	29	712	1014.9	105.4
12	29	707	1014.9	105.4
13	22.4	617	662.3	58.6
14	19.1	662	290	40.7

The 10 Kb minimum seed assembly produced the largest N50 but generated ~200 contigs more than the other assemblies, suggesting over-assembly of the genome may have occurred whereby erroneous assembly of multiple haplotypes had resulted in extra contigs being represented in the assembly. The increase in contig number is also represented by little increase in genome size when compared to the next largest seed input (11 Kbp). This could indicate many smaller contigs being added to the assembly. On this basis, this assembly was rejected from analysis. The remaining two assemblies (11 Kbp – 12 Kbp minimum seed) generated genomes very similar to the one with an automatically generated (11.5 Kbp) seed. The 11 Kbp minimum seed assembly was rejected on the basis that it had an N50 10 Kbp smaller than those in the automatically generated assembly and the 12 Kbp minimum seed assembly. The remaining two assemblies (automatically generated seed and 12 Kbp seed) were very similar with the 12 Kbp seed assembly generating 5 fewer contigs and a 41 Kbp smaller genome. These analyses showed that the automatically calculated seed provided a good result that was not substantially improved upon by setting the seed manually, therefore this assembly was chosen.

The final HGAP assembly, here on in referred to as the HGAP assembly, was generated by HGAP2 using default parameters with a expected genome size of 28 Mbp.

2.3.3.2. Comparison of four long read assembler outputs

Four assembler programs were chosen to assemble the PacBio reads: HGAP, Canu, Falcon and Miniasm. The trimmed sub-reads produced from the PacBio sequencing were used as the input for each. With the exception of the HGAP assembler, which was run multiple times as outlined in section 2.3.3.1, all assemblers were run using default parameters and with an estimated genome size of 28 Mbp (if this parameter was required). The HGAP2 assembly run with default parameters and an estimated genome size of 28 Mbp (section 2.3.3.1) was compared to the other three assemblers (hence setting the estimated genome size to 28 Mbp for the other assemblers, to aid comparison).

Assemblies are summarised in Table 2.3.5. They ranged from 19.7 to 29.1 Mbp, assembled in between 314 and 803 contigs. N50s varied across the four assemblers with HGAP, Canu and Falcon producing similar N50s of 105.4 Kbp, 97.2 Kbp and 138.6 Kbp, respectively and Miniasm producing a much lower N50 of 57 Kbp. The largest contig produced by the assemblers followed a similar pattern with HGAP, Canu and Falcon producing maximum contigs of 1.02 Mbp, 732.5 Kbp and 759.5 Kbp, respectively and Miniasm producing a much smaller longest contig of 288.9 Kbp.

Table 2.3.5. Comparison of gene metrics produced by different long read assemblers

Data were assembled using four assemblers (HGAP, Canu, Falcon, Miniasm), with default assembly parameters and an estimated genome size of 28 Mbp.

Feature	HGAP Assembly	Canu Assembly	Falcon Assembly	Miniasm Assembly
Contigs	712	432	314	803
Size (bp)	29,007,650	25,984,130	19,690,672	29,062,976
GC Content (%)	24.2	24	24.23	25.13
N50 Length (bp)	105,410	97,159	138,598	56,960
Mean Length (bp)	40,741	60,148	62,709	36,193
Longest Sequence (bp)	1,014,864	732,495	759,500	288,800
Shortest Sequence (bp)	1,991	8,657	125	18
Gaps	0	0	0	0

2.3.4. Assessing the quality of *de novo* whole genome assemblies produced by different assemblers

Having created a range of *de novo* assemblies, criteria for defining the best overall *Entamoeba histolytica* HM-1:IMSS assembly were required. A range of metrics used as measures of genome quality were calculated for the different assemblies.

2.3.4.1. Analysing the distribution of assembled contig sizes

The N50 and NG50 values of the different assemblies were calculated to determine which were composed of more large contigs, which would be useful downstream when analysing gene organisation and the structure of gene families (Fig 2.3.2).

N50 and NG50 values varied across and within the assemblies. N50 values ranged from 138.6 Kbp in the Falcon assembly to 57.0 Kbp in the Miniasm assembly with the HGAP and Canu assemblies in the middle of this range with N50s of 105.4 Kbp and 97.2 Kbp respectively. NG50 values ranged from 117.0 Kbp in the HGAP assembly to 59.0 Kbp in the Miniasm. The Canu assembly had an NG50 of 116.9 Kbp, similar to that produced by HGAP. Falcon performed poorly, with an NG50 value of 77.1 Kbp.

Generally, N50 and NG50 values of the same assembly did not differ dramatically. The exception was the Falcon assembly, which had a 61.5 Kbp difference between the N50 and NG50 value due to the assembler producing an assembly much smaller than the estimated genome size. A lesser extreme is the Canu assembly where there is a 19.6 Kbp difference between these two values. In both cases, N50 was greater than the NG50 value.

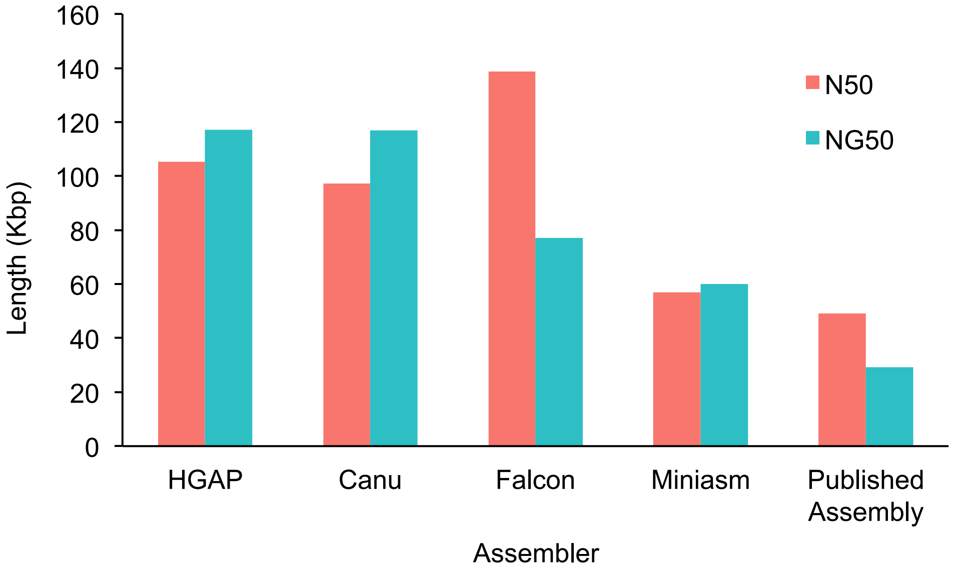


Figure 2.3.2. Comparison of N50 and NG50 across different assemblies. N50 and NG50 values were calculated for assemblies produced by different assembler programs. NG50 values were calculated using an estimated genome size of 28 Mbp.

To observe the distribution of contig lengths across the assemblies, the NG(X) length (using an estimated genome size of 28Mb) was calculated for a range of values (NG10 to NG100 in 10% intervals) in each assembly, to produce an NG graph (Figure 2.3.3). This NG graph allowed visual comparison of contig length distribution among the assemblies. The graph shows that HGAP and Miniasm assemblies were larger than the estimated genome size and that Canu and Falcon assemblies were smaller than the estimated genome size. The Falcon and Canu assemblies meet the x-axis at 70.36% and 92.86%, respectively (not shown) and this directly represents how long the assemblies were as a proportion of estimated genome size. NG(X) values for the HGAP and Canu assemblies were consistently higher than those of Falcon and Miniasm assemblies, which is indicative of HGAP and Canu assemblies being composed of longer contigs across the entire assembly.

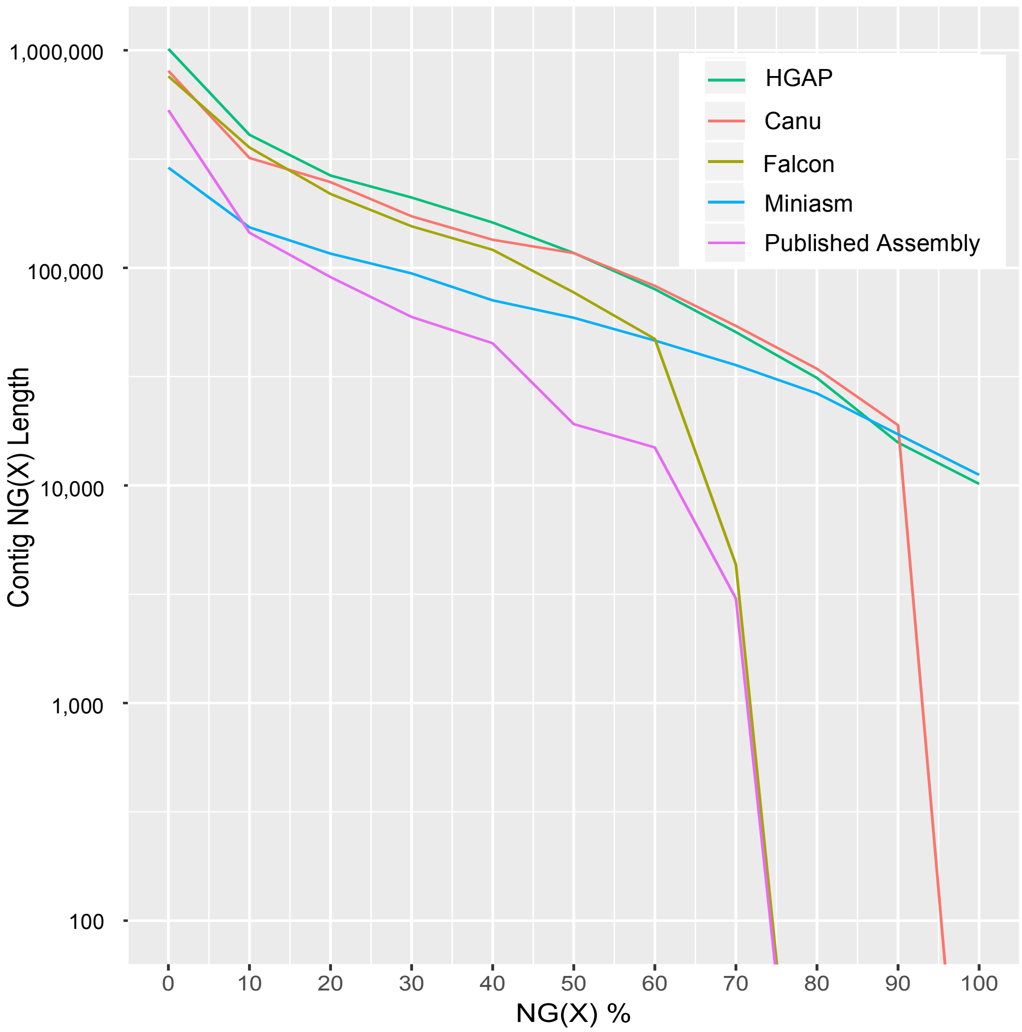


Figure 2.3.3. NG graph comparing assembly contig/scaffold length distribution. The NG contig/scaffold length, calculated in integer values of 10 (10%-100%) and the contig/scaffold length that each particular threshold is passed on the y-axis (bp). For example, to calculate the NG50 value for an assembly, all contig/scaffold lengths are cumulatively added together from longest to shortest. The NG50 value is that length at which the sum length accounts for 50% of the estimated genome size (28 Mbp). The first data point displays the longest scaffold in the assembly and where a series touches the x-axis (contig NG(X) length = 0), this is indicative of the assembly being smaller than the estimated genome size. If a series never touches the x-axis, this is indicative of an assembly being larger than the estimated genome size.

2.3.4.2. The proportion of assemblies represented by gene-sized scaffolds

To discover which assembly would be most useful in downstream applications such as gene prediction and annotation, the proportion of 'gene-sized scaffolds' in each assembly was calculated. A gene-sized scaffold was determined as a contig/scaffold equal to, or longer than, the average gene size reported for *Entamoeba histolytica* HM-1:IMSS (1,280 bp). All of the assemblies were almost entirely composed of gene-length scaffolds with 100% of contigs generated in the HGAP and Canu assemblies and 99.8% and 97.5% of contigs in the Miniasm and Falcon assemblies, respectively, meeting this criterion.

2.3.4.3. Identifying the presence of Benchmarking Universal Single-Copy Orthologues (BUSCOs)

To assess the 'completeness' of the assemblies, BUSCO [224] was applied using the publicly available Eukaryota data set within the BUSCO package (Figure 2.3.4 and Table 2.3.6). BUSCO identified 177 (58.4%) genes in both the CANU and HGAP assembly. 159 (52.5%) genes and 142 (46.9%) genes were identified in the Miniasm and Falcon assemblies, respectively. The Canu and HGAP assemblies had slightly fewer missing BUSCOs than the published assembly however, they also had slightly fewer single copy BUSCOs too. This could suggest that the Canu and HGAP genomes are more repetitive; this could be real and the published assembly has collapsed repeated regions. Alternatively, the new HGAP and Canu assemblers could be over-splitting alleles leading to false gene duplications. In addition the missing BUSCOs in both the HGAP and Canu assemblies are the same as those missing from the published assembly.

Table 2.3.6. Number of BUSCOs identified in the HGAP, Canu, Falcon and Miniasm *Entamoeba histolytica* HM-1:IMSS assemblies. The full list of Eukaryota orthologues contains 303 genes.

Assembler	BUSCOs identified	Complete Single Copy BUSCOs	Duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs
HGAP	177	112	50	15	126
Canu	177	110	53	14	126
Falcon	142	96	27	19	161
Miniasm	159	108	19	32	144
Published Assembly	171	123	30	18	132

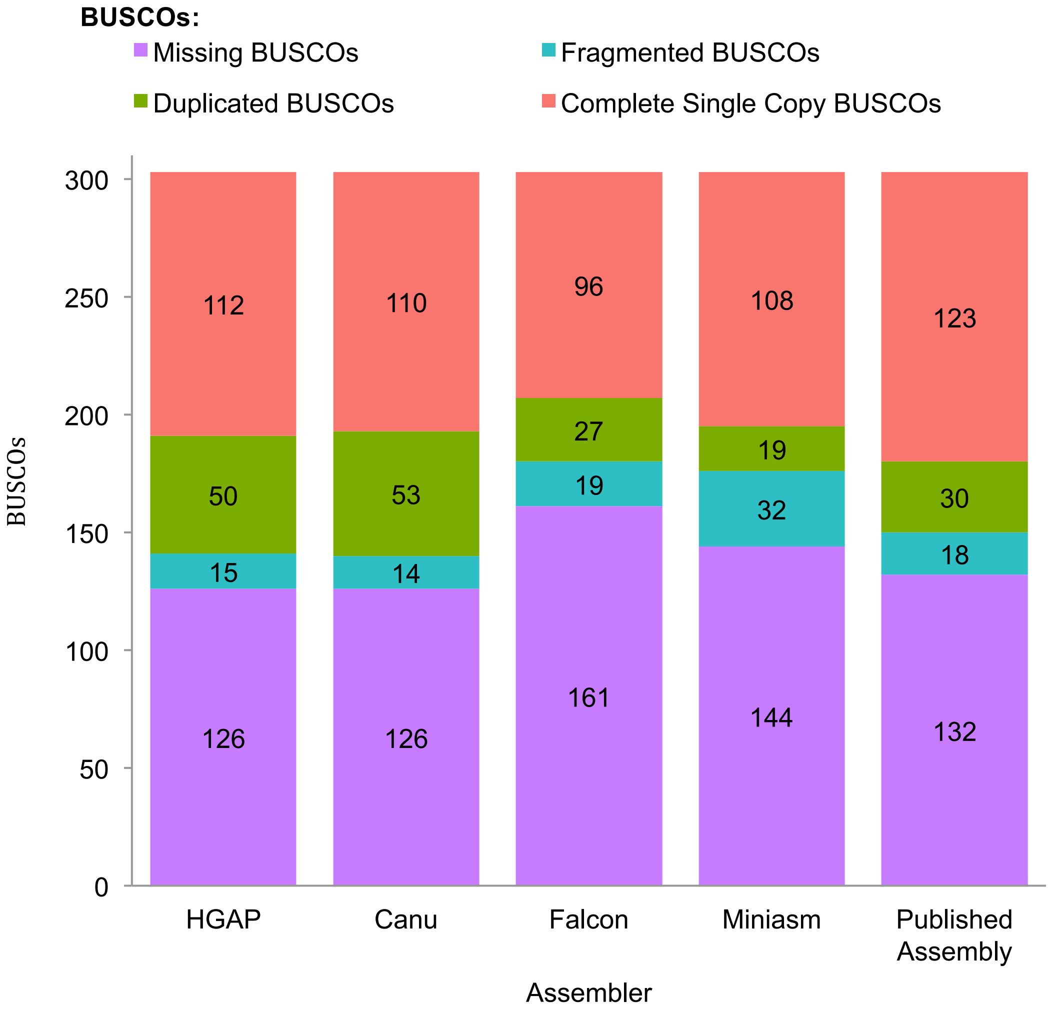


Figure 2.3.4. BUSCO scores of the assemblies. BUSCO scores were calculated for each assembly using the Eukaryota data set available within the BUSCO v3 package. Different assemblers are shown on the x-axis. The maximum number of genes available for detection is 303. Genes detected by BUSCO are those represented by the Complete Single Copy, Duplicated and Fragmented BUSCO proportions and missing genes are represented by the Missing BUSCOs section of each stack (purple).

2.3.5. Producing a final *de novo Entamoeba histolytica* HM-1:IMSS assembly

Based on the comparative analyses outlined in 2.3.4, the HGAP and Canu assemblies were chosen as the best assemblies. Further comparisons of these two assemblies were carried out to determine how to create a final genome assembly for downstream applications and analysis.

2.3.5.1. Identifying repetitive features in the genome assembly

Transfer RNAs (tRNAs) arrays were detected using tRNA-scan-SE. Subsequent manual inspection identified tRNA genes arranged in array units, which have been described previously [4]. Likewise, the rDNA sequence [241] was used in a BLASTN search to identify putative rDNA episomes. Contigs containing the rDNA sequence were manually inspected to check length and *in silico* restriction mapping confirmed that the restriction digestion pattern of these molecules were consistent with that previously reported [241]. A large proportion of the contigs in both assemblies comprised repetitive tRNA sequence or rDNA episomes. These were initially removed from both assemblies to allow for a more accurate comparison of the contigs comprising the core genome (though they were included in the final assembly).

Both assemblies contained contigs entirely composed of tRNA array units (called 'tRNA-only contigs'). They both also contained contigs with both tRNA arrays and other, non-repetitive, sequence (called 'tRNA-genic' contigs). The HGAP assembly contained 137 tRNA-only contigs and 21 tRNA-genic contigs. The Canu assembly contained only 14 tRNA-only contigs and 21 tRNA-genic contigs. The HGAP assembly also contained more contigs representing the rDNA episomes (150 contigs) than did the Canu assembly (9 contigs).

When tRNA arrays and rDNA episomes were removed from the two genomes to produce a 'core genome' assembly, both assemblies became more comparable. The HGAP 'core' assembly consisted of 25.5 Mbp across 425 contigs and the Canu 'core' assembly consisted of 25.5 Mbp across 409 contigs. A more thorough comparison of the original and 'core' assemblies is shown in Table 2.3.7.

Table 2.3.7. Assembly statistics of the total and 'core' HGAP and Canu assemblies.

*-Refined assemblies have had tRNA and rDNA regions removed.

Feature	HGAP Assembly	Refined HGAP Assembly*	Canu Assembly	Refined Canu Assembly*
Contigs	712	425	432	409
Size (bp)	29,007,650	25,643,432	25,984,130	25,513,136
GC Content (%)	24.2	24	24	24
N50 Length (bp)	105,410	136,819	97,159	98,095
Mean Length (bp)	40,741	60,337	60,148	62,379
Longest Sequence (bp)	1,014,864	1,014,864	732,495	732,495
Shortest Sequence (bp)	1,991	1,991	8,657	8,657
Gaps	0	0	0	0

2.3.5.2. Merging of the HGAP2 and Canu assemblies

Based on the outcomes of section 2.3.5.1, it was determined that the best approach in creating a final assembly would be to merge the Canu and HGAP assemblies together. The HGAP assembly managed to assemble more of the tRNA array units whereas many reads containing tRNA array sequence were unassembled in the Canu assembly. However, the Canu assembly managed to assemble a contiguous rDNA episome sequence whereas this sequence in the HGAP assembly was fragmented.

The nine putative rDNA episome contigs in the Canu assembly were comprised of near identical sequence repeated tandemly in a range of different lengths. This was thought to result from the episomal DNA being circular, so each contig was split into individual complete and partial rDNA episome sequences. These sequences were then aligned and a consensus determined. This consensus rDNA episome sequence was added to the 'core' HGAP assembly (with putative rDNA episome contigs of the HGAP assembly removed). The merged assembly is summarised in Table 2.3.8.

Table 2.3.8. Assembly statistics of the *Entamoeba histolytica* HM-1:IMSS genome assembly produced by HGAP, Canu and the merged HGAP/Canu assembly. Comparison of the merged HGAP/Canu assembly to the original HGAP and Canu assemblies.

Feature	HGAP Assembly	Canu Assembly	Merged HGAP and Canu Assembly
Contigs	712	432	563
Size (bp)	29,007,650	25,984,130	27,452,180
GC Content (%)	24.2	24.0	24.1
N50 Length (bp)	105,410	97,159	117,635
Mean Length (bp)	40,741	60,148	48,762
Longest Sequence (bp)	1,014,864	732,495	1,014,895
Shortest Sequence (bp)	1,991	8,657	1,991
Gaps	0	0	0

The merged assembly contains 27.4 Mbp of sequence across 563 contigs. It is almost entirely comprised of the HGAP assembly however, is 1.6 Mbp smaller and has 149 fewer contigs. The removed sequence represents the misassembled rDNA episomal sequences. Consequently, it has an N50 that is 12 Kbp larger than the original HGAP assembly. The only non-HGAP produced sequence is the rDNA episomal sequence, which was entirely assembled by the Canu assembler; contigs comprised entirely of tRNA arrays are those produced by the HGAP assembly.

2.3.5.3. Error correction of the final assembly

The merged assembly was processed using Pilon [233] to correct any uncorrected errors in the PacBio assembly. Reads produced from the *Entamoeba histolytica* HM-1:IMSS Illumina TruSeq library were used to correct the more error-prone PacBio data.

Pilon made 2,618 changes to the merged assembly: 146 deletions, 908 insertions and 1,564 single base substitutions, which are described in Table 2.3.10. Single base substitutions made up the majority of the changes introduced by Pilon with 1,526 incidences. Differences to the assembly are summarised in Table 2.3.9. It is important that any errors, especially indels, are corrected prior to annotation (Chapter 3) as they can produce false frame shifts in coding sequence and cause changes to the predicted protein sequence during gene annotation.

Table 2.3.9. Pilon-induced changes to assembly metrics of the merged assembly. The merged HGAP/Canu assembly was processed using Pilon. Gene metrics were altered and differences between the assemblies were summarised using custom perl scripts.

Feature	Merged Assembly	Error Corrected Merged Assembly	Difference
Contigs	563	563	0
Size (bp)	27,453,180	27,437,923	-15,257
GC Content (%)	24.1	24.1	0
N50 Length (bp)	117,635	117,638	+3
Mean Length (bp)	48,762	48,735	-28
Longest Sequence (bp)	1,014,864	1,014,864	0
Shortest Sequence (bp)	1,991	1,991	0
Gaps	0	0	0

Table 2.3.10. Summary of changes introduced by Pilon to the merged assembly. The merged HGAP/Canu assembly was processed using Pilon. 350 bp paired-end Illumina reads for *Entamoeba histolytica* HM-1:IMSS were used as input.

* $N_{(>1)}$ represents a string of nucleotides (A/T/G/C) greater than 1 bp. A dot indicates that no sequence is present.

Correction type	Original Sequence	Corrected Sequence	Frequency
Insertion	.	A	380
		C	62
		G	61
		T	357
		$N_{(>1)}$	48
Deletion	A	.	20
	C		12
	G		6
	T		19
	$N_{(>1)}$		89
Substitution	A	C	95
		G	168
		T	235
	C	A	79
		G	31
		T	150
	G	A	198
		C	22
		T	62
	T	A	221
		C	173
		G	92
	A/C/G/T	$N_{(>1)}$	0
	$N_{(>1)}$	A/C/G/T	0
	$N_{(>1)}$	$N_{(>1)}$	38
		Total	2,618

Pilon removed ~15 Kbp of sequence from the merged assembly. Though some of this was comprised of single nucleotide deletions and some small deletions (up to 3 nucleotide deletions), the majority of deletions to the assembly occurred across 28 of the contigs comprised entirely of tRNA arrays. The structure and variation of the tRNA arrays in *E. histolytica* are explored in Chapter 4 however briefly, the tRNA genes occur in mixed sets, each gene separated by DNA that contains short tandem repeats (STRs); the entire set forms a repeat unit that is tandemly duplicated in many copies [4]. The deletions represented situations where an entire tRNA array unit was removed from a tRNA array.

2.3.6. Attempts to scaffold the *Entamoeba histolytica* HM-1:IMSS genome assembly

Though using third generation sequencing technologies has largely improved the *Entamoeba histolytica* reference genome, the assembly was still not a complete, whole chromosome-scale assembly. Therefore, further approaches were used in an attempt to scaffold the assembly into larger scaffolds approaching whole chromosomes.

2.3.6.1. Scaffolding of the current published reference assembly using SSPACE

An attempt to utilise the PacBio assembly and the existing reference assembly [98] was performed. SSPACE-Long Read was used in an attempt to scaffold the published assembly scaffolds together using the HGAP assembly from the PacBio sequencing as pseudo-reads. Any gaps that were introduced in this process were attempted to be filled using PB Jelly [250].

The resulting assembly was compared to the previous published assembly and the *de novo* PacBio Assembly in Table 2.3.11.

Table 2.3.11. Hybrid approach to assemble the *Entamoeba histolytica* HM-1:IMSS genome. Comparison to the original *E. histolytica* reference genome and the new assembly produced using only third-generation sequence data.

Feature	SSPACE assembly (Sanger/PacBio data: 2016)	Published assembly [98] (Sanger data: 2010)	Merged HGAP and Canu Assembly (PacBio data: 2016)
Contigs	1014	1496	563
Size (bp)	24,579,421	20,799,072	27,452,180
GC Content (%)	21.1	24.2	24.1
N50 Length (bp)	180,785	49,118	117,635
Mean Length (bp)	24,240	13,903	48,762
Longest Sequence (bp)	643,157	530,629	1,014,895
Shortest Sequence (bp)	235	235	1,991
Gaps	1,153	643	0

The hybrid assembly approach produced an assembly of 24.6 Mbp arranged in 1,014 scaffolds. 1,153 gaps were introduced into the assembly amounting to 3.26 Mbp of Ns (13.3% of the assembly). The assembly had an N50 of 180.8 Kbp, an improvement of 131.7 Kbp over the original Lorenzi assembly and increased the largest contig from 530.6 Kbp to 643.2 Kbp.

2.3.6.2. Scaffolding the assembly using optical map data

The final HGAP assembly was mapped to an existing optical map using Soma (version 2.0) [243]. The existing optical map contained 21 linkage groups ranging from 0.3 Mbp to 2.08 Mbp totalling a small amount of the genome size. The map was generated with the restriction enzyme, HindIII.

271 contigs from the HGAP assembly contained 2 HindIII restriction sites and 34 contigs contained >2 HindIII restriction sites. These contigs were deemed suitable for mapping. Of these, 57 contigs were placed onto the optical map with a total length of 2.61 Mbp combined. These contigs covered 10% of the optical map and were generally not distributed in clusters. There were 3 incidences where two contigs were placed within 10 Kbp of each other and no incidences where three or more contigs were aligned next to one another with gaps of <10 Kbp between neighbouring contigs.

Owing to the difficulties encountered using the existing optical map, it was decided a new optical map would be generated using the newer BioNano technology and using two restriction enzymes to improve placing accuracy of contigs. *In silico* digestion of the final PacBio assembly with the enzymes available on the BioNano platform indicated that only one enzyme, BspQ1, produced a nicking density within the recommended range. DNA extracted from agarose plugs did not pass the QC length threshold of 100-200 Kbp. Most molecules of DNA extracted from the agarose plugs measured around 30-50 Kbp. Also, the quantity threshold of 4-8 µg of sample material per plug was not met. Five *E. histolytica* plugs were processed together producing a 1 µg of DNA for processing. Despite this, DNA was loaded on to the instrument for optical map generation however no molecules passed through the channels of the BioNano chip and no data were generated.

2.3.6.3. Scaffolding of the assembly using Hi-C data

A Hi-C dataset generated at the Institut Pasteur, Paris by the Gullién Lab using the Illumina HiSeq 2000 platform (Gullién, N. & Koszul, R., 2017, Pers. Comms.) and was used in an attempt to scaffold the final PacBio assembly using the long-range sequence information provided by the Hi-C sequencing. Scaffolding and analyses were performed at the Pasteur Institut (Paris, France) by Dr Romain Koszul using software developed internally. The Hi-C data produced very little signal and not enough 3D contacts between the PacBio contigs were made to allow for successful scaffolding (Koszul, K., 2017, Pers. Comms.).

2.4 Discussion

2.4.1. The HGAP2 and Canu assemblers outperformed other assemblers and merging the outputs of each produced an assembly that balanced the assembly of non-repetitive and repetitive and extra-chromosomal sequences

HGAP2 and Canu produced assemblies that consistently performed well compared to other assemblers across a variety of quality indicators.

The Falcon assembly was discarded from further analysis due to its small assembly size. The Falcon assembler produced a genome approximately two-thirds the size of the genomes produced by other assemblers. This may indicate approximately one third of the genome is difficult to assemble and is consistent with previous estimates which predict that upwards of 20% of the *Entamoeba histolytica* HM-1:IMSS genome is repetitive [97,98]. This is supported further by our own analysis that approximately 6.2% of the genome assembly is comprised of repetitive tRNA arrays and 23.61% of the genome is identified as transposons (Chapter 4).

The assembly produced by Miniasm was discarded due to low N50 and NG50 values. When compared to HGAP and Canu assemblies, these were consistently >40 Kbp smaller in length and therefore, would be less useful downstream when analysing gene structure and organisation of the genome. The assembly produced by Miniasm also contained fewer BUSCOs than the HGAP and Canu assemblies suggesting that it is less complete despite its similar genome size. This may be due to the assembler not containing a consensus step and relying solely on one round of read/sequence correction leading to fewer errors being corrected. If these errors, especially indels, occur in a region where a BUSCO lies, it is possible that the error may cause the gene not to be detected. Another possibility is that Miniasm has over-assembled regions of the genome leading to a similar genome size to those produced by HGAP and Canu. It is also possible that owing to the nature of the Miniasm assembler (i.e. no consensus correction step), assembly is less stringent and two reads representing the same region of the genome may be assembled into two separate contigs if indels or other errors are present in one, or both, of the reads.

Ultimately, the HGAP and Canu assemblers both produced good quality assemblies, with large N50 and NG50, a large proportion of BUSCOs and a reasonable total genome size.

The ~3 Mbp genome size difference between the HGAP and Canu assemblies was due to different handling of repetitive tRNA and rDNA episomes by the assemblers. If these were removed, both performed similarly (section 2.3.5.1). This suggests that both assemblers produce fairly consistent reconstructions of the core, non-repetitive portion of the genome and give fairly consistent estimates of the size of this portion of the genome (i.e. excluding extra-chromosomal contigs and tRNA array structures), that broadly match the previous estimates [97,98] of approximately 24 Mbp. Further supporting this consistency is the BUSCO analysis (Section 2.3.4.2) in which both assemblies contained the same 177 BUSCOs.

Both HGAP and Canu assemblies had different strengths. HGAP was better able to assemble contigs that terminated with large strings of repetitive sequence such as the tRNA arrays whereas Canu was better able to assemble circular molecules into a single sequence such as the rDNA episome. For this reason, the episomal rDNA sequence from the CANU assembly was used in place of those from the HGAP assembly. The final assembly is almost entirely composed of the HGAP assembly, with the misassembled rDNA episomes replaced by a fully assembled rDNA episome from the Canu assembly.

The assembly produced is a major improvement upon the previous genome assembly. However, the assembly does not yet have telomere-to-telomere contiguity. Subsequent improvements to the Pacific Biosciences sequencing chemistry have been released since *E. histolytica* was sequenced for this thesis and therefore, it is possible that further PacBio sequencing could produce a 20

Kbp insert library (double the size of the 10 Kbp insert library generated for this PacBio sequencing of *E. histolytica)*. Although, this would still remain a challenge for *E. histolytica* owing to the difficulties involved in extraction of high molecular weight genomic DNA from this organism. Improvements to assembly algorithms and the emergence of new software may also ultimately improve the output of the raw PacBio data. However, at the time of sequencing few programs were available to assemble long sequencing reads and therefore, only a limited number of assemblers could be tested. Comparing the output of these long-read assemblers can also be challenging. Metrics useful for comparing assemblies produced by short read technologies might not be directly transferable into analysis of long-read assemblies. The most evident of these is the comparison of gene-sized scaffolds across the assemblies. The reason for this being the majority of reads produced by third–generation technologies can be longer than the average gene length meaning that the majority of all contigs, regardless of assembler, are also longer than this value making it hard to draw comparisons between different assemblers. This is observed in our dataset where 90.1% (2,355,043/2,613,934 reads) of the reads produced by the PacBio sequencing are longer than average gene-size and as a result, the different assemblies show little difference in their proportions of gene-sized scaffolds. It is possible though, that this metric may still be useful for third-generation sequencing and assembly of large vertebrates and also invertebrates, such as mammals and insects, whose genomes are large and complex. However, when assembling single cell eukaryote genomes such as that of *E. histolytica*, it becomes more relevant to compare the N50 and NG50 values of resulting assemblies, especially if subsequent analyses regarding the structure and organisation of genes and gene families are likely to be performed (Chapter 3). This is because spatial organisation of genes to one another (or to certain structural features) relies on them being placed on the same contig or scaffold. For example, genes in close proximity to the telomeres can only be assessed if an assembly contains telomeric sequence and gene sequences within the same contiguous sequence.

2.4.2. Hybrid assembly approaches produced poorer assemblies than those produced with PacBio data alone

Scaffolding of the previous reference assembly of *Entamoeba histolytica* HM-1:IMSS with the assembly produced by the PacBio sequencing did produce an improved genome when compared to the published assembly; The scaffold number was reduced by approximately a third and increases in N50 length and genome size were observed. This is perhaps unsurprising as the long read information that third-generation reads provide may span repetitive regions that are longer than reads produced by Illumina sequencing. As the *Entamoeba* genomes are highly repetitive, many repeats are likely to be resolved and therefore, there is a large improvement to the original assembly when scaffolded with the PacBio assembly.

More surprising is the observation that a hybrid approach utilising both Sanger and PacBio data did not outperform assemblies produced using PacBio data alone. The hybrid assembly contained nearly double the number of contigs of the final PacBio assembly and also was missing ~3 Mbp of sequence that the non-hybrid assembly contained; this is consistent with other reports that non-hybrid assemblies outperform hybrid assemblies at higher coverages above 50X [216,251,252]. It was concluded that this could be due to errors in the original *E. histolytica* HM-1:IMSS assembly. Large regions of existing assembly had been scaffolded together and therefore, it is likely that some incorrect joins were made (evidence of this is presented in Chapter 3). If these joins were close to the end of a published assembly scaffold it is not unlikely that it would not be scaffolded by the PacBio data. This is because the order of the joined contigs in a published assembly scaffold may not match the corresponding region in the contiguous and more accurate PacBio contigs and therefore, scaffolding is prohibited as SSPACE is not aware that these two sequences are in reality, the same region of the genome.

2.4.3. Error-correction of the assembly made few corrections to the third-generation sequence data.

Despite PacBio sequencing being regarded as an error-prone and many approaches recommending polishing of a PacBio genome with more accurate short NGS reads, polishing of the *Entamoeba histolytica* HM-1:IMSS assembly resulted in changes to only a small proportion of the genome. Assuming all the errors corrected by polishing were in fact real errors, 17,643 bp of original sequence (0.06%) was corrected to a different base pair, or removed, indicating 99.94% of the assembly produced by PacBio data alone was accurate before polishing. This most likely owed to the assembly having deep coverage (~200x) meaning the alignment of reads already produced an accurate consensus. The changes introduced by polishing were largely made up of deletions of singular tRNA array units from long tRNA arrays and it is possible that these represent real differences between the PacBio library (used for assembly) and the Illumina paired end library (used for genome polishing). The PacBio dataset and NGS Illumina dataset were produced at different time points almost 12-18 months apart. Both libraries were produced from a number of pooled organisms and therefore, the variation in the PacBio and NGS sequence data may be real and not indicative of original errors in the PacBio sequencing.

Single base pair insertions to the assembly are indicative of the original PacBio read containing a single base pair deletion. The observation of these is consistent with other findings that claim insertions and deletions are the most common errors found in PacBio reads [178,253] and can most likely be explained by incorporation errors in two ways. Firstly, incorporation events, or the interval between them, can be too short to be reliably detected resulting in no base being called. Secondly, errors can be introduced by unlabelled nucleotide contamination (dark nucleotides) whereby a nucleotide is introduced without emitting a detectable signal and therefore, the output sequence does not include this base. Deletion rate in PacBio reads has been reported as up to 7.8% in raw PacBio reads [178] however, the final *E. histolytica* HM-1:IMSS assembly only contained a deletion error rate of 0.003%, supporting the idea that corrected reads produced for the assembly had eliminated the majority of deletion and insertion errors without the need for polishing with NGS reads.

In conclusion, it was determined that PacBio sequencing alone is able to produce a highly accurate assembly if coverage is deep and circular consensus sequencing is utilised. Overall, the PacBio-only assembly produced a high rate of accuracy and it is debatable whether polishing using a NGS dataset is required especially as the majority of changes to the PacBio genome assembly after polishing were regarding the structure of these tRNA repeats and did not largely affect the core genome, including gene coding regions. This is in keeping with the recent observation that Illumina-polished long-read assemblies have a reduced number of structural variants compared to non-polished long-read assemblies [254].

2.4.4. Acquisition of high quality gDNA is a limiting factor in further improving the *Entamoeba histolytica* HM-1:IMSS genome

Despite the PacBio data producing a longer, more contiguous assembly, the *Entamoeba histolytica* HM-1:IMSS genome was still not assembled into a telomere-to-telomere, whole-chromosome assembly. Further techniques such as optical mapping and Hi-C were applied, though largely these were limited by the ability to collect large amounts of high molecular weight DNA from *E. histolytica* HM-1:IMSS trophozoites.

To generate both an optical map and Hi-C data, a large amount of concentrated high molecular weight DNA is required. This is very difficult to achieve with *Entamoeba* DNA for many reasons. Firstly, *Entamoeba* trophozoites are exceedingly large for the amount of DNA they contain and a single 15 mL culture containing ~ 100,000 cells will only yield ~100 ng of DNA (~1 pg per cell). As a result, the DNA needs extensive processing to concentrate the DNA and as a result, DNA can become fragmented. This was the biggest problem in creating the optical map; Pooling of many plugs meant the final volume needed to be reduced using a SpeedVac and this agitation of the DNA may have contributed to the DNA degradation. It is hypothesized that this factor is also the reason for poor Hi-C outcome. The Hi-C library was generated during early emergence of the technology and little was known about the amounts of DNA that would be needed to create high quality Hi-C libraries (Marbouty, M., 2017, Pers. Comms.). As such, the Hi-C library was created from approximately 200,000 – 300,000 cells. It is now known that the library ideally would need to be generated from 10 million cells (Koszul, R., 2017, Pers. Comms.) and the reason the Hi-C data yielded low results was likely to have been directly due to the difficulty in generating enough *Entamoeba* gDNA.

It is also predicted that the high levels of lytic enzymes in *E. histolytica* trophozoites are released when cells are lysed during DNA extraction. These enzymes may have degraded long fragments of DNA resulting in a reduced fragment length in both the optical mapping and Hi-C attempts.

Further contributing to these difficulties, *Entamoeba* cells are carbohydrate rich, necessitating additional DNA clean up, processing that can cause further degradation of genomic DNA.

2.5. Conclusions

This chapter presents a new reference genome for *Entamoeba histolytica* HM-1:IMSS and comparisons of the different assembly tools and approaches for assembling long-read sequencing data. No single assembler was able to produce a final assembly that managed to both assemble extra-chromosomal molecules and assemble long stretches of repetitive sequences. The Canu assembler was the only assembler that managed to fully assemble the well-characterised circular rDNA episome of *E. histolytica.* However, it struggled to assemble long repetitive structures within the genome such as the tRNA arrays. On the other hand, HGAP2 assembled long stretches of these tRNA array units but was unable to assemble complete episomal molecules. It was decided that complementing the HGAP2 assembly with the rDNA episome assembled by Canu would produce the most balanced, representative assembly that would be most useful for downstream analysis of genome structure.

Assemblies produced by all the programs tested (HGAP2, Canu, Miniasm and Falcon) vary dramatically in genome size and N50 lengths, as well an in the number of core genes (BUSCOs) identifiable. This emphasizes the need to explore assembler options when assembling long read data as not all assemblers perform equally well for a given genome. It is not possible to say whether the best performing assemblers in this analysis (HGAP2 and Canu) are the best performing in general but it can confidently be concluded that at least for highly repetitive, small, single-cell eukaryote genomes, it would appear that these are the best contenders of those tested. As third-generation sequencing becomes more affordable, and with the development of other long-read sequencing methods such as Nanopore sequencing, more tools and assemblers are becoming available (that were not available at the time of this analysis). It would be useful to reassemble the data using newer versions of HGAP2 and Canu alongside new programs such as MECAT [255]. MECAT has been reported as performing very well when assembling *Plasmodium falciparum* long read data despite the highly biased (AT-rich) nucleotide composition, as also seen in *E. histolytica* [256].

Comparisons of the assemblies also highlighted the lack of BUSCOs present in both the existing published reference and the new PacBio assembly of the *E. histolytica* genome. Though the analysis highlights how identifying and comparing the number BUSCOs present between different assembler outputs can be a useful metric, its raises the question as to whether using BUSCOs is a good indicator of predicting genome completeness in highly divergent eukaryotes. Only 58.4% (177 genes) of the core gene set for Eukaryota were represented in the HGAP2 and Canu assemblies which means the remaining 31.6% (or 126 genes) are either absent from the assembly or do not exist in the *E. histolytica* genome. It is hard to determine which is correct, however it is clear from a range of other protist genomes that the pattern is not unique for *Entamoeba histolytica*. Other intestinal protists genomes such as those for *Giardia intestinalis* and *Cryptosporidium parvum* are missing 57.4% (174 BUSCOs) and 34.7% (105 BUSCOs) of the core set of conserved Eukaryota genes, respectively (Data not shown). It should be noted here that the *Entamoeba* genomes are very fluid and it is not rare to find large regions of the genome that have been duplicated [98] and therefore, duplicated BUSCOs may not always be representative of over-assembly or erroneous assembly of haplotypes.

Further to this, the Eukaryote BUSCO gene set is also troublesome as the point at which the set was defined was after the point of divergence of the *Entamoeba* lineage and hence the genomes of many basal organisms, including *Entamoeba* and other parasites such as the *Giardias* and *Trypanosomas*, have not been included when determining orthologous gene sets which have been conserved. As such, BUSCO scores cannot be used for scoring genome completeness and instead, can only be used as a metric to rapidly compare assemblies to conclude which contains the most representative gene set.

The BUSCO results for the *E. histolytica* genomes also highlight high levels of BUSCOs that are apparently duplicated; it is largely regarded that BUSCOs are single copy genes. Analysis of other amoeba genomes (Data not shown), including the high quality genome for the model organism, *Dictyostelium discoideum*, shows similar levels of BUSCO duplication. The *D. discoideum* genome contains 43 duplicated BUSCOs (14.2% of BUSCOs) and the *Acanthamoeba castellanii* genome (of much poorer assembly quality then *D. discoideum*) contains 65 duplicated BUSCOs (21.5% of BUSCOs); the 50 duplicated BUSCOs in the HGAP assembly and 53 duplicated BUSCOs in the Canu assembly are in line with the levels of duplication in these other amoeba species. Both the observation of low levels of BUSCOs across other intestinal protists and the higher levels of BUSCO duplication amongst other amoeba species support the idea that the BUSCOs identified in the new PacBio assembly are a realistic representation of the *E. histolytica* genome and that the missing BUSCOs may in fact not be present in the genome.

Finally, polishing of the final assembly with high-coverage, accurate Illumina short reads made some corrections to the sequence, though these largely affected repetitive, non-protein-coding regions of the genome and raises the questions of whether true structural variants are been masked by this process, as has been suggested elsewhere [254] and explained in section 2.4.3. It would appear that the high PacBio coverage (~200x) produces a high level of accuracy in the final assembly, largely obviating the need for polishing with short-read data.

Overall, the genome produced is a valuable resource and forms the basis for the work described in Chapters 3, 4 and 5.

Chapter 3: Genome Annotation of the new *Entamoeba histolytica* HM-1:IMSS genome assembly and analysis of virulence gene families in their genomic context

3.1. Introduction

Genome structure and organisation is very important in understanding how genes (and gene families) have evolved and how they are regulated. As mentioned in Chapter 1, the structure of gene families and their regulation can confer an advantage to the survival of the organism; mono-allelic expression of genes families is used by many organisms to regulate the variation of surface proteins and is often facilitated by gene families being situated in the sub-telomeric regions of a genome.

The Apicomplexan parasite *Plasmodium falciparum* utilises mono-allelic expression in a highly effective immune evasion mechanism by differentially expressing different surface proteins known as *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) [185,186]. PfEMP1s are encoded by the (approximately) 60-member *var* gene family and are associated with virulence due to their role in immune evasion and intravascular parasite sequestration [87,187–189]. These *var* genes are categorised into three types determined by which upstream sequence (upsA, upsB or upsC) they are associated with. UpsA and upsB-type *var* genes are located in sub-telomeric regions while upsC-type *var* genes are located in chromosomal clusters [87]. In all types of *var* genes the close proximity of the genes to one another is thought to facilitate antigen switching, which is mediated epigenetically [191].

Another antigen switching mechanism is observed in the Kinetoplast parasite *Trypanosoma brucei,* in which Variant Surface Glycoproteins (*VSGs*) are differentially regulated to mediate immune evasion. They provide a protective cell surface coat to the parasite throughout the mammalian infectious cycle [192–194]. The underlying mechanism of successful immune evasion is clone-specific, singular *VSG* expression combined with regular switching of one *VSG* to another to continuously expose the host immune system to novel surface antigens [257]. The expressed *VSG* genes (*VSG* expression sites) are found to be adjacent to a telomere [258,259]. It is thought that ~80% of the *VSG* genes reside on the telomeres of *Trypanosoma brucei's* multiple mini chromosomes which appear to be almost exclusively dedicated to containing these *VSG* gene 'reservoirs' [259–261]. The remaining *VSG* genes are thought to be located adjacent to telomeres of the intermediate and megabase chromosomes that compromise the *T. brucei* genome [262]. The *VSG* genes are flanked by 70-bp repeats upstream and highly conserved elements within the 3'-untranslated Region (3'-UTR) [88]. These sequences facilitate recombination of unexpressed *VSG* genes in the *VSG* expression sites resulting in a constant turnover of different surface proteins on the cell surface [97,98].

3.1.1. Problems with the current *Entamoeba histolytica* gene set

In the first published *Entamoeba histolytica* HM-1:IMSS genome assembly 9,938 genes were predicted within the genome [97]. Subsequent re-assembly and re-annotation reduced this to 8,333 genes [98]. The majority (55%) of annotated genes encode proteins of unknown function. This is not uncommon in protists. Two other gut parasites, *Cryptosporidium parvum* and *Giardia lamblia,* both have large portions of their gene content annotated as 'hypothetical protein' (40% of 4,367 genes and 75% of 9,747 genes, respectively). This large number of uncharacterised genes presents a problem for genome-wide analyses (e.g. of gene expression) because the majority of genes of interest are of unknown function.

In addition to this, very little is known about the organisation of genes and the expansions of gene families within the *E. histolytica* HM-1:IMSS genome. The repetitive nature of the genome has meant that despite continuous sequencing efforts, the assembly remains fragmented and a chromosome level assembly is yet to be reached. The current published reference assembly contains 1,496 scaffolds. Single molecule sequencing and assembly (described in Chapter 2) has improved the *E. histolytica* HM-1:IMSS assembly.

E. histolytica contains a range of virulence gene families that have been previously identified however, no information on gene family organisation has been performed as often members of the same gene family occur in multiple scaffolds (AmoebaDB, release 35 data). In addition, the lack of information regarding the wide-scale structure of *E. histolytica* HM-1:IMSS, including the structure of the telomeres, has made it impossible to understand whether telomeres are enriched for particular gene families or gene function as is observed in *P. falciparum* and *T. brucei.* Further, little is known about how gene families have become expanded in the *E. histolytica* genome. Expanded gene families are often associated with important functions and complex processes within organisms. *Entamoeba histolytica* contains a large number of multi-gene families [98]. Specifically, this chapter will focus on a number of those gene families that have been associated with virulence in *E. histolytica* HM-1:IMSS and were described in Chapter 1. Specifically the organisation of the AIG-1, Ariel-1, BspA, cysteine protease (CP), Gal/GalNAc lectin and STIRP families has been investigated.

3.1.2. Annotation of eukaryotic pathogen genomes

Reliably identifying coding regions within *de novo* assembled genomes can be difficult. Gene finding is usually facilitated through known orthologues, properties indicative of a genic sequence, or a combination of the two [263]. If the organism has already been sequenced or has a sequenced close relative, it is possible to transfer annotation by aligning protein sequences from that genome to the new genome assembly.

However, when a genome is re-sequenced to a higher quality it is likely that novel genes will also be present in the new assembly, especially if the new assembly is larger than the previous one. In these cases, *ab initio* gene finding is also required. Algorithms such as Genscan [264] detect genomic regions likely to contain coding sequences based on detection of both signals and content properties. Such signals can include transcription/translation start and termination sites and donor/acceptor splice sites [265]. Compositional properties that can indicate genic regions include those shared by exons, introns and intergenic regions [263]. Many programs, such as Ensembl [266] and AUGUSTUS [226], are able to synthesise orthologue alignments to inform *ab initio* predictions, further improving their accuracy and reliability.

Despite these available tools, annotation of non-model organisms remains difficult. Until recently, very few tools were dedicated to the annotation of eukaryotic parasites and few pipelines were written for eukaryote annotation [267]. Companion (COMprehensive Parasite ANnotatION) has been recently developed solely for the purpose of annotating eukaryotic pathogen genomes using a reference-based approach [267]. Companion delivers a usable annotation of features in the target genome, formatted for submission to public databases. It also contains several extra features for identifying differences between the reference and the new assembly, such as orthologous clusters, species specific singleton genes and missing core genes present in larger reference species sets [267].

The Companion workflow uses a combination of homology-based and *ab initio* methods to produce a set of protein-coding genes. Transferring of highly conserved gene models with no modification to the reference set is done by RATT [268]. Further gene models are predicted by *ab initio* gene prediction programs SNAP [269] and AUGUSTUS [226]. This process utilises extrinsic evidence such as ESTs and RNA-seq data. At the end of the structural annotation, a final set of gene models is determined by merging the outputs of the gene finding software. Functional annotation of the genes is transferred from annotations associated with orthologous genes in the reference gene set that have function defined by OrthoMCL [270]. If no orthologues are found for a query gene, the best Pfam-A [117] hit is used to predict function. If GO-terms are available, these are also transferred to the annotation. Non-coding RNAs are identified *ab initio* by ARAGORN (tRNAs) [271] and INFERNAL [272] (rRNA and other ncRNAs from the Rfam database) [273].

3.1.3. Aims of Chapter

The published reference assembly for *E. histolytica* was highly fragmented and as a result the structure and organisation of genes and gene families in the *E. histolytica* genome remains unknown. Similarly, the genes that are in close proximity to features such as the telomeres and transposable elements could not be identified using the published assembly. This is important as the enrichment of genes in the subtelomeric regions in other protists has been associated with virulence and parasite-host interactions. Further, owing to the fragmented nature of the assembly means that it is likely the genome is incomplete and genes were missing when annotating the assembly produced using short Sanger reads. This chapter specifically aims to:

- Produce a high quality annotation of the PacBio *Entamoeba histolytica* HM-1:IMSS assembly using Companion.
- Compare gene content between the PacBio assembly and the existing reference assembly to identify missing and novel genes
- Analyse the predicted functions of novel genes to identify new members of existing genes families, as well as single-copy novel genes.
- Investigate the organisation of genes within a range of virulence gene families.
- Investigate gene content and predicted function of all genes within close proximity to structural features such as transposable elements and putative telomeres

3.2 Materials and Methods

3.2.1. Comparison of the new PacBio assembly and published assemblies

Details of how the new PacBio assembly was generated are described in Chapter 2. For comparison, the published *Entamoeba histolytica* genome assembly (in FASTA format) and annotation (in GFF3 format) were obtained from AmoebaDB (Release 35, released November 2017). Genome metrics were determined using custom perl scripts (described in Chapter 2.2.6). Gene annotation of the PacBio genome is described in section 3.2.2.

As an illustrative example, the longest contig from the published assembly was aligned to the PacBio assembly using BLASTn, to produce a 'crunch' format file that was visualised using the Artemis Comparison Tool (ACT; release 16.0.0) [274], with GFF3 tracks displayed. This was inspected manually for assembly and gene content differences.

3.2.2. Companion annotation

The PacBio assembly described in Chapter 2 was annotated using the Companion web platform [267]. The current genome annotation for *Entamoeba histolytica* HM-1:IMSS on AmoebaDB (Release 35) was used as a reference from which to transfer the annotation to the new assembly via the Rapid Annotation Transfer Tool (RATT) [268] step within the Companion pipeline. Aside from this specification, all other parameters were left as default.

The largest contig in the new *Entamoeba histolytica* HM-1:IMSS genome was manually inspected in Artemis [275] with predicted genes displayed to manually inspect any obvious defects in the gene models predicted by Companion.

3.2.3. Identification of missing and novel genes in the new *Entamoeba histolytica* HM-1:IMSS genome assembly

The companion GFF3 output was parsed to extract successfully transferred gene IDs from the AmoebaDB annotation (v35). This list was compared to the full list of *E. histolytica* HM-1:IMSS gene IDs from AmoebaDB to identify non-transferred genes (i.e. missing from the new assembly).

To ensure none of the non-transferred genes has been incorrectly annotated as novel genes, novel genes predicted by the Companion pipeline were extracted from the Companion GFF3 file. These were used in a BLASTN search against the non-transferred genes. Any *de novo* annotated genes that matched a non-transferred AmoebaDB gene were identified and a true version of non-transferred (missing) genes in the new assembly was produced.

3.2.4. Determining the genomic distribution of large gene families in the *Entamoeba histolytica* HM-1:IMSS genome

Gene sequences for a set of virulence gene families in *Entamoeba histolytica* HM-1:IMSS (AIG-1, Ariel, Amoebapore, BspA, Cysteine protease, Gal/GalNAc lectin and STIRP gene families) were obtained from AmoebaDB (v35) and split into family-specific FASTA files. For each gene family, the AmoebaDB sequences were used in a BLASTN search against the new *E. histolytica* HM-1:IMSS genome. An e-value of 0.05 and a query cut off length was used. Any identified regions were cross-referenced against the Companion file to ensure that all of the genes had been detected and annotated by Companion. The Companion file was also parsed for any orthologous genes to the AmoebaDB IDs and any *de novo* annotated genes that had been annotated as being part of the same family. Contig co-ordinates were extracted for the genes within each gene family and manually inspected to determine the distribution of each gene family.

3.2.5. Identification of genes distributed within putative sub-telomeric regions or close to transposable elements

In Chapter 4, evidence was presented to support the theory that tRNA arrays form the telomeres in *Entamoeba* species. In this chapter, tRNA:genic contigs were also identified (i.e. contigs that contains a tRNA array that terminates the chromosome at one end and genic content at the other). The gene information for the 100 Kbp flanking region of the tRNA arrays on these contigs was extracted from the Companion GFF3 output and manually inspected for any gene families that contained two or more members in the putative sub-telomeres. GO terms for the genes were extracted from the Companion output and summarised and visualised using REVIGO [276].

Retro-transposons and other transposable elements in the new *E. histolytica* HM-1:IMSS genome were identified using RepeatMasker (Version 4.0.7) [277] specifying crossmatch for the search engine flag. Annotated *Entamoeba* elements already exist in the RepeatMasker database and therefore, the species flag was used specifying *Entamoeba* as the target species.

BED files were created specifying coordinates 1 Kbp upstream and downstream of detected transposable elements (TEs). These sequences were extracted from the new *E. histolytica* HM-1:IMSS genome assembly using the BEDtools getfasta tool [278]. Sequences were used in a BLASTN search against *E. histolytica* HM-1:IMSS coding sequences (CDS) available on AmoebaDB (v35). Manual inspection of the gene functions associated in these regions was used to identify any gene family enrichment in these areas. The BED files specifying the co-ordinates 1 Kbp upstream and downstream of putative TEs were used as input for Homer [279] to extract any annotation of the regions from the Companion GFF3 output. The genes identified were cross-referenced with those identified through the BLASTN search to compile a list of genes associated with (within 1 Kbp of) TEs.

To assess clustering of the gene families identified based on contig number or TE type they were associated with, phylogenetic analysis was performed on gene families which were enriched in the regions surrounding the TEs (AIG-1, Cysteine Protease, Ariel-1 gene families). For each gene family, sequences were aligned in the SeaView GUI (Version 4.6) [280] using the MUSCLE algorithm. BMGE (Block Mapping and Gathering with Entropy; Version 1.12) [281] was used to remove ambiguously aligned regions from the alignment, effectively trimming the alignment into conserved blocks of sequence. To determine the best model of protein evolution of the alignment, the trimmed alignment was then used as the input for ProtTest3 (Version 3.2) [282] using default parameters and using the all-distributions flag to display how the best model selection is affected under different scenarios. Once the most appropriate model was identified, PhyML (Version 20120412) [283] was used to determine a gene phylogeny of members of the gene family. The parameters used by PhyML are displayed in Appendix 3 (Table S3.1.).

The output form PhyML was visualised in MEGA7 (Release #7180411-i386) [284]. Phylogenetic trees were manually inspected for quality and for any clades specific to contig number or to a single type of transposable element. Trees were also inspected for any distinct separation of gene family members that occur in association with a transposable element with those that are not associated with transposable elements.

3.3. Results

3.3.1. The new *Entamoeba histolytica* reference genome improves on the existing reference assembly

The *Entamoeba histolytica* genome produced using PacBio sequencing in Chapter 2 is a major improvement on the existing reference genome assembly. The PacBio assembly contains 933 fewer contigs than the published assembly and has an N50 over double its size (Table 3.3.1). The PacBio assembly contains an extra 6.7 Mbp of sequence compared to the published assembly, although approximately 1.4 Mbp of this sequence is comprised of tRNA sequences and rDNA episomal sequence (removed from previous published assemblies). Removal of these regions from the PacBio assembly leaves a 'core' genome of approximately ~26 Mbp, ~5 Mbp larger than the published assembly.

Table 3.3.1. Comparison of the published and the PacBio-produced *Entamoeba histolytica* genome assemblies.

^{A} Core genome size = Total genome excluding telomeric and episomal sequences

Feature	Published Assembly (Lorenzi, 2010)	PacBio Assembly (2018)
Contigs	1,496	563
Size (bp)	20,799,072	27,407,923
Core genome size$^{A\ (bp)}$	20,799,072	25,984,130
GC content (%)	24.2	24.1
N50 length (bp)	49,118 (scaffolds)	117,638 (contigs)
Mean length (bp)	13,903 (scaffolds)	48.682 (contigs)
Coverage	12.5x	200x
Longest sequence (bp)	530,629 (scaffold)	1,014,895 (contig)
Shortest sequence (bp)	235	1,991

To identify some specific differences between the new and the published assemblies, the largest scaffold of the published assembly was aligned to the new assembly and manually inspected to identify differences. The published assembly consists of scaffolds created from many contigs, joined by ‘N’ bases. The new PacBio assembly contains no such gaps. In addition, it resolved regions where the published assembly scaffold had incorrectly joined two smaller contigs together (Figure 3.3.1).

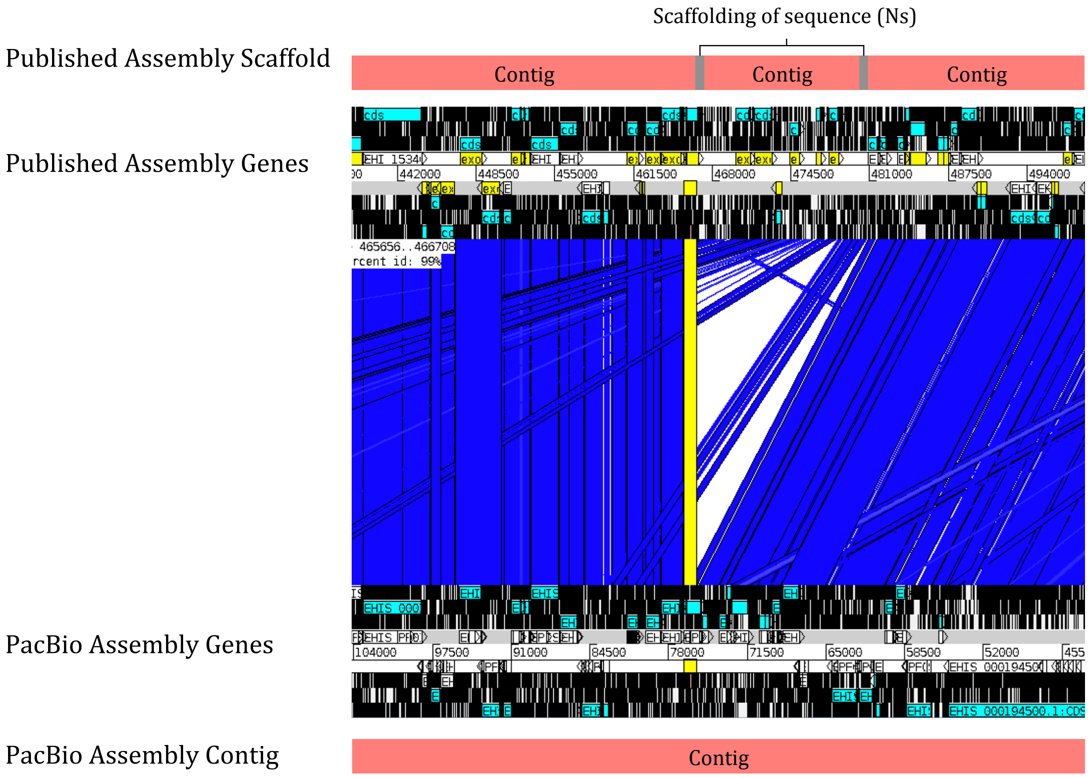


Figure 3.3.1. Assembly differences between the published assembly and the PacBio assembly. The longest scaffold in the published assembly consists of several contigs scaffolded together using Ns**.** The corresponding region in the PacBio sequence is a single contig and highlights a scaffolding error in the published assembly.

The PacBio assembly contains a single contig (contiguous sequence) corresponding to the longest scaffold in the published assembly. The PacBio contig does not contain one of the contigs in the published scaffold. The genes on this contig are found elsewhere in the PacBio assembly, suggesting that the published scaffold contained a misassembly.

3.3.2. Companion annotated more genes than previously identified in the reference *Entamoeba histolytica* HM-1:IMSS genome

The Companion software predicted 10,164 protein-coding genes in the new *Entamoeba histolytica* HM-1:IMSS genome assembly (Table 3.3.2). Of these 1,889 genes were predicted *de novo* by AUGUSTUS and the remainder transferred from the previous annotation. In addition to these genes, 373 pseudogenes, 3,946 tRNA genes, 230 rRNA genes and 6 snRNA genes were identified; rRNA and tRNA genes were not extensively annotated in the published assembly as reads containing ribosomal or tRNA sequences were removed before assembly [97,98].

Table 3.3.2. Genome statistics and gene comparison. Comparison of previous *Entamoeba histolytica* HM-1:IMSS genome assemblies compared to the new PacBio assembly.

Genome	**Loftus *et al* (2005)** [97]	**Lorenzi *et al* (2010)** [98]	**PacBio (2017)**
Size	23,361,983	20,799,072	27,407,923
Number of Genes	9,985	8,333	10,164
Mean Gene Length (bp)	1,170.7	1,260.9	1,216.3
Gene Density (Genes/10 Kbp)	4.3	3.9	3.94
Longest Gene (bp)	15,210	15,210	15,210
Shortest Gene (bp)	96	147	123
Percent Coding (%)	50	49.7	47.5

An increase in gene number was observed in the new assembly compared to the reference assembly (A.K.A. Lorenzi assembly). 1,831 more genes were identified in the new PacBio assembly than the currently used reference assembly however other metrics remain relatively consistent. The average gene length is ~1,200 bp with an average gene density of four genes per 10 Kbp of sequence. The longest gene remains consistent at 15,210 bp and the shortest gene now becomes 123 bp in length based on a 100 bp cut-off. The percentage of the genome that encodes for genes has decreased slightly compared to previous estimates of ~50% of the genome.

3.3.3. A small subset of genes were not transferred to the new *Entamoeba histolytica* HM-1:IMSS genome assembly

As part of the Companion pipeline, reference gene sequences are transferred to the new genome. Gene sequences for *Entamoeba histolytica* HM-1:IMSS, available from AmoebaDB (v35) were transferred to the new genome assembly. 58 genes were not transferred to the PacBio assembly from the published reference assembly. 38 of the 58 missing genes are annotated as hypothetical genes with no InterProScan annotation so it is hard to infer their function or if they are real genes. The remaining 20 missing genes with functional annotation are outlined in Table 3.3.3.

Table 3.3.3. Missing AmoebaDB genes in the new *Entamoeba histolytica* HM-1:IMSS single molecule assembly. 58 genes were identified as missing from the new assembly that were annotated in previous assembly attempts. 20 of these genes have an assigned function.

AmoebaDB Gene ID	Product Description
EHI_046900	4-alpha-glucanotransferase
EHI_075660	CAAX prenyl protease
EHI_075700	Casein kinase II regulatory subunit family protein
EHI_077260	DNA repair helicase
EHI_076880	DNA replication licensing factor
EHI_073780	DNAJ homolog subfamily A member 2
EHI_077230	Geranylgeranyl transferase beta subunit
EHI_053090	Leucine rich repeat and phosphatase domain containing protein
EHI_130360	Modulator of drug activity B homolog
EHI_077000	Pre-mRNA cleavage factor | 25 kDa subunit
EHI_077220	Pre-mRNA cleavage factor | 25 kDa subunit
EHI_053130	Protein kinase 2
EHI_026690	Protein kinase domain containing protein
EHI_115050	Protein kinase
EHI_075640	Protein phosphatase domain-containing protein
EHI_053150	Rab family GTPase
EHI_185010	Ribosomal protein L10
EHI_053170	RNA-binding protein
EHI_077240	Transcription initiation factor TFIID family protein
EHI_167840	WH2 motif domain containing protein

Non-transferred genes occur on 27 scaffolds of the published assembly. Of these, 19 are small scaffolds (1.0 to 2.7 Kbp) where the non-transferred gene is the only annotated gene on the scaffold. Three scaffolds in the published assembly contain the majority of the non-transferred genes; DS571229, DS571238 and DS571394 contain 10, 9 and 9 of the missing genes, respectively.

The non-transferred genes appear as three clusters across the three published scaffolds and therefore, to check whether regions of these scaffolds are missing in the new assembly the published and new assemblies were aligned using Mummer and visualised for areas of no alignment. The sequence corresponding to these three clusters did not align to any sequence in the new PacBio genome (Figure 3.3.2A). To further validate the absence of these regions from the new PacBio assembly, short-read paired end sequencing data (described in Chapter 2) was mapped to the published assembly with duplicate reads removed. The genes clusters were inspected for mapping; reads were found to map uniquely to all three non-transferred clusters from the old assembly with an average depth of ~300x (Figure 3.3.2B). However, regions of missing genes often contained drops in coverage indicative of where contigs had been scaffolded together on either side of the non-transferred gene clusters.

To ensure that the missing genes were not present in the PacBio sequencing data but not assembled into the genome, a tBLASTn search was performed. The missing gene sequences were used in a tBLASTn query against the corrected PacBio reads. None of the 58 non-transferred genes were detected in the raw PacBio reads indicating these gene sequences were either present in the genome but not sequenced or the that these genes genuinely do not exist in the *E. histolytica* HM-1:IMSS strain that is kept at University of Liverpool.

It is worth noting here that the Illumina paired end sequencing data was generated from *E. histolytica* HM-1:IMSS cell stocks maintained at the London School of Hygiene and Tropical Medicine (LSHTM) and the discrepancies in the presence of the missing genes between the PacBio reads and the Illumina reads may indicate real biological differences between the two cell stocks.

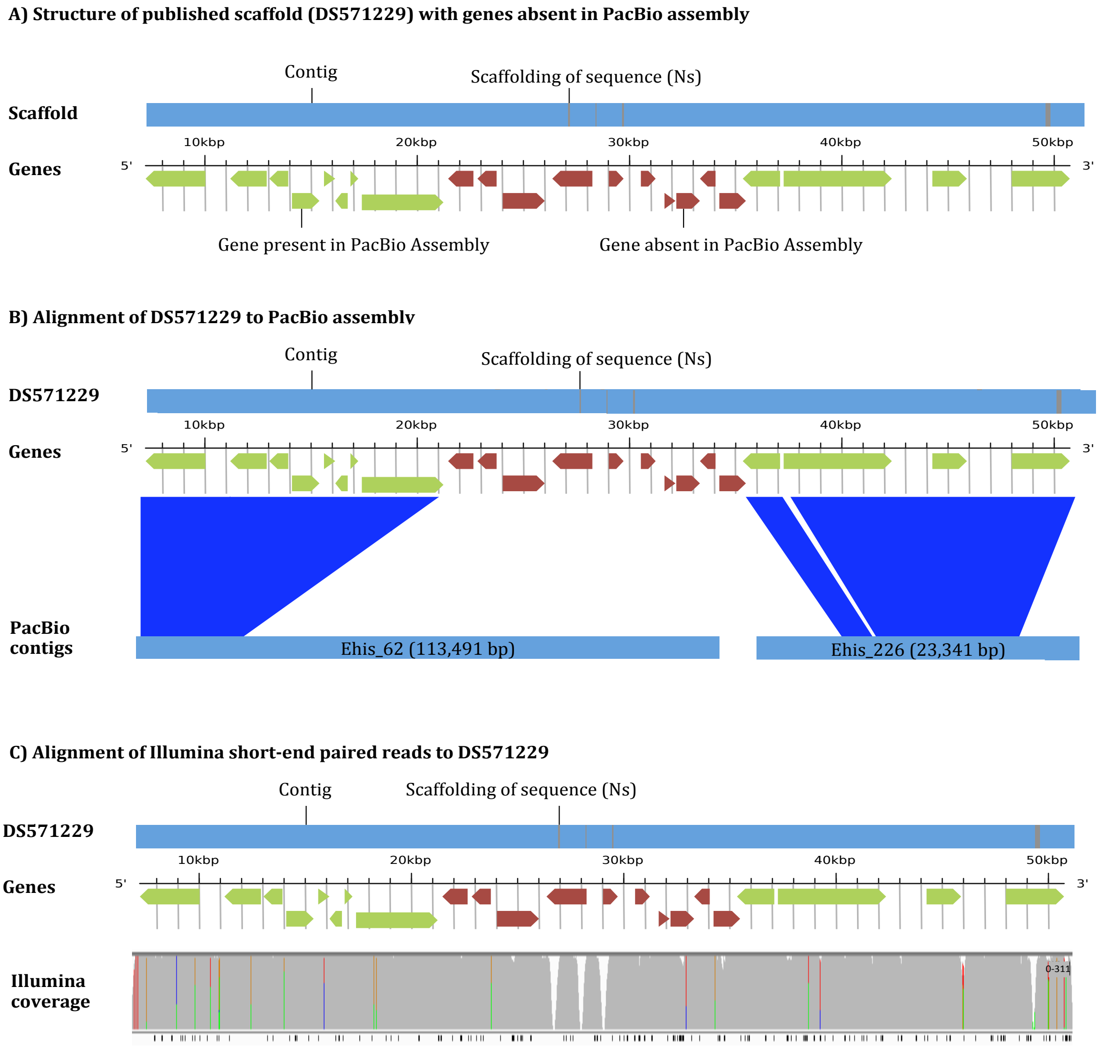

A) Structure of published scaffold (DS571229) with genes absent in PacBio assembly
Contig
Scaffolding of sequence (Ns)
Scaffold
10kbp
20kbp
30kbp
40kbp
50kbp
Genes
5'
3'
Gene present in PacBio Assembly
Gene absent in PacBio Assembly
B) Alignment of DS571229 to PacBio assembly
Contig
Scaffolding of sequence (Ns)
DS571229
Genes
PacBio contigs
Ehis_62 (113,491 bp)
Ehis_226 (23,341 bp)
C) Alignment of Illumina short-end paired reads to DS571229
Contig
Scaffolding of sequence (Ns)
DS571229
Genes
Illumina coverage
0-311

Figure 3.3.2. Example of assessment of missing gene regions in the new PacBio *Entamoeba histolytica* HM-1:IMSS genome. Genes that were present in the current *E. histolytica* HM-1:IMSS reference genome but were missing from the new PacBio reference genome were identified. 10/59 missing genes were present in a region from a single scaffold in the reference assembly (DS571229). **A. Structure of DS571229.** The scaffold is displayed schematically with grey regions representing strings of 'Ns' where contigs had been joined together to create the scaffold. Genes present and absent in the PacBio assembly are shown in green and red, **B. Alignment of DS571229 to the PacBio assembly.** The region of genes in DS571229 that are absent in the PacBio assembly do not align to any contigs in the PacBio assembly. The remaining regions align to two separate contigs in the PacBio assembly. **C. Alignment of short-read paired end Illumina data to DS571229.** High-depth short-read data, generated from the same strain of *E. histolytica*, was mapped to the published assembly. Mapping was observed across the missing gene regions.

3.3.4. A set of genes was identified in the PacBio assembly that were absent from the published assembly

Companion identified 1,889 *de novo* predicted genes in the PacBio genome assembly. The GFF3 file produced by Companion was parsed to identify which novel genes had an orthologue available in AmoebaDB (v35). Putative functions were assigned to genes during annotation with Companion. Genes generally could be separated into 2 groups, those with an orthologue in the published reference assembly and those without. Those gene sets could be split again based on whether they had a putative function assigned to them or not during the annotation step (Table 3.3.4).

Table 3.3.4. Classification of novel gene predicted in the *E. histolytica* HM-1:IMSS PacBio genome. Novel genes predicted by Companion were grouped into four categories as displayed in the matrix. AmoebaDB version: v35.

	Number of genes with putative function assigned by Companion	**Number of genes with no functional information assigned by Companion**
Novel genes with AmoebaDB orthologue	937	399
Novel genes without AmoebaDB orthologue	147	406

1,336/1,889 (70.7%) of novel genes were orthologous to at least one gene in the published assembly. 937 novel genes with an orthologue in the published assembly were assigned a putative function during the annotation process. These genes were grouped by function and functions represented by five or more genes are shown in Table 3.3.5. 399 novel genes with an orthologue present in the published assembly were not assigned any functional information during the annotation step and these genes remain hypothetical. Of these, 358/399 were orthologous to a single copy gene in the published assembly and 41/399 were orthologous to two of more genes in the published assembly.

Many virulence gene families (BspA, DEAD/DEAH box helicases, heat shock proteins) and groups of genes with other functions (Protein kinase, Protein kinase domain containing protein, Rab family GTPase) are represented by the gene functions of the novel genes. 168 novel genes were annotated as being orthologous to Trichohyalin, which is annotated as a single-copy gene in published assembly. These genes all form one gene family and are located across 123 contigs, all of which contain functionally annotated *E. histolytica* genes, with a maximum number of four trichohyalin genes located on a singular contig; copies do not appear to be tandemly arrayed. The trichohyalin protein sequences identified in the *E. histolytica* PacBio genome were used in a tBLASTn search against the published reference to check whether they had been missed during the published annotation attempt. 107/168 PacBio trichohyalin sequences matched a region of the published assembly (% identity > 95%, % length > 95%). Analysis of these regions in the published assembly annotation GFF file revealed no annotation of these regions. 61 PacBio trichohyalin genes did not produce a hit against the published assembly.

Table 3.3.5. Functional analysis of novel genes with an available AmoebaDB orthologue in the new *Entamoeba histolytica* reference genome. Novel genes were extracted from the Companion annotation *novel* genes with an available AmoebaDB orthologue were extracted and grouped by predicted protein function.

Gene Function	**Genes in published assembly** (values from AmoebaDB v.35)	**Number of novel genes annotated with function**
Trichohyalin	1	168
Leucine rich repeat protein, BspA family	109	17
Domain of unknown function containing protein	4	16
Protein kinase domain containing protein	95	12
WD domain containing protein	50	11
DEAD/DEAH box helicase	24	8
Protein kinase	233	8
RNA recognition motif domain containing protein	36	7
Heat shock protein 101	6	6
Rab family GTPase	82	5
Nucleosome assembly protein	10	5
Ubiquitin carboxyl-terminal hydrolase domain containing protein	23	5

553/1,889 novel genes were not orthologous to any genes in the published assembly. To ensure none of these genes had been previously identified, gene sequences for *E. histolytica* HM-1:IMSS were extracted from NCBI. The 553 novel non-orthologous genes were used in a Blast search against the NCBI gene set. No hits were identified for the novel predicted genes and therefore, they were assumed to be truly novel. The function of these genes was predicted by

Companion. 147 novel, non-orthologous genes had functional annotation predicted and these genes were grouped by function. Table 3.3.6 displays functions represented by five or more genes. 406 novel, non-orthologous genes had no functional information assigned to them during annotation and were defined as hypothetical.

Table 3.3.6. Functional analysis of novel genes with no AmoebaDB orthologue in the new *Entamoeba histolytica* reference genome. Novel genes were extracted from the Companion annotated. *De novo* genes with no available AmoebaDB orthologue were extracted and grouped by predicted protein function (as produced by Companion).

Gene Function	Number of novel genes annotated with function
Reverse transcriptase (RNA-dependent DNA polymerase)	57
AIG1 family/50S ribosome-binding GTPase	12
Leucine Rich Repeat (LRR) containing protein	10
Domain of unknown function containing protein	9
RNA recognition motif containing protein	5
StAR-related lipid-transfer (START) domain containing protein	5

A large family of retro-transposon reverse transcriptases was identified in the new genome that were probably masked in the published assembly (as transposable element sequences were masked before gene annotation). There are a number of novel genes predicted as being part of the AIG1 family despite these members not having an orthologue to an existing AIG1 gene in the published assembly. In addition, two domain-containing sets of genes were identified in the novel gene set. Five novel genes formed were annotated as StAR-related lipid-transfer (START) domain containing proteins and nine genes were annotated as possessing a 'domain of unknown function'. Five proteins were also annotated as RNA recognition motif domain containing proteins. Five START-domain containing proteins and 36 RNA-recognition motif-containing proteins already exist in the AmoebaDB reference set of genes.

3.3.5. Virulence gene families are largely not organised within close proximity to one another

A number of *Entamoeba histolytica* HM-1:IMSS virulence gene families have been previously identified during the original sequencing project [98,131]. The most highly studied of these virulence families include the surface antigens (Ariel-1, BspA and STIRPs), GTPases (AIGs) and other proteins that interact with the host cell (Cysteine proteases degrade host cell extracellular matrix, Gal/GalNAc lectins are involved in host cell adhesion and Amoebapores are associated with lysis of host cells). Almost all virulence gene families that have previously been identified in *E. histolytica* HM-1:IMSS appear more expanded in the new PacBio genome assembly (Table 3.3.7). The cysteine proteases (CPs), AIG and BspA gene families are the most expanded out of the seven virulence gene families being analysed. The number of CP genes has doubled in size from 43 members to 87 members. The AIG1 gene family also more than doubled in size in the new assembly and the BspA family increased by 20%.

Table 3.3.7. Distribution of virulence gene families in the PacBio *Entamoeba histolytica* HM-1:IMSS genome.

Gene family members were identified in the output of the Companion annotation of the new *E. histolytica* HM-1:IMSS genome. Identified genes were analysed to identify the genomic distribution of the gene family members.

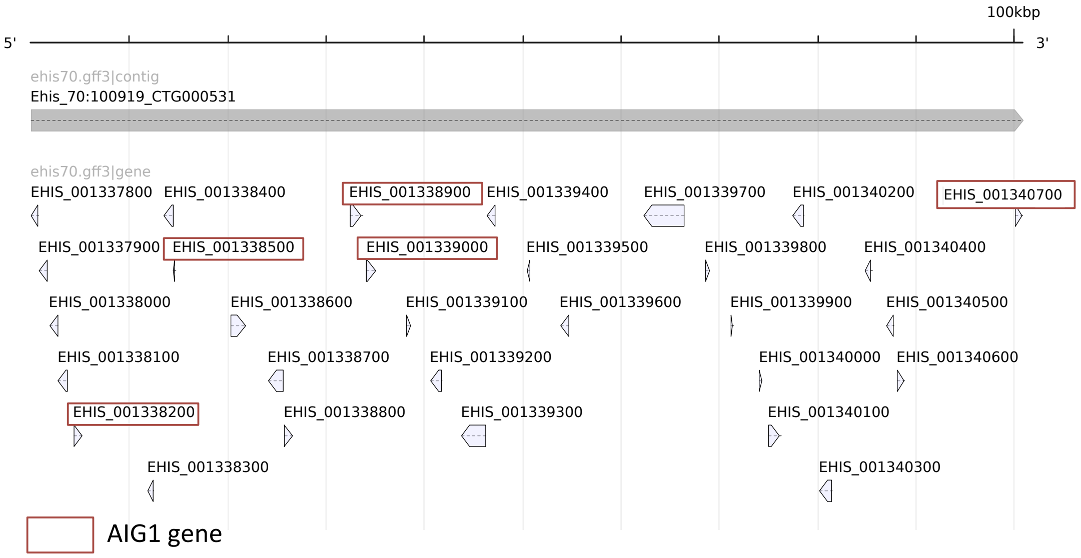
Protein Family	Members in published assembly	Members identified in PacBio assembly	Number of contigs family distributed across	Number of contigs with >2 members	Max number of members on one contig
AIGs	32	73	43	17	5
Amoebapores	4	5	3	0	1
Ariel-1	13	12	10	1	3
BspA	109	130	101	19	7
Cysteine Proteases	43	87	65	19	3
Gal/GalNAc	14	18	15	2	3
STIRPs	4	4	4	0	1

All of the virulence gene families under analysis showed very little clustering of gene families within one contig or a particular region of the genome. The smallest gene families, STIRPs and Amoebapores, showed no cases of multiple gene family members located on the same contig in the new assembly.

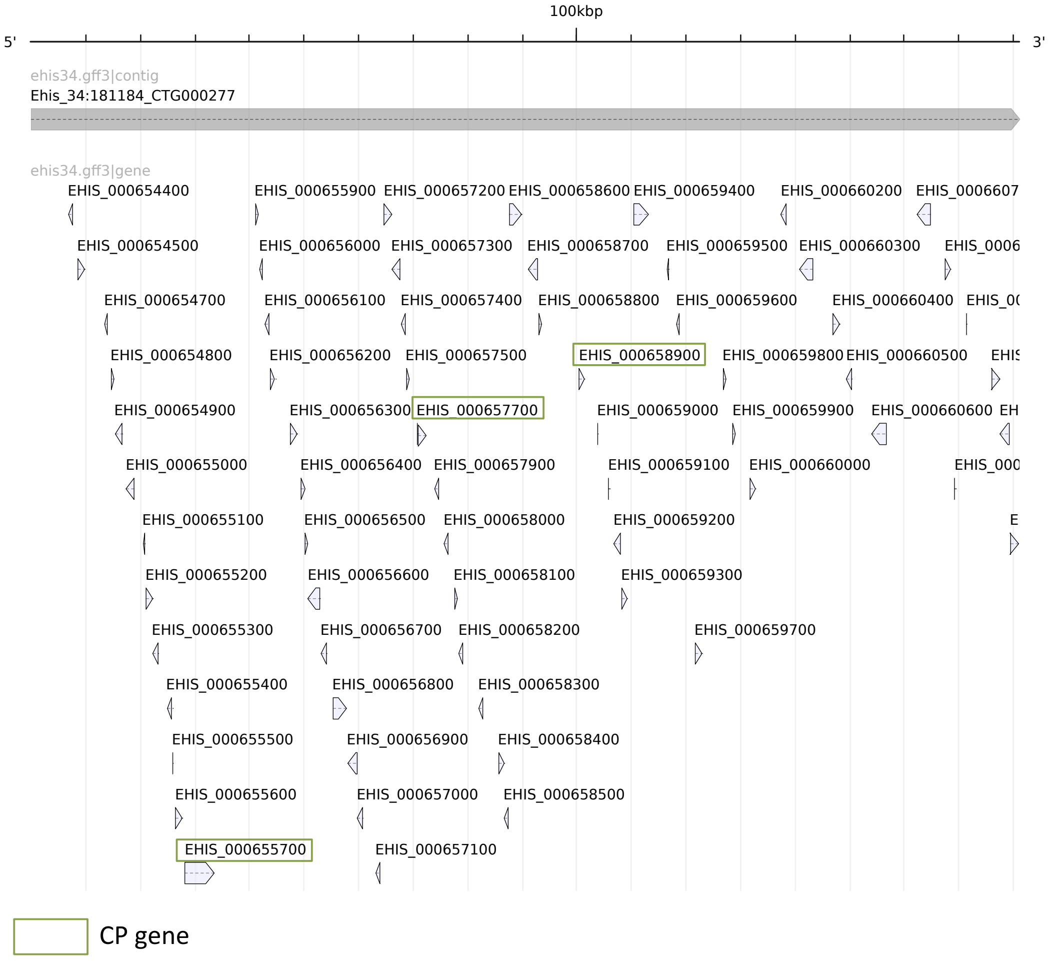
The remaining gene families contained small numbers of members in the same contig. For each of these families, the contig containing the most family members was analysed to identify if members shared sequence similarity and to assess the distribution of the genes along the contig, to determine whether they resulted from tandem duplication. In all five families, gene length varies across members of the same family on a single contig (Figure 3.3.3; Contigs containing the largest numbers of BspA genes and Gal/GalNAc genes not visualised due to size of the contig). The sequence is also not highly conserved outside of domain regions. The gene members are not organised within close proximity to one another or form clusters. Instead, it appears that length of a contig correlates with an increased number of genes from a particular family; this suggests that more members of a gene family occur on particular contigs due to their contig length, not due to any pattern of distribution.

Protein sequences for each gene family were also aligned and a phylogenetic tree was created for each gene family. In all cases, gene family members on the same contig did not cluster together or form contig-specific clades (Data not shown; Trees presented in 3.3.5/3.3.6).

A) AIG1 Family: 5 AIG1 genes on Contig 70 (100,919 bp)



B) Cysteine Protease (CP) Family: 3 CP-1 genes on Contig 34 (181,184 bp)



C) Ariel-1 Family: 3 Ariel-1 genes on Contig 6 (395,765 bp)

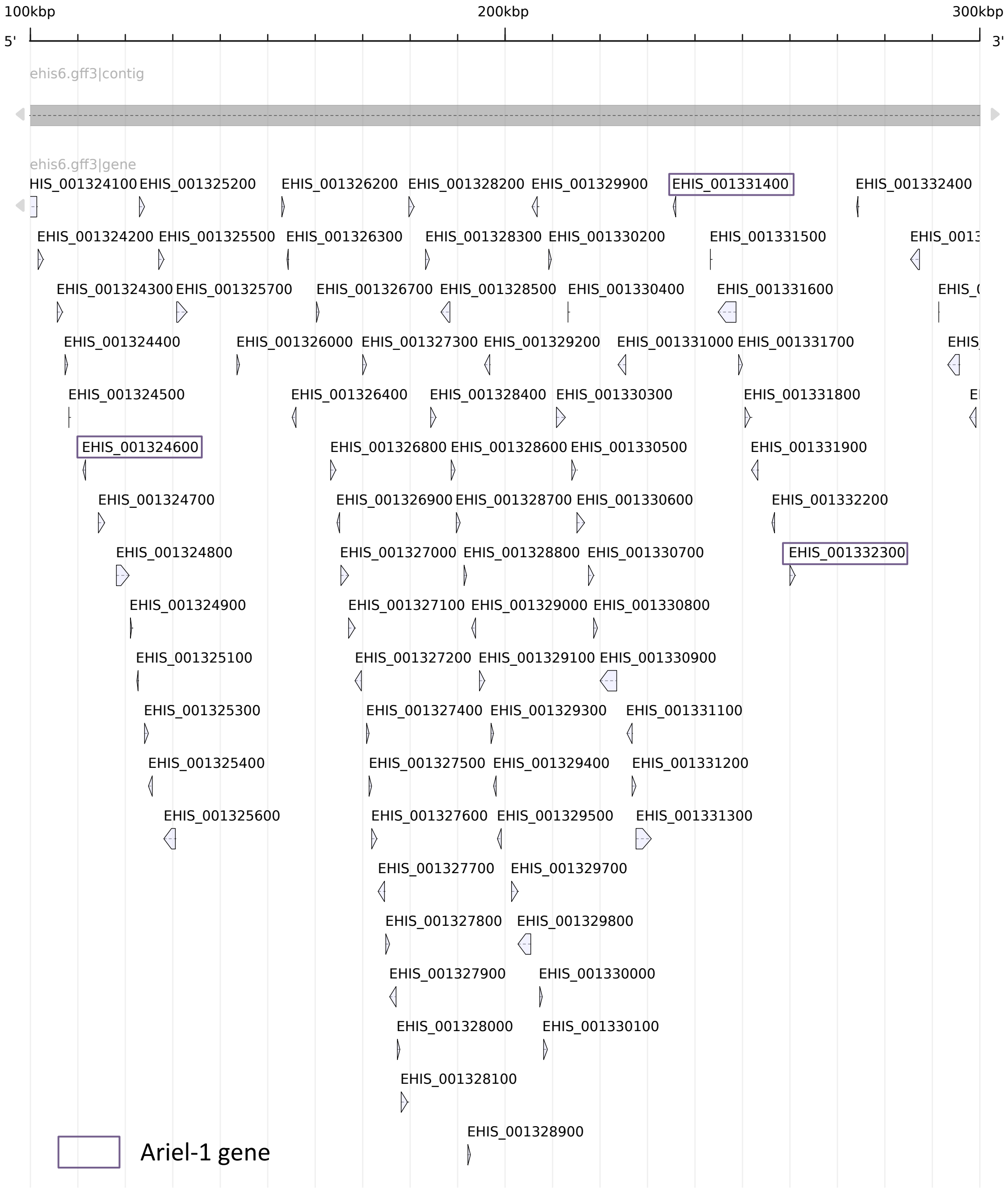


Figure 3.3.3. Gene organisation of the multi-gene virulence families in *Entamoeba histolytica* HM-1:IMSS. Gene families are distributed across many contigs in the *E. histolytica* HM-1:IMSS genome. The contig with the largest number of members for each virulence gene family (AIG1, Ariel-1, Cysteine Protease) are displayed in each panel.

3.3.6. Sub-telomeric genes have a wide range of functions and are expressed at a variety of life cycle stages

Arrays of tRNA genes separated by short tandem repeats (STRs) may act as *Entamoeba* telomeres (see Chapters 4 and 5). 21 contigs in the PacBio assembly are comprised of genic content and tRNA arrays (tRNA-genic contigs); in all cases the contigs are terminated by the tRNA array and tRNA arrays are never seen between two genic regions (Chapter 4). As such, it is assumed that the tRNA arrays in *Entamoeba* species act as telomeres in an analogous mechanism to what are seen in *Drosophila* and *Dictyostelium* where arrays of retro-transposons and rRNA genes form telomeres, respectively.

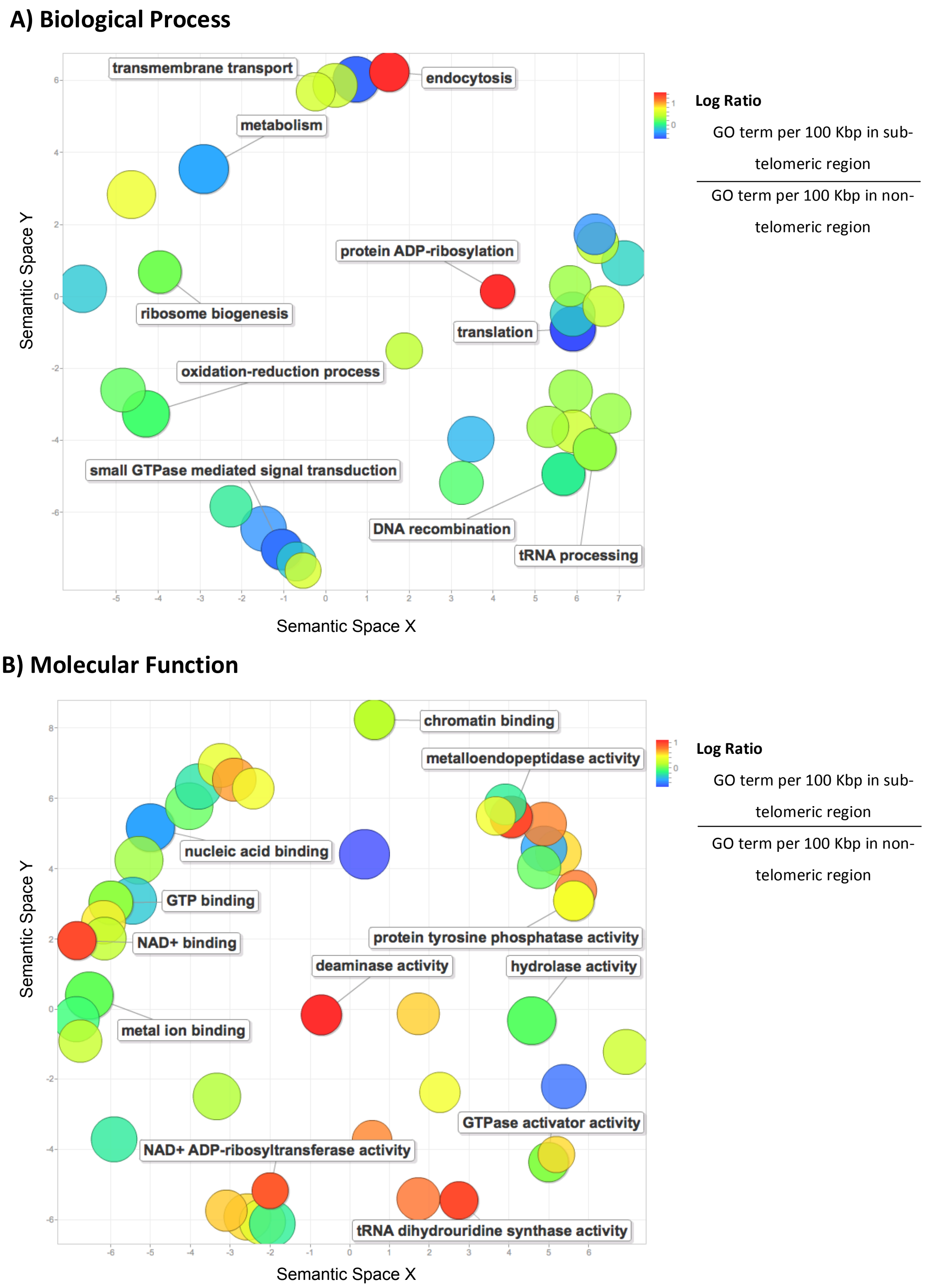
The gene content was extracted from the annotation output corresponding to the 100 Kbp of sequence that directly flanks the tRNA on these tRNA:genic contigs. A 100 Kbp cut off was chosen as the sub-telomeric regions of other well studied protozoans are reported to be up to this length. In *Plasmodium falciparum*, the sub-telomeric regions are reported as stretching larger than 100 Kbp [285]. If the contig was smaller than 100 Kbp, then all genic information for this contig was extracted. 327 genes were annotated in the sub-telomeric regions of the 21 contigs. The putative function of the genes annotated in the sub-telomeric regions were analysed to detect any enrichment for gene families. 164 genes (50%) were annotated as hypothetical with no InterProScan information and hence no function for these genes can be inferred Of the remaining 163 genes, five gene families had more than five members in putative sub-telomeric regions (Table 3.3.8).

Table 3.3.8. Gene families enriched in the sub-telomeric region of the new *Entamoeba histolytica* HM-1:IMSS assembly. Gene families with five or more members in sub-telomeric regions are represented in the table. Significantly enriched telomeric gene families are highlighted in bold

Family	Sub-telomeric copies	Genomic copies	Sub-telomeric density (Copies/10Kbp)	Genomic density (Copies/10Kbp)	Ratio $\frac{\frac{Subtelomeric}{frequency}}{\frac{Genomic}{frequency}}$	Fishers Exact Test (Significance)
AIG-1	11	62	0.13	0.026	**5.07**	**<0.00001 (p<0.05)**
Trichohyalin	11	159	0.13	0.064	**2.07**	**0.019 (p<0.05)**
BspA	7	123	0.084	0.050	1.70	0.119 (p>0.05)
Protein Kinase	7	142	0.084	0.057	1.48	0.227 (p>0.05)
Domain of unknown function containing-protein	5	73	0.060	0.029	2.05	0.086 (p>0.05)

Sub-telomeric gene families were analysed to determine whether they occurred at a higher frequency in the sub-telomeric regions compared to other genomic regions. All of the gene families found in the sub-telomeric regions were found to occur at a higher frequency in the sub-telomeric regions compared to the remaining genome regions. The Trichohyalin and AIG-1 gene families show a significantly greater density in the putative sub-telomeric regions (Fisher's exact test, $p<0.05$). Notably, the AIG-1 genes had a density of more than five times in the sub-telomeric regions than in the rest of the genome.

Genes in the sub-telomeric regions were analysed to see if they were enriched for any broad functions, processes or cellular locations using Gene Ontology (GO) enrichment analysis. GO-terms for the annotated 'sub-telomeric genes' were available for 275/327 (84.1%) genes. These were summarised and visualised using REVIGO (Figure 3.3.4).



A) Biological Process

B) Molecular Function

Figure 3.3.4. GO term enrichment of genes in sub-telomeric regions of the *Entamoeba histolytica* HM-1:IMSS genome. GO terms for genes were extracted from genes found in sub-telomeric regions, summarised and visualised using REVIGO [276].

In REVIGO, the x and y co-ordinates are derived from a multi-dimensional scaling and acts so that similar GO terms cluster close together. The colour scale represents a custom calculated metric to determine enrichment of a specific GO term in the sub-telomeric region (GO term per 100 Kbp in sub-telomeric region/GO term per 100 kbp in non-telomeric region). The frequency of each GO term was calculated for both the sub-telomeric region and the remaining genomic region (instances of GO term/100 Kbp). A ratio was calculated between the two regions and log-transformed to aid visualisation. Red represents GO-terms more frequently seen in the sub-telomeric regions, green represents those equally likely to occur in both regions and blue represent those which occur at higher frequency in non-telomeric regions.

The analysis showed enrichment for endocytosis, transmembrane transport and protein ADP-ribosylation. Analysis of cellular component showed enrichment for the proton-transporting two-sector ATPase complex. Analysis of molecular function revealed enrichment for NAD+ binding, NAD+ ADP-ribosyltransferase activity, metalloendopeptidase activity, deaminase activity and tRNA dihydrouridine synthase activity.

3.3.7. A number of protein families commonly occur in close proximity to transposable elements in the genome

RepeatMasker was used to identify retrotransposons and other transposable elements in the PacBio *E. histolytica* HM-1:IMSS assembly. The 1 Kbp flanking sequence on either side of each transposable element was used in a BLASTn search against the *Entamoeba histolytica* HM-1:IMSS gene set from AmoebaDB. The output was inspected to remove overlapping hits resulting from closely located TEs containing the same flanking sequence.

For a number of protein families, large numbers of their members occur in close proximity to transposable elements in the new *E. histolytica* HM-1:IMSS genome. Table 3.3.9 shows gene families with >30% of members associated with TEs.

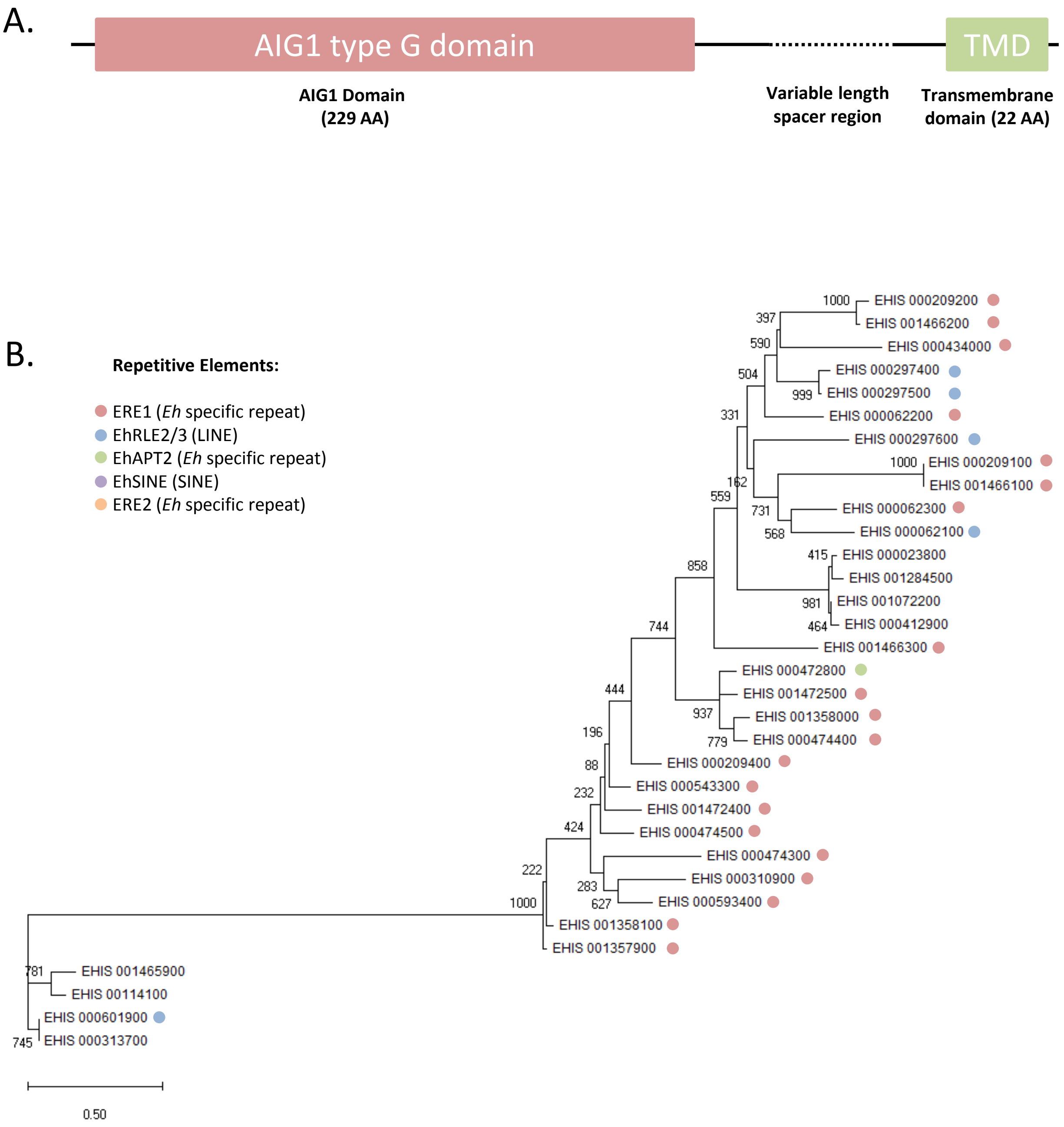
Table 3.3.9. ***Entamoeba histolytica*** **gene families showing high association (5 or more members) with repetitive elements.**

Protein family name	Number of associated elements	Number of members in gene family	Percentage of Association
Chitobiosyldiphosphodolichol beta-mannosyltransferase family	5	5	100
Elongation factor 1-alpha family	4	4	100
Ariadne-1 family	3	3	100
D-glycerate dehydrogenase family	4	4	100
Elongation factor 2 family	3	3	100
Regulator of nonsense transcripts family	9	10	90
Transketolase family	4	5	80
AIG1 family protein	51	73	70.00
Ariel-1 family	8	12	66.67
BspA-like family protein	85	130	65.38
RNA pseudouridylate synthase family	5	8	62.5
Chaperonin 1 family	5	10	50.00
Heat shock protein 70 family	25	58	43.10
NADPH-dependent FMN reductase domain containing protein	5	12	41.67
DEAD/DEAH box helicase family	13	37	35.14
Leucine rich repeat containing protein family	10	30	33.33
Cysteine protease family	27	87	31.03

For some gene families, including elongation factors 1 and 2, Ariadne proteins and Chitobiosyldiphosphodolichol beta-mannsyltransferase families, all members occurred in close proximity to a TE. Several gene families associated with virulence, including AIG-1, BspA, Ariel-1 and Cysteine protease families, also show a high proportion of genes associated with TEs.

Generally, members of the same gene family were associated with the same type of repetitive element. For example, all members of the Chitobiosyldiphosphodolichol beta-mannsyltransferase family were associated with the *Entamoeba* specific repetitive element, ERE1, whereas all members of the elongation factor 1 and 2 families were associated with the Dong-R4 type LINEs (EhRLE2/EhRLE3). Eight of the nine members of the 'regulators of non-sense transcript family were also associated with EhRLE2/EhRLE3 LINEs. Some of the larger gene families appear to be associated with a range of elements. AIG1 members associated with a repetitive element were mostly associated with ERE1 (35/51 genes) followed by EhRLE2/EhRLE3 (12/51). However, ERE2, EhSINE1 and EhAPT2 elements were associated with one, two and one AIG1 genes, respectively.

To test whether AIG1 genes clustered on the basis of the repetitive element they were associated with, a phylogenetic tree was produced (Figure 3.3.5). Three distinct AIG1 clusters can be seen in the new *Entamoeba histolytica* HM-1:IMSS genome. In all clusters, clades largely cluster based on the type of repetitive element the members are associated with. Generally, members with no associated repetitive element cluster together or appear as singletons, forming their own clade/group.


A.
AIG1 type G domain
TMD
AIG1 Domain (229 AA)
Variable length spacer region
Transmembrane domain (22 AA)
B.
Repetitive Elements:
● ERE1 (*Eh* specific repeat)
● EhRLE2/3 (LINE)
● EhAPT2 (*Eh* specific repeat)
● EhSINE (SINE)
● ERE2 (*Eh* specific repeat)
1000
EHIS 000209200 ●
397
EHIS 001466200 ●
590
EHIS 000434000 ●
504
EHIS 000297400 ●
999
EHIS 000297500 ●
331
EHIS 000062200 ●
EHIS 000297600 ●
1000
EHIS 000209100 ●
162
EHIS 001466100 ●
559
731
EHIS 000062300 ●
568
EHIS 000062100 ●
415
EHIS 000023800
858
EHIS 001284500
981
EHIS 001072200
744
464
EHIS 000412900
EHIS 001466300 ●
EHIS 000472800 ●
444
EHIS 001472500 ●
937
EHIS 001358000 ●
196
779
EHIS 000474400 ●
88
EHIS 000209400 ●
232
EHIS 000543300 ●
EHIS 001472400 ●
424
EHIS 000474500 ●
222
EHIS 000474300 ●
283
EHIS 000310900 ●
1000
627
EHIS 000593400 ●
EHIS 001358100 ●
EHIS 001357900 ●
781
EHIS 001465900
EHIS 00114100
EHIS 000601900 ●
745
EHIS 000313700
0.50

C.

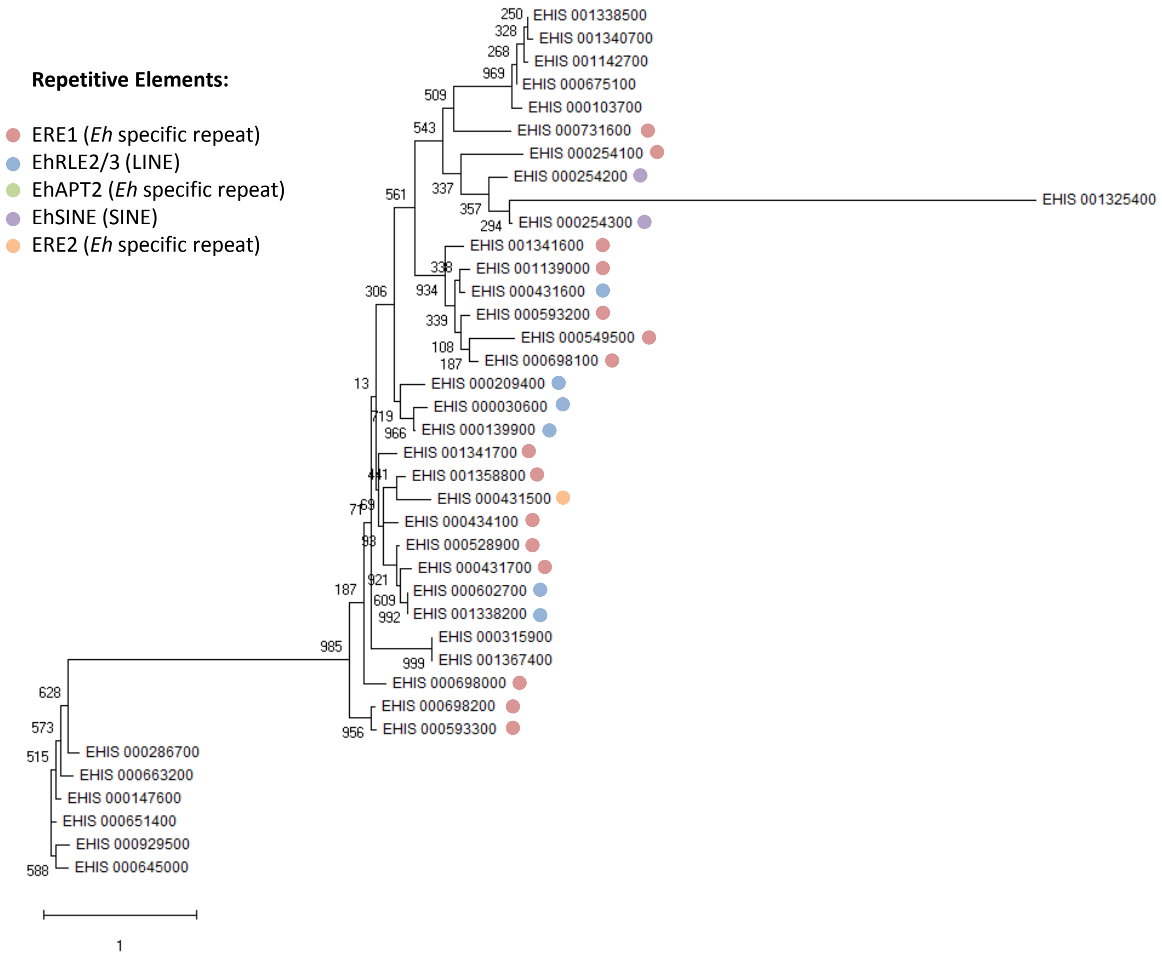


Figure 3.3.5. Phylogeny of the AIG1 gene family in the *Entamoeba histolytica* HM-1:IMSS genome. A) Protein domain structure of AIG1 genes. B&C) Phylogenetic trees of the two AIG1 gene family clusters. Gene members associated with repetitive elements are shown by a coloured circle; members with no coloured circle represent AIG1 members that are not associated with repetitive elements. Bootstrapping was performed for 1,000 replicates and values are shown. 204 amino acids were aligned across the conserved AIG1 type G domain and transmembrane domain.

A similar pattern is seen for the Ariel-1 family (Figure 3.3.6), where 12 members form 4 main groups with the largest having all members associated with an EhSINE1 element. Copies of the Ariel-1 genes that are not associated with a TE cluster away from the EhSINE1 group. It is worth noting that, in this family, members on the same contig do not cluster close together and this reflects a pattern seen in all of the gene families analysed.

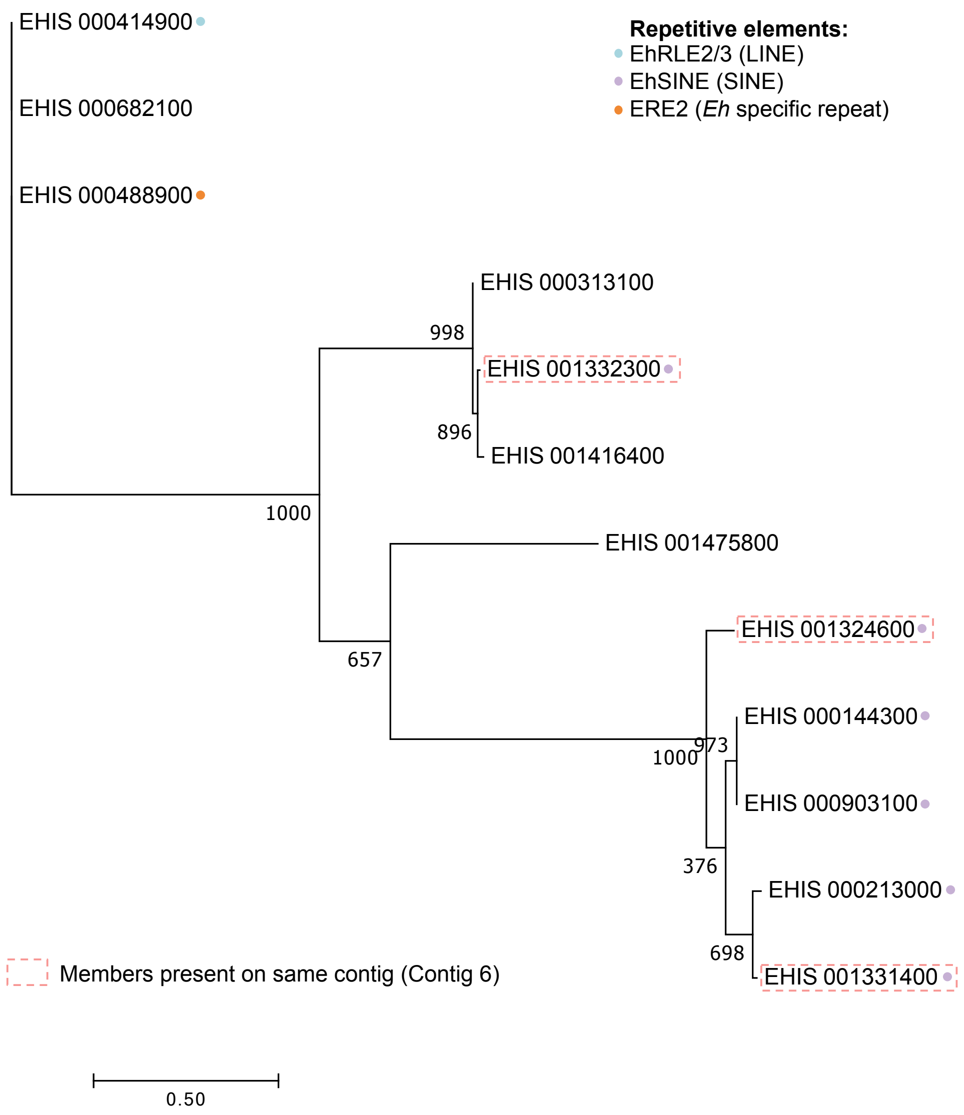


Figure 3.3.6. Phylogeny of the Ariel-1 gene family in the *Entamoeba histolytica* HM-1:IMSS genome. Genes associated with repetitive elements are shown by a coloured circle; those with no coloured circle are not associated with a repetitive element. Genes on the same contig are represented by a dashed border. Bootstrap branch support (1,000 replicates) values are shown.

3.4. Discussion

3.4.1. The *Entamoeba histolytica* HM-1:IMSS genome assembly produced using single-molecule sequencing contains more genes than the published genome assembly

The new *Entamoeba histolytica* HM-1:IMSS reference genome contains 1,831 more genes than previously reported [98]. The majority of these novel genes (70.7%) consist of additional copies of known genes or new members of known gene families. This is unsurprising as the majority (55%) of *Entamoeba histolytica* genes form part of one of 897 gene families. The expansion of these families in the new assembly is likely a result of more accurate assembly of the genome, resulting in a larger genome size than previously reported. The smaller gene family sizes in the published assembly suggest that these genes may have been unassembled or that true homologues were collapsed. Collapsing of genes in the older assembly is more likely for tandemly duplicated genes. The previous Sanger reads may not have been long enough to span each copy of a gene in regions where tandem duplications have occurred. It is also possible that there has been a real biological change in the gene content of the *E. histolytica* HM-1:IMSS cell line over time. The strain was sequenced 13 years ago and it is possible that the cell line has undergone gene loss or gene gain during this time. Organisms grown *in vitro* are subject to increased selection pressures of some genes and loss of constraint on others, and differences in gene content could reflect this. For instance, in *Salmonella*, gene loss in cultured populations may be adaptive as superfluous genes confer a fitness cost [286]. Similarly, the *Plasmodium falciparum* laboratory strain 3D7 has lost the ability to differentiate into its sexual forms (gametocytes), which are not required to complete the life cycle in continuous blood-stage culture, suggesting a lack of selective constraint to maintain functional copies of these genes [287,288]. However, these examples would suggest gene loss is more likely than gene gain, as seen here. Therefore, technical improvements in assembly are more likely to account for the differences seen.

A small subset of novel genes (29.3% of 1,889 novel genes) contain no orthologues to any known gene in the *E. histolytica* genome although some definitely contain Pfam domains and InterPro identifiers which allow for them to be identified as members of existing gene families. For example, 10 novel genes with no *E. histolytica* 'orthologue' were annotated by InterPro as AIG1 family/50S ribosome-binding GTPases. The AIG1 family in *E. histolytica* is diverse in sequence and members can differ greatly in size whilst still containing the defining domains of the AIG1 family. These novel AIG1 genes without a published orthologue also occur on the same contig as known published genes suggesting they are part of the *E. histolytica* genome and not a result of contamination. It is likely that these novel genes form part of the AIG1 family despite not having an orthologue present in the published assembly. In addition, four of the ten genes form their own clade that is nested in the middle of one of the AIG1 clusters, further suggesting these genes are genuinely members of the AIG1 family. It is likely that there are novel singleton genes too however, a large proportion of novel genes that were identified were characterised as hypothetical as the annotation pipeline could not detect any putative function for these genes. It would be interesting to look at each of these novel genes, as well as the large amount of hypothetical genes that also exist in the published assembly, to experimentally predict the function or pathways that these genes are involved in.

In one striking example, a single gene, Trichohyalin (EHI_077870), in the old assembly appears massively expanded (to 168 copies) in the new assembly. This suggests that expansion has occurred since the published *E. histolytica* genome was sequenced. None of the members have orthologues to hypothetical proteins in the old assembly, ruling out the prospect that the gene family had been previously identified but functionally unannotated. 107/170 (64%) trichohyalin genes in the PacBio assembly can be identified in the published assembly sequence, however they have not been annotated, suggesting that their existence is real and not result of a technical assembly error. Further suggesting their genuine existence is the observation that the members occur across 123 contigs and are positioned on contigs with functionally annotated *E.*

histolytica genes suggesting they are not resultant of contamination. Trichohyalin is an intermediate filament-associated protein involved in cross-linking. In humans, this protein associates between keratin intermediate filaments (KIFs) in the hair follicle and the granular layer in the epidermis [289]. Interestingly, trichohyalin-like proteins are expressed in *Trypanosoma cruzi* during the infective trypomastigote life stage, where they are predicted to reflect specialised capacities linked to host-cell recognition, signalling and invasion [290–292]. Further work should be performed on the *E. histolytica* HM-1:IMSS trichohyalins to determine at what life cycle stages these proteins are expressed and to test for any interaction between trichohyalin and the host cell. Nevertheless, this trichohyalin gene family in *E. histolytica* HM-1:IMSS is an interesting candidate for further investigation.

Although 99.3% of genes from the old assembly were transferred to the new assembly, 58 genes remain absent from the new assembly. It appears that this is not a result of errors in the annotation pipeline. Instead, analysis of where these genes occur in the old assembly points to regions of three scaffolds that are not present in the new PacBio assembly. The absent genes were not present in the raw PacBio reads, suggesting that the missing genes do not result from the genome assembly process but that they did not exist in the genome to begin with (i.e. a biological difference). However, mapping of short-read Illumina data across the absent gene regions in the published assembly provided an average coverage depth of ~300x suggesting that these regions do indeed exist in the *E. histolytica* genome. This discrepancy may reflect genuine biological differences between the organisms used for the PacBio and Illumina sequencing. The *E. histolytica* cell used for the PacBio sequencing were revived from cell stocks at the University of Liverpool whereas, the *E. histolytica* cells used for the Illumina sequencing were revived from cell stocks from the London School of Hygiene and Tropical Medicine. It is possible that the gene differences represent real biological differences between the *E. histolytica* HM-1:IMSS cell lines held at each institution, which would highlight the high levels of genome plasticity that have previously been reported within *E. histolytica* [175]. Another explanation for the missing genes in the PacBio reads is that these genes were present in the genomic DNA but coverage was not high enough to cover all regions of the genome. This is highly unlikely as average coverage depth of the genome was 200x.

3.4.2. *Entamoeba histolytica* HM-1:IMSS sub-telomeric regions are not enriched for the virulence AIG-1 gene family and the Trichohyalin gene family

Assuming that the tRNA arrays (Chapter 4) form a telomere analogue in *Entamoeba,* the genes flanking these regions do not appear to contain larger numbers of members of the known gene families associated with virulence in *Entamoeba* species. Though the AIG-1 and trichohyalin gene families form an exception to this observation as these two families were significantly enriched in the putative sub-telomeric regions. Further to this, gene families do not appear to have members within close proximity to each other across the entire genome; few members of the same family occur on the same contig and often when they appear on the same contig the gene sequences can be divergent and occur far away from one another. This mirrors a recent report on the AIG1 virulence gene family, where AIG1 genes close together spatially clustered into different phylogenetic clades based on sequence similarity [149].

GO term enrichment analysis of the genes present in the sub-telomeric regions suggests a range of functions is represented in these regions. These include metabolism of various compounds and DNA/RNA interactions, however some of the enriched functions could be associated with virulence of the parasite. Biological processes that are most enriched include endocytosis and protein ADP-ribosylation. Interestingly, ADP-ribosylation has been implicated as an important process for bacterial toxicity. Protein ADP-ribosylation involves the addition of an ADP-ribose moiety to a protein involved in cell signalling, DNA repair or/and gene regulation. This process forms the basis of toxicity in bacterial compounds such as the cholera toxin, diphtheria toxin and other bacterial ADP-ribosylating exotoxins (bAREs). Further, ADP-ribosylated proteins have been implicated as playing an important role in the survival of the *Entamoeba histolytica* parasite and interaction with host molecules [293] .

Metalloendopeptidase activity was an enriched molecular function. The major surface proteins (MSPs) in *Leishmania* are enriched for metalloendopeptidase activity and contribute significantly to virulence [294]. In *E. histolytica* strains that show resistance to Metronidazole, the main drug used to treat amoebic dysentery, show increased expression of genes with metalloendopeptidase activity [295]. This suggests that metalloendopeptidase proteins may be involved in drug resistance. A metalloendopeptidase surface protein gene family, similar to those seen in *Leishmania*, could exist in the *Entamoeba histolytica* genome.

Overall, the majority of the genes found in the sub-telomeric regions were of unknown function, mirroring the genome as a whole. Further functional studies of some of the vast number of hypothetical genes in the *Entamoeba* genome could go some way to helping elucidate the functions of the sub-telomeric genes and reveal any functional enrichment of genes in these regions.

3.4.3. A large number of virulence gene families are associated with repetitive elements in the *Entamoeba histolytica* HM-1:IMSS genome

Transposable elements (TEs) are dynamic elements that can reshape host genomes by generating rearrangements with the potential to create or disrupt genes, shuffle existing genes and modulate their expression pattern [296–298]. In the genomes of parasites, TEs have been identified that are likely to have been maintained throughout evolution as they confer some benefit to organism through their contribution to gene function or gene expression. In parasite, TEs can frequently be inserted inside a coding sequence or within the 3’-untranslated regions (UTR) of genes and domestication of these TEs has proven beneficial to a range of parasites. For example, in *Plasmodium yoelii yoelii*, the insertion of a TE into the open reading frame (ORF) of a putative *yir3* is suggested as being associated with immune evasion of the human host. This is because, the *yir3* family in *P. y. yoelii* is analogous to the *var* genes in *P. falciparum* (Chapter 1), which play an important role the antigen switching that generates the antigenic diversity of the parasite infective schizont stage [299].

In addition to regulating gene expression, models of transposable elements mediating gene family expansion and diversification have been suggested. TEs are substrates for recombination events that can generate genomic rearrangements and duplications. Significant associations between retrotransposons (LINEs and LTRs) and the expansion of gene families in the mouse and human genomes have been identified with the LINEs associated with gene duplication [300]. For example, LTRs and LINEs are implicated in the gene expansion of the mouse Androgen-binding protein (Abp) gene family. The presence of ERVII (LTR) and L1 (LINE) repeat families in high densities in the mouse and rat Abp gene regions with corresponding depletion of other families suggested a functional role for ERVII and L1 in the two Abp gene family expansions [301].

A number of protein families appear to be physically linked to transposable elements in the PacBio genome. Many of these genes are members of previously identified virulence gene families in *E. histolytica* and include members of the surface protein virulence families (Ariel-1, BspA), the GTPase virulence family AIG1, the proteolytic cysteine protease (CP) enzyme virulence gene family and the virulence heat shock 70 gene family. For example, 43% of the 58 member Hsp70 protein family have a transposon within 1 Kbp of the gene. All of the Hsp70 gene associated with a transposon are of the cytosolic Hsp70 type and no mitochondrial or endoplasmic reticulum (ER-like) Hsps appear to be associated with a transposon. The Hsp70s act as molecular chaperones and aid a range of protein folding processes. The family is highly conserved and gene expression induced under stress conditions [302]. In *Drosophila*, the insertion of transposable elements near Hsp70 gene promoters has been associated with attenuated expression; insertions occur frequently and it is thought the sequence of the *Drosophila* Hsp70 promoter regions is a suitable target for transposon insertion [303]. However, while most insertions result in reduced thermo-tolerance of the *Drosophila*, cases of exceptional thermo-tolerance have been reported despite reduced Hsp70 expression [304] suggesting the insertion of TEs could be playing an adaptive role in producing novel alleles and manipulating the expression of genes critical for parasite fitness [304].

Another gene family with associated transposable elements is the large AIG1 GTPase family. It comprises 73 members distributed across two large and one small sequence similarity-based clusters, of which 51 are within 1 Kbp of a repetitive element. AIG1 genes within *E. histolytica* are thought to be involved with adaption to the host environment [146,147] and adherence to host cells [149]. In comparison to *E. histolytica,* the expression of AIG1 proteins, as well as heat shock proteins, is significantly lower in *E. dispar* (*E. histolytica's* non-virulent sister species that also parasitizes the human gut) [141]. LINEs and SINEs have already been proven to be involved in genome rearrangements that catalyse genomic evolution [122] and from this observation it could be hypothesised that the insertion of these repetitive elements into the neighbouring regions of the *E. histolytica* AIG1 genes could lead to further expansion of the AIG1 family in the genome. This is because as the transposable elements (TEs) propagate themselves, the flanking sequence can also be copied. If this flanking sequence includes an AIG1 gene, this gene can become copied across the genome. In addition, it is also possible that the close proximity of the TEs to the AIG1 genes may effect gene expression of these genes and could lead to altered expression and hence virulence in *E. histolytica* compared to *E. dispar*. AIG1 genes exist as a gene family in *E. dispar* however, fewer members have been annotated compared to those found in the PacBio *E. histolytica* genome. It remains unresolved whether the further expansion of AIG1 in *E. histolytica* HM-1:IMSS has been propagated by the presence of repetitive elements however, the evidence that similar AIG1 sequences are often accompanied by the same repetitive element type suggests that this could be the case.

The *E. histolytica* specific family, Ariel-1, is not found in *E. dispar* [105]. This family occurs as 12 members in the new reference genome, 8 of which are associated with a repetitive element, 6 of these with EhSINE1. Ariel-1 genes encode surface proteins, but their function is unknown. Nonetheless, it is interesting that the majority of members are associated with EhSINE1 and it raises the question whether TEs are responsible for the existence of Ariel-1 genes through their ability to produce novel paralogues and also through their ability to amplify gene families throughout a genome. The phylogeny of the Ariel-1 family would definitely support the theory that EhSINE1 has catalysed amplification of the family, as the Ariel-1 genes associated with this element appear the most similar in sequence when compared to Ariel-1 singletons. Clustering of members based on the contig they occur on suggests members are not being propagated by tandem repeat events but instead members have been transposed with a EhSINE1 at different time points during the evolution of the *E. histolytica* genome. It would be interesting in the future to perform functional analysis to determine if Ariel-1 plays a role in host-parasite.

Similarly, the cysteine protease gene family members and the BspA gene family members that are associated within 1 Kbp of a TE also cluster based on the TE type they are associated with. This again, suggests that virulence gene families are associated with TE propagation. Alternatively, it is possible that the expansion of the TE-associated gene families occurred prior to the insertion of the TE sequences in these regions. If this is the case, the TEs may be affecting the expression of the nearby gene families producing differential expression of key virulence genes between different *Entamoeba* species.

3.5. Conclusions

This chapter presents an overview of the improvements made to the *E. histolytica* reference genome and its annotation. The larger assembly contains more genes than the published assembly. This is largely due to more members of gene families being annotated, though a small number of genes appear to be truly novel. The most striking observation in the new annotation is the expansion of the trichohyalin gene family, which had only a single copy in the previous assembly.

A small number of genes previously present are absent from the new PacBio assembly. While these appear not to exist in the *E. histolytica* HM-1:IMSS cell lines at the University of Liverpool (used for PacBio sequencing), Illumina data generated from the same cell line stored at the London School of Hygiene and Tropical Medicine (LSHTM) suggests they exist in that one. This indicates genuine differences between these two different cell lines and should be tested for the two cell lines.

The assembly of fewer, larger contigs has helped facilitate the analyses of the genomic distribution of virulence gene families and has linked putative telomeric sequences to the core genome such that putative sub-telomeric regions could be identified. Genes in these sub-telomeric regions could be analysed and GO term enrichment performed on the genes that have functional annotation assigned.

Analysis into the genes present in the putative sub-telomeric regions revealed that AIG1, BspA and trichohyalin gene family members occurred at a higher frequency in the sub-telomeres compared to the core genome (result only significant for the AIG1 and trichohyalin gene families, $p<0.05$). Further life cycle stage analysis would be useful to determine whether members of these families are expressed together or mono-allelically as is seen in sub-telomeric gene families of *P. falciparum* and *T. brucei*. However, this analysis is limited currently as an *in vitro* for the full life cycle of *E. histolytica* HM-1:IMSS does not yet exist as *E. histolytica* trophozoites cannot be made to encyst in culture. As a result, life-cycle analyses can only be performed in *E. invadens*, an *Entamoeba* species that causes amoebiasis in reptiles. *E. invadens* is distantly related to *E. histolytica* and not all *E. histolytica* genes have an orthologue in *E. invadens.* Of these orthologous genes, only a few show synteny between the two species. Therefore, it is likely that the expression pattern in *E. histolytica* may not the same in *E. invadens* and therefore, it is hard to perform life-cycle expression analysis of *E. histolytica* despite the existence of the *E. invadens* life-cycle model.

Transposable elements occur in close proximity to many genes. Members of several gene families cluster based on the transposable element they are in close proximity to suggesting that propagation of these virulence gene families in *E. histolytica* could result from transposable element translocation events. Members of virulence genes families rarely appear on the same contig and those on the same contig vary in length suggesting that gene family expansions of these families are not catalysed by tandem duplication events.

Overall, the gene annotation demonstrates an increase in gene number, largely facilitated by the increase in genome size of the PacBio assembly compared to the published assembly. A large number of these genes remain functionally unannotated and large numbers of functional studies and manual curation will be required to bring the *E. histolytica* HM-1:IMSS genome up to the same standard as other protists such as *P. falciparum* [305].

Chapter 4: Analysis of *Entamoeba histolytica* repetitive DNA features

4.1 Introduction

4.1.1. *Entamoeba* transfer RNA genes occur in long multi-gene, multi-copy arrays

The transfer RNA (tRNA) genes in *Entamoeba histolytica* show an unusual organisation. The genes occur in sequence units containing one to five tRNA genes, each separated by DNA that contains short tandem repeats (STRs). The tRNA genes and STRs form a unit that is tandemly duplicated in many copies, to form a tRNA array. 25 different tRNA array units have been detected in *Entamoeba histolytica* HM-1:IMSS. Four of these units also contain copies of the 5S small ribosomal subunit gene alongside the tRNA genes [4,5]. The 25 array types are named according to the tRNA isoacceptor genes and 5S rRNA genes present. For example, the [R5] array contains one arginine tRNA isoacceptor and one 5S rRNA gene, while the [SPPCK] array unit contains single serine, cysteine and lysine isoacceptor genes and two proline tRNA isoacceptors (though the codons these correspond to are different). A list of all 25 isoacceptor types in *E. histolytica,* and the tRNA and 5S rRNA genes they contain, can be found in the appendix (S4.1, Appendix 4). Schematic representations of four examples are shown in Figure 4.1.1.

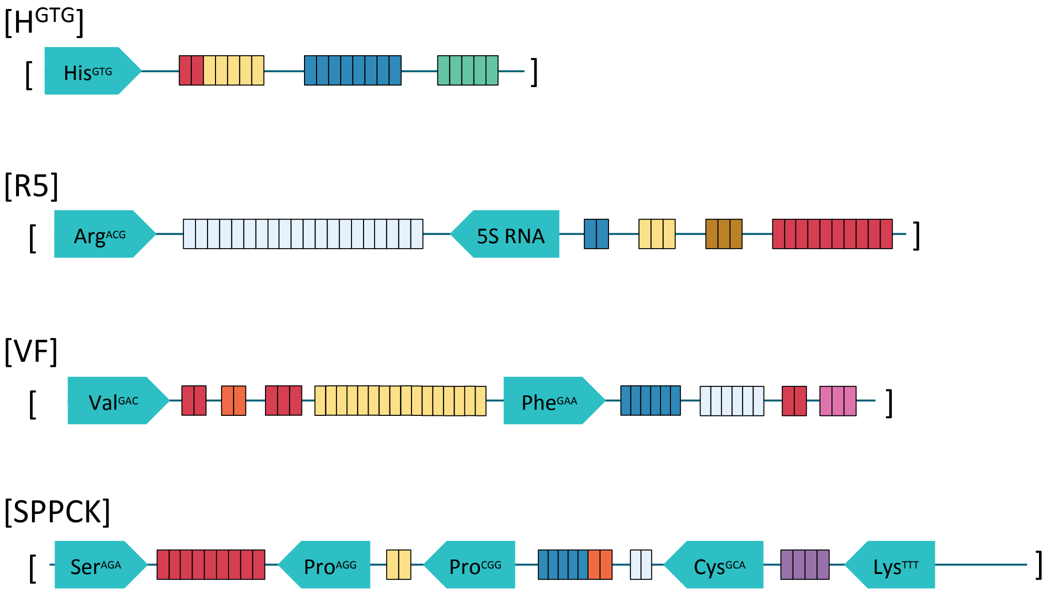


Figure 4.1.1. Schematic representations of *Entamoeba histolytica* tRNA array units. Four representative units are shown. Orientation is indicated by arrow direction. STRs are indicated by coloured boxes, each colour indicating a distinct STR sequence. STR copy number is as shown but size is not to scale. Figure redrawn from Clark *et al*, 2006 [4].

Though clustering of tRNA genes has been observed in a number of eukaryote genomes, this tandem array structure appears unique to *E. histolytica.* In *Dictyostelium discoideum,* the closest sequenced relative of *Entamoeba histolytica* (though still distantly related), tRNA genes are distributed throughout the genome [103]. The origin, evolution and function of these arrays are not yet known, but they appear to be common to species in the *Entamoeba* genus and have been observed in *E. nuttalli*, *E. dispar, E. moshkovskii*, *E. invadens* and *E. terrapinae.* Some tRNA array units are shared between closely related species but generally, each *Entamoeba* species has its own set of tRNA arrays and in *E. moshkovskii* these do not contain STRs [306]. A structural role has been predicted for the tRNA arrays; all but nine of the *E. histolytica* tRNA arrays have been identified as containing Scaffold/Matrix Attachment Regions (S/MARs) and have been implicated in nuclear matrix binding and providing a structural role in the nucleus [306]. However, these S/MAR sequences are not conserved between species and hence doubt has been cast on their matrix-binding role.

Where the arrays are located in the genome is not known but some evidence suggests that they occur in sub-telomeric or telomeric regions [306] and they have been hypothesised as acting as telomeres in an analogous mechanisms to those seen in *Dictyostelium* [103] and *Drosophila* [307,308] where tandemly repeated rDNA genes and retro-transposons, respectively, form the telomeres. No evidence of classical telomere sequence, or sequences present in other organisms, emerged from the original sequencing of the *E. histolytica* genome [97].

4.1.2. *Entamoeba* ribosomal DNA occurs on extra-chromosomal circular DNA episomes

The *Entamoeba histolytica* rDNA genes are carried on circular episomes, multiple copies of which exist in the nucleus [110]. Two different circular episomes have been described. EhR1, an rDNA circle of *Entamoeba histolytica* HM-1:IMSS is 24.5 Kbp in size and contains two inverted copies of an rDNA transcription unit (Figure 4.1.2A). The rDNA transcription unit encodes the 18S, 5.8S and 28S rRNAs. Several short tandem repeats are located in the intergenic spacers (IGS) upstream and downstream of the rDNA [110]. EhR2 is 14.1 Kbp derivative of EhR1 formed by intra-chromosomal recombination (marked by arrows in Figure 4.1.2B). EhR2 contains a single rDNA transcription unit and a range of short tandem repeat families in the IGS [124]. The promoter of the rDNA genes has been mapped to 2.6 Kbp upstream of the mature 18S rRNA between the *Ava*II and *Hinf*1 repeats [309].

Extra-chromosomal rDNAs are seen in some other species. In *Dictyostelium discoideum,* a chromosomal ‘master-copy’ of the rDNA genes generates many linear extra-chromosomal molecules encoding these genes [103,126]. A chromosomal rDNA master-copy remains to be found in *Entamoeba*. No chromosomal rDNA genes were identified in the published genome assembly but, given the incompleteness of the assembly, their existence cannot be ruled out.

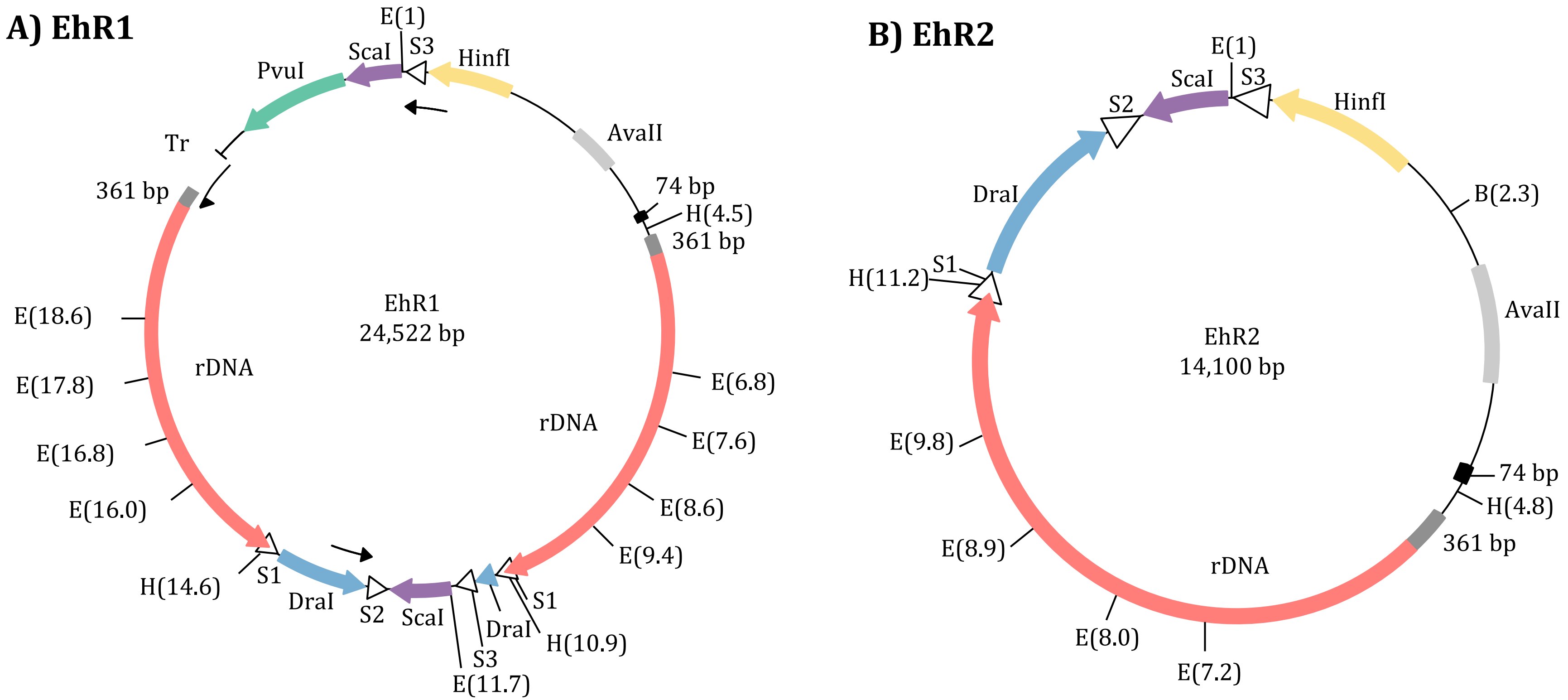


Figure 4.1.2. Ribosomal DNA (rDNA) episomes in *Entamoeba histolytica* HM-1:IMSS. HindIII and EcoR1 restriction sites are indicated with H and E (Position of cut site displayed in brackets), Red arrows show the rDNA genes and short tandem repeat families are indicated in other colours. *Tr* is a transcriptional unit only transcribed by EhR1 and can be used for distinguishing between the two units. Two black arrows in EhR1 (A) show the sites of homologous recombination that produced EhR2 (B). (Images redrawn from [108,122]).

4.1.3. tRNA array STRs and rDNA episomes are phylogenetic and population genetic markers

Accurate genotyping methods are crucial for correctly identifying species and strains of *Entamoeba*, to distinguish between virulent (*E. histolytica*) and avirulent (*E. dispar*) species or to identify virulent strains within a species.

The rDNA episome has been used for genotyping different *Entamoeba* species. As well as containing the rDNA genes, the rDNA episomes also contain the *Tr* region, which occurs in the upstream region of the episome and is transcribed into a polyadenylated 0.7 Kbp RNA [125,241,310–312]. This gene is absent from *Entamoeba dispar* and is useful marker for usage in the *Tr*-present *E. histolytica*. The *Tr* region contains tandem repeats which differ in copy number between strains and can be detected using PCR amplification [313]. The 18 SSU rRNA gene can be used to differentiate between different *Entamoeba* species. PCR and sequencing of specific 18S rRNA gene regions can identify the *Entamoeba* species found in stools [314]. However, this method assumes that there is no differentiation between the many hundreds of copies of the RNA episomes present in each *Entamoeba* cell.

Genotyping *Entamoeba* strains has been hampered by a general lack of microsatellites in the genome. A few PCR-based DNA typing methods have been developed which utilise the existence of polymorphic repeats in coding regions. The repeat-containing protein-coding chitinase [315–317] and the Gal/GalNAc lectin [318] demonstrate polymorphic repetitive regions between different strains but the extent of polymorphisms is limited. Similarly, the serine rich *E. histolytica* protein (SREHP) gene contains tandemly repeated 12 bp and 8 bp sequences that differ in copy number between strains [319,320]. Further to this, sequence polymorphisms and altered restriction sites have been observed between *E. histolytica* strains [315,316,321–323] and one study has observed that these polymorphic patterns are different between intestinal and amoebic liver abscess strains [321].

PCR amplification of tRNA array STRs is currently widely used for genotyping *E. histolytica*. This is because the STR sequences differ between different *Entamoeba* strains. Species specific primers exist that amplify selected tRNA array unit STRs however, this method does not consider any variation within the STRs in a single population [230]. Variation between single tRNA array units in a tandem array have not yet been studied as the sequencing technologies were not able to produce long contiguous sequences that span the length of multiple tRNA array units. As such, it is possible that there may be variation present in the STR sequences within a population that have not yet been detected. These variations may interfere with the tRNA-based genotyping method as expansions or contractions of STR copy number could result in a PCR product that is larger or smaller than expected and lead to incorrect genotyping.

4.1.4. Aims of chapter

Sequences within tRNA array units and the rDNA episomes play an important role in genotyping and distinguishing *Entamoeba* species and strains from one another. Though this relies on their stable presence in the genome. Currently, analyses in to the variation within copies of these structures in a population have been prohibited owing to the lack of long-read technologies and no contiguous assembly of multiple tRNA arrays or entire rDNA episome molecules. In addition, the tRNA arrays have been implicated in having a structural role though, their large-scale structure and position in the genome is unknown. Again this was a result of the lack of long reads technology in the published assembly, which meant the tRNA array lengths could not be determined nor, could the genomic location of these structures be resolved. Specifically this chapter aims to elucidate some of these unknowns by performing the following:

- Identify genomic locations of tRNA arrays
- Identify putative function as to why the tRNA genes are arrayed in tandemly duplicated structures

- Identify variation between tRNA array repeat units within the same array and evaluate the effectiveness of using tRNA STR as genotyping markers
- Identify rDNA episomes in the new assembly and assess the core genome for the existence of a master-copy of rDNA genes, which may be acting as a master copy to produce extra-chromosomal rDNA molecules (as is seen in *Dictyostelium discoideum*)

4.2. Materials and Methods

4.2.1. Alignment and visualisation of repeat units of tRNA arrays

Identification of transfer RNA genes (tRNAs) was described in section 2.2.7. Contigs were grouped based on the tRNA array unit they contained. The tRNA arrays were extracted from the assembly using custom-made BED files. These arrays were then split into individual tRNA array units, that were aligned and visualised in the sequence alignment editor, SeaView (version 4.6), using the MUSCLE alignment algorithm [280]. Alignments were inspected to identify variation within the tRNA gene and the STRs of each array type.

4.2.2. PCR amplification and sequencing of short tandem repeats separating tRNA genes in tRNA arrays

To see if the inter-unit variability seen in the PacBio data was visible using standard methods used for distinguishing species, standard PCR and Sanger sequencing were carried out on DNA from the same source. DNA was isolated as described in section 2.2.2.3. Polymerase chain reaction (PCR) amplification was performed using published primers designed to amplify the tRNA Short Tandem Repeats (STRs) [324]. Target STRs were split into two PCR groups based on annealing temperature. All PCRs were carried out using KAPA Biosystems HiFi Hotstart PCR ReadyMix (KAPA Biosystems, Massachusetts, USA) with 5 ng of *E. histolytica* HM-1:IMSS genomic DNA. Temperature for annealing was reduced by 5 °C as per the PCR ReadyMix instructions. PCR group 1 (Average annealing temperature: 61 °C, STRs: R-T, M-E, P-P) and PCR group 2 (Average annealing temperature: 55 °C, STRs: A-A, H-H, R-M, R-R, Y-E, N-K were subjected to 98 °C for 2 minutes followed by 20 cycles of 95 °C for 20 seconds, group-specific annealing temperature minus 5 °C for 15 seconds and 70 °C for 30 seconds.

PCR products were separated by electrophoresis on 1.5% agarose gels (150V, 1 hour). Gels were stained using ethidium bromide (1 uL per 100 mL of gel) and visualised using a UV transilluminator. PCR product bands were cut from the gels and purified using the ThermoFisher GeneJET PCR Purification Kit (Thermo Scientific, Wilmington, DE, USA) following the manufacturer's instructions. Purified PCR products were Sanger sequenced (GATC Biotech, Konstanz, Germany) and electropherograms inspected visually for evidence of sequence variation (mixed electropherogram peaks).

4.2.3. Short tandem repeat (STR) and codon usage identification

Tandem Repeat Finder [325] was used to identify the STRs between adjacent tRNAs within an array unit.

To calculate codon usage in the PacBio *Entamoeba histolytica* HM-1:IMSS, predicted CDSs (Chapter 3) were processed on the command line (S4.2, Appendix 4) to identify the abundance of each codon.

Mapping of the Illumina Truseq 350 bp paired end library to the *E. histolytica* HM-1:IMSS tRNA genes was performed using the Burrows Wheeler Aligner (BWA version 7.12) [235]. Strict mapping parameters were applied and reads with only 100% sequence identity were mapped to eliminate inaccurate mapping of reads to similar tRNA gene sequences. Average sequencing depths for each tRNA gene were calculated using the SAMTools pileup tool [236]. Average sequencing depth of each tRNA gene was plotted against the abundance of the corresponding codon in the CDS sequences and an R^2 value calculated using the Pearson correlation coefficient.

4.2.4. Identification of the Pro^{TGG} array unit

Entamoeba histolytica transfer RNA (tRNA) array units have previously been published [241]. The Pro^{TGG} array gene sequence (accession number BK005669) was used as a query in a BLASTn search against the PacBio assembly generated by the single molecule sequencing data. An E-value threshold of 0.01 was applied to the BLAST query [239]. BLAST alignments were visualised and manually inspected in SeaView (version 4.6) [280].

To confirm Pro^{TGG} array units occur in strings of more than one unit, PacBio sub-reads were processed using tRNAScan-SE [240] to identify tRNA genes on single sub-reads. AWK and Grep were used to extract sub-reads containing >5 Pro^{TGG} array units and these were manually inspected.

4.2.5. Identification of putative telomeric repeat sequence

The first 150 bp and last 150 bp of every contig in the *E. histolytica* PacBio assembly was extracted using custom made BED files followed by BEDTools (version 2.16.2) getfasta function [278]. Sequences were then analysed using Tandem Repeat Finder (version 4.07b) to identify repeat sequences [325]. The output was manually inspected to identify any common repeat units.

4.2.6. Analysis of the EhR2 episome

Identification of rDNA episomal sequences has been previously described in section 2.2.7. Contigs with BLAST hits (section 2.2.7) to the previously identified *E. histolytica* HM-1:IMSS rDNA sequence were extracted. A 14 Kbp repeat containing the rDNA sequence was identified across multiple contigs. The 14 Kbp repeat arrays were split into individual repeat units and aligned using the Cyclic DNA Sequence Aligner (http://kdbio.inesc-id.pt/~csa/). The consensus alignment was digested *in silico* using NEBCutter [326] with restriction enzymes previously reported in the rDNA restriction maps of the *E. histolytica* rDNA episomes [241].

4.2.7. Confirmation of the absence of EhR1 episome

Previous restriction maps for EhR1 and EhR2 were manually inspected to identify unique regions in the EhR1 episome that would distinguish it from the EhR2 episome. The restriction sequence for PvuI was identified as only occurring in the EhR1 episome. The restriction site for PvuI (CGATCG) was used as a query in a string match search against the original PacBio sub-reads

4.3. Results

4.3.1. tRNA genes are abundant in the new *E. histolytica* HM:1-IMSS genome assembly

tRNAscan-SE was used to scan the new *Entamoeba histolytica* HM-1:IMSS genome assembly. 4,436 tRNA genes were identified with 213/563 of the contigs containing at least one tRNA gene. 333 Kbp of the genome (1.2%) is made up solely of tRNA gene sequence with an average tRNA length of 75 bp. These tRNA genes are largely arranged into 25 distinct array units, with a small number dispersed as individual genes within the genome. The tRNA arrays account for 1.9 Mbp of the genome assembly (7.2%) and occur across two types of contigs defined as tRNA-only contigs (n=137) and tRNA-genic contigs (n=21). tRNA-only contigs are entirely composed of tRNA array units from end-to-end whereas contigs in which tRNA arrays are present with non-array sequence are referred to as tRNA-genic contigs.

Most tRNA isoacceptor genes occur as single-exon genes. $\mathrm{Leu}^{\mathrm{AAG}}$, $\mathrm{Ile}^{\mathrm{TAT}}$, and $\mathrm{Tyr}^{\mathrm{GTA}}$ all have two exons and one intron. $\mathrm{Asn}^{\mathrm{GTT}}$ and $\mathrm{Leu}^{\mathrm{TAA}}$ both exist as two forms in the genome; one copy contains one exon and the other contains two exons with an intron.

tRNA abundance (i.e. depth of mapping of reads to tRNA genes) is not correlated with codon usage in the *Entamoeba histolytica* HM-1:IMSS genome (R^2=0.01, p=0.45). One of the most abundant tRNA isoacceptor types, $\mathrm{Ile}^{\mathrm{TAT}}$, accounts for 3.41% of codons in CDSs however only exists as a few dispersed copies across the assembly and has an average coverage depth of 397x. Conversely, one of the least used tRNA isoacceptors, $\mathrm{Asp}^{\mathrm{CGC}}$, accounts for only 0.09% of codons in CDSs but exists as multiple copies within an array and has an average coverage depth of 4018x.

For many amino acids, the corresponding codon can exist in multiple forms (synonymous codons). These synonymous codons usually have different nucleotides at the last base pair position (degenerate positions). As the *E. histolytica* genome is very AT–rich, degenerate third codon positions are often biased towards A or T bases. This is observed in the CDSs where 86.0% of the codons end in one of these bases. It may be expected that due to this, there may be a bias in the number of tRNA isoacceptor genes that associate with these codons. tRNA isoacceptors that correspond to codons with an A or T in a degenerate position have an average coverage of 2,459x and tRNA isoacceptors that correspond to codons with a G or T in a degenerate position have an average coverage of 2,620x, suggesting no correlation between the copy number of the tRNA isoacceptor genes and the base pair present in the degenerate site of a codon.

4.3.2. Dispersed tRNA genes are more abundant than previously reported

Previous estimates reported that 30 tRNA genes were present in the genome without forming part of an array unit and no tRNAs that appeared within a tRNA array also appeared dispersed in the genome [4]. In the new PacBio assembly, seven tRNA isoacceptor types were found to be dispersed in small numbers throughout the genome totalling 63 dispersed tRNA genes. Of these seven tRNA isoacceptor types (Table. 4.3.1.), three were also encoded in a tRNA array. Generally, the tRNA genes that occur as part of an arrayed unit elsewhere appear at a lower frequency than those tRNA genes that exist solely as dispersed copies (Table. 4.3.1.).

Table 4.3.1. Dispersed tRNA isoacceptor types in the PacBio *Entamoeba histolytica* HM-1:IMSS genome assembly. tRNAscan-SE was used to identify tRNA genes in the genome assembly. Manual inspection of the output determined which tRNA genes were encoded in a non-arrayed unit.

tRNA isoacceptor type	Number dispersed in *E. histolytica* HM-1:IMSS PacBio assembly	Number containing an intron [A]	Array type [B]
Leu^{TAG}	13	-	#
Ile^{TAT}	35	35	#
Gly^{CCC}	7	-	#
Arg^{CCG}	1	-	#
Arg^{ACG}	2	-	R5
Leu^{TAA}	1	1	ALL
Asn^{GTT}	2	2	NK
Total	63		

[A] tRNA isoacceptor types that do not have an intron-containing sequence variant are denoted as a dash (-)

[B] tRNA isoacceptor types that do not exist as an arrayed unit are denoted by a hash (#)

As previously mentioned, two isoacceptor types (Asn^{GTT}, and Leu^{TAA}) exist as two distinct sequence types with one containing an intron. In both of these cases, the dispersed copies of the tRNA contain an intron. The 52 other exon-only copies of Leu^{TAA} gene and the 96 remaining copies of the Asn^{GTT} gene are arranged in arrayed units. None of these contain an intron, meaning that the intron-containing versions of these two genes are exclusive to the non-array regions of the genome.

4.3.3. Genetic variability among tRNA array repeat units in the same genome

The remaining 4,436 tRNA genes are arranged in arrays of units consisting of one or more tRNA separated by short tandem repeats (STRs). Several different STRs can occur between adjacent tRNA genes in a unit. Units are repeated in tandem to create long arrays (the longest assembled being 43,389 bp). 25 array unit types have been previously identified [4,306]. The PacBio assembly contained all 25 of these, with no new tRNA array types identified. Individual array unit lengths varied between 471 bp and 1,777 bp, arranged into arrays of up to 43,389 bp (excluding the Pro^{TGG} array which did not form a multi-unit array).

Most of the arrays showed evidence of variation among units in the same array (expanded in 4.3.2). The modal length for each array was calculated and compared to the array unit lengths previously report by Clark *et al* [4] and are reported in Table 4.3.2. They were largely consistent with those previously reported, with two exceptions: the ASD array units are generally 9 bp smaller than has been previously reported and the WI array units are generally 24 bp larger. This is due to differences in the STR regions. In the case of the ASD array, the previously reported ASD sequence contains an extra copy of an 8 bp tandem repeat sequence followed by a 1bp indel in the STR region between the Asp^{GTC} and Ser^{GCT} tRNA isoacceptor genes. The PacBio assembled WI array contained one extra copy of two separate STR sequences of 8 bp and 16 bp, respectively (Figure 4.3.1).

The Pro^{TGG} array was not assembled into a multi unit array in the new *E. histolytica* HM-1:IMSS genome. tRNAscan-SE identified only one full-length Pro^{TGG} isoacceptor gene and one partial Pro^{TGG} isoacceptor gene. These were positioned at the end of a contig with the partial Pro^{TGG} isoacceptor gene terminating the contig sequence. The two genes were 761 bp apart, consistent with the Pro^{TGG} array unit length reported previously [4]. The putative array unit was aligned to the published Pro^{TGG} array unit sequence (accession number

BK005669) and visualised using SeaView [280]. Sequences showed 100% identity. The existence of the Pro^{TGG} was confirmed using PCR amplification of the array unit. To confirm that the Pro^{TGG} array units exist as array structures in the new PacBio assembly, unassembled reads generated from the PacBio sequencing were analysed using tRNAscan-SE [240] to identify tRNA genes. 2,013 Pro^{TGG} isoacceptor genes were identified across 489 reads. The largest number of Pro^{TGG} isoacceptor genes on a single read was 16 copies, positioned at intervals of 761 bp. This confirmed that the Pro^{TGG} isoacceptor array unit found in the assembly also exists as tandemly arrayed units in the PacBio genome. A product approximately 800 bp in length was visualised before being extracted and sequenced using Sanger sequencing. The sequence was identified as the *Entamoeba histolytica* Pro^{TGG} array unit using BLASTn (BK005669, 99% identity, e-value = 0.0).

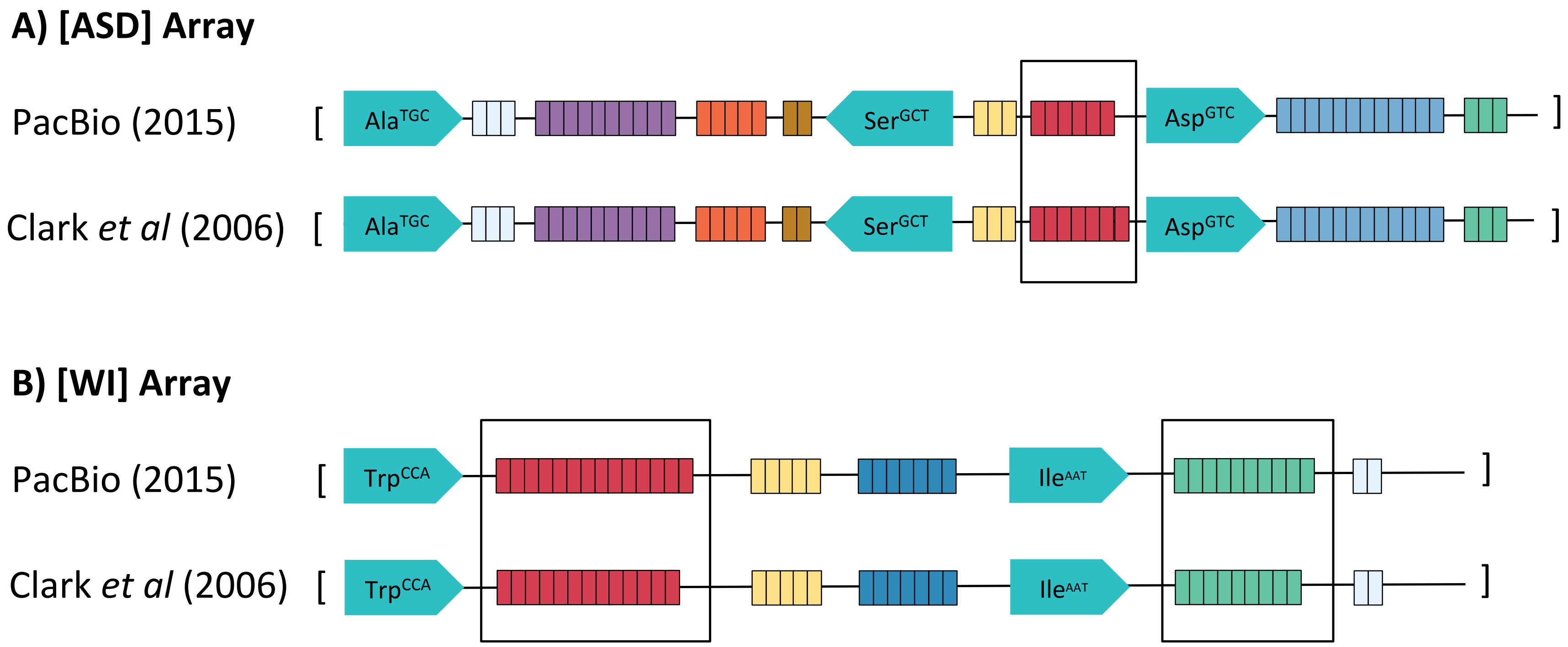


Figure 4.3.1. Schematic representation of tRNA array units that differ between the previous assembly and the new PacBio assembly. The ASD (A) and WI (B) arrays differed between the new PacBio assembly and those reported elsewhere [4]. The orientation of the genes is indicated by the arrow direction. Coloured boxes indicate STRs; each colour in each array indicates a distinct STR sequence. The STR copy number is as shown but the size of the unit is not to scale. Differences are shown in boxes. STRs were detected using Tandem repeat finder (version 4.09).

Table 4.3.2. tRNA Array types in the PacBio *Entamoeba histolytica* HM-1:IMSS genome assembly. The colour scale represents variability among array units (Low=green, High=red). Partial/full deletion of a tRNA gene in one unit was observed in the WI, ASD and V5 arrays. Only one copy of the $\mathrm{P^{TGG}}$ isoacceptor was assembled in the genome therefore variation of this array could not be assessed.

Array	No. contigs containing array	No. of full array units	Array unit min. (bp)	Array unit max. (bp)	Range (bp)	Modal unit length (bp)	Previously reported length [4] (bp)	Change from previously reported length (bp)	Variation index $(\frac{Range}{Modal\ Length} \times 100)$	Variation class (Low/Med/High)
$\mathrm{A^{AGC}}$	1	21	559	559	0	559	559	0	0	Low
ALL	5	45	1109	1154	45	1154	1154	0	3.899480069	Low
ASD	8	69	1098	1157	59	1157	1166	-9	5.099394987	Med
$\mathrm{G^{GCC}}$	5	55	791	813	22	813	813	0	2.70602706	Low
$\mathrm{G^{TCC}}$	6	147	471	498	27	490	490	0	5.421686747	Med
$\mathrm{H^{GTG}}$	7	122	585	659	74	634	634	0	11.22913505	High
LS	5	63	920	962	42	961	961	0	4.365904366	Low
LT	6	80	905	936	31	935	935	0	3.311965812	Low
MR	4	68	932	1033	101	1031	1031	0	9.777347531	Med
NK1	2	37	990	1006	16	1006	1006	0	1.590457256	Low
NK2	6	49	1234	1251	17	1251	1251	0	1.35891287	Low
$\mathrm{P^{TGG}}$	1	1	N/A	N/A	N/A	761	761	0	N/A	N/A
R5	10	102	766	850	84	846	846	0	9.882352941	Med
RT	3	47	960	965	5	964	964	0	0.518134715	Low
$\mathrm{R^{TCT}}$	7	130	634	694	60	686	686	0	8.645533141	Med
SD	6	76	750	774	24	767	767	0	3.100775194	Low
SPPCK	6	39	1712	1777	65	1775	1775	0	3.65785031	Low
SQCK	10	85	1340	1407	67	1403	1403	0	4.761904762	Low
TQ	2	53	748	787	39	767	767	0	4.955527319	Low
TX	3	26	1098	1107	9	1107	1107	0	0.81300813	Low
V5	1	44	795	819	24	812	812	0	2.93040293	Low
VF	14	163	899	973	74	965	965	0	7.605344296	Med
VME5	9	72	1334	1522	188	1466	1466	0	12.3521682	High
WI	24	265	1080	1167	87	1159	1135	24	7.455012853	Med
YE	7	79	843	953	110	938	938	0	11.54249738	High

Variation in unit length among array repeat units was seen in nearly all of the array types though some array units showed more variation in length than others. A variation index was calculated for each array type using the formula $\frac{R}{L} \times 100$, where R is the range in length among the different array units and L the modal length of the array units. Variation was classified into three categories based on how much variation was observed when multiple array units from the same array were aligned. Array units with a variability index of 0-5, 5-10 and 10-15 were classified as having low, medium and high variability, respectively (Table 4.3.2). The majority of array types (14/25) show low inter-unit variation. Of the remaining 11, seven showed medium variation and three showed high variation. Variation for the Pro^{TGG} array could not be calculated as only one copy of the array was assembled into the PacBio genome.

Inter-unit variation occurs almost exclusively in the STR regions between the tRNA genes. The length variation in the all of the arrays classified as having low variation (Table 4.3.2) was explained by small indels (1-3 bp) in regions of the STR where a string of adenines (poly-A regions) occurs. Arrays with medium and high variation also contain these small indels. However, the arrays with higher variation also have different STR copy numbers between units in the same array. On three occasions, the deletion or partial deletion of a whole tRNA gene was observed. This was observed once in the WI, ASD and V5 arrays and however, the deletion of tRNA gene was not correlated with the variation index as the WI, ASD and V5 array units were classified as all having a medium variation index. With the exception of these rare deletions, the tRNA gene sequences were highly conserved between array units.

Despite some variation among STR regions within a single array type, alignment of multiple reads from the same array and manual inspection showed that most copies of an array unit are homogenous in length and sequence. This was confirmed by PCR amplification across a range of different array types (2 low, 2 medium and 3 high variability arrays) followed by Sanger sequencing of the PCR product. Manual inspection of the Sanger chromatograms showed evenly spaced single peaks and did not show any instances of mixed peaks indicative of

SNPs or indels in the tRNA array units (Figure 4.3.2). This suggests that inter-unit variability is below the level that it can be detected by Sanger sequencing and the majority unit type determines the genotype measured by PCR and sequencing.

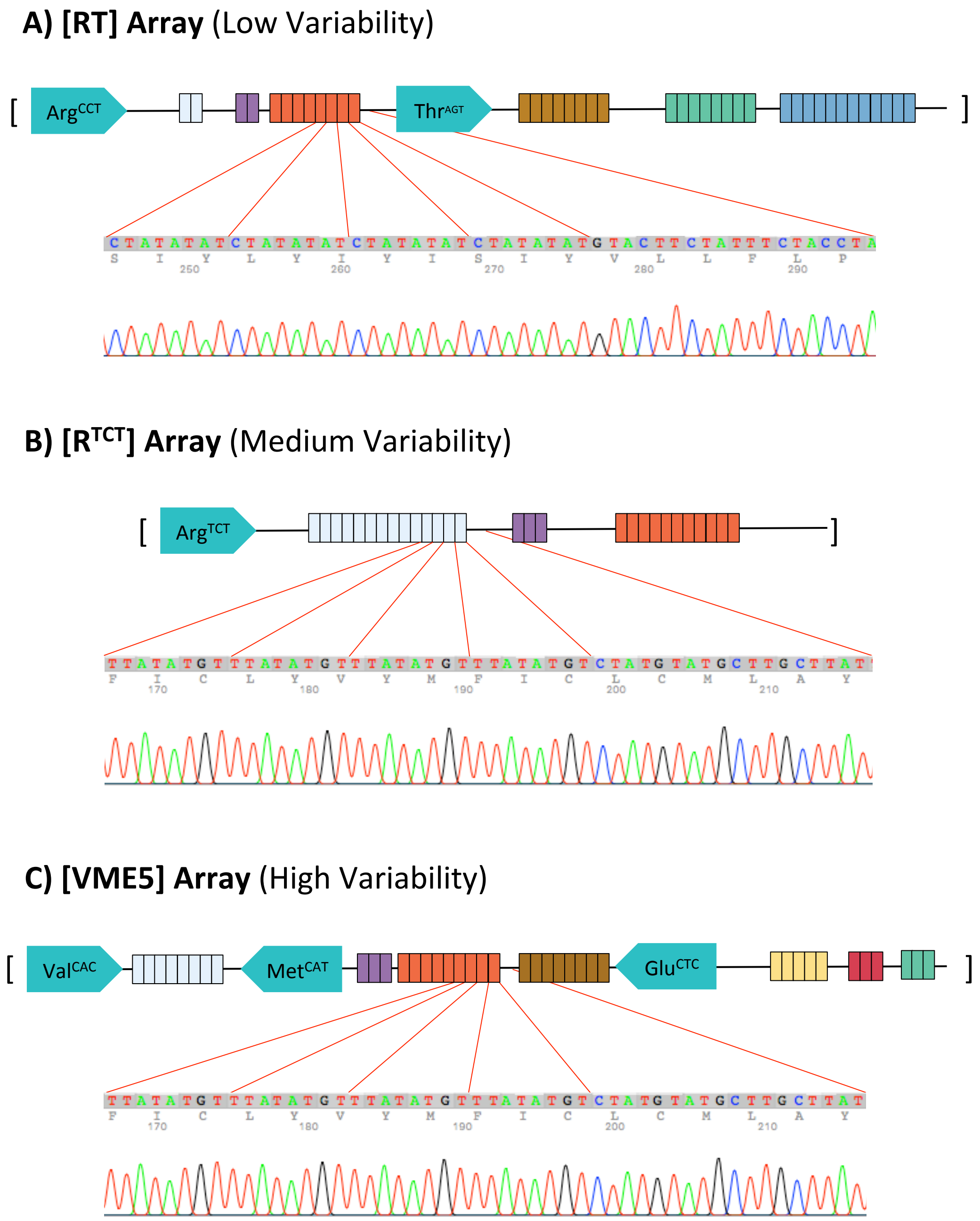


Figure 4.3.2. Sanger sequencing of intergenic STR regions in the tRNA array units. tRNA arrays were classified as having low, medium or high sequence variability ($\frac{\textit{Range between longest and shortest array unit}}{\textit{Modal length}} \times 100$)) . The intergenic regions of a range of arrays were sequenced using Sanger sequencing and analysed to determine if sequence variability could be detected. Traces of STR sequences from each variability category are shown (Low = [RT], Medium = [R^TCT], High = [VME5]). No variation or mixed bases were detected in any of the STR traces.

4.3.4. Sequence evidence is consistent with the occurrence of tRNA arrays at the end of chromosomes

It has been suggested that tRNA arrays act as telomeres, capping the chromosomes in an mechanism analogous to that seen in *Dictyostelium discoideum* [306]. Manual inspection of contigs containing tRNA array sequence and genic sequence (tRNA-genic contigs) were analysed to identify where the arrays were located within the contigs. 21 tRNA-array contigs were identified and are summarised in Table 4.3.3.

All of the tRNA arrays within the tRNA-contigs were flanked exclusively at one end of the array. There were no occasions where a tRNA array was flanked on both ends by genic sequence. Nearly all of the 21 tRNA-genic contigs contained a distinct tRNA array with the only exception to this being contigs Ehis_175 and Ehis_212 that are both terminated by a VF array orientated in opposite directions, meaning the array could occur within a chromosome (Figure 4.3.3).

Five tRNA array types (LS, NK1, R^{TCT}, TQ and TX) were identified only in contigs entirely comprised of tRNA array units. Therefore, the genomic location of these arrays could not be determined.

In other protists, the repeat unit that forms the telomeres is often a 5-8 bp unit similar to the telomeric sequence identified in humans (TTAGGG, Data from the Telomerase Database, http://telomerase.asu.edu). The repeat is common to all the telomeres in an organism however, the length of these repeats can vary. To test whether any repetitive sequences (outside of those that exist as tRNA arrays) were present at the terminal ends of the contigs in the PacBio assembly, 150 bp sequences from both ends of every contig were analysed using Tandem Repeat Finder. 25 sequences contained a small repeat (<20 bp per unit) however, none of these repeats occurred at the end of two or more contigs.

Table 4.3.3. tRNA-genic contigs summary. tRNAscan-SE was used to identify tRNA genes in the genome assembly. Manual inspection of the output determined which contigs contained both genic sequence and tRNA sequence (tRNA-genic contigs).

Contig	Contig Length (bp)	tRNA array type terminating the contig	Length of terminating tRNA array (bp)	tRNA Array location [A]
Ehis_002	486,879	SPPCK	9,419	Terminal
Ehis_013	274,584	WI	27,459	Terminal
Ehis_044	159,583	V5	36,008	Terminal
Ehis_064	108,674	VME5	9,757	Terminal
Ehis_071	99,721	SD	3,981	Terminal
Ehis_119	61,257	ALL	12,267	Terminal
Ehis_130	53,025	P^{TGG}	795	Terminal
Ehis_164	40,748	YE	11,091	Terminal
Ehis_175[B]	38,079	VF	7,785	Terminal
Ehis_187	34,451	SQCK	11,104	Terminal
Ehis_190	34,029	A^{AGC}	861	Terminal
Ehis_212[B]	28,385	VF	16,174	Terminal
Ehis_222	25,995	G^{GCC}	5,965	Terminal
Ehis_224	25,606	G^{TCC}	11,100	Terminal
Ehis_234	23,661	H^{GTG}	6,058	Terminal
Ehis_241	23,049	ASD	1,757	Terminal
Ehis_253	21,689	LT	16,985	Terminal
Ehis_255	21,287	RT	12,683	Terminal
Ehis_295	15,011	NK2	11,799	Terminal
Ehis_334	11,898	MR	1,013	Terminal
Ehis_405	5,311	R5	1,859	Terminal

[A] tRNA array location was determined as being terminal (i.e. tRNA array flanked at one end only by genic sequence) or intergenic (tRNA array flanked on both ends by genic sequence)

[B] Two arrays were found to contain the same tRNA type (VF). These are highlighted in the table

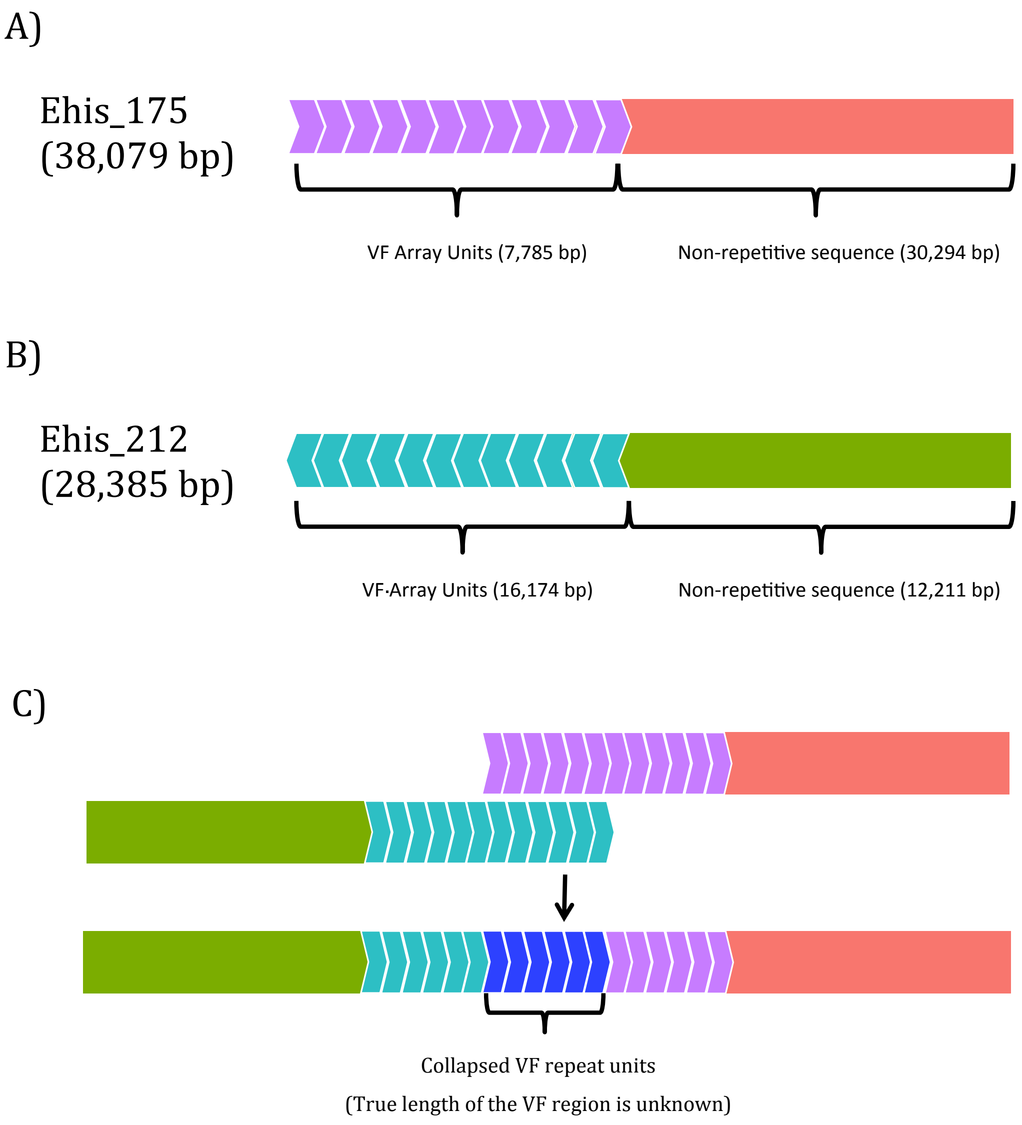
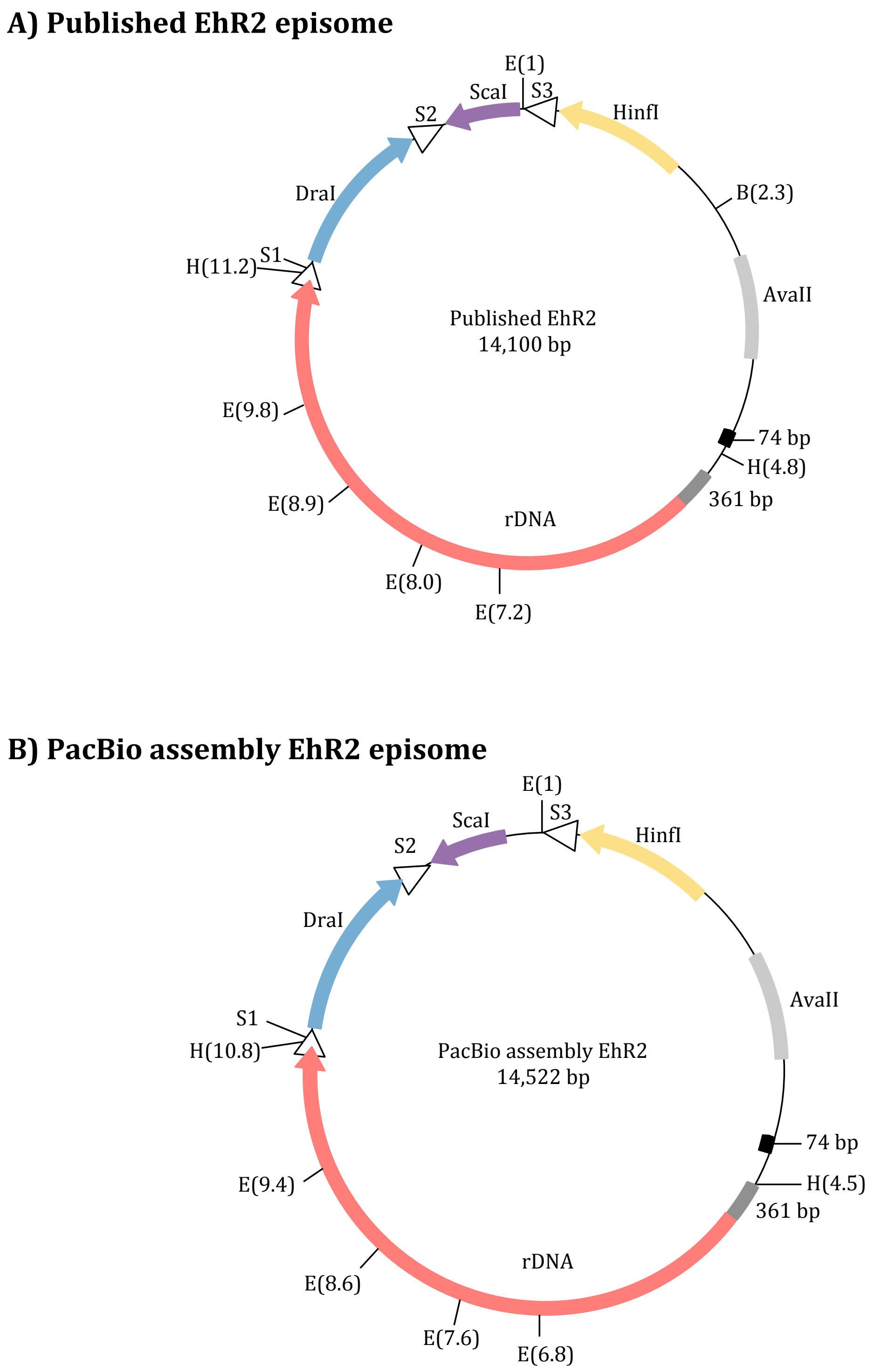

A)
Ehis_175
(38,079 bp)
VF Array Units (7,785 bp)
Non-repetitive sequence (30,294 bp)
B)
Ehis_212
(28,385 bp)
VF-Array Units (16,174 bp)
Non-repetitive sequence (12,211 bp)
C)
Collapsed VF repeat units
(True length of the VF region is unknown)

Figure 4.3.3. The 'VF' tRNA array may be a single array flanked by non-repetitive chromosomal sequence. Two tRNA-genic contigs terminate with the VF array: Ehis_175 (Panel A) and Ehis_212 (Panel B). The VF arrays of these two scaffolds are orientated in directions that would allow them to be aligned such that the resulting contig would contain an internalised VF array (Panel C). The true size of the resultant internalised VF array cannot be determined.

4.3.5. Sequence evidence suggests loss of the EhR1 rDNA episome from the *E. histolytica* HM-1:IMSS cell line sequenced in this study

The rDNA genes have been previously found exclusively on extra-chromosomal plasmids that contain either 1 (EhR2) or 2 transcriptional rDNA units (EhR1) [125,241,311]. To detect these sequences in the new *E. histolytica* HM-1:IMSS genome assembly, the rDNA sequence was used in a BLASTN search to identify contigs containing rDNA genes. These were then restriction digested *in silico* to check if the restriction pattern corresponded with the patterns observed in EhR1 and EhR2 [110].

No chromosomal rDNA genes were found in the new PacBio assembly. Multiple contigs were assembled in the preliminary assemblies containing a 14 Kbp section of sequence repeated multiple times. Splitting the contigs into individual 14 Kbp repeats and then aligning these to create a consensus sequence was used for analysis. All of the rDNA genes were found in this 14 Kbp sequence in the same orientation and distance from each other seen in the reported EhR2 episome. *In silico* digestion of the 14 Kbp sequence also identified that the same short tandem repeat families found in the published EhR2 episome were present in the 14 Kbp sequence fragment (in the same orientation). Finally, *in silico* digestion of the 14 Kbp fragment using HindIII and EcoR1 resulted in a very similar restriction enzyme digestion pattern that has previously been determined (Figure 4.3.4), further suggesting the 14 Kbp fragment is the EhR2 episome.



A) Published EhR2 episome

B) PacBio assembly EhR2 episome

Figure 4.3.4. Sequence organization of the *Entamoeba histolytica* HM-1:IMSS EhR2 episomes as previously identified EhR2 (Panel A) and characterized from PacBio sequencing (Panel B). Restriction enzyme sites indicated on the circles are EcoR1 (E), HindIII (H) and BamHI (B). Short tandem repeat families are marked as PvuI, ScaI, HinfI, AvaII, 74 bp and DraI with orientation indicated by arrows.

The EhR1 episome could not be found within the new PacBio assemblies. To ensure that sequence reads from the EhR1 episome were not misassembled into the similar EhR2 episome, raw reads from the PacBio were used in a string match query on command line search using a unique region of the EhR1 as a query. The EhR1 episome contains a PvuI short tandem repeat family that has been lost from the EhR2 episome during recombination. This restriction site (5'-CGAT^CG-3') was not found within the raw reads, suggesting that the EhR1 episome really was absent from the cell line.

Mapping of short paired end Illumina reads generated for *E. histolytica* HM-1:IMSS revealed an average depth of coverage across the extra-chromosomal EhR2 episome to be approximately 530,000x compared to an average depth of coverage across the genomic portion of the genome of 2,579x. This suggests that the EhR2 episome has a relative copy number of approximately 200 copies (530000/2579 = 205.5).

4.4. Discussion

4.4.1. Intra-genome variation occurs within the tRNA short tandem repeat (STR) sequence

The tRNA STRs are used as population genetic markers for strain typing of *Entamoeba* species [230]. Researchers have attempted to identify specific STR genotypes linked to phenotypes, such as virulence. Such work relies on STR markers that consistently produce a single signal for an individual genome/strain. A mixed signal for an STR within a strain would invalidate a marker for genotyping analyses. The tRNA STRs are potentially poor markers because they exist in multiple copies per genome, so that appreciable differences among different copies could create a mixed signal for a marker. Previous analysis predicted that most copies of an array unit are homogenous in length and sequence [4]. Here, it was possible to investigate this further, owing to the long reads produced from SMRT sequencing and the assembly of multi-unit tRNA arrays.

Intra-genome, and intra-array, variation among the tRNA STRs was observed, to different degrees for different arrays. Variation among array units on the same read is evidence of intra-genome variation (within a single trophozoite). Variation seen among reads (and among assembled contigs) could indicate inter-genome variation, as the assembly was generated from a pool of trophozoites. Despite the variation observed, PCR and Sanger sequencing showed that it was not sufficient to have an effect on the consensus sequence produced (producing a mixed genotype), meaning that the tRNA STRs are probably reliable markers (in this case). However, the arrays are likely to be evolving by concerted evolution, whereby duplications of a STR sequence that is different from the majority sequence occurs and eventually replaces the majority sequence by slipped-strand mis-pairing and gene conversion to become the new stable STR sequence. Given that this is a different molecular evolutionary process than both the stepwise mutation of microsatellites and the mis-incorporation and substitution of single nucleotides, care should be taken in how the markers are used.

4.4.2. Frequencies of tRNA genes in arrays are not correlated with codon usage bias

The PacBio *Entamoeba histolytica* HM-1:IMSS genome assembly suggests that at least some copies of each distinct tRNA gene within the tRNA arrays must be functional. This is because only seven distinct tRNA isoacceptor genes are assembled outside of the tRNA arrays. The tRNA genes within the tRNA arrays are also highly conserved and show almost no variation between copies of the same tRNA gene within and between array units suggesting that selection is occurring on these genes to conserve their sequence and function.

The tRNAs genes that exist in arrays occur in huge numbers (~4300 genes) in the array units and evidence large sequence depth when reads were mapped back the gene sequences of the tRNA genes. The few tRNA genes that exist as dispersed copies on the genome exist as far fewer copies and have lower coverage depth. Despite the range in abundances of individual tRNA genes, no correlation exists with the codon usage in protein coding sequences. Further, no bias was observed in coverage of those tRNA genes that encode codons with an adenine or thymine at the degenerate base position (wobble base) despite the genome being very AT-rich. Owing to the redundancy seen in the tRNA genes and the observation of no correlation between abundance and usage, means there is no evidence for strong selection on copy number caused by codon usage bias (or selection on codon usage caused by the copy number of tRNA gene types). The lack of strong selection may suggest that the tRNA arrays serve another purpose within the *Entamoeba histolytica* HM-1:IMSS genome such as providing a structural role or regulatory role as was suggested through previous identification of S/MARs in many of the tRNA arrays [306].

4.4.3. tRNA array sequence data is consistent with the evidence that tRNA arrays act as telomeres

It has been proposed that tRNA arrays may act as telomeres, capping the ends of chromosomes and protecting them from degradation [306]. The PacBio genome assembly and annotation is consistent with this model. No telomerase gene was identified during gene annotation of the PacBio genome (Chapter 3) and no common repeat was identified within the first or last 150 bp of sequence across any of the contigs in the PacBio assembly, suggesting that a short (<20bp) repeat does not form the telomeres in *E. histolytica*. The assembled tRNA arrays were never flanked on both ends by non-repetitive DNA, but a small set (21/563) of contigs was flanked on one end by non-repetitive DNA (tRNA-genic contigs). With the exception of two of these contigs, all are terminated with a different array type. If this was due to the assembler being unable to assemble a complete internal array and both flanking, non-array regions, we would expect to see more cases like that of the VF array, where two different tRNA-genic contigs end with the same tRNA array type, in orientations which could allow them to be scaffolded in such a way that an internalised tRNA array exists. We see this only once, suggesting that the other 19 arrays occur at the ends of chromosomes. It is proposed that the *E. histolytica* HM-1:IMSS contains 14 chromosomes (hence, 28 telomeric ends). 25 tRNA array units have been previously identified (all of which were confirmed in the PacBio assembly), potentially accounting for 12 chromosomes (+1 end). The new assembly has 21 tRNA-genic contigs that could account for 9 chromosomes (+1 end) and one internal array, or 10 chromosomes (+1 end). It remains to be seen if the 4 arrays that do not occur in tRNA-genic contigs are anchored to the non-array genome, or possibly occur as episomes or mini-chromosomes. A final piece of circumstantial evidence for a telomeric role is that many of the STRs and DNA regions surrounding the tRNA genes are highly methylated, accounting for the majority of methylation in the *E. histolytica* genome (Chapter 5). DNA methylation has been associated with telomere length and stability and may contribute to the formation of heterochromatic regions in the genome and transcriptional silencing [327]. Telomeres are constitutively heterochromatic, and the presence of DNA methylation of the tRNA STRs may suggest that the tRNA arrays are organised into a similar heterochromatic state [328]. Methylation of the tRNA array units is discussed in greater detail in Chapter 5.

4.4.4. rDNA episome sequences differ from those previously identified in *Entamoeba histolytica* HM-1:IMSS

The absence of the any *PvuI* cut sites in any of the sequence reads indicates that EhR1 has been lost from the *Entamoeba histolytica* HM-1:IMSS genome whilst it has been in culture. This is because both the strain cultured at the University of Liverpool and the strain analysed for rDNA sequence at Jawaharlal Nehru University (where much of the work on the rDNA episomes has been carried out) are from the same original HM-1:IMSS strain. This is significant as it highlights the plasticity of the rDNA episomes, suggesting that the rDNA episomes may be redundant and the loss of one of these from *E. histolytica* is not detrimental to the survival of *E. histolytica* trophozoites *in vitro.* This is of particular importance as often regions from EhR1 are used to screen for *Entamoeba histolytica* positive samples [329,330]. Though some of the marker and probes generated from EhR1 are represented in the EhR2 episomes, some markers are not such as those generated from the upstream region of the rDNA that is lost when EhR2 is generated from recombination of EhR1. This upstream region contains many unique sequences including the *Tr* region that is transcribed into a polyadenylated 0.7 Kbp RNA detectable by northern blots [110]. Detection of this *Tr* region using PCR has been suggested as a method for diagnosis of *E. histolytica* in a clinical setting [331]. If the loss of EhR1, and the *Tr* region, *in vitro* is not an adaptation to culture and is resultant of rDNA gene redundancy between EhR1 and EhR2, it is not unrealistic to assume that rDNA episomes can be lost from populations *in vivo*. If the EhR1 episome was lost from populations *in vivo* they would not be detected using *Tr*-based genotyping leading to false-negative diagnoses and under-reporting of *E. histolytica* infections.

The working model of how EhR2 was produced suggests that intra-molecular recombination of direct repeats in EhR1 resulted in two half molecules of which only one was retained by the cell (EhR2). EhR1 contains two copies of the rDNA (rDNA I and rDNA II) of which only rDNA I is retained in EhR2 after recombination. The EhR2 molecule contains all of the DNA sequence required for successful transcription of the rDNA genes including the rDNA sequence itself as well as promoters and enhancer sequences that have been previously determined [332]. The lost PvuI short tandem repeat family occurs upstream of the rDNA II transcriptional unit on EhR1 and a HinfI short tandem repeat family occurs upstream of the rDNA I unit. The HinfI repeats have been suggested as acting as enhancers which may be required for the efficient transcription of the rDNA I unit [241,333]. This suggests that the majority of the rDNA molecules transcribed in the *Entamoeba* cell may arise from the transcription of rDNA I. As it is rDNA I and HinfI repeats that are maintained in the EhR2 episome it could be hypothesised that EhR2 has the capacity to produce enough copies of the rDNA genes to support the cell despite only containing one copy of the rDNA. This is because it has retained the more transcriptionally active copy.

4.5 Conclusions

In this chapter, the repetitive DNA features (tRNA arrays and rDNA episomes) have been analysed in the new PacBio assembly. There was previously no knowledge of the lengths of the tRNA arrays and no quantitative information on variation between units in the same arrays. This was because, before SMRT sequencing, the longest sequence reads producible were limited to around 1 Kbp, not long enough to span multiple array units. PacBio assembly allowed for array units to be assembled into longer array structures (up to 43.3 Kbp here), which allows for variation between units in a single array to be quantified. Variation was quantified for each array type and demonstrated that the different tRNA arrays show variable levels of intraspecific variation in the STR regions. However, amplification of these variable regions followed by Sanger sequencing suggested that one major STR type is present within each array and therefore, markers that are designed to target these STR regions are most likely accurate.

Further investigation into the function of the arrayed structure was explored, building on the hypothesis that the tRNA arrays act as telomeres in the *Entamoeba* genomes. No telomerase or common short repeat (indicative of telomeric sequence) was observed in the PacBio genome, suggesting an alternative telomere structure is present in *E. histolytica.* Before the generation of the PacBio genome, tRNA array structures had not been linked to any sequence from the 'core' genome. The PacBio assembly linked 21 tRNA arrays directly to non-array regions of the genome providing a set of genes that occur in close proximity to the tRNA arrays and facilitated analyses in Chapter 3. The tRNA arrays that were attached to protein coding sequences were always found to terminate the contigs and there was only one occasion where two tRNA arrays could be orientated in such a way that would facilitate scaffolding of an internalised tRNA array. This novel discovery provides further support to the tRNA telomere theory previously proposed and is explored further in Chapter 5.

Analyses of the rDNA episomes performed on the PacBio assembly concluded that one of the rDNA episomes has been lost *in vitro.* Only EhR2 was fully assembled in the PacBio assembly and no EhR1-specific sequences were found, further confirming the loss of EhR1 *in vitro*. Relative copy number analyses determined that the depth of coverage of the EhR2 episome was ~200 times as deep as the coverage for the chromosomal portion of the genome, indicating a relative copy number of ~200 for the EhR2 episome which is consistent with previous reports.

Chapter 5 – Genome-wide study of 5-Methyl-cytosine methylation in *Entamoeba*

5.1 Introduction

5.1.1. Epigenetics

Epigenetics is the study of heritable changes to DNA that do not involve a sequence change but do alter transcription and protein expression. Epigenetic modifications have been identified in a wide range of living organisms from simple prokaryotes and eukaryotes to multicellular organisms including plants, animals and humans. Epigenetic modifications (or 'marks') include DNA methylation [334] and modifications (phosphorylation, ubiquitination, acetylation and methylation) to histone proteins, which condense the DNA [335,336]. Collectively, these modifications result in changes to the chromatin structure that affect its accessibility to transcription factors (TFs) [171] and other proteins, such as methyl-binding domain proteins that interact with the DNA [172], to modify gene transcription.

DNA methylation is the addition of a methyl group ($-CH_3$) to a cytosine or adenine base in DNA. It occurs in bacteria, plants, fungi and animals [327,337]. Cytosine methylation (commonly 5-methylcytosine, in which the methyl group is added to the 5th atom of 6 in the cytosine ring) represents an important epigenetic mark that affects gene expression in a range of species [327,338]. Although DNA methylation is phylogenetically widespread, genomic patterns of methylation (a possible function) show considerable variation [337]. For example, vertebrate genomes show extensive DNA methylation; 3-8% of cytosines in mammals are normally methylated, generally these occur in a CpG context, where a methylated cytosine is followed by a guanine [339]. Conversely, many invertebrate genomes display low or no DNA methylation

[327,337,340]. This variation in genome methylation patterns across different taxa suggests that the role of DNA methylation may vary among species.

DNA methylation has largely been associated with gene silencing and with the control of transposons [341,342]. DNA methylation can silence genes (and transposons) by recruiting methylated CpG binding domain (MBD) proteins which interact with histone deacetylase to condense chromatin around the gene (or transposon), repressing gene expression [343]. In some species, novel functions have been suggested for DNA methylation. DNA methylation in the Honey Bee, *Apis mellifera* [344], appears to be directly associated with the differentiation of castes in this social species [345,346] and the down-regulation of a key DNA methyltransferase (Dnmt3) results in profound changes in caste development trajectories in this organism. As such, DNA methylation may represent an important mechanism in facilitating the evolution of certain social systems [347].

Recent work in protist parasites has suggested that epigenetics is an important factor in virulence, differentiation and lifecycle control in *Toxoplasma gondii, Plasmodium falciparum* and *Trypanosoma brucei* [200–203]. It has been hypothesised that DNA methylation could be responsible for the different transcriptomic profiles seen in virulent (e.g. HM-1:IMSS) and avirulent (e.g. Rahman) *E. histolytica* strains [205].

5.1.2. Identification of 5-methylcytosine

5-cytosine methyltransferase proteins (m5C-MTase) catalyse the attachment of a methyl group to the 5th atom of 6 in the cytosine; the resulting molecule is known as 5-methylcytosine (5-MeC). The mammalian DNA methylation machinery consists of three DNA methyltransferases (MTases), Dnmt1, Dnmt3a and Dnmt3b. Dnmt1 acts as a maintenance DNA MTase, maintaining the methylation of hemi-methylated DNA regions following mitotic events [348,349]. Dnmt3a and Dnmt3b are *de novo* DNA MTases and act on unmethylated DNA [339]. A fourth DNA MTase, Dnmt2, is the most conserved of all the MTases and has been identified in all species from yeasts to humans [350]. It is generally considered as having weak DNA methylating activity and is regarded as an RNA methyltransferase (RNA MTase) [351] however, more recently it has been discovered that Dnmt2 catalyses all of the DNA methylation in Dnmt2-only organisms such as *Drosophila* and *Dictyostelium* [352–355].

Bisulphite sequencing (Figure 5.1.1) is a useful tool for studying the methylation status of 5-MeCs and is the gold standard for analysing 5-MeC DNA methylation. Sodium bisulphite treatment of the DNA deaminates unmethylated cytosine residues converting then to uracils. Methylated cytosines are unaffected and therefore when amplified using PCR and subsequently sequenced, these sites are represented by a cytosine whereas unmethylated cytosines are represented by a thymine [356]. This allows effective discrimination of each cytosine residue by analysing the proportion of mapped reads that contains either a cytosine (a methylated site) or a thymine (an unmethylated site) at each cytosine site in the reference.

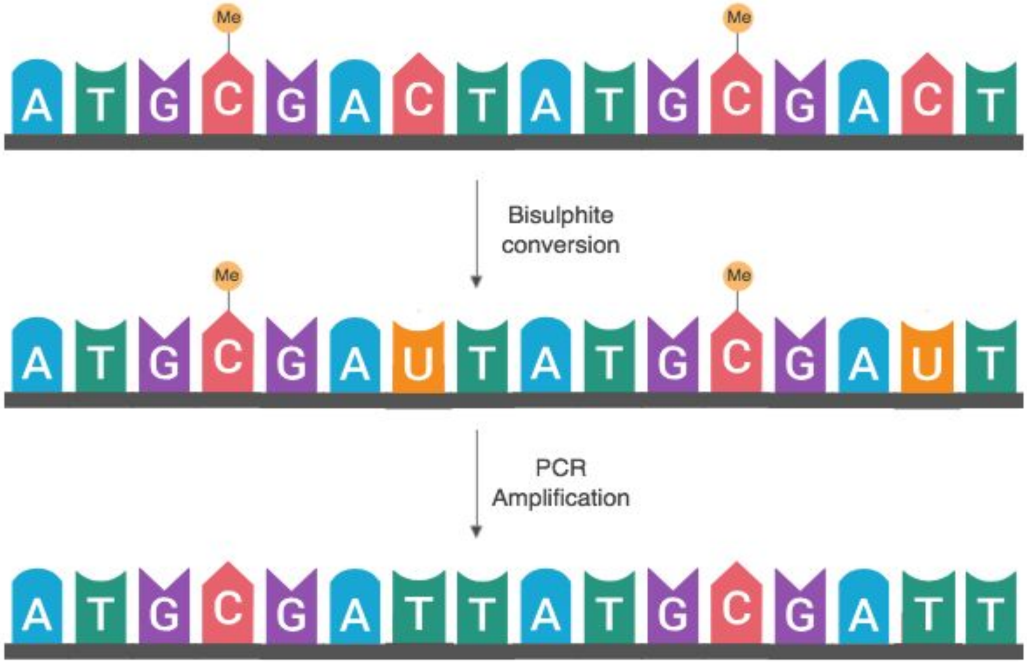

Me
Me
A T G C G A C T A T G C G A C T
Bisulphite conversion
Me
Me
A T G C G A U T A T G C G A U T
PCR Amplification
A T G C G A T T A T G C G A T T

Figure 5.1.1. Bisulphite conversion of DNA and subsequent analyses. Treatment of DNA with sodium hydroxide converts unmethylated cytosines to uracils. Methylated cytosines remain unchanged. Following PCR amplification, unmethylated cytosines are represented by a thymine and methylated cytosines remain as cytosines.

The bisulphite treated DNA can be used as a template for PCR and Sanger sequencing to identify DNA methylation in the specific region amplified or, whole genome sequencing (WGS) of the bisulphite treated DNA can be performed to attain a global picture of genome methylation.

SMRT sequencing allows the detection of DNA methylation without the need for bisulphite conversion [177,357]. In PacBio SMRT sequencing, DNA polymerase-catalysed incorporation of fluorescently labelled nucleotides is recorded in real time, so the arrival times and durations of the resulting fluorescence pulses yield information about polymerase kinetics. This allows direct detection of modified nucleotides in the DNA template, as different modified bases have a distinct effect on these kinetics. Detectable modifications include 5-methylcytosine, N6-methyladenine and 5-hydroxymethylcytosine [177,357].

5.1.3. Epigenetics in *Entamoeba*

As mentioned above, the mammalian DNA methylation machinery consists of four DNA methyltransferases, each with distinct roles. However, some organisms have the entirety of their DNA methylation catalysed by a single DNA methyltransferase, Dnmt2, despite this being largely regarded as an RNA methyltransferase.

E. histolytica is thought to belong to this group of Dnmt2-only organisms and not contain any of the canonical DNA methyltransferases (Dnmt1 and Dnmt3). The Dnmt2 in *Entamoeba histolytica* (EhMeth) has been identified as a genuine DNA MTase and methylated DNA has been identified via methylated DNA immunoprecipitation (MedIP) using 5-MeC antibodies [173]. More recently, high pressure liquid chromatography (HLPC) and mass spectrometry (MS) has estimated low levels of 5'-methylcytosines (5-MeCs) at around 0.05% of *E. histolytica* DNA [173]. Results from the MedIP suggested some of the methylated sequences were ribosomal DNA (*rDNA*), heat-shock genes (HSP70 and HSP100), and retro-transposons [358,359]. The effect of EhMeth DNA methylation on gene expression is still not understood. A correlation between

DNA methylation of HSP100 and gene silencing was reported [359] however, treatment of the parasite with 5-azacytidine (a compound that blocks MTase proteins) had little effect on gene expression of *E. histolytica* [360]. Interestingly, the treatment of *E. histolytica* with 5-azacytidine did reduce the trophozoite's ability to form liver abscesses in infected hamsters, suggesting that EhMeth may have a role in controlling virulence of the parasite [361].

One function of DNA methylation in higher eukaryotes is to protect the organism from transposable elements [362]. DNA methylation is thought to occur in the transposable elements that litter the *E. histolytica* genome [97,98]. It is known from the original sequencing efforts that many of the transposable elements have lost their transcriptase ability and it has been suggested that the accelerated deamination that occurs to methylated cytosines may accelerate this process [174]. Dnmt2-mediated control of retro-transposons has already been demonstrated in other Dnmt2-only organisms such as *Entamoeba's* closest sequenced relative, *Dictyostelium discoideum* [123], as well as in *Schistosoma mansoni* [363] *and Drosophila* spp [354].

5.1.4. Aims of Chapter

Although methylation is believed to be a functional mechanism in *Entamoeba histolytica* and is thought to be similar to those seen in mammalian cells and model systems, protozoan parasites have consistently shown diverse and unique mechanisms that control epigenetic gene regulation [364–366]. This chapter will build on previous observations of low levels of DNA methylation in *E. histolytica* by determining where specifically this 5-methylcytosine methylation occurs in the genome. Specifically this chapter aims to:

- Detect the level of 5-methylcytosine methylation and characterize its distribution across the *E. histolytica* genome.

- Determine if any correlation exists between methylation and gene expression utilizing *E. histolytica* RNA-seq data and life cycle RNA-seq data from *Entamoeba invadens* IP-1.
- Identify if any methylated genes are involved in virulence of the parasite
- Determine if methylation is protective to the genome by methylating the abundant transposable elements.
- Identify if DNA methylation occurs in other *Entamoeba* spp (*E. moshkovskii* and *E. invadens*) and if so, to what level does this occur and are orthologous genes across *Entamoeba* spp differentially methylated?

5.2 Materials and Methods

5.2.1 Generation of gDNA, library preparation and sequencing of *E. histolytica* HM-1:IMSS, *E. invadens* IP-1 and *E. moshkovskii* Laredo

Entamoeba histolytica HM-1:IMSS genomic DNA was isolated as described in section 2.2.2.3. *E. invadens* IP-1 genomic DNA was prepared from trophozoites by Dr. Gretchen Ehrenkaufer (Singh Lab, Stanford University, USA) and permission to use this DNA in bisulphite sequencing was granted by Prof. Upinder Singh (Stanford University, USA). *E. moshkovskii* Laredo genomic DNA was prepared from trophozoites by Dr. Gareth Weedall (Liverpool John Moores University, Liverpool, UK) and he granted permission for use.

Genomic DNA (*E. histolytica, E. invadens* and *E. moshkovskii*) was submitted to the CGR (Liverpool, UK) for library preparation and sequencing. Briefly, samples were subjected to bisulphite conversion using Zymo EZ DNA Methylation-GoldTM Kit (Irvine, CA, USA). The product was used to produce a sequencing library using an Illumina DNA methylation kit (San Diego, CA, USA). This converts the single stranded DNA (ssDNA) into a next generation sequencing (NGS) library for sequencing. Bisulphite treated ssDNA is randomly primed using a polymerase able to read uracil nucleotides to synthesize DNA strands with a specific tag sequence. The 3'-ends are tagged with another specific tag sequence. These tags enable enrichment of sequences through polymerase chain reaction (PCR) (10 cycles). The final library was checked for quantity, purity and size before being pooled and sequenced on an Illumina HiSeq4000 to generate 2 x 150 bp paired-end reads.

Truseq Nano paired end libraries for *Entamoeba invadens* IP-1 and *Entamoeba moshkovskii* Laredo with an insert size of 350 bp were also produced and sequenced as a comparison for the bisulphite-treated libraries (one had already been sequenced for *E. histolytica*). Sequencing was performed on an Illumina HiSeq4000 to generate 2 x 150 bp paired-end reads. The protocol is outlined in section 2.2.5.

Sequence reads were initially processed by CGR. Illumina adapter sequences were trimmed from the raw reads using Cutadapt (Version 1.2.1.) [367]. The option -O 3 was used, so the 3'-end of any reads that match the adapter sequence for 3 bp or more are trimmed. Reads were further trimmed using Sickle (Version 1.200) [368] with a minimum window quality score of 20. Reads shorter than 20 bp after trimming were removed.

5.2.2. Mapping methyl-seq reads and bias detection using Bismark

Mapping of the bisulphite reads to the references were performed using Bismark (Version 0.18.1) [369], SAMTools (Version 0.1.18-r580) [236] and Picard Tools (Version 1.85) [370]. Formatting of the output was performed using Perl scripts (provided by Dr. Laura Gardiner, Earlham Institute, Norwich, UK) and Awk commands on the command line and is outlined in Figure 5.2.1 (Command line details in S5.2. Methylation_protocol.sh, Appendix 5). The *E. histolytica* HM-1:IMSS bisulphite reads were mapped to the new PacBio reference described in Chapter 2. The *E. invadens* IP-1 and *E. moshkovskii* datasets were mapped to the latest version of each genome available on AmoebaDB (AmoebaDB.org) at the time of mapping (AmoebaDB, Release 35 for both species).

Reference genomes were indexed using the Bismark Genome Preparation step. Bismark was then run with default parameters, using Bowtie2 as the specified mapping program. The BAM file produced was sorted and indexed using SAM Tools and duplicate reads removed using the Picard Toolkit Markduplicates program to produce a new BAM file with duplicate reads removed. This file is sorted and indexed as before and used as input to the Bismark Methylation Extractor.

The Bismark Methylation extractor was run using default parameters and the sensitivity flag (-s) as ‘comprehensive’ to identify the state of each cytosine in each read. Cytosines are divided into three files based on flanking nucleotides;

these are CpG_sites.out, CHG_sites.out and CHH_sites.out where H represents any non-G base (A, T, C) and CpG represents cases were a methylated cytosine is followed by a guanine (CG). The Methylation Extractor also produced an M-bias plot. The M-bias plot shows the methylation proportion across each possible position in the read. Library construction of standard directional BS-Seq samples consists of several steps including sonication, end-repair, A-tailing and adapter ligation. End repair of directional BS-Seq libraries results in artificial hypermethylation at the end of read 2 of paired-end BS-Seq libraries. This will add spurious hypomethylated calls if not removed. The M-bias plots display any hypermethylation across the reads. M-bias plots were manually inspected and the Methylation Extractor rerun using the –ignore flag if any bias was detected to the ends of the read. For all three BS-Seq libraries, the Methylation Extractor was re-run using this flag (–ignore 8) to ignore the first 8 bp of the reads as these showed bias when calling methylated cytosines.

Perl scripts produced by Laura Gardiner (IBM Research, Daresbury, UK) was used to quantify the number of 5-methylated cytosines mapped to each cytosine in the reference (S5.3. Are_SNP_reads_methylated.pl, Appendix 5). AWK scripts were then used to extract cytosines that had a coverage depth of at least 10x for *E. histolytica* HM-1:IMSS and *E. invadens* IP-1 cytosines and 5x coverage for *E. moshkovskii.* The coverage threshold for *E. moshkovskii* was lowered to 5x owing to the library output being smaller leading to lower average mapped coverage.

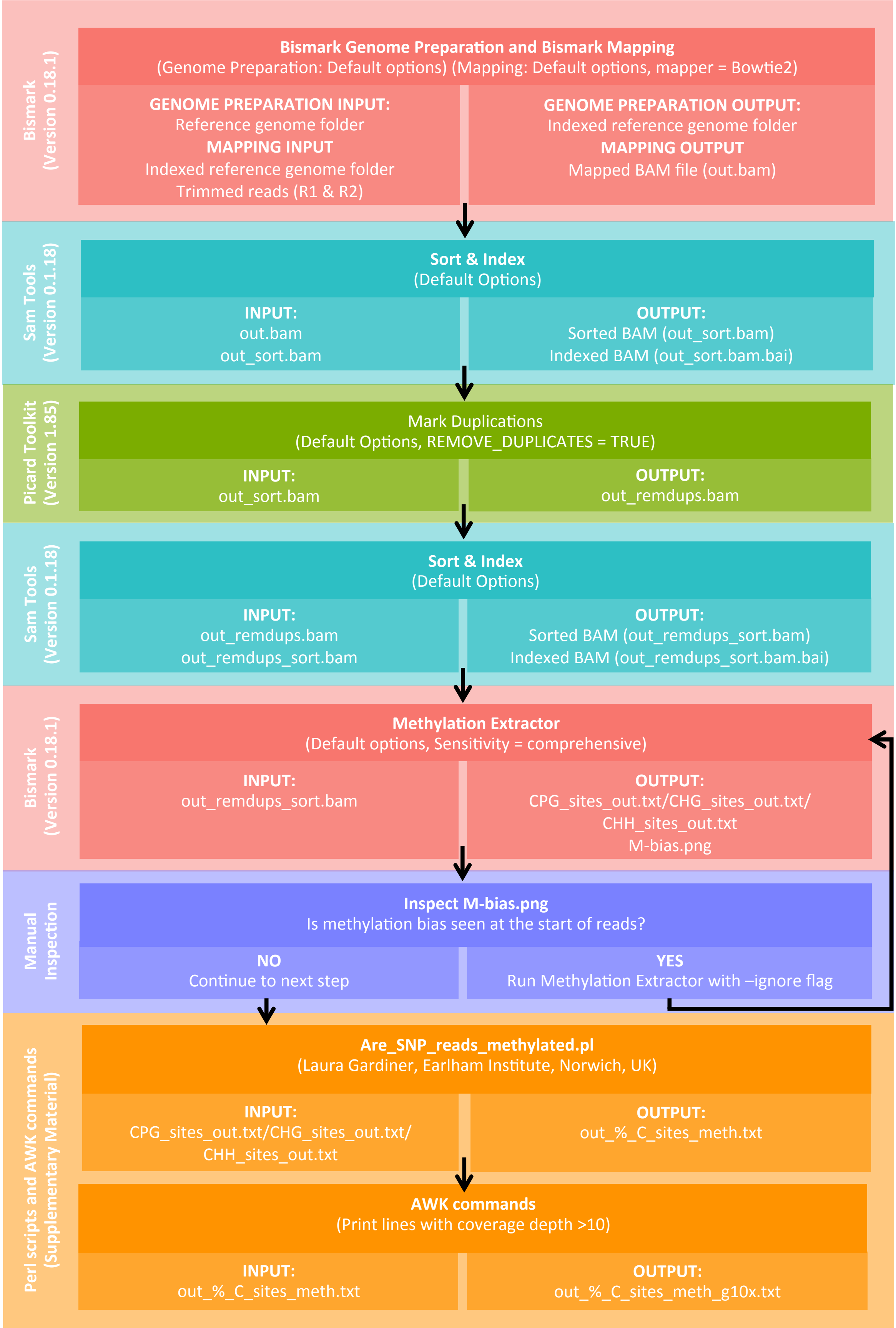

Bismark (Version 0.18.1)
Bismark Genome Preparation and Bismark Mapping
(Genome Preparation: Default options) (Mapping: Default options, mapper = Bowtie2)
GENOME PREPARATION INPUT:
Reference genome folder
MAPPING INPUT
Indexed reference genome folder
Trimmed reads (R1 & R2)
GENOME PREPARATION OUTPUT:
Indexed reference genome folder
MAPPING OUTPUT
Mapped BAM file (out.bam)
Sam Tools (Version 0.1.18)
Sort & Index
(Default Options)
INPUT:
out.bam
out_sort.bam
OUTPUT:
Sorted BAM (out_sort.bam)
Indexed BAM (out_sort.bam.bai)
Picard Toolkit (Version 1.85)
Mark Duplications
(Default Options, REMOVE_DUPLICATES = TRUE)
INPUT:
out_sort.bam
OUTPUT:
out_remdups.bam
Sam Tools (Version 0.1.18)
Sort & Index
(Default Options)
INPUT:
out_remdups.bam
out_remdups_sort.bam
OUTPUT:
Sorted BAM (out_remdups_sort.bam)
Indexed BAM (out_remdups_sort.bam.bai)
Bismark (Version 0.18.1)
Methylation Extractor
(Default options, Sensitivity = comprehensive)
INPUT:
out_remdups_sort.bam
OUTPUT:
CPG_sites_out.txt/CHG_sites_out.txt/
CHH_sites_out.txt
M-bias.png
Manual Inspection
Inspect M-bias.png
Is methylation bias seen at the start of reads?
NO
Continue to next step
YES
Run Methylation Extractor with –ignore flag
Perl scripts and AWK commands (Supplementary Material)
Are_SNP_reads_methylated.pl
(Laura Gardiner, Earlham Institute, Norwich, UK)
INPUT:
CPG_sites_out.txt/CHG_sites_out.txt/
CHH_sites_out.txt
OUTPUT:
out_%_C_sites_meth.txt
AWK commands
(Print lines with coverage depth >10)
INPUT:
out_%_C_sites_meth.txt
OUTPUT:
out_%_C_sites_meth_g10x.txt

Figure 5.2.1. Pipeline for mapping bisulphite treated reads to the reference genome. Bisulphite reads were mapped to the reference genome using Bismark [369], SAMTools [236] and the Picard Toolkit [370]. The alignment was formatted for analysis using Perl scripts and AWK commands.

5.2.3. Setting methylation thresholds and detecting methylated regions

Frequency charts were produced using R (Version 3.1.1.) plotting the percentage of methylated cytosines mapped to each cytosine in the genome. These were inspected to determine the thresholds for methylation calling. If 75-100% of reads mapped to a cytosine in the reference were methylated the cytosine was classified as highly methylated. If 50-75% or 25-50% or reads mapped were methylated the site was classified as medium or low methylated, respectively. If 0-25% of reads mapped to a cytosine were methylated, the site was called as un-methylated to account for incomplete cytosine conversion in the library preparation step and to avoid inaccurate calling of methylated cytosines due to this incomplete conversion. AWK commands were used to create files containing lists of cytosines in each category.

The highly methylated cytosines for each species were extracted and corresponding regions in the genome identified using Homer (AnnotatePeaks tool) [279]. GTF files for genes in *E. histolytica* HM-1:IMSS were produced by Companion (Chapter 3). Gene GTF files for *E. moshkovskii* Laredo and *E. invadens* IP-1 were downloaded from AmoebaDB (Release 35). *E. histolytica* HM-1:IMSS GTF files were manually curated for the tRNA genes and rDNA using the output of the tRNA and rDNA gene identification (Chapter 4). tRNA arrays for *Entamoeba invadens* IP-1 were obtained from NCBI (EF421262-80) and identification of tRNA arrays for *E. moshkovskii* Laredo was carried out using the same methods outlined in section 4.2.1. The output was then manually curated into species-specific GTF files in the same way as those for *E. histolytica* HM-1:IMSS. Homer was used to cross-reference the location of a methylated cytosine with the location of a feature in the GTF files. Homer was run using default parameters specifying the species-specific GTF files as the feature databases for each species.

To create a transposable element GTF file for *E. histolytica* HM-1:IMSS, *Entamoeba* transposable element sequences were obtained from RepBase [371,372]. Sequences were downloaded in FASTA format and used as queries for BLASTN searches of the PacBio *E. histolytica* HM-1:IMSS assembly. Default parameters were used with an E-value cut-off of 0.05. Output was then organised into GTF format for use in Homer searches as previously described.

5.2.4. Mapping RNA-seq data

An RNA-seq library has previously been generated and permission to use the data was provided by Dr Kanok Preativatanyou (Chulalongkorn University, Bangkok, Thailand). Reads were mapped using STAR, an RNA-seq aligner, using default parameters [373]. Counting of reads mapped to each gene was performed using the HTSeq [374] with default parameters. The GFF3 file produced from Companion (Chapter 3) was used as the list of genes provided to HTSeq. FPKM values were calculated using the formula:

$$\frac{Reads\ mapped\ to\ specific\ feature}{Length\ of\ feature\ (Kb) * Total\ number\ of\ mapped\ reads\ (millions)}$$

Histograms of FPKM values were produced with R [237] using the ggplot2 package [238].

5.2.5. Life cycle expression analyses of methylated genes

Life cycle expression data only exists for *E. invadens,* as this species is the only species that can be made to perform its entire life cycle *in vitro.* Methylated genes detected in *E. invadens* were analysed on the AmoebaDB platform (which hosts the life cycle expression data for *Entamoeba* species).

To investigate whether the expression of any methylated genes in *E. histolytica* or *E. moshkovskii* changes over the course of the life cycle, lists of methylated gene were created for *Entamoeba histolytica* HM-1:IMSS and *Entamoeba moshkovskii* were searched on AmoebaDB (AmoebaDB.org) and the record for each gene manually inspected. Known orthologues are listed on the gene record page. These invadens orthologues were also analysed on AmoebaDB and transcription information across the life cycle stages was accessed from these gene record pages.

5.2.6 Identification of *de novo* and maintenance DNA methyl-transferases in *Entamoeba histolytica* HM-1:IMSS

The PacBio *Entamoeba histolytica* HM-1:IMSS assembly annotation (Chapter 3) was searched for any reference to DNA methyltransferases, methylating domains or hits to EhMeth (known Dnmt2 homolog in *Entamoeba* spp). Putative DNA MTase sequences were validated using BLAST. Sequences for *Homo sapiens* and *Arabidopsis thaliana* DNA methyltransferases (DNA MTases) were obtained from UniProt (accession numbers in Table S5.1, Appendix 5; Uniprot.org; Accessed: August 2017). These were Dnmt1, Dnmt2, Dnmt3a and Dnmt3b. These were used in a tBLASTn search against the new PacBio genome for *Entamoeba histolytica* generated in Chapter 2. An e-value cut off of 0.05 was applied and any hits were cross-referenced to the output of the companion genome annotation GFF produced in Chapter 3.

5.3 Results

5.3.1 Cytosine methylation across the *Entamoeba* genomes is sparse

For each BS-seq library, the number of reads produced and mapped using Bismark can be seen in Table 5.3.1. 90.1% of bisulphite treated reads mapped to the *E. histolytica* PacBio genome, consistent with previous estimates of Bismark mapping efficiency [375]. 58.4% of reads mapped uniquely (to a single genome location). Bisulphite sequencing reads often contain high levels of duplicate reads owing to a large loss of DNA during library preparation followed by PCR amplification [376]. Consistent with this, after duplicate removal only 5% of reads remained uniquely aligned. Cytosines in the reference genome with at least 10x coverage were analysed to identify 5-methyl cytosines (5-MeC). Cytosines were considered highly methylated if more than 75% of reads contained a cytosine at this position (i.e. methylation at this position had inhibited C to U conversion). By this criterion, only 0.4% of cytosines with 10x coverage were highly methylated in *E. histolytica* HM-1:IMSS (Table 5.3.1). Even lower proportions of methylated cytosines were detected in *E. invadens* IP-1 and *E. moshkovskii* Laredo, with 0.006% and 0.03% of cytosines with 10x coverage classified as highly methylated in each species, respectively (Table 5.3.1).

To check that complete conversion of the DNA had taken place, each cytosine position was analysed to calculate how many of the reads contained an untransformed (i.e. methylated) cytosine and the frequencies of methylated reads plotted (Figure 5.3.1). A bimodal distribution was observed. The majority of cytosines were <25% methylated, with a much smaller peak at 100% methylated. This bimodal distribution is indicative of complete bisulphite conversion of the DNA and accurate calling of methylated cytosines (incomplete bisulphite conversion produces a flattened distribution without the bimodal distribution).

Table 5.3.1. Bismark mapping statistics of the bisulphite treated libraries. Mapping was performed using the bowtie mapper within the Bismark program. Only uniquely mapped reads are reported.

Sample	Number of reads (millions)	Uniquely mapped reads (millions)	Mapped reads after duplicate removal (millions)	Number of cytosines in reference	Cytosines with 1x coverage (% of all cytosines)	Cytosines with 10x coverage (% of all cytosines)	Cytosines with 10x coverage and >75% methylation (% of 10x cytosines)
Entamoeba histolytica HM-1:IMSS	857.6	501.3 (58.5 % of all reads)	9.35 (1.1 % of all reads)	3,278,194	3,121,861 (95.2%)	1,471,244 (44.9%)	6,211 (0.4%)
Entamoeba invadens IP-1	137.5	72.1 (52.4 % all reads)	4.79 (3.5% of all reads)	6,109,569	5,837,147 (95.6%)	2,418,750 (39.6%)	141 (0.006%)
Entamoeba moshkovskii Laredo	115.0	44.4 (38.6% of all reads)	2.40 (2.1% of all reads)	3,382,334	2,844,288 (84.1%)	999,454 (29.5%)	347 (0.03)

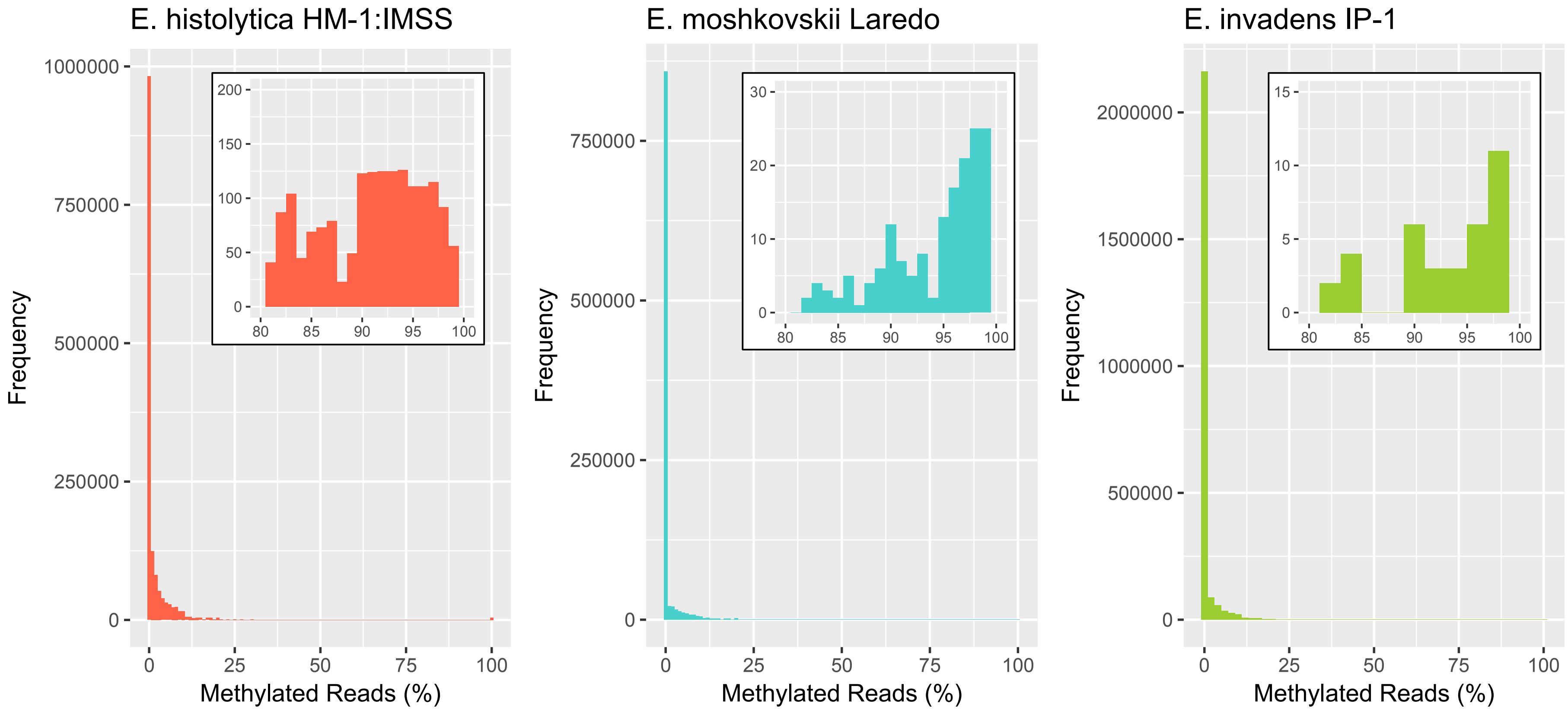


Figure 5.3.1. Frequencies of methylated reads mapped to cytosine positions in three *Entamoeba* genomes. Cytosines with 10x coverage of BS-seq data were analysed to determine the number of reads with a methylated cytosine (plotted on the x-axis) at that position. The frequency of cytosine positions in each methylation percentile are plotted on the y-axis. Bimodal distribution was observed with one peak observed at around 0% methylation and another at 100% methylation (inset for each figure).

Cytosines with <25% methylation were determined as being effectively unmethylated to account for noise. Cytosines with >75% methylation were classified as highly methylated. These highly methylated cytosines were observed in CpG, CHG and CHH sites (C=cytosine, G=guanine, H=non-guanine nucleotide). Only 0.76% (47/6211) of 5-MeCs occurred in CpG. This accounts for 0.06% of the total CG sites in the genome surveyed with 10x coverage. 0.96% (60/6211 of 5-MeCs occurred in CHG cites where a methylated cytosine is followed by a non-G base then a guanine and the remaining 98.3% (6104/6211) of 5-MeCs occurs at CHH sites where H is equivalent to a non-G base.

5.3.2. Methylation of genes is limited to a few genes in each *Entamoeba* species

Highly methylated cytosines were analysed to identify those occurring within the protein coding regions of genes. The Homer AnnotatePeaks function was used to compare a list of methylated cytosine positions to a GTF file of genomic features to determine where the methylated cytosines occur [279]. Highly methylated cytosines in the regions upstream and downstream of genes were not analysed unless these occurred <21 bp upstream of an exon. This is owing to the fact that 5' untranslated regions (5'UTR) in *E. histolytica* are very short (0-21 bp) [151]. The promoter regions of *Entamoeba* are not well characterised, with only very few promoter sequences determined, therefore methylation of these features was not analysed [377].

Nine genes within the *E. histolytica* HM-1:IMSS genome contained one or more 5-MeCs within an exon (Table 5.3.2.). Seven of these genes encode hypothetical proteins and two have functional annotation: one encodes a papain family cysteine protease and another a cysteine protease binding protein. No genes were associated with 5-MeCs in a putative 5'-UTR. Four of the methylated genes in *E. histolytica* contained an orthologous gene in *E. moshkovskii* and *E. invadens.* However, none of the methylated genes in *E. histolytica* were also methylated in *E. moshkovskii* or *E. invadens*.

Table 5.3.2. Methylated genes in the *Entamoeba histolytica* HM-1:IMSS genome.

Gene ID	AmoebaDB ID	Gene Length (bp)	Methylated sites	Methylated Cs in gene (%)	Putative function	Protein features
EHIS_000308900	EHI_107170	1433	32	17.7	Hypothetical protein	Terminal signal peptide
EHIS_000400200	EHI_107170	1433	29	16.0	Hypothetical protein	Terminal signal peptide
EHIS_000047600	EHI_146150	721	10	7.1	Hypothetical protein	Interpro ID: IPR006476 (Conserved hypothetic plant protein)
EHIS_000734900	EHI_127670	914	7	5.4	Hypothetical protein	-
EHIS_000085300	EHI_087660	2616	17	5.2	Cysteine protease binding protein family	Terminal signal peptide
EHIS_000301500	EHI_035760	770	5	4.2	Hypothetical protein	Terminal signal peptide 4 x Transmembrane domains
EHIS_000253500	EHI_158600	1736	6	3.5	Hypothetical protein	Terminal signal peptide 1 x Transmembrane domains
EHIS_000350000	EHI_138460	1436	4	1.7	Papain family CP domain containing protein	Interpro ID: IPR000668 (Papain family cysteine protease)
EHIS_000096300	EHI_099320	2667	1	0.3	Hypothetical protein	2 x Transmembrane domains

RNA-seq data for *E. histolytica* HM-1:IMSS (Dr Kanok Preativanyou. Pers. Comms.) was mapped to the new PacBio *E. histolytica* HM-1:IMSS reference and FPKMs calculated for each gene. Methylated genes were assessed to identify how highly expressed they were in comparison to other genes in the genome (Figure 5.3.2A-B). The median FPKM value for the gene set was 13.92. The majority of methylated genes (6/9) had FPKMs lower than the median (Figure 5.3.2C-K).

Life cycle RNA-seq data has been previously collected for *Entamoeba invadens* IP-1 as it is the only *Entamoeba* species that can be induced into encystation in culture [84]. To determine whether the methylation of the nine *E. histolytica* HM-1:IMSS genes was correlated with stage-specific expression (i.e. repression in trophozoites, expression during encystation), *E. invadens* orthologues for the highly methylated *E. histolytica* genes were identified. Six of the methylated genes have an orthologue in *E. invadens* IP-1 (AmoebaDB, May 2018). The expression profiles of these six orthologues were examined on AmoebaDB (Figure S5.4, Appendix 5). Three of the six orthologues were associated with lower expression during the trophozoite life stage compared to the encysting and excysting time points. The three remaining orthologues showed a wide range of expression at the trophozoite stage therefore in some trophozoites the transcript expression was reduced.

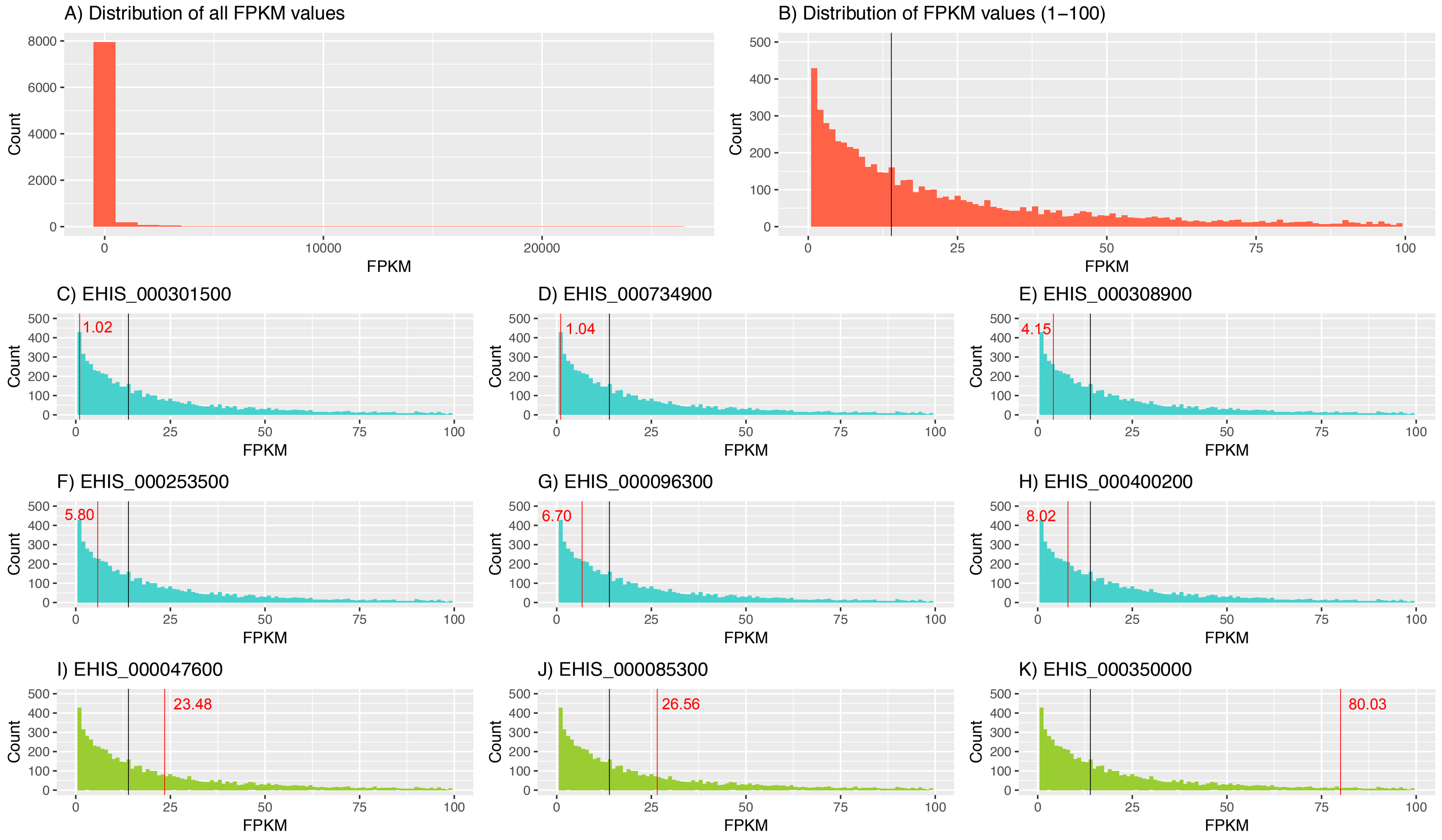


Figure 5.3.2. FKPM analysis of methylated (5-MeC) *Entamoeba histolytica* HM-1:IMSS genes. A: Distribution of FKPMs across all genes in the *E. histolytica* genome. **B:** Subset of plot A showing genes with FKPMs of 1-100. **C-H:** FPKMs of methylated genes that have a lower FKPM than the majority of genes. **I-K:** FKPMs of methylated genes that have a higher FKPM than the majority of genes. **N.B.** Black lines represent the median FKPM for the dataset and red lines represent the FKPM of the methylated *E. histolytica* HM-1:IMSS genes.

Five genes in the *Entamoeba moshkovskii* Laredo genome were detected as containing one or more 5-MeCs in an exon (Table 5.3.3). Most of these genes contained one or two methylated sites. The exception to this was EMO_133490 that contained 14 methylated sites at a density of 3.56 per 100 bp. The gene function is unknown, it has no orthologues identified within AmoebaDB and no functional annotation existed.

Table 5.3.3. Methylated genes in the *Entamoeba moshkovskii* genome.
Ei: Entamoeba invadens Eh: Entamoeba histolytica

Gene ID	Gene Length (bp)	Methylated sites	Methylated Cs in gene (%)	Putative function	Protein features
EMO_133490	393	14	4.67	Unspecified product	-
EMO_011960	1682	2	1.30	Unspecified product	Orthologous to serine/threonine kinase in *Eh* and *Ei*
EMO_052010	2814	2	0.67	Hypothetical protein	-
EMO_037420	1616	1	0.43	Hypothetical protein	Interpro ID: IPR001849 (PH domain profile)
EMO_128410	1209	1	0.41	RhoGAP domain containing protein	Interpro ID: IPR000198 (Rho GTPase-activating proteins domain)

Four of the five methylated *E. moshkovskii* Laredo genes contained an orthologue in *E. invadens* IP-1 and the expression profile of these orthologues across the life cycle was observed in AmoebaDB (Figure S5.5, Appendix 5). Three of these orthologues were associated with reduced expression in *E. invadens* IP-1 trophozoites when compared to other life cycle stages. The remaining orthologue showed variable expression in trophozoites with evidence of a wide range of expression across replicate experiments in the trophozoite stage. No methylation was detected in the 21 bp upstream region of any genes suggesting no methylation of 5'UTRs is present in this species.

Only one gene was detected as being methylated in the *Entamoeba invadens* IP-1 genome. The gene (EIN_282370) contained one methylated site in its 703 bp gene length (accounting for 0.93% of cytosines in the gene). The gene is functionally unannotated however orthologues suggest it may encode a myosin heavy chain gene. Expression of the transcript was analysed on AmoebaDB but it was difficult to assess true expression as few reads map to the gene uniquely. The gene appears to form part of a myosin heavy chain orthologue group in *E. invadens* IP-1 whose expression is associated with encystation and transcription is significantly reduced in the trophozoite stage.

5.3.3. tRNA arrays are methylated in three *Entamoeba* species

The largest proportion of 5-MeCs in the *E. histolytica* HM-1:IMSS genome was observed in the tRNA array units, with 1,858/6,211 (30%) 5-MeC occurring in these regions. The methylation occurred largely in the short tandem repeats (STRs) between the tRNA genes in the array units (Initially outlined in figure 4.3.1). 19 highly methylated cytosines were observed within the tRNA genes themselves across all of the arrays (Figure 5.3.3-5.3.8). These methylated sites within the tRNA genes were always located at the end of the gene and as part of a cluster of high methylated cytosines upstream or downstream of the tRNA gene. Similarly in *E. moshkovskii* Laredo and *E. invadens* IP-1 few cytosines within the tRNA genes themselves were methylated and the methylation was mostly within the spacer regions between the tRNA genes (in *E. invadens* these are STRs, in *E. moshkovskii* they are non-repetitive) however, these cytosines were not as highly methylated as observed in *E. histolytica* HM-1:IMSS. In *E. moshkovskii* Laredo tRNA arrays, most 5-MeC were up to 80% methylated (Figures S5.6-S5.8, Appendix 5) and in *E. invadens* IP-1, most 5-MeC were only around 30-40% methylated (Figures S5.9-S5.13, Appendix 5), suggesting *E. invadens* has a considerably lower level of methylation of the tRNA arrays compared with *E. histolytica* HM-1:IMSS and *E. moshkovskii* Laredo.

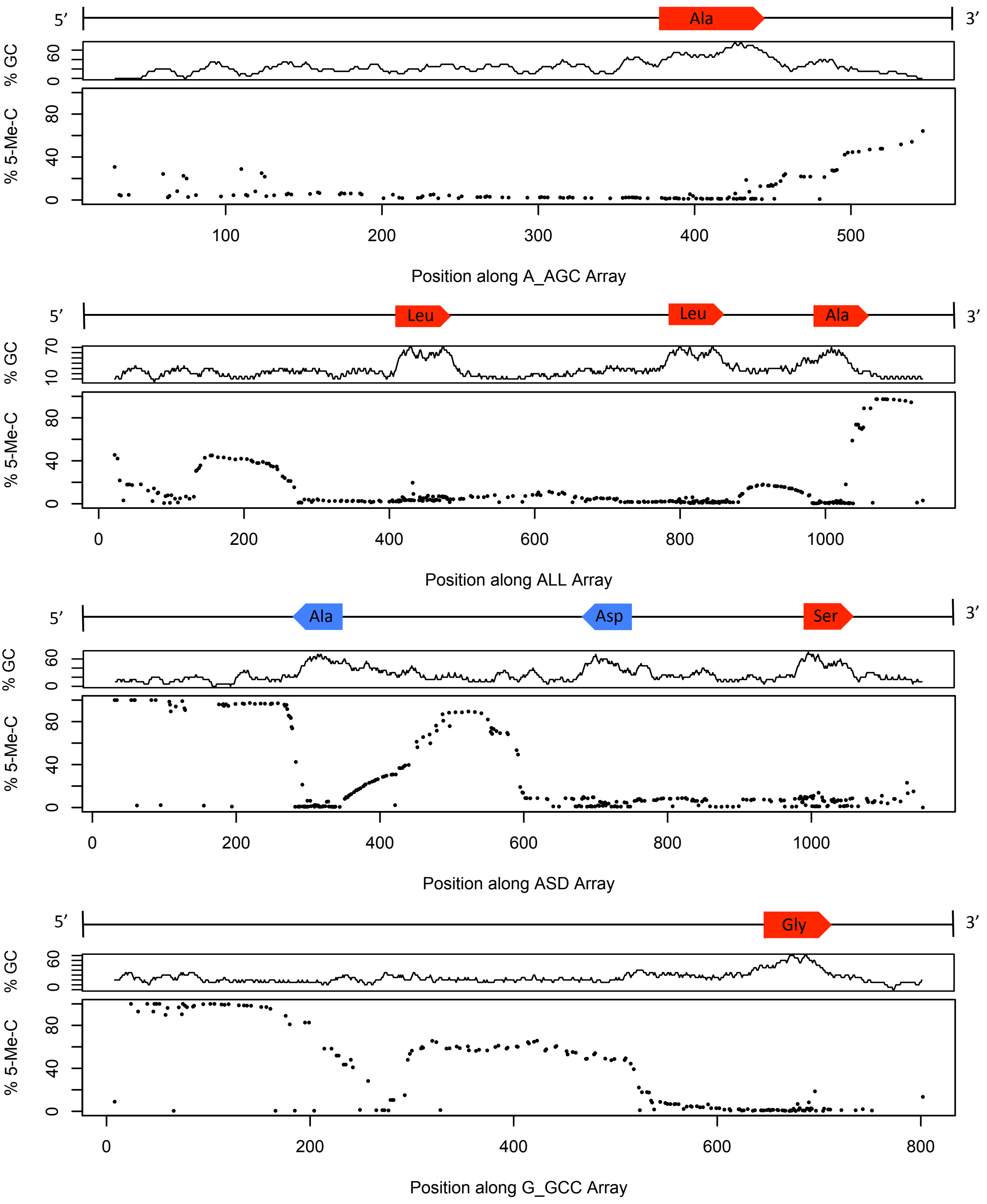


Figure 5.3.3. Methylation of *Entamoeba histolytica* HM-1:IMSS tRNA array units A^{AGC}, ALL, ASD and G^{GCC}. Total % methylation of cytosine bases in all array units is shown for a single tRNA array unit in each case. Points represent the percentage of methylated reads at cytosine positions (with at least 10X coverage).

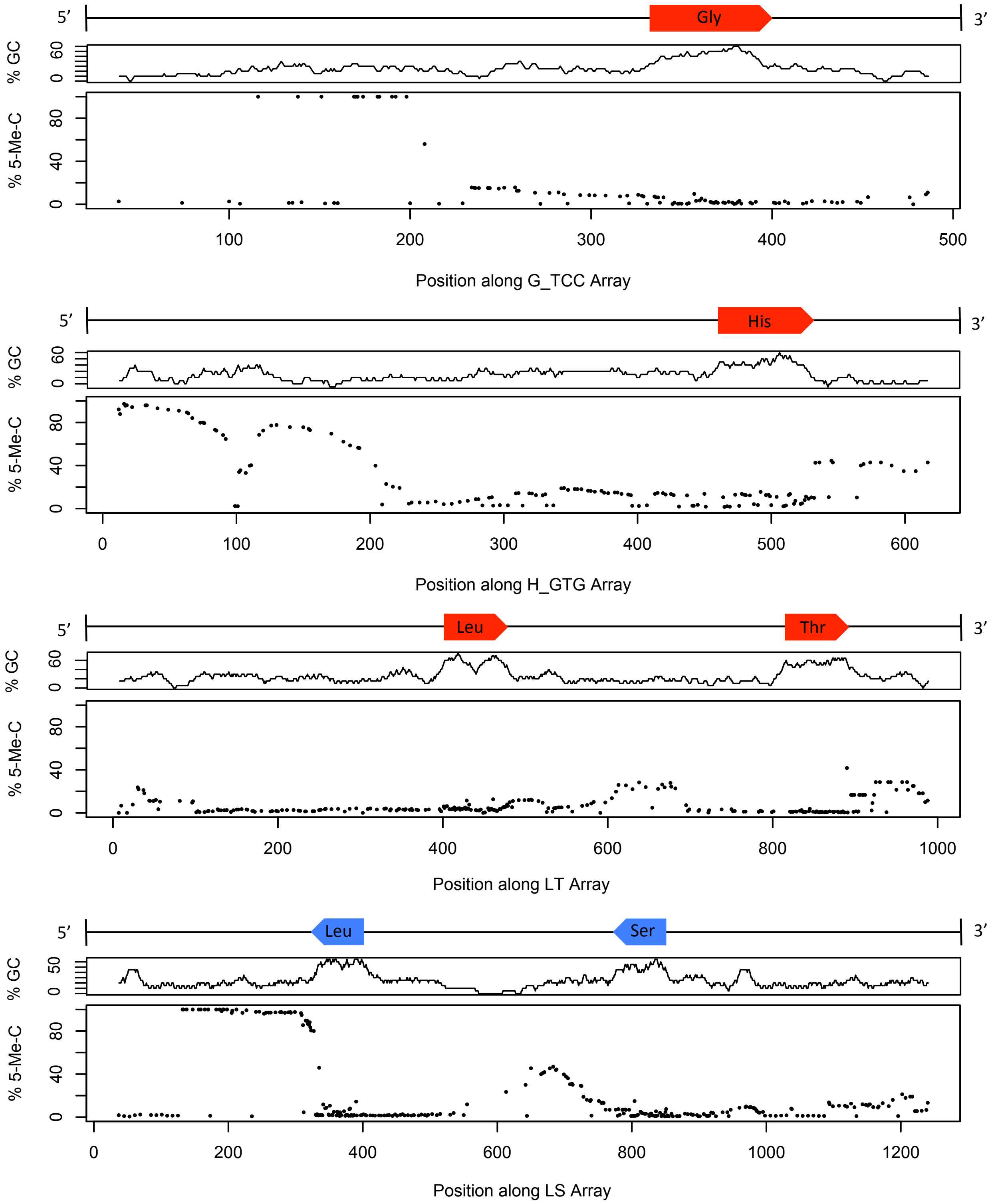


Figure 5.3.4. Methylation of *Entamoeba histolytica* HM-1:IMSS tRNA array units G^{TCC}, H^{GTG}, LT and LS. Total % methylation of cytosine bases in all array units is shown for a single tRNA array unit in each case. Points represent the percentage of methylated reads at cytosine positions (with at least 10X coverage).

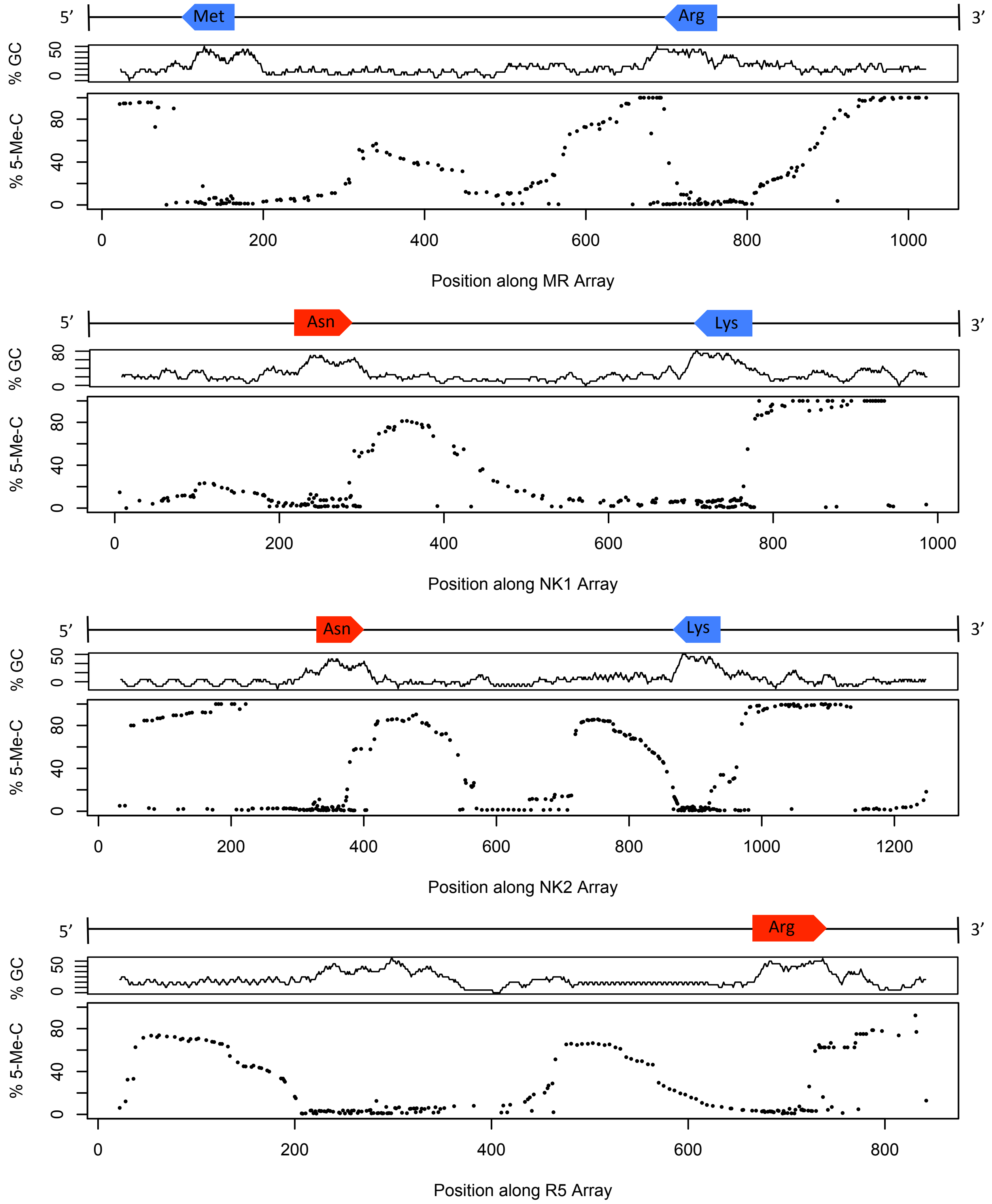


Figure 5.3.5. Methylation of *Entamoeba histolytica* HM-1:IMSS tRNA array units MR, NK1, NK2 and R5. Total % methylation of cytosine bases in all array units is shown for a single tRNA array unit in each case. Points represent the percentage of methylated reads at cytosine positions (with at least 10X coverage).

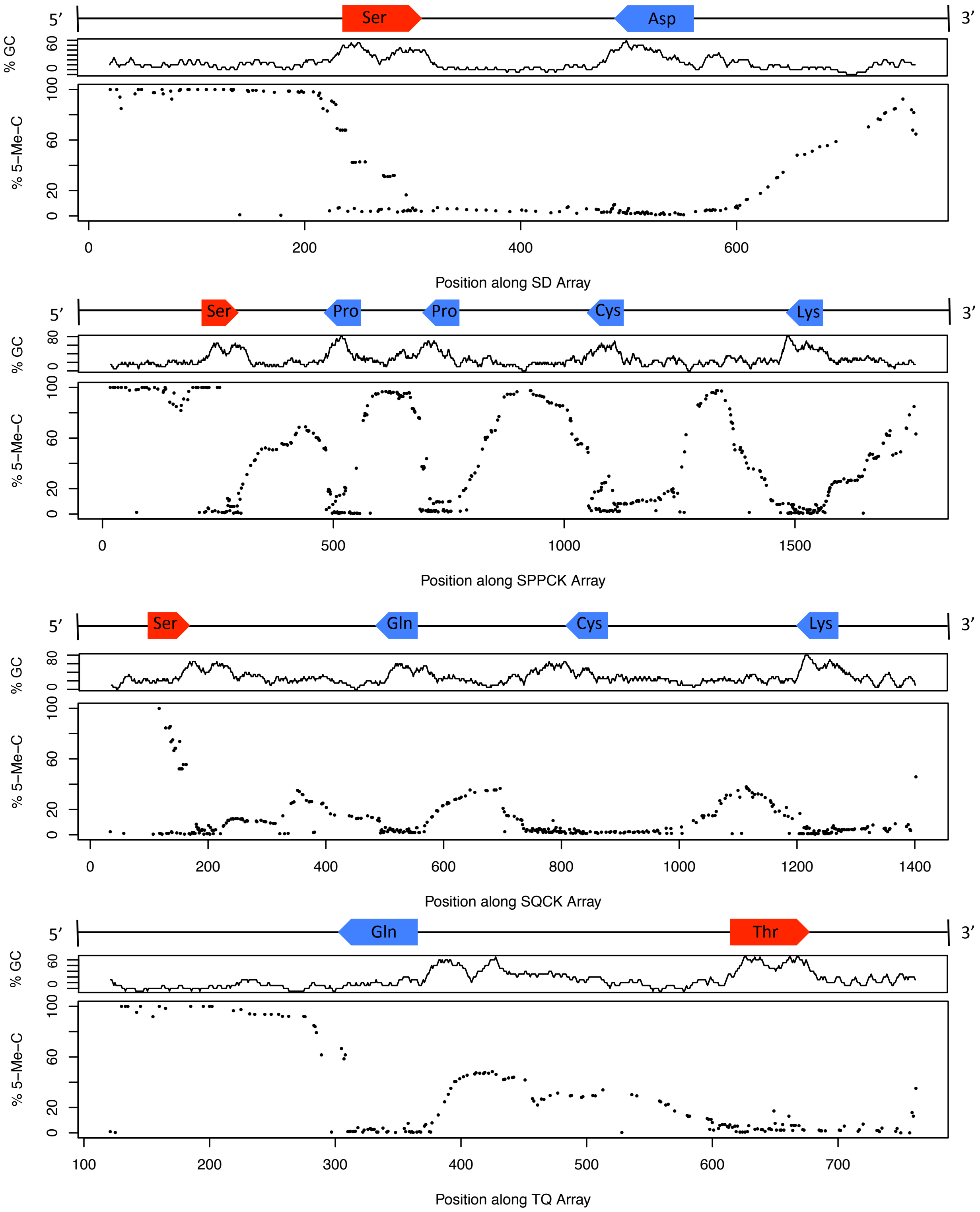


Figure 5.3.6. Methylation of *Entamoeba histolytica* HM-1:IMSS tRNA array units SD, SPPCK, SQCK and TQ. Total % methylation of cytosine bases in all array units is shown for a single tRNA array unit in each case. Points represent the percentage of methylated reads at cytosine positions (with at least 10X coverage).

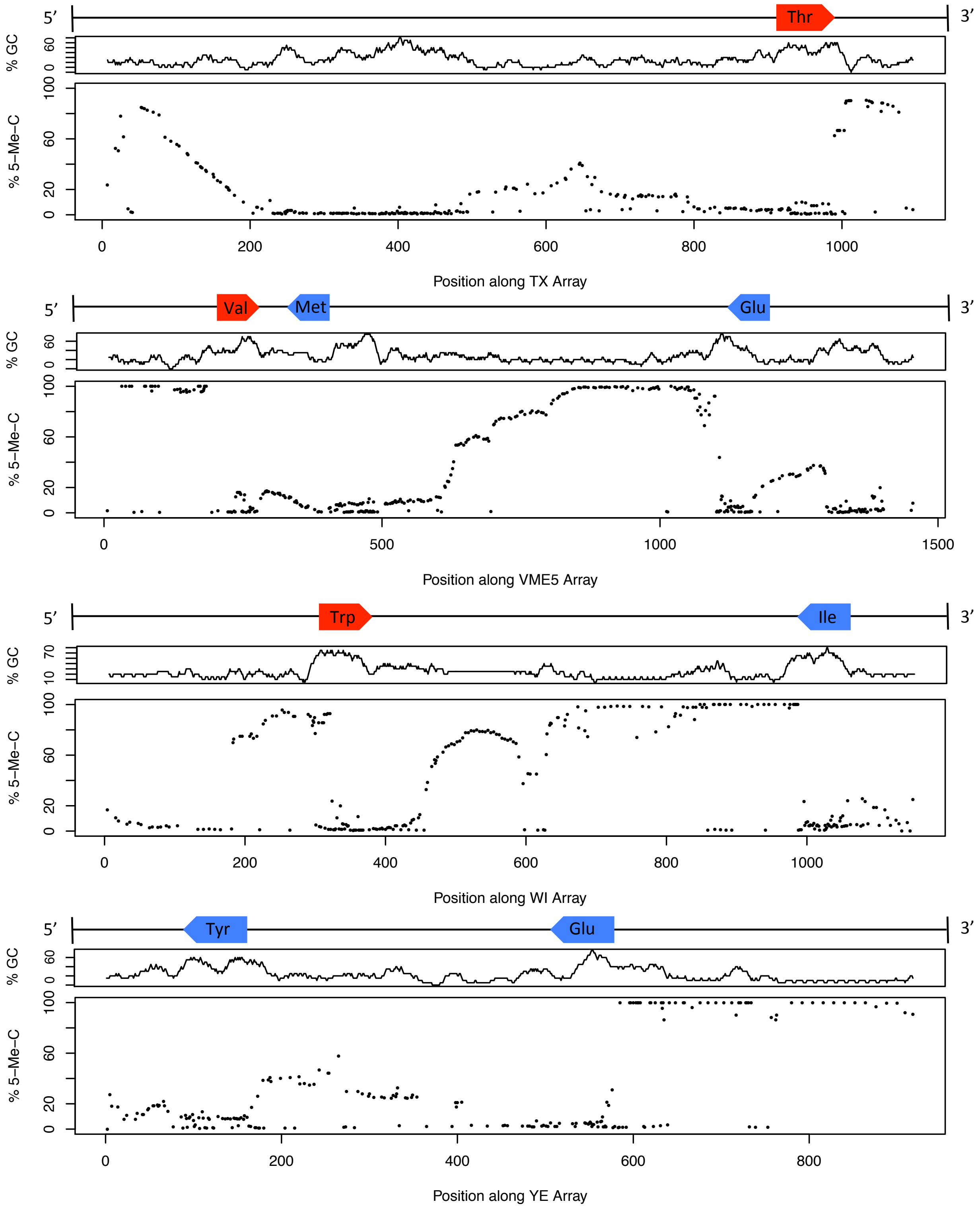


Figure 5.3.7. Methylation of *Entamoeba histolytica* HM-1:IMSS tRNA array units TQ, VME5, WI and YE. Total % methylation of cytosine bases in all array units is shown for a single tRNA array unit in each case. Points represent the percentage of methylated reads at cytosine positions (with at least 10X coverage).

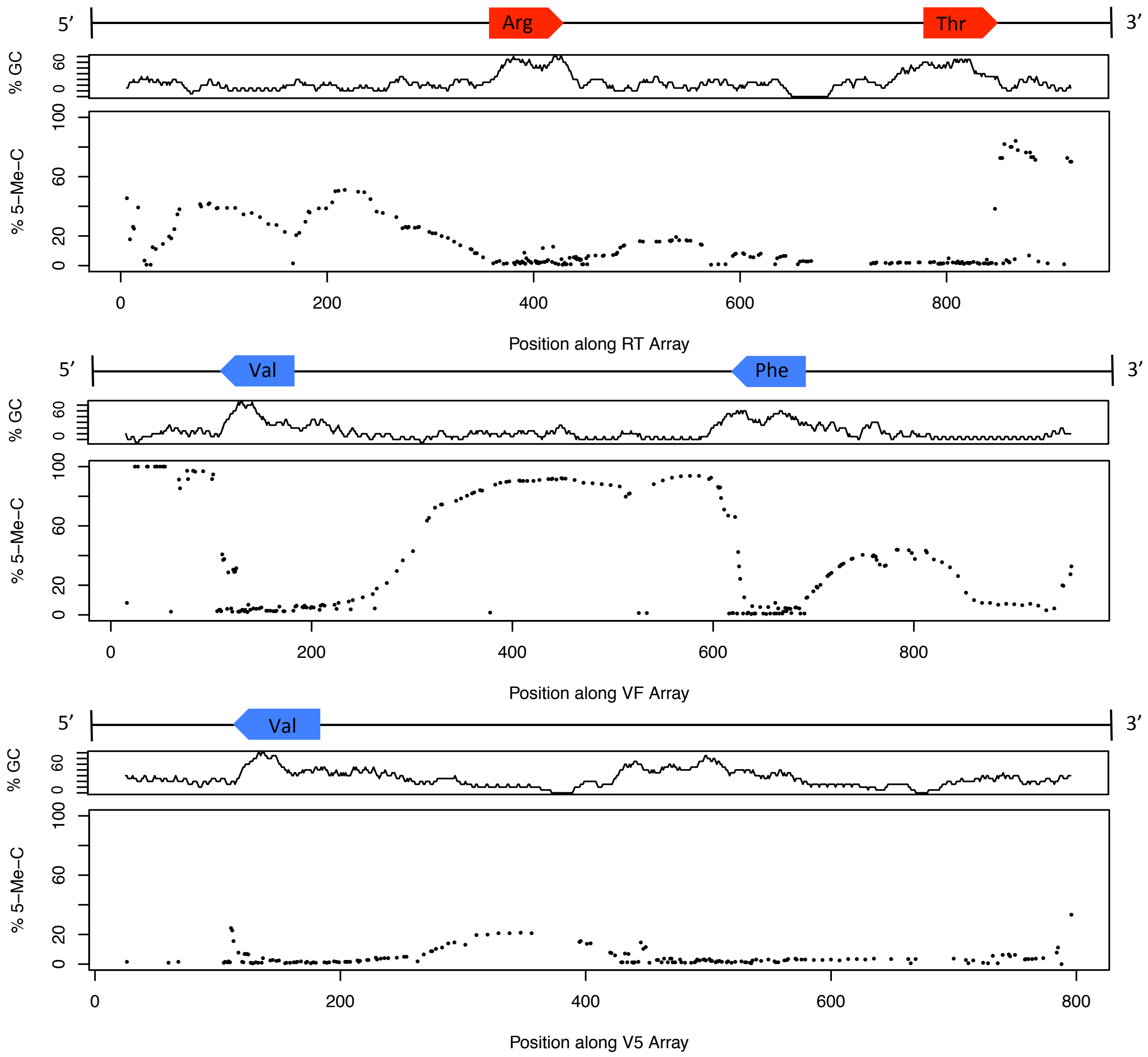


Figure 5.3.8. Methylation of *Entamoeba histolytica* HM-1:IMSS tRNA array units RT, VF and V5. Total **%** methylation of cytosine bases in all array units is shown for a single tRNA array unit in each case. Points represent the percentage of methylated reads at cytosine positions (with at least 10X coverage).

Spikes of methylation are observed both upstream and downstream of the tRNA genes however for the majority of tRNA genes the most extreme methylation occurs immediately downstream of the tRNA genes. The extent of methylation differs between the flanking regions of individual tRNA genes in the same array units and between array units, with some array types (V5, LT, RT) showing low levels of methylation and others (SPPCK, NK2, WI) showing high levels of methylation in the STRs.

To test if these methylation patterns are associated with the copy number of the large number of tRNA array genes, the *E. histolytica* HM-1:IMSS TruSeq paired end library was mapped to a single copy of each tRNA gene sequence. No mismatches were allowed, to eliminate mapping of reads to similar tRNA gene sequences. It was hypothesised that tRNA genes with high mapping coverage (i.e. more copies in the genome) would be more likely to be methylated and therefore repressed, as over-expression of these redundant genes could be wasteful to the organism. No such correlation was observed; the tRNA genes in the tRNA arrays with the highest level of methylation, V5, LT and RT, are in the top 25% of tRNA genes when ranked by coverage.

An alternative theory is that methylation may be correlated with codon usage as the most commonly used codons need to be transcribed more often than rarely used codons. Hence, it could be hypothesised that rarely used codons would be associated with the higher levels of methylation to repress transcription of these sequences. This correlation was also not observed; the some of the most commonly used tRNA isoacceptor types (Lys^{TTT}, Glu^{TTC}, Ile^{AAT}) are associated with highly methylated flanking regions in their specific tRNA array (SPPCK, YE and WI respectively). Conversely, the some of the most rare tRNA isoacceptor types (Pro^{CGG} and Ser^{CGA}) occur in tRNA arrays (SQCK and LS respectively) with lower levels of methylation in the flanking regions.

5.3.4. Methylation occurs in other repetitive regions of the *Entamoeba histolytica* HM-1:IMSS genome, including transposable elements

In other eukaryotes, CHH and CHG methylation has been associated with the methylation of non-coding regions and involved in roles such as silencing transposable elements [378]. The *Entamoeba histolytica* genome contains a large number of transposable elements (TEs) including SINEs, LINEs and unclassified transposable elements. To test for methylation in these regions, sequences for 11 well-characterised transposable elements were used to detect copies in the *E. histolytica* HM-1:IMSS PacBio genome (Table 5.3.4). In total, 894 copies of these transposable elements were detected accounting for approximately ~2 Mbp (~6%) of the genome sequence. The average depth of BS-seq coverage for these elements was low, at 5.8x. Therefore, cytosines with at least 5x coverage were analysed to detect methylation.

Table 5.3.4. Methylation of transposable elements (TEs) in the *Entamoeba histolytica* HM-1:IMSS genome. Cytosines in each TE were analysed for the percentage of reads mapping that contained a methylated cytosine at that position (Coverage threshold: 5x).

TE type	Total copies	Copies with ≥1 low methylated site[A]	Copies with ≥1 medium methylated site[A]	Copies with ≥1 high methylated site[A]	Total copies with 5-MeCs[B]	Proportion of TE type methylated (%)
EHAPT2	125	3	3	9	13	10.40
EhINV1	7	0	1	2	2	28.57
EhINV2	7	0	0	1	1	14.29
EhRLE2	22	6	7	17	18	81.82
EhRLE3	142	69	54	112	119	83.80
ERE1	50	17	29	49	49	98.00
ERE2	316	16	27	246	247	78.16
RLEEh2	151	42	45	78	89	58.94
RLEEh3	70	14	12	23	27	38.57
RLEEh4	3	0	0	1	1	33.33
RLEEh5	1	1	1	1	1	100
Total	**894**	168	179	539	**567**	**63.42**

[A] Levels of methylation are defined by the percentage of reads mapped to a cytosine are methylated at that position (25-50% = Low, 50-75% = medium, 75-100% = high)

[B] Total copies does not equal the sum of low, medium and high methylated sites as some TEs have multiple categories of methylated cytosines in the same TE.

567 TEs contain at least one methylated site (regardless of methylation classification: low, medium or high), accounting for 63.42% off all TEs in the genome. All types of elements contain some methylation however the proportion of each type that is methylated differs. High proportions of the ERE1, ERE2 and RhRLE2/3 elements are methylated (>75% of elements in each family are methylated). Lower levels of methylation are observed in the RLEEh2/3/4/5 elements, the EhINV1/2 elements and the EHAPT2 element.

5.3.5. No *de novo* or maintenance DNA methyl-transferases were detected in the *Entamoeba histolytica* HM-1:IMSS genome

Entamoeba histolytica contains the Dnmt2 DNA MTase (EhMeth) but no other DNA MTases have been annotated in the genome. As the previous reference genome was fragmented and incomplete [97,98], the existence of other DNA MTase genes could not be ruled out. Some eukaryotes possess only Dnmt2 and lack the other DNA methyltransferases (DNA MTase) often associated with the methylation machinery (Dnmt1, Dnmt3a and Dnmt3b). These so called Dnmt2-only organisms have all of their methylation performed and maintained by Dnmt2 and include the amoeba *Dictyostelium discoideum* and *Drosophila* [352–355].

Annotation of the new *Entamoeba histolytica* HM-1:IMSS genome was performed using Companion (Chapter 3). This program performs both annotation transfer of already known genes and *ab initio* prediction of novel genes in the genome. The companion annotation confirmed one copy of the EhMeth gene (EHI_069140) in the new assembly, but no additional DNA MTs (Dnmt1, Dnmt3a or Dnmt3b) were found, nor any proteins with putative DNA MTase domains. These results were confirmed using tBLASTn. As DNA MTases tend to be relatively well conserved across the eukaryotes, human and plant DNA MTases were used to detect the presence of DNA MTases in the new *E. histolytica* HM-1:IMSS genome assembly. Aside from the one copy of EhMeth, no other regions of the genome shared significant (e-value < 0.05) homology with Dnmt1, Dnmt2. Dnmt3a or Dnmt3b.

5.4 Discussion

5.4.1. Sparse methylation of genic regions suggests methylation does not play a large role in the control of gene expression

A very small proportion of cytosines in the *Entamoeba histolytica* HM-1:IMSS genome are methylated, with only 0.4 % of cytosines methylated across the entire genome. Though it should be noted that this number may also include a number of false positive 5-MeC reads arising from incomplete conversion of un-methylated cytosines to uracils in the library preparation and therefore, even this small number of 5-MeCs may be an overestimate. This low level of methylation is not distributed randomly across the genome but mainly occurs in distinct regions such the tRNA arrays, suggesting that the bisulphite conversion of the DNA was successful and the methylation was real (as incomplete conversion would affect all unmethylated cytosines, resulting in a signal of 'partial methylation' throughout the genome). As *E. histolytica* does not contain a known unmethylated region (e.g. a mitochondrial genome) a thorough assessment of the extent of conversion could not be carried out.

Despite this, the large majority of methylation occurs at CHH sites (where H is equivalent to a non-guanine base) and very few 5-MeCs occur at CpG sites (0.06% of the CpG sites covered by 10x coverage); a stark inversion of the trend seen in other eukaryotes. Up to 60-90% of CpG sites, which often occur as CpG islands, can be methylated in mammals [378] while in hymenopteran insects, around 0.5-0.7% of CpGs are methylated [379,380]. The sparse 0.06% of methylated CpG sites detected in *Entamoeba histolytica* HM-1:IMSS is, therefore, at least one order of magnitude lower than the methylation levels found in some eukaryotes. Not all organisms have genomes that contain CpG islands that are methylated, as seen in mammals. Unlike vertebrates, most invertebrates exhibit mosaically methylated genomes comprising alternating methylated and non-methylated domains [381,382]. Though DNA methylation is not always necessary for transcriptional silencing, it is generally thought to render the methylated region transcriptionally inactive. In particular, DNA methylation appears critical for regulating gene expression, in particular gene silencing, and has been demonstrated in fungi, plants and animals [383,384]. The observation of few methylated CpGs and the fact that only a handful of genes are methylated suggests that 5-MeC in *Entamoeba histolytica* may not play a large role in the regulating the expression of genes as a whole. However, most of *E. histolytica* HM-1:IMSS methylated genes are associated with down-regulated expression in *Entamoeba invadens* IP-1 orthologues and also are less expressed than the majority of genes in the transcriptome and therefore, methylation may be involved in transcription repression of the small number of genes that are methylated.

5.4.2. No association can be drawn between DNA Methylation and virulence as seen in other protists

Of the small number of methylated genes in the *Entamoeba histolytica* HM-1:IMSS genome, only one could be associated with parasite virulence. This gene, EHIS_00035000 (equivalent to EHI_138460 in AmoebaDB), is a part of a family of cysteine proteases in the *E. histolytica* genome that have been implicated in the virulence of the parasite [385]. A further methylated gene was annotated as a cysteine protease binding protein however, if this gene is involved in virulence is unknown though, it does interact with the virulent cysteine protease genes. The remaining genes could not be associated with any function. The observation of signal-peptides in the many of the methylated genes suggests these genes may be secreted or exported to the cell membrane. Secreted proteins have been implicated with virulence and the defence mechanisms of a range of parasites including *Entamoeba* species. Ultimately, further validation of the function of these methylated genes needs to be done before DNA methylation in *E. histolytica* HM-1:IMSS can be associated with any specific function or phenotype though the concentration of 5-MeCs in these select genes suggest they are not a result of random noise from sequencing but instead, genuinely are targeted for methylation.

However, that is not to say that the methylation of virulence genes isn't associated with their gene expression, it just suggests that 5-MeC DNA methylation may not be the main mechanism controlling the expression of virulence genes. Other types of methylation exist and it would be interesting to look at levels of 4-Methyl-cytosine (4-MeC) and 6-Methyl-Adenine (6-MeA) methylation which have been associated with virulence attenuation when blocked in other pathogens (both prokaryotes and eukaryotes) [386,387].

5.4.3. Dense methylation of the tRNA array units suggests a role for the tRNA arrays in telomere formation

The methylation of tRNA array units was proven not to be associated with either codon usage (i.e. tRNA demand) or with putative copy number and therefore, it seems unlikely that the DNA methylation is involved in the dynamic regulation of tRNA gene transcription. This is consistent with previous studies that have proven that treatment of *Entamoeba histolytica* trophozoites with 5-azacytidine (an agent that blocks DNA methylation) does little to affect the levels of transcription of most genes and suggests that DNA methylation may not play an important role in the transcriptional regulation of genes [360].

Alternatively, DNA methylation may serve another purpose. High levels of methylation in the tRNA arrays is consistent with a putative role of these arrays as telomeres in the *Entamoeba* genomes (Chapter 4). DNA methylation can lead to the condensation of chromatin. If this is occurring in the tRNA arrays, the majority of the tRNA genes are likely being repressed by Dnmt2 (EhMeth), the sole methyltransferase in *E. histolytica* [364]. Dnmt2-mediated methylation has been implicated in controlling telomere length in other eukaryotic species, where down-regulation of Dnmt2 results in the shortening or loss of telomeres in mice and *Drosophila* [354,388]. Dnmt2 plays a similar role in both *Drosophila* and mice despite them having very different telomere sequences: mice telomeres consisting of the hexamer 'TTTAGG' and *Drosophila* telomeres consisting of tandemly repeated retro-transposons. An analogous mechanism could be occurring in *E. histolytica* HM-1:IMSS, whereby Dnmt2-mediated cytosine methylation of tRNA arrays confers stability to the DNA by modifying the tRNA STRs in a way that allows them to act as telomeres. The observation of increased levels of methylation in the tRNA arrays of other *Entamoeba* species (*E. moshkovskii* Laredo and *E. invadens* IP-1) suggests this mechanism may common to multiple species within the *Entamoeba* genus.

5.4.4. The presence of DNA methylation suggests a protective role against retro-transposons and other mobile elements

63% of the transposable elements (567/894 TEs) contain some level of methylation, with nearly all (539/567 TEs) of these methylated elements containing at least one highly methylated site. Many of the transposable elements demonstrated low alignment coverage with BS-seq reads, most likely due to their highly repetitive nature that makes unique mapping of reads difficult. The average mapping depth of the most abundant TE group, ERE2 (35% of all TE sequences), was 5.97x and therefore, the majority of cytosines in these sequences would have been lost in the 10x coverage cut-off used for the main methylation analysis. To include these regions, cytosines with at least 5x coverage were analysed. Despite lower coverage, a large number of elements were detected as being methylated and therefore, it is highly likely that transposable elements are being silenced by the DNA methylation machinery to prevent disruption of genes caused by the integration of mobile elements into coding regions. The number of methylated transposable elements is likely to be even larger, as some were omitted from analysis due to low coverage depth (<5x coverage). To address this problem, further bisulphite sequencing, possibly with longer read lengths, may be required to increase the coverage of these regions and other mapping algorithms may need to be explored to ensure the maximum number of reads are being mapped to the genome.

Contrary to this, some groups of TEs showed good mapping coverage, EhRLE2 for example had an average mapping depth of 127x. In these groups, not all of the TE copies were methylated suggesting some copies of these elements are genuinely unmethylated. The outcome of the original genome sequencing reported that many of the transposable element sequences in the *Entamoeba histolytica* HM-1:IMSS genome were degenerate and had lost their ability to transpose themselves [97,98]. However, some copies do remain with functional transposase ability and it would be interesting in the future to check whether it is these functional copies that are being methylated in order to effectively silence them within the genome and whether degenerate sequences have lost their methylation.

5.4.5. The absence of *de novo* or maintenance DNA methyltransferases suggests the RNA methyltransferase EhMeth may be able to methylate DNA

DNA methylation of *Entamoeba histolytica* HM-1:IMSS, *Entamoeba invadens* IP-1 and *Entamoeba moshkovskii* Laredo has been proven to exist despite not having the DNA methyltransferases (DNA MTases) supposedly required for DNA methylation. *E. histolytica* contains only a Dnmt2 gene ('EhMeth'), which has been reported to be an RNA MTase. Nevertheless, DNA methylation does occur Therefore, it is proposed that EhMeth is a methyltransferase of both DNA and RNA, a phenomenon proposed for other Dnmt2-only organisms such as *Drosophila* and the amoeba *Dictyostelium* [339,389].

5.5. Conclusions

The lack of studies investigating the DNA methylation of *Entamoeba* species has prohibited a detailed understanding of the evolution and function of early eukaryotic DNA methylation. *Entamoeba histolytica* has a unique phylogenetic position; that is, it diverged soon after the plant and animal lineages split [2] and as such, is closely related to both animal and plant kingdoms that have organisms with largely present or absent methylation systems (For example, mammals and fungi, respectively). Bisulphite sequencing has shown that *Entamoeba histolytica* HM-1:IMSS had a rudimentary DNA methylation system as evidenced by low levels of genome methylation. The findings support earlier observations that the *Entamoeba histolytica* genome is largely devoid of DNA methylation [364,390]. This work extends these previous studies by identifying the locations of methylated cytosines, genome-wide. It showed that a very small number of genes show high levels of methylated sites, but that many transposons are methylated. Extensive methylation of the tRNA arrays is also observed and this may play a role in telomere formation and protection of chromosomes ends. The distribution of methylated cytosines was broadly similar in *Entamoeba moshkovskii* (Laredo) and *Entamoeba invadens* (IP-1) although to a lesser extent.

63.4% of transposable elements appear to be methylated suggesting a methylation-mediated mechanism of protecting the DNA from deleterious mutations caused by the transposition of the retro-transposons and mobile genetic elements that litter the *Entamoeba histolytica* HM-1:IMSS genome. However, this analysis was limited by the poor coverage of these repetitive areas.

It has been suggested that the expression of EhMeth can become decreased over several generations of sub-culturing [173]. Therefore, DNA methylation may be more active during actual infections of individuals. It would be useful to bisulphite-sequence trophozoites directly from infections, or after as few sub-cultures as possible, to elucidate a better picture of DNA methylation in these species. This is limited largely by the fact large numbers of trophozoites need to be collected in order to generate enough genomic DNA for sequencing, which would be difficult to collect from one individual. Also, purifying only *Entamoeba* from the gut/stool would be a great challenge.

Overall, combining these new data, with the observation that there is a functional Dnmt2 homolog (EhMeth), possibly with both DNA and RNA methylating capacity, opens up many exciting questions about DNA methylation in *Entamoeba*. Is the DNA methylation system being lost from some species such as *E. invadens*? Are methylation patterns conserved across generations of trophozoites? Are any methylated genes associated with specific phenotypes? How can an active mechanism of transposon repression be reconciled with the huge number of transposons in *Entamoeba* genomes?

Chapter 6 – General Discussion

Entamoeba histolytica is an obligate parasite of humans and is the causative agent of the disease amoebiasis [7]. The genome of *E. histolytica* has been studied in depth over the past decade with much being learned about its structure, as well as the genes and proteins it contains [4,97,98]. However, genomic studies have been limited in *Entamoeba* species owing to the fragmented and incomplete nature of the reference assembly for *E. histolytica* HM-1:IMSS and the large number of functionally unannotated genes and gene families [98,112]. This project sought to utilize SMRT sequencing and improved genomic assembly and gene annotation techniques to produce an improved reference genome from which further analyses regarding genome structure, gene content, and gene organisation and regulation could be performed.

6.1. SMRT sequencing, assembly and annotation of the *Entamoeba histolytica* reference genome

A telomere-to-telomere contiguous genome assembly is the goal of any genome assembly project. In common with most eukaryote genome assemblies, the current published reference genome for *E. histolytica* falls a long way short of this gold standard. It consists of 1,498 scaffolds with a large majority ending with repetitive sequence, indicative of assembly problems associated with short-read lengths. A new assembly and annotation of the genome of *E. histolytica* HM-1:IMSS was presented in this thesis. The new PacBio *E. histolytica* genome is approximately 29 Mbp in 563 contigs (i.e. fully contiguous sequence not containing assembly gaps). Excluding contigs containing tRNA arrays and rDNA arrays (these were excluded from the published reference assembly) the number of contigs is 432, representing a decrease in contig number of over two thirds in comparison with the published assembly. This large improvement to the reference genome is reflected in an increased N50 and average contig length in the PacBio assembly.

Though SMRT sequencing of the *E. histolytica* genome has greatly improved the genome assembly, the gold standard of telomere-to-telomere contiguity is yet to be reached and further work to improve the assembly will need to be performed.

6.1.1. Utility of existing and emerging long read assemblers on a challenging genome

Chapter 2 demonstrated the differences in assembly quality that can result from different assembler programs and highlighted how the assembly of long-read data was still in relative infancy at the time of the SMRT sequencing of *E. histolytica.* Guidelines on best practice when assembling long read data and thorough comparisons of the genome assembly software were not available at the time of assembling the *E. histolytica* HM-1:IMSS reference genome. Long read assemblers have rapidly evolved in the time since this assembly was done and it is likely that reassembly of the *E. histolytica* genome using the newest releases of the assemblers outlined in Chapter 2, as well as testing some of the newer emerging long-read assemblers, may produce an even more improved assembly.

6.1.2. Improvements to extraction of high molecular weight gDNA from *Entamoeba*

Many of the contigs produced in the final PacBio genome still terminated with repetitive sequence, suggesting that read-length may still be a limiting factor in assembling the *E. histolytica* genome. The genome is largely repetitive and contains a variety of repetitive elements such as transposable elements and tRNA arrays, some of which span lengths up to 40 Kbp (Chapter 4). The average insert size of the PacBio assembly was 3.8 Kbp, which is much longer than the average read length of 750 bp in the published assembly, but still too short to span all of the repetitive regions of the *E. histolytica* genome and facilitate construction of a fully contiguous genome assembly. Further investigation into the generation of *E. histolytica* genomic DNA may prove useful before further

SMRT sequencing of the parasite is performed. *Entamoeba* genomic DNA is very carbohydrate-rich (akin to plant DNA) and contains an abundance of lytic enzymes that make extraction of high molecular weight DNA a challenge. As SMRT sequencing becomes ever more popular, protocols describing improved DNA isolation methods are emerging [391] and those designed especially for the extraction of carbohydrate-rich plant DNA may be useful to explore for *E. histolytica* cells [392,393]. The generation of high molecular weight DNA for *E. histolytica* would likely improve the genome assembly quality as it could facilitate further PacBio sequencing with a longer insert sizes (20 kb inserts and longer are possible, but require a lot of gDNA) or facilitate genome scaffolding technologies such as optical mapping or Hi-C sequencing (Chapter 2).

6.1.3. Validation of gene models and assigning putative gene function

The PacBio *E. histolytica* HM-1:IMSS genome contained 10,164 genes, an increase of 1,831 genes compared to the current published reference assembly ([98]; AmoebaDB data – Release 39; August 2018). It is likely, as was concluded in the previous sequencing attempts [97], that this number of predicted genes may be an over-estimate of the true number. Manual inspection of the gene models contained on the longest contig of the PacBio assembly (~1 Mbp) revealed no obvious errors in the gene annotation however, gene models will need to be inspected further before official release of the assembly and annotation. Putative gene function for many genes (both existing and novel) remains unknown and further experimental investigation into the function of these genes will need to be performed.

6.1.4. Further comparative analyses of *E. histolytica* with avirulent *Entamoeba* species

The *E. histolytica* PacBio genome produced by SMRT sequencing allowed analyses of the structure and wide-scale organisation of this parasite including analysis in to the organisation of a range of gene families previously associated with virulence (Chapter 3). This analysis was permitted owing to an improved assembly produced by the long read sequencing outlined in Chapter 2. Previous studies into the existing *Entamoeba* genomes revealed large differences in coverage depth among genes, indicating differences in copy number between genomes. Of particular note is the observation that a large proportion of these genes that showed differential copy number were implicated in virulence in the *Entamoeba* parasites [175].

The new PacBio *E. histolytica* genome resolved previously collapsed regions of the genome revealing tandem duplications of regions of the genome including tandem duplication of a range of genes. Further interrogation into the function of these genes will be needed in the future to confirm whether the tandemly duplicated genes that have been resolved are associated with virulence or members of existing virulence gene families. Further long read sequencing of a non-virulent *Entamoeba* species would also complement this analysis by allowing an in-depth comparison of not only CNVs between virulent and non-virulent *E. histolytica* species, but would also permit further research into the importance of the organisation of virulence gene families. An ideal candidate for long read sequencing would be *E. moshkovskii* or *E. dispar* due to their close evolutionary relationship to *E. histolytica,* despite these two organisms being largely accepted as avirulent.

The existing genomes for *E. moshkovskii* and *E. dispar* are both challenging to perform research on owing to their fragmented nature with 4,607 contigs and 3,312 contigs, respectively (AmoebaDB data – Release 39, August 2018). It has already been proven that gene prediction on fragmented draft genomes can produce extensive errors [394] and therefore, it is possible that the current annotation of these two species could be flawed. New *E. moshkovskii* and *E. dispar* genomes produced by long-read sequencing would improve the fragmented nature of these genomes and aim to produce a more contiguous assembly on which gene prediction can be performed more accurately. Once an accurate gene set has been produced for *E. moshkovskii* and *E. dispar,* RNA sequencing could be performed on these two species (existing data also available) to analyse the differential expression of orthologous genes between the virulent and avirulent *Entamoeba* species. Current analyses suggest a low level of nucleotide diversity between these three species [175] and therefore, it would be interesting to analyse any differential expression between the avirulent species and the parasitic *E. histolytica* to understand whether the emergence of virulence in *E. histolytica* is due to transcriptional changes between the species rather than changes at the nucleotide level.

6.2. Genome structure and organization of genes within the *Entamoeba histolytica* genome

The presence of gene families has previous been reported in the *E. histolytica* genome, however many of these remain functionally unannotated [97,98,112]. Many virulence gene families have been identified and characterised though the organisation of members of these families in the genome is poorly understood resulting from the fragmented published genome assembly. Chapter 3 utilised the new genome produced in Chapter 2 to investigate the organisation of a range of virulence gene families. Members of the same virulence families were not seen to form tandemly duplicated arrays but instead, the organisation of many gene families was associated with the location of transposable elements in the genome. In addition, some enrichment was seen for virulence gene family members in the putative sub-telomeric regions (formed by the tRNA arrays as described in Chapter 4). Though the organisation of a select few virulence gene families was determined, questions still remain surrounding the organisation of other gene families as well as the regulation of those associated with virulence.

6.2.1. *Entamoeba histolytica* gene families

Gene annotation of the PacBio genome generated for *E. histolytica* annotated a large number of novel genes, as well as transferring the annotation of genes predicted in the published assembly. Many of the annotated genes were characterised as hypothetical (i.e. functionally unannotated). 583 gene families in the new PacBio genome are functionally unannotated and the organisation and function of these families were not explored in this thesis. Characterisation of the function and organisation of these gene families will be important in fully understanding the biology of *E. histolytica*, especially as expanded gene families have long been reported to play an important role in the biology of other parasitic protists such as *Trypanosoma* and *Plasmodium* [87,259,285,395].

6.2.2. Regulation of virulence gene families

The PacBio genome determined that the putative subtelomeres are slightly enriched for virulence gene families and the flanking regions of the TEs are extremely enriched for virulence genes (Chapter 3). However, very little methylation was observed in these genes (Chapter 5) suggesting that DNA methylation is not a major mechanism controlling the expression of virulence genes. It is therefore likely that another mechanism controls the regulation of the virulence gene families. Investigations into alternative mechanisms of gene regulation in *E. histolytica* will help to elucidate how the virulence gene families are regulated. In addition, experimental studies may help to determine if the individual proteins encoded by virulence gene families are differentially expressed across the life cycle of the parasite in a mono-allelic way, similar to those observed in *Plasmodium var* genes and the *Trypanosoma VSG* genes.

However, investigations into virulence genes assume all *E. histolytica* trophozoites have the ability to cause invasive disease. Evidence is emerging that *E. histolytica* may in fact be better termed a pathobiont, that is, an organism that has the ability to be parasitic but under normal circumstances is a symbiont. Supporting this theory is the observation that the presence of *E. histolytica* in the microbiome of pygmy hunter gatherers from both farming and fishing populations in Southwest Cameroon was correlated with increased gut diversity and reduced signatures for autoimmune disorders [49]. The study also observed that grazing style and diet played a key role in whether *E. histolytica* infections became invasive in certain populations in Peru and Tanzania [49]. Observations of this kind suggest that the presence of *E. histolytica* under normal conditions could be protective, or in fact a symbiont, and that it is only under abnormal conditions that the parasite takes the opportunity to parasitise its host. Investigations into the microbiomes of affected individuals may help to understand whether this hallmark of increased gut diversity is maintained in individuals with invasive amoebiasis. A large proportion of individuals infected with invasive *E. histolytica* infections are prone to re-infection. Observations of infected individuals microbiome after successful treatment of amoebiasis may also help to understand how the host environment can influence such a high rate of re-infection in populations prone to *E. histolytica* infections.

Investigations into how diet affects the microbiome of populations where invasive *E. histolytica* infections are endemic may also be helpful. By investigating any differential compositions of gut diversity between populations where *E. histolytica* causes a high proportion of symptomatic disease and populations where most *E. histolytica* infections are asymptomatic, the role the host environment plays in activating the switch of asymptotic to symptomatic *E. histolytica* colonization may be revealed.

6.2.3. Support for the tRNA array telomere hypothesis

Chapter 4 provides evidence to support the tRNA array telomere hypothesis [4] in the *E. histolytica* genome. The tRNA arrays were exclusively observed to occur at the end of contigs with only one instance where contigs could be orientated in such a way as to facilitate the internal scaffolding of the array (Chapter 4). Future work would ideally experimentally validate the existence of the tRNA arrays at the end of the chromosomes. The observation that the chromosomes of *Entamoeba* species do not condense has prohibited the use of conventional fluorescent in-situ hybridization (FISH) experiments. Techniques such as fiber-FISH [396] may provide an alternative for imaging the tRNA arrays experimentally. Fiber-FISH on deproteinized, stretched DNA prepared by *in situ* extraction of whole cells immobilized on microscope glass slides allows the visualization of individual genes or other small DNA elements on chromosomes. If a protocol for the extraction of high molecular weight *E. histolytica* can be developed (Section 6.1.2.) it is possible that fiber-FISH could experimentally demonstrate the location of the tRNA arrays along stretches of DNA and validate the findings of Chapter 4.

6.2.4. Confirmation of the loss of the rDNA episome EhR1 from the *E. histolytica* genome in axenic culture

The rDNA genes of *Entamoeba* species are known to exist on extra-chromosomal molecules called EhR1 and EhR2, differentiated by the number of copies of the rDNA gene they contain (EhR1 has two copies, EhR2 has one) [110,125]. The PacBio *E. histolytica* assembly contained a fully assembled EhR2 (Chapter 2) however, no sequence was identified, in both the assembled genome or the raw reads, that corresponded to the unique regions of EhR1, suggesting it has been lost from the *E. histolytica* HM-1:IMSS cell stocks at the University of Liverpool. The experimental validation of this loss, using PCR and restriction digestion, will determine definitively whether this is the case. If experimentally validated, the findings will contribute to how the community genotypes stool samples when performing epidemiological studies as currently, unique regions from the EhR1 episome (including the *Tr* region) are used in epidemiological surveys for the disease [329,330].

6.3. Genome-wide bisulphite sequencing of *Entamoeba histolytica* HM-1:IMSS

Only a small proportion of the *E. histolytica* genome is methylated, as determined by HPLC coupled with mass spectrometry [364]. Chapter 5 presented whole genome bisulphite sequencing and genome-wide identification of methylated regions of three *Entamoeba* species; *E. histolytica* HM-1:IMSS, *E. moshkovskii* Laredo and *E. invadens* IP-1. All three genomes demonstrated low levels of methylation, with very few genes being methylated in each genome; the orthologues of these genes were never observed to be methylated in the other *Entamoeba* species analysed. The majority of the methylation was targeted to the non-coding portion of the genomes in all three species. In *E. histolytica,* the tRNA arrays showed high levels of methylation in the STR regions that separate the tRNA genes in the tRNA arrays. This pattern of methylation was also observed in *E. moshkovskii* and *E. invadens* however the levels of methylation were not as extreme in these other species. This may be explained by the lower depths of BS-seq coverage obtained for the *E. invadens* and *E. moshkovskii*, compared to *E. histolytica*. The lower level of coverage in the non-*histolytica* species meant that a lower proportion of cytosines were covered by 10x coverage and, as a result, genome-wide methylation in these two species may be underestimated. Further bisulphite sequencing of *E. moshkovskii* and *E. invadens* would allow for a greater coverage depth and improved confidence in the identification of methylated sites in these species.

6.3.1. Methylation of transposable elements in *Entamoeba histolytica*

Transposable elements (TEs) within the *E. histolytica* HM-1:IMSS genome were found to be highly methylated, with a large number of the TEs containing at least one highly methylated site. The analysis was not extended to *E. moshkovskii* and *E. invadens,* as the transposable elements of these species are not so well characterised. Methylation is assumed to silence the harmful TEs that can disrupt the genome however, if the virulence gene families are expanded through the propagation of TEs throughout the *E. histolytica* genome (Chapter 3) then some copies of the TEs must not be silenced or at least, be allowed to exist unmethylated in the genome for long enough for expansion to occur. It would be interesting to determine if the same pattern of TE methylation is present across all three species especially as the observation of many virulence gene families are suggested as being propagated by the TEs moving around the genome (Chapter 3).

6.4. Concluding remarks

Although a telomere-to-telomere genome assembly was not produced from the PacBio sequencing of *E. histolytica*, the new PacBio genome is a major improvement on the published assembly. This new assembly and subsequent annotation (Chapter 2 and 3) allowed for further analyses of the genome that were performed in Chapters 4 and 5. The analyses revealed a lot of information regarding the structure of the *E. histolytica* genome as well as the organisation of virulence gene families and a genome-wide study of methylation. The analysis of virulence gene families revealed their correlation with the position of TEs in the genome. The analyses reported that many gene families had increased in size, but perhaps most striking was the observation that the trichohyalin gene, which exists as a single copy in the published reference, is expanded in the PacBio genome assembly and forms a gene family with over 100 members. This observation highlights how third generation sequencing allows for better understanding of genome organisation (though it is possible this gene family underwent expansion within culture, subsequent to the original sequencing attempt).

The analyses of variation in the tRNA arrays provided a novel insight in to the length and structure of these arrays and provided additional evidence that these structures form the telomeres in *E. histolytica*. The genome-wide methylation analysis also confirmed that these regions were highly methylated and this DNA hyper-methylation may contribute to changing the conformation of these regions into condensed protective structures, though this will need to be experimentally validated. The sparse DNA methylation across the protein-coding regions of the three representative *Entamoeba* species' genomes suggests DNA methylation may not be a major regulator of gene expression, as has previously been suggested.

Overall, the new PacBio reference genome provides a platform for future studies of the biology, genetics and evolution of *Entamoeba* parasites.

References

1 Bapteste E, Brinkmann H, Lee JA, Moore D V, Sensen CW, Gordon P, Duruflé L, Gaasterland T, Lopez P, Müller Ml, Philippe, H: **The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba.** *Proc. Natl. Acad. Sci. U. S. A.* 2002, **99**:1414–9.

2 Song J, Xu Q, Olsen R, Loomis WF, Shaulsky G, Kuspa A, Sucgang R: **Comparing the Dictyostelium and Entamoeba Genomes Reveals an Ancient Split in the Conosa Lineage** *PLoS Comput. Biol.* 2005, **1**:e71.

3 Silberman JD, Clark CG, Diamond LS, Sogin ML: **Phylogeny of the genera Entamoeba and Endolimax as deduced from small- subunit ribosomal RNA sequences** *Mol. Biol. Evol.* 1999, **16**:1740–1751.

4 Clark CG, Ali IKM, Zaki M, Loftus BJ, Hall N: **Unique organisation of tRNA genes in Entamoeba histolytica.** *Mol. Biochem. Parasitol.* 2006, **146**:24–9.

5 Clark CG, Kaffashian F, Tawari B, Windsor J, Twigg-Flesner A, Davies-Morel MC, Biessmann H, Ebert F, Peschel B, Le Van A, Jackson, C. J, Macfarlane, L, Tannich, Egbert: **New insights into the phylogeny of Entamoeba species provided by analysis of four new small- subunit rRNA genes.** *Int. J. Syst. Evol. Microbiol.* 2006, **56**:2235–2239.

6 Walsh JA: **Problems in recognition and diagnosis of amebiasis: estimation of the global magnitude of morbidity and mortality.** *Rev. Infect. Dis.* 1986, **8**:228–38.

7 Bonner M, Amard V, Bar-Pinatel C, Charpentier F, Chatard J-M, Desmuyck Y, Ihler S, Rochet J-P, Roux de La Tribouille V, Saladin L, Verdy M, Gironès N, Fresno M, Santi-Rocca J: **Detection of the amoeba *Entamoeba gingivalis* in periodontal pockets** *Parasite* 2014, **21**:30.

8 Lyons T, Sholten T, Palmer JC: **Oral amoebiasis: a new approach for the general practitioner in the diagnosis and treatment of periodontal disease.** *Oral Health* 1980, **70**:39–41.

9 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**:340–4.

10 Stanley SL: **Amoebiasis.** *Lancet* 2003, **361**:1025–34.

11 Garcia LS: **Protozoa: Intestinal and Urogenital Amebae, Flagellates and Ciliates** In *Infectious Diseases*. Elsevier; 2017:1725–1733

12 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** *Protist* 2011, **162**:525–541.

13 Clark CG, Diamond LS: **Intraspecific variation and phylogenetic relationships in the genus Entamoeba as revealed by riboprinting.** *J. Eukaryot. Microbiol.* 1997, **44**:142–54.

14 Weedall GD, Hall N: **Evolutionary genomics of Entamoeba.** *Res. Microbiol.* 2011, **162**:637–45.

15 Silva MTN, Santana JV, Bragagnoli G, Marinho AM da N, Malagueño E: **Prevalence of Entamoeba histolytica/Entamoeba dispar in the city**

of Campina Grande, in northeastern Brazil. *Rev. Inst. Med. Trop. Sao Paulo* 2014, **56**:451–4.

16 Mortimer L, Chadee K: **The immunopathogenesis of Entamoeba histolytica** *Exp. Parasitol.* 2010, **126**:366–380.

17 Garcia LS, Bruckner DA: **Diagnostic medical parasitology**. ASM Press; 1997.

18 Abd-Alla MD, Jackson TG, Ravdin JI: **Serum IgM antibody response to the galactose-inhibitable adherence lectin of Entameoba histolytica.** *Am. J. Trop. Med. Hyg.* 1998, **59**:431–4.

19 Gathria V, Jackson TF: **A longitudinal study of asymptomatic carriers of pathogenic zymodemes of Entamoeba histolytica.** *South African Med. J.* 1987, **72**:669–72.

20 Haque R, Huston CD, Hughes M, Houpt E, Petri WA: **Amebiasis** *N. Engl. J. Med.* 2003, **348**:1565–1573.

21 Tanyuksel M, Petri WA, Jr.: **Laboratory diagnosis of amebiasis.** *Clin. Microbiol. Rev.* 2003, **16**:713–29.

22 Haque R, Ali IKM, Akther S, Petri WA: **Comparison of PCR, Isoenzyme Analysis, and Antigen Detection for Diagnosis of Entamoeba histolytica Infection.** *J. Clin. Microbiol.* 1998, **36**:449-452.

23 Haque R, Uddin Mollah N, Karim Ali IM, Alam K, Eubanks A, Lyerly D, Petri WA: **Diagnosis of Amebic Liver Abscess and Intestinal Infection with the TechLab Entamoeba histolytica II Antigen Detection and Antibody Tests.** *J. Clin. Microbiol.* 2000, **38**:3235-3239.

24 Speelman P, McGlaughlin R, Kabir I, Butler T: **Differential clinical features and stool findings in shigellosis and amoebic dysentery.** *Trans. R. Soc. Trop. Med. Hyg.* 1987, **81**:549–51.

25 Katz M, Despammier D, Gwadz R: **Parasitic Diseases**. Springer-Verlag; 1989.

26 Salles JM, Moraes LA, Salles MC: **Hepatic amebiasis** *Brazilian J. Infect. Dis.* 2003, **7**:96–110.

27 Adams E, MacLeod I: **Invasive amebiasis. II. Amebic liver abscess and its complications.** *Med.* 1977, **56**:325–34.

28 Thompson JE, Glasser AJ: **Amebic abscess of the liver. Diagnostic features.** *J. Clin. Gastroenterol.* 1986, **8**:550–4.

29 Nazir Z, Moazam F: **Amebic liver abscess in children.** *Pediatr. Infect. Dis. J.* 1993, **12**:929–32.

30 Maltz G, Knauer CM: **Amebic liver abscess: a 15-year experience.** *Am. J. Gastroenterol.* 1991, **86**:704–10.

31 Petri WA, Singh U: **Diagnosis and Management of Amebiasis** *Clin. Infect. Dis.* 1999, **29**:1117–1125.

32 Maldonado-Barrera CA, Campos-Esparza M del R, Muñoz-Fernández L, Victoria-Hernández JA, Campos-Rodríguez R, Talamás-Rohana P, Ventura-Juárez J: **Clinical case of cerebral amebiasis caused by E. histolytica** *Parasitol. Res.* 2012, **110**:1291–1296.

33 Lamps LW: **Infectious Causes of Appendicitis** *Infect. Dis. Clin. North Am.* 2010, **24**:995–1018.

34 Bumb RA, Mehta RD: **Amoebiasis cutis in HIV positive patient.** *Indian J. Dermatol. Venereol. Leprol.* 2006, **72**:224–6.

35 Quach J, St-Pierre J, Chadee K: **The future for vaccine development against *Entamoeba histolytica*** *Hum. Vaccin. Immunother.* 2014, **10**:1514–1521.

36 Ayed L Ben, Sabbahi S: **Entamoeba histolytica**. In *Global Water Pathogen Project*. Michigan State University; 2015:3–36.

37 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**:833–40.

38 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**:26051–6.

39 Forsgren A, Forssman L: **Metronidazole-resistant Trichomonas vaginalis.** *Br. J. Vener. Dis.* 1979, **55**:351–3.

40 Voolmann T, Boreham P: **Metronidazole resistant Trichomonas vaginalis in Brisbane.** *Med. J. Aust.* 1993, **159**:490.

41 Grossman JH, Galask RP: **Persistent vaginitis caused by metronidazole-resistant trichomonas.** *Obstet. Gynecol.* 1990, **76**:521–2.

42 Schmid G, Narcisi E, Mosure D, Secor WE, Higgins J, Moreno H: **Prevalence of metronidazole-resistant Trichomonas vaginalis in a gynecology clinic.** *J. Reprod. Med.* 2001, **46**:545–9.

43 McAuley JB, Juranek DD: **Luminal Agents in the Treatment of Amebiasis** *Clin. Infect. Dis.* 1992, **14**:1161–1162.

44 Venable S, Peterson A: **Unit VI : Pharmacotherapy for Gastrointestinal Tract Disorders: Parasitic Infections**. In *Pharmacotherapeutics for Advanced Practice: A Practical Approach.* Lippincott Williams and Wilkins; 2006:430–452.

45 World Health Organisation: **WHO Model Prescribing Information**. 1993.

46 Debnath A, Parsonage D, Andrade RM, He C, Cobo ER, Hirata K, Chen S, García-Rivera G, Orozco E, Martínez MB, Gunatilleke SS, Barrios AM, Arkin MR, Poole LB, McKerrow JH, Reed SL: **A high-throughput drug screen for Entamoeba histolytica identifies a new lead and target** *Nat. Med.* 2012, **18**:956–960.

47 Capparelli E V, Bricker-Ford R, Rogers MJ, Mckerrow JH, Reed SL: **Phase I Clinical Trial Results of Auranofin, a Novel Antiparasitic Agent** *Antimicrob. Agents. Chemother.* 2016, **61:**e01947-16

48 Fotedar R, Stark D, Beebe N, Marriott D, Ellis J, Harkness J: **Laboratory diagnostic techniques for Entamoeba species.** *Clin. Microbiol. Rev.* 2007, **20**:511–32.

49 Ximénez C, Morán P, Rojas L, Valadez A, Gómez A, Ramiro M, Cerritos R, González E, Hernández E, Oswaldo P: **Novelties on amoebiasis: a neglected tropical disease.** *J. Glob. Infect. Dis.* 2011, **3**:166–74.

50 Ximénez C, Morán P, Rojas L, Valadez A, Gómez A: **Reassessment of the epidemiology of amebiasis: state of the art.** *Infect. Genet. Evol.* 2009,

9:1023–32.

51 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**:578–83.

52 Ali IKM, Hossain MB, Roy S, Ayeh-Kumi PF, Petri WA, Haque R, Clark CG: **Entamoeba moshkovskii infections in children, Bangladesh.** *Emerg. Infect. Dis.* 2003, **9**:580–4.

53 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**:1141–3.

54 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**:549–52.

55 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**:106–10.

56 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.

57 Kannathasan S, Murugananthan A, Kumanan T, Iddawala D, de Silva NR, Rajeshkannan N, Haque R: **Amoebic liver abscess in northern Sri Lanka: first report of immunological and molecular confirmation of aetiology.** *Parasit. Vectors* 2017, **10**:14.

58 Andersson Y, de Jong B: **An Outbreak of Giardiasis and Amoebiasis at a Ski Resort in Sweden** *Water Sci. Technol.* 1989, **21**:143–146.

59 Chen KT, Chen CJ, Chiu JP: **A school waterborne outbreak involving both Shigella sonnei and Entamoeba histolytica.** *J. Environ. Health* 2001, **64**:9–13, 26.

60 Barwick RS, Uzicanin A, Lareau S, Malakmadze N, Imnadze P, Iosava M, Ninashvili N, Wilson M, Hightower AW, Johnston S, et al.: **Outbreak of amebiasis in Tbilisi, Republic of Georgia, 1998.** *Am. J. Trop. Med. Hyg.* 2002, **67**:623–31.

61 Clark CG, Diamond LS: **The Laredo strain and other "Entamoeba histolytica-like" amoebae are Entamoeba moshkovskii.** *Mol. Biochem. Parasitol.* 1991, **46**:11–8.

62 Jackson TF: **Entamoeba histolytica and Entamoeba dispar are distinct species; clinical, epidemiological and serological evidence.** *Int. J. Parasitol.* 1998, **28**:181–6.

63 WHO, PAHO, UNESCO: **WHO/PAHO/UNESCO report. A consultation with experts on amoebiasis. Mexico City, Mexico 28-29 January, 1997.** *Epidemiol. Bull.* 1997, **18**:13–4.

64 Feng M, Cai J, Min X, Fu Y, Xu Q, Tachibana H, Cheng X: **Prevalence and genetic diversity of Entamoeba species infecting macaques in southwest China** *Parasitol. Res.* 2013, **112**:1529–1536.

65 Guan Y, Feng M, Cai J, Min X, Zhou X, Xu Q, Tan N, Cheng X, Tachibana H: **Comparative analysis of genotypic diversity in Entamoeba nuttalli**

isolates from Tibetan macaques and rhesus macaques in China *Infect. Genet. Evol.* 2016, **38**:126–131.

66 Tachibana H, Yanagi T, Akatsuka A, Kobayashi S, Kanbara H, Tsutsumi V: **Isolation and characterization of a potentially virulent species Entamoeba nuttalli from captive Japanese macaques** *Parasitology* 2009, **136**:1169-77.

67 Tachibana H, Yanagi T, Lama C, Pandey K, Feng M, Kobayashi S, Sherchand JB: **Prevalence of Entamoeba nuttalli infection in wild rhesus macaques in Nepal and characterization of the parasite isolates** *Parasitol. Int.* 2013, **62**:230–235.

68 Tachibana H, Yanagi T, Feng M, Bandara KBAT, Kobayashi S, Cheng X, Hirayama K, Rajapakse RPVJ: **Isolation and Molecular Characterization of *Entamoeba nuttalli* Strains Showing Novel Isoenzyme Patterns from Wild Toque Macaques in Sri Lanka** *J. Eukaryot. Microbiol.* 2016, **63**:171–180.

69 Tuda J, Feng M, Imada M, Kobayashi S, Cheng X, Tachibana H: **Identification of *Entamoeba polecki* with Unique 18S rRNA Gene Sequences from Celebes Crested Macaques and Pigs in Tangkoko Nature Reserve, North Sulawesi, Indonesia** *J. Eukaryot. Microbiol.* 2016, **63**:572–577.

70 Takano J, Tachibana H, Kato M, Narita T, Yanagi T, Yasutomi Y, Fujimoto K: **DNA characterization of simian Entamoeba histolytica-like strains to differentiate them from Entamoeba histolytica** *Parasitol. Res.* 2009, **105**:929–937.

71 Levecke B, Dreesen L, Dorny P, Verweij JJ, Vercammen F, Casaert S, Vercruysse J, Geldhof P: **Molecular Identification of Entamoeba spp. in Captive Nonhuman Primates**. *J. Clin. Microbiol.* 2010, **48**:2988–2990.

72 Levecke B, Dorny P, Vercammen F, Visser LG, Van Esbroeck M, Vercruysse J, Verweij JJ: **Transmission of Entamoeba nuttalli and Trichuris trichiura from Nonhuman Primates to Humans.** *Emerg. Infect. Dis.* 2015, **21**:1871–2.

73 Shibayama M, Dolabella SS, Silva EF, Tsutsumi V: **A Brazilian species of Entamoeba dispar (ADO) produces amoebic liver abscess in hamsters.** *Ann. Hepatol.* 2007, **6**:117–8.

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

75 Neal RA: **Studies on the morphology and biology of Entamoeba moshkovskii Tshalaia, 1941.** *Parasitology* 1953, **43**:253–68.

76 Tshalaia L: **A Species of Entamoeba detected in Sewage.** *Meditsinskaya Parazitol. i Parazit. Bolezn.* 1941, **10**:244–252.

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

78 Shimokawa C, Kabir M, Taniuchi M, Mondal D, Kobayashi S, Ali IKM, Sobuz SU, Senba M, Houpt E, Haque R, Petri 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.

79 Balamuth W: **Effects of some Environmental Factors upon Growth and Encystation of Entamoeba invadens** *J. Parasitol.* 1914, **48**:101–9.

80 Sanchez L, Enea V, Eichinger D: **Identification of a developmentally regulated transcript expressed during encystation of Entamoeba invadens.** *Mol. Biochem. Parasitol.* 1994, **67**:125–35.

81 Kojimoto A, Uchida K, Horii Y, Okumura S, Yamaguch R, Tateyama S: **Amebiasis in four ball pythons, Python reginus.** *J. Vet. Med. Sci.* 2001, **63**:1365–8.

82 Ehrenkaufer GM, Weedall GD, Williams D, Lorenzi HA, Caler E, Hall N, Singh U, Ehrenkaufer, G. M., Weedall, G. D., Williams, D., Lorenzi, H. A., Caler, E., Hall, N., & Singh, U: **The genome and transcriptome of the enteric parasite Entamoeba invadens, a model for encystation.** *Genome Biol.* 2013, **14**:R77.

83 Ehrenkaufer GM, Suresh S, Solow-Cordero D, Singh U: **High-Throughput Screening of Entamoeba Identifies Compounds Which Target Both Life Cycle Stages and Which Are Effective Against Metronidazole Resistant Parasites.** *Front. Cell. Infect. Microbiol.* 2018, **8**:276.

84 Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, 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: **EuPathDB: the eukaryotic pathogen database.** *Nucleic Acids Res.* 2013, **41**:D684–91.

85 Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S Paulsen, Ian T, James K, Eisen JA, Rutherford K, Salzberg SL, Craig A, Kyes S, Chan M, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian GM, Mungall C, Venter, CJ, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser C, Barrell B: **Genome sequence of the human malaria parasite Plasmodium falciparum.** *Nature* 2002, **419**:498–511.

86 Berriman M, Ghedin E, Hertz-Fowler C., Blandin G, Renauld H, Bartholomeu D, Lennard N, Caler E, Hamlin N, Haas B, Böhme U, Hannick L, Aslett MA, Shallom J, Marcello L, Hou L, Wickstead B, Alsmark UC, Arrowsmith C, Atkin RJ, Barron AJ, Bringaud F, Brooks K, Carrington M, Cherevach I, Chillingworth TJ, Churcher C, Clark LN, Corton CH, Cronin A, Davies RM, Doggett J, Djikeng A, Feldblyum T, Field MC, Fraser A, Goodhead I, Hance Z, Harper D, Harris BR, Hauser H, Hostetler J, Ivens A, Jagels K, Johnson D, Johnson J, Jones K, Kerhornou AX, Koo H, Larke N, Landfear S, Larkin C, Leech V, Line A, Lord A, Macleod A, Mooney PJ, Moule S, Martin DM, Morgan GW, Mungall K, Norbertczak H, Ormond D, Pai G, Peacock CS, Peterson J, Quail MA, Rabbinowitsch E, Rajandream MA, Reitter C, Salzberg SL, Sanders M, Schobel S, Sharp S, Simmonds M, Simpson AJ, Tallon L, Turner CM, Tait A, Tivey AR, Van Aken S, Walker D, Wanless D, Wang

S, White B, White O, Whitehead S, Woodward J, Wortman J, Adams MD, Embley TM, Gull K, Ullu E, Barry JD, Fairlamb AH, Opperdoes F, Barrell BG, Donelson JE, Hall N, Fraser CM, Melville SE, El-Sayed NM: **The Genome of the African Trypanosome Trypanosoma brucei** *Science.* 2005, **309**:416–422.

87 El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, Ghedin E, Worthey EA, Delcher AL, Blandin G, Westenberger SJ, Caler E, Cerqueira GC, Branche C, Haas B, Anupama A, Arner E, Aslund L, Attipoe P, Bontempi E, Bringaud F, Burton P, Cadag E, Campbell DA, Carrington M, Crabtree J, Darban H, da Silveira JF, de Jong P, Edwards K, Englund PT, Fazelina G, Feldblyum T, Ferella M, Frasch AC, Gull K, Horn D, Hou L, Huang Y, Kindlund E, Klingbeil M, Kluge S, Koo H, Lacerda D, Levin MJ, Lorenzi H, Louie T, Machado CR, McCulloch R, McKenna A, Mizuno Y, Mottram JC, Nelson S, Ochaya S, Osoegawa K, Pai G, Parsons M, Pentony M, Pettersson U, Pop M, Ramirez JL, Rinta J, Robertson L, Salzberg SL, Sanchez DO, Seyler J, Sharma R, Shetty J, Simpson AJ, Sisk E, Tammi MT, Tarleton R, Teixeira S, Van Aken S, Vogt C, Ward PN, Wickstead B, Wortman J, White O, Fraser CM, Stuart KD, Andersson B: **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease.** *Science* 2005, **309**:409–15.

88 Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream M-A, Adlem E, Aert R, Anupama A, Apostolou Z, Attipoe P, Bason N, Bauser C, Beck A, Beverley S, Bianchettin G, Borzym K, Bothe G, Bruschi C, Collins M, Cadag E, Ciarloni L, Clayton C, Coulson R, Cronin A, Cruz A, Davies R, De Gaudenzi J, Dobson D, Duesterhoeft A, Fazelina G, Fosker N, Frasch A, Fraser A, Fuchs M, Gabel C, Goble A, Goffeau A, Harris D, Hertz-Fowler C, Hilbert H, Horn D, Huang Y, Klages S, Knights A, Kube M, Larke N, Litvin L, Lord A, Louie T, Marra M, Masuy D, Matthews K, Michaeli S, Mottram J, Müller-Auer S, Munden H, Nelson S, Norbertczak H, Oliver K, O'neil S, Pentony M, Pohl T, Price C, Purnelle B, Quail M, Rabbinowitsch E, Reinhardt R, Rieger M, Rinta J, Robben J, Robertson L, Ruiz J, Rutter S, Saunders D, Schäfer M, Schein J, Schwartz D, Seeger K, Seyler A, Sharp S, Shin H, Sivam D, Squares R, Squares S, Tosato V, Vogt C, Volckaert G, Wambutt R, Warren T, Wedler H, Woodward J, Zhou S, Zimmermann W, Smith D, Blackwell J, Stuart K, Barrell B, Myler P: **The genome of the kinetoplastid parasite, Leishmania major.** *Science* 2005, **309**:436–42.

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

315:207–12.

90 Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, Olsen GJ, Best AA, Cande WZ, Chen F, Cipriano MJ, Davids BJ, Dawson SC, Elmendorf HG, Hehl AB, Holder ME, Huse SM, Kim UU, Lasek-Nesselquist E, Manning G, Nigam A, Nixon, JEJ, Palm D, Passamaneck NE, Prabhu A, Reich CI, Reiner DS, Samuelson J, Svard SG, Sogin ML: **Genomic Minimalism in the Early Diverging Intestinal Parasite Giardia lamblia** *Science.* 2007, **317**:1921–1926.

91 Franzén O, Jerlström-Hultqvist J, Castro E, Sherwood E, Ankarklev J, Reiner DS, Palm D, Andersson JO, Andersson B, Svärd SG: **Draft Genome Sequencing of Giardia intestinalis Assemblage B Isolate GS: Is Human Giardiasis Caused by Two Different Species?** *PLoS Pathog.* 2009, **5**:e1000560.

92 Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson, William R, Dear PH, Bankier AT, Peterson DL, Abrahamsen KV, Tzipori S, Buck G: **The genome of Cryptosporidium hominis.** *Nature* 2004, **431**:1107–12.

93 Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, Deng M, Liu C, Widmer G, Tzipori S, Buck, Gregory A, Xu P, Bankier AT, Dear PH, Konfortov BA, Spriggs HF, Iyer L, Anantharaman V, Aravind L, Kapur V.: **Complete genome sequence of the apicomplexan, Cryptosporidium parvum.** *Science* 2004, **304**:441–5.

94 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.

95 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, Squares R, Whitehead S, Quail M, Rabbinowitsch E, Norbertczak H, Price C, Wang Z, Guillén N, Gilchrist C, Stroup S, Bhattacharya S, Lohia A, Foster P, Sicheritz-Ponten T, Weber C, Singh U, Mukherjee C, El-Sayed N, Petri W, Clark C, Embley T, Barrell B, Fraser C, Hall N: **The genome of the protist parasite Entamoeba histolytica.** *Nature* 2005, **433**:865–8.

96 Lorenzi HA, Puiu D, Miller JR, Brinkac LM, Amedeo P, Hall N, Caler E V: **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.

97 Meerovitch E, Ghadirian E: **Restoration of virulence of axenically cultivated *Entamoeba histolytica* by cholesterol.** *Can. J. Microbiol.* 1978, **24**:63–65.

98 Thibeaux R, Weber C, Hon C-C, Dillies M-A, Avé P, Coppée J-Y, Labruyère E, Guillén N: **Identification of the virulence landscape essential for Entamoeba histolytica invasion of the human colon.** *PLoS Pathog.* 2013, **9**:e1003824.

99 Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat. Methods* 2008, **5**:16–18.

100 Monaco AP, Larin Z: **YACs, BACs, PACs and MACs: Artificial chromosomes as research tools**. *Trends Biotechnol.* 1994, **12**:280–286.

101 Eichinger L, Pachebat JA, Glöckner G, Rajandream M-A, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov B, Rivero F, Bankier A, 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, Hemphill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, van Driessche N, Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Madan Babu M, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Mungall K, Oliver K, Price C, Quail M, Urushihara H, Hernandez J, Rabbinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shaulsky G, Schleicher M, Weinstock G, Rosenthal A, Cox E, Chisholm R, Gibbs R, Loomis W, Platzer M, Kay R, Williams J, Dear P, Noegel A, Barrell B, Kuspa A: **The genome of the social amoeba Dictyostelium discoideum** *Nature* 2005, **435**:43–57.

102 Willhoeft U, Tannich E: **The electrophoretic karyotype of Entamoeba histolytica** *Mol. Biochem. Parasitol.* 1999, **99**:41–53.

103 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**:787–9.

104 Riveron a M, Lopez-Canovas L, Baez-Camargo M, Flores E, Perez-Perez G, Luna-Arias JP, Orozco E: **Circular and linear DNA molecules in the Entamoeba histolytica complex molecular karyotype.** *Eur. Biophys. J.* 2000, **29**:48–56.

105 Dhar SK, Choudhury NR, Bhattacharaya A, Bhattacharya S: **A multitude of circular DNAs exist in the nucleus of Entamoeba histolytica.** *Mol. Biochem. Parasitol.* 1995, **70**:203–6.

106 Lioutas C, Schmetz C, Tannich E: **Identification of Various Circular DNA Molecules in Entamoeba histolytica** *Exp. Parasitol.* 1995, **80**:349–352.

107 Báez-Camargo M, Riverón AM, Delgadillo DM, Flores E, Sánchez T, Garcia-Rivera G, Orozco E: **Entamoeba histolytica: gene linkage groups and relevant features of its karyotype.** *Mol. Gen. Genet.* 1996, **253**:289–96.

108 Bhattacharya S, Som I, Bhattacharya A: **The ribosomal DNA plasmids of Entamoeba.** *Parasitol. Today* 1998, **14**:181–5.

109 Dear PH, Cook PR: **Happy mapping: linkage mapping using a physical analogue of meiosis.** *Nucleic Acids Res.* 1993, **21**:13–20.

110 Weedall GD: **The Genomics of Entamoebae: Insights and Challenges**. In *Amebiasis*. Edited by Nozaki T, Bhattacharya A. Springer; 2015:27–47.

111 Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK: **Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping.** *Science* 1993, **262**:110–4.

112 Mullikin JC, Ning Z: **The Phusion Assembler** *Genome Res.* 2003, **13**:81–90.

113 Cawley SE, Wirth AI, Speed TP: **Phat--a gene finding program for Plasmodium falciparum.** *Mol. Biochem. Parasitol.* 2001, **118**:167–74.

114 Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders** *Bioinformatics* 2004, **20**:2878–2879.

115 Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer, Erik LL, Tate J, Punta M: **Pfam: the protein families database** *Nucleic Acids Res.* 2014, **42**:D222–D230.

116 Roberts M, Hunt BR, Yorke JA, Bolanos RA, Delcher AL: **A Preprocessor for Shotgun Assembly of Large Genomes** *J. Comput. Biol.* 2004, **11**:734–752.

117 Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, Anson EL, Bolanos RA, Chou HH, Jordan CM, Halpern AL, Lonardi S, Beasley EM, Brandon RC, Chen L, Dunn PJ, Lai Z, Liang Y, Nusskern DR, Zhan M, Zhang Q, Zheng X, Rubin GM, Adams MD, Venter JC: **A whole-genome assembly of Drosophila.** *Science* 2000, **287**:2196–204.

118 Allen JE, Majoros WH, Pertea M, Salzberg SL: **JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions** *Genome Biol.* 2006, **7**:S9.

119 Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell R, Wortman JR: **Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments** *Genome Biol.* 2008, **9**:R7.

120 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.

121 Kuhlmann M, Borisova BE, Kaller M, Larsson P, Stach D, Na J, Eichinger L, Lyko F, Ambros V, Söderbom F, Hammann C, Nellen W: **Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi** *Nucleic Acids Res.* 2005, **33**:6405–6417.

122 Ghosh S, Satish S, Tyagi S, Bhattacharya A, Bhattacharya S: **Differential use of multiple replication origins in the ribosomal DNA episome of the protozoan parasite Entamoeba histolytica** *Nucleic Acids Res.* 2003, **31**:2035–2044.

123 Bhattacharya S, Bhattacharya A, Diamond LS, Soldo AT: **Circular DNA of Entamoeba histolytica Encodes Ribosomal RNA** *J. Protozool.* 1989, **36**:455–458.

124 Sucgang R, Chen G, Liu W, Lindsay R, Lu J, Muzny D, Shaulsky G, Loomis W, Gibbs R, Kuspa A: **Sequence and structure of the**

extrachromosomal palindrome encoding the ribosomal RNA genes in Dictyostelium. *Nucleic Acids Res.* 2003, **31**:2361–8.

125 Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler E V, Fischer S, Gajria B, Gao X, Gingle A, Grant G, et al.: **AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species.** *Nucleic Acids Res.* 2011, **39**:D612–9.

126 Clark CG, Alsmark UCM, Tazreiter M, Saito-Nakano Y, Ali V, Marion S, Weber C, Mukherjee C, Bruchhaus I, Tannich E, et al.: **Structure and Content of the Entamoeba histolytica Genome** *Adv. Parasitol.* 2007, **65**:51–190.

127 Davis PH, Zhang Z, Chen M, Zhang X, Chakraborty S, Stanley SL, Jr.: **Identification of a family of BspA like surface proteins of Entamoeba histolytica with novel leucine rich repeats.** *Mol. Biochem. Parasitol.* 2006, **145**:111–6.

128 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.

129 Wilson IW, Weedall GD, Hall N: **Host-Parasite interactions in Entamoeba histolytica and Entamoeba dispar: what have we learned from their genomes?** *Parasite Immunol.* 2012, **34**:90–9.

130 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.

131 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.

132 Meléndez-López SG, Herdman S, Hirata K, Choi M-H, Choe Y, Craik C, Caffrey CR, Hansell E, Chávez-Munguía B, Chen YT, et al.: **Use of Recombinant Entamoeba histolytica Cysteine Proteinase 1 To Identify a Potent Inhibitor of Amebic Invasion in a Human Colonic Model Downloaded from** *Eukaryot. Cell* 2007, **6**:1130–1136.

133 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.

134 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.

135 Moncada D, Keller K, Chadee K: **Entamoeba histolytica-secreted products degrade colonic mucin oligosaccharides.** *Infect. Immun.* 2005, **73**:3790–3.

136 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.

137 Ravdin JI, Stanley P, Murphy CF, Petri WA: **Characterization of cell surface carbohydrate receptors for Entamoeba histolytica adherence lectin.** *Infect. Immun.* 1989, **57**:2179–86.

138 Ravdin JI, Croft BY, Guerrant RL: **Cytopathogenic mechanisms of Entamoeba histolytica.** *J. Exp. Med.* 1980, **152**:377–90.

139 MacFarlane RC, Singh U: **Identification of differentially expressed genes in virulent and nonvirulent Entamoeba species: potential implications for amebic pathogenesis.** *Infect. Immun.* 2006, **74**:340–51.

140 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**:5703–10.

141 Hirt RP, Harriman N, Kajava A V, 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** *Mol. Biochem. Parasitol.* 2002, **125**:195–199.

142 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.

143 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**:353–5.

144 Biller L, Davis PH, Tillack M, Matthiesen J, Lotter H, Stanley SL, Tannich E, Bruchhaus I, 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.

145 Gilchrist CA, Houpt E, Trapaidze N, Fei Z, Crasta O, Asgharpour A, Evans C, Martino-Catt S, Baba DJ, Stroup S, et al.: **Impact of intestinal colonization and invasion on the Entamoeba histolytica transcriptome** *Mol. Biochem. Parasitol.* 2006, **147**:163–176.

146 Reuber TL, Ausubel FM: **Isolation of Arabidopsis Genes That Differentiate between Resistance Responses Mediated by the RPS2 and RPM1 Disease Resistance Genes** *Plant Cell Online* 1996, **8**:241–249.

147 Nakada-Tsukui K, Sekizuka T, Sato-Ebine E, Escueta-de Cadiz A, Ji D, Tomii K, Kuroda M, Nozaki T: **AIG1 affects in vitro and in vivo virulence in clinical isolates of Entamoeba histolytica** *PLOS Pathog.* 2018, **14**:e1006882.

148 Zamorano A, López-Camarillo C, Orozco E, Weber C, Guillen N, Marchat LA: **In silico analysis of EST and genomic sequences allowed the prediction of cis-regulatory elements for Entamoeba histolytica mRNA polyadenylation** *Comput. Biol. Chem.* 2008, **32**:256–263.

149 Bruchhaus I, Leippe M, Lioutas C, Tannich E: **Unusual gene organization in the protozoan parasite Entamoeba histolytica.** *DNA Cell Biol.* 1993, **12**:925–33.

150 Hon C-C, Weber C, Sismeiro O, Proux C, Koutero M, Deloger M, Das S, Agrahari M, Dillies M-A, Jagla B, Coppee J, Bhattacharya A, Guillen N: **Quantification of stochastic noise of splicing and polyadenylation**

in Entamoeba histolytica *Nucleic Acids Res.* 2013, **41**:1936–1952.

151 Ketting RF: **The Many Faces of RNAi** *Dev. Cell* 2011, **20**:148–161.

152 Agrawal N, Dasaradhi PVN, Mohmmed A, Malhotra P, Bhatnagar RK, Mukherjee SK: **RNA interference: biology, mechanism, and applications.** *Microbiol. Mol. Biol. Rev.* 2003, **67**:657–85.

153 Ghildiyal M, Zamore PD: **Small silencing RNAs: an expanding universe** *Nat. Rev. Genet.* 2009, **10**:94–108.

154 Hutvagner G, Simard MJ: **Argonaute proteins: key players in RNA silencing** *Nat. Rev. Mol. Cell Biol.* 2008, **9**:22–32.

155 Kuhn C-D, Joshua-Tor L: **Eukaryotic Argonautes come into focus** *Trends Biochem. Sci.* 2013, **38**:263–271.

156 Joshua-Tor L: **The Argonautes.** *Cold Spring Harb. Symp. Quant. Biol.* 2006, **71**:67–72.

157 Wilson RC, Doudna JA: **Molecular Mechanisms of RNA Interference** *Annu. Rev. Biophys.* 2013, **42**:217–239.

158 Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications.** *Cell Res.* 2011, **21**:381–95.

159 Zhang H, Ehrenkaufer GM, Pompey JM, Hackney JA, Singh U: **Small RNAs with 5′-Polyphosphate Termini Associate with a Piwi-Related Protein and Regulate Gene Expression in the Single-Celled Eukaryote Entamoeba histolytica** *PLoS Pathog.* 2008, **4**:e1000219.

160 Zhang H, Pompey JM, Singh U: **RNA interference in *Entamoeba histolytica*: implications for parasite biology and gene silencing** *Future Microbiol.* 2011, **6**:103–117.

161 Pak J, Fire A: **Distinct Populations of Primary and Secondary Effectors During RNAi in C. elegans** *Science.* 2007, **315**:241–244.

162 Zhang H, Alramini H, Tran V, Singh U: **Nucleus-localized Antisense Small RNAs with 5′-Polyphosphate Termini Regulate Long Term Transcriptional Gene Silencing in *Entamoeba histolytica* G3 Strain** *J. Biol. Chem.* 2011, **286**:44467–79.

163 Pompey JM, Foda B, Singh U: **A Single RNaseIII Domain Protein from Entamoeba histolytica Has dsRNA Cleavage Activity and Can Help Mediate RNAi Gene Silencing in a Heterologous System** *PLoS One* 2015, **10**:e0133740.

164 Huguenin M, Bracha R, Chookajorn T, Mirelman D: **Epigenetic transcriptional gene silencing in Entamoeba histolytica: insight into histone and chromatin modifications** *Parasitology* 2010, **137**:619-27.

165 Morf L, Pearson RJ, Wang AS, Singh U: **Robust gene silencing mediated by antisense small RNAs in the pathogenic protist Entamoeba histolytica** *Nucleic Acids Res.* 2013, **41**:9424–9437.

166 Pearson RJ, Morf L, Singh U: **Regulation of H2O2 stress-responsive genes through a novel transcription factor in the protozoan pathogen Entamoeba histolytica.** *J. Biol. Chem.* 2013, **288**:4462–74.

167 Zhang H, Ehrenkaufer GM, Manna D, Hall N, Singh U: **High Throughput Sequencing of Entamoeba 27nt Small RNA Population Reveals Role in Permanent Gene Silencing But No Effect on Regulating Gene Expression Changes during Stage Conversion, Oxidative, or Heat Shock Stress** *PLoS One* 2015, **10**:e0134481.

168 Zhang H, Ehrenkaufer GM, Hall N, Singh U: **Small RNA pyrosequencing in the protozoan parasite Entamoeba histolytica reveals strain-specific small RNAs that target virulence genes.** *BMC Genomics* 2013, **14**:53.

169 Ramakrishnan V: **Histone Structure and the Organisation of the Nucleosome** *Annu. Rev. Biophys. Biomol. Struct.* 1997, **26**:83–112.

170 Ballestar E, Paz MF, Valle L, Wei S, Fraga MF, Espada J, Cigudosa JC, Huang TH-M, Esteller M: **Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer.** *EMBO J.* 2003, **22**:6335–45.

171 Kirschenbaum M, Ankri S: **Entamoeba histolytica: Bridging the Gap Between Environmental Stress and Epigenetic Regulation**. In Amebiasis. Edited by Nozaki T, Bhattacharya A. Springer; 2015:171–185.

172 Harony H, Bernes S, Siman-Tov R, Ankri S: **DNA methylation and targeting of LINE retrotransposons in Entamoeba histolytica and Entamoeba invadens** *Mol. Biochem. Parasitol.* 2006, **147**:55–63.

173 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 Biol.* 2012, **13**:R38.

174 Gilchrist CA, Ali IKM, Kabir M, Alam F, Scherbakova S, Ferlanti E, Weedall GD, Hall N, Haque R, Petri 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.

175 McCarthy A: **Third generation DNA sequencing: Pacific Biosciences' single molecule real time technology.** *Chem. Biol.* 2010, **17**:675–6.

176 Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323**:133–8.

177 Shi L, Guo Y, Dong C, Huddleston J, Yang H, Han X, Fu A, Li Q, Li N, Gong S, Lintner K, Ding Q, Wang Z, Hu J, Wang D, Wang F, Wang L, Lyon G, Guan Y, Shen Y, Evgrafov O, Knowles J, Thibaud-Nissen F, Schneider V, Yu C, Zhou L, Eichler E, So K, Wang K: **Long-read sequencing and de novo assembly of a Chinese genome** *Nat. Commun.* 2016, **7**:12065.

178 Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, Hastie A, Cao H, Yun J-Y, Kim J, Kuk J, Park GH, Kim J, Ryu H, Kim, J Roh M, Baek J, Hunkapiller MW, Korlach J, Shin J, Kim C: **De novo assembly and phasing of a Korean human genome** *Nature* 2016, **538**:243–247.

179 Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin C, Korlach J, Wilson RK, Eichler EE: **Discovery and genotyping of structural variation from long-read haploid**

genome sequence data *Genome Res.* 2017, **27**:677–685.

180 Pasini EM, Böhme U, Rutledge GG, Voorberg-Van der Wel A, Sanders M, Berriman M, Kocken CH, Otto TD: **An improved Plasmodium cynomolgi genome assembly reveals an unexpected methyltransferase gene expansion** *Wellcome Open Res.* 2017, **2**:42.

181 Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C-S,Guill K, Regulski M, Kumari S, Olson A, Gent J, Schneider KL, Wolfgruber TK, May MR, Springer NM, Antoniou E, McCombie RW, Presting GG, McMullen M, Ross-Ibarra J, Dawe RK, Hastie A, Rank DR, Ware D: **Improved maize reference genome with single-molecule technologies.** *Nature* 2017, **546**:524.

182 Kaur P, Bayer PE, Milec Z, Vrána J, Yuan Y, Appels R, Edwards D, Batley J, Nichols P, Erskine W, Doležel J: **An advanced reference genome of *Trifolium subterraneum* L. reveals genes related to agronomic performance.** *Plant Biotechnol. J.* 2017, **15**:1034–1046.

183 Biggs BA, Goozé L, Wycherley K, Wollish W, Southwell B, Leech JH, Brown G V: **Antigenic variation in Plasmodium falciparum.** *Proc. Natl. Acad. Sci. U. S. A.* 1991, **88**:9171–4.

184 Roberts DJ, Craig AG, Berendt AR, Pinches R, Nash G, Marsh K, Newbold CI: **Rapid switching to multiple antigenic and adhesive phenotypes in malaria** *Nature* 1992, **357**:689–692.

185 Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA, Wellems TE: **The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of Plasmodium falciparum-infected erythrocytes.** *Cell* 1995, **82**:89–100.

186 Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, Pinches R, Newbold CI, Miller LH: **Switches in expression of Plasmodium falciparum var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes.** *Cell* 1995, **82**:101–10.

187 Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ: **Cloning the P. falciparum gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes.** *Cell* 1995, **82**:77–87.

188 Kraemer SM, Smith JD: **A family affair: var genes, PfEMP1 binding, and malaria disease** *Curr. Opin. Microbiol.* 2006, **9**:374–380.

189 Deshmukh AS, Srivastava S, Dhar SK: **Plasmodium falciparum: Epigenetic Control of var Gene Regulation and Disease** In *Epigenetics: Development and Disease*. Springer, Dordrecht; 2013:659–682.

190 Cross GA: **Crossreacting determinants in the C-terminal region of trypanosome variant surface antigens.** *Nature* 1979, **277**:310–2.

191 Cross GA: **Identification, purification and properties of clone-specific glycoprotein antigens constituting the surface coat of Trypanosoma brucei.** *Parasitology* 1975, **71**:393–417.

192 Blum ML, Down JA, Gurnett AM, Carrington M, Turner MJ, Wiley DC: **A structural motif in the variant surface glycoproteins of Trypanosoma brucei** *Nature* 1993, **362**:603–609.

193 Nanavaty V, Sandhu R, Jehi SE, Pandya UM, Li B: **Trypanosoma brucei RAP1 maintains telomere and subtelomere integrity by suppressing TERRA and telomeric RNA:DNA hybrids** *Nucleic Acids Res.* 2017, **45**:5785–5796.

194 Kaur G, Lohia A: **Inhibition of gene expression with double strand RNA interference in Entamoeba histolytica** *Biochem. Biophys. Res. Commun.* 2004, **320**:1118–1122.

195 Solis CF, Guillén N: **Silencing Genes by RNA Interference in the Protozoan Parasite Entamoeba histolytica** *Methods Mol. Biol.* 2008, **442**:113–128.

196 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.

197 Linford AS, Moreno H, Good KR, Zhang H, Singh U, Petri WA, Jr: **Short hairpin RNA-mediated knockdown of protein expression in Entamoeba histolytica.** *BMC Microbiol.* 2009, **9**:38.

198 Cortés A, Carret C, Kaneko O, Yim Lim BYS, Ivens A, Holder AA: **Epigenetic Silencing of Plasmodium falciparum Genes Linked to Erythrocyte Invasion** *PLoS Pathog.* 2007, **3**:e107.

199 Dixon SE, Stilger KL, Elias E V, Naguleswaran A, Sullivan WJ, Jr.: **A decade of epigenetic research in Toxoplasma gondii.** *Mol. Biochem. Parasitol.* 2010, **173**:1–9.

200 Alsford S, duBois K, Horn D, Field MC: **Epigenetic mechanisms, nuclear architecture and the control of gene expression in trypanosomes** *Expert Rev. Mol. Med.* 2012, **14**:e13.

201 Croken MM, Nardelli SC, Kim K: **Chromatin modifications, epigenetics, and how protozoan parasites regulate their lives** *Trends Parasitol.* 2012, **28**:202–213.

202 Gilchrist CA, Petri WA: **Using differential gene expression to study Entamoeba histolytica pathogenesis** *Trends Parasitol.* 2009, **25**:124–131.

203 Løbner-Olesen A, Skovgaard O, Marinus MG: **Dam methylation: coordinating cellular processes** *Curr. Opin. Microbiol.* 2005, **8**:154–160.

204 Heithoff D, Sinsheimer R, Low D, Mahan M: **An Essential Role for DNA Adenine Methylation in Bacterial Virulence**. *Science..* 1999, **284**:967–970.

205 Marinus MG, Casadesus J: **Roles of DNA adenine methylation in host–pathogen interactions: mismatch repair, transcriptional regulation, and more** *FEMS Microbiol. Rev.* 2009, **33**:488–503.

206 Korlach J: **Perspective - Understanding Accuracy in SMRT Sequencing** *Pacific Biosciences.* 2013.

207 Detter JC, Johnson SL, Bishop-Lilly KA, Chain PS, Gibbons HS, Minogue TD, Sozhamannan S, Van Gieson EJ, Resnick IG: **Nucleic acid sequencing for characterizing infectious and/or novel agents in complex samples** In *Biological Identification*. Elsevier; 2014:3–53.

208 Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J: **Nonhybrid, finished microbial genome assemblies from long-read**

SMRT sequencing data. *Nat. Methods* 2013, **10**:563–9.

209 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.** *Genome Res.* 2017, **27**:722–736.

210 Phillippy A, Koren S, Walenz BP: **Canu: A single molecule sequence assembler for genomes large and small** *Github* 2015.

211 Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC: **Phased diploid genome assembly with single-molecule real-time sequencing.** *Nat. Methods* 2016, **13**:1050–1054.

212 Li H: **Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.** *Bioinformatics* 2015, **32**:2103-10

213 Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM: **Reducing assembly complexity of microbial genomes with single-molecule sequencing** *Genome Biol.* 2013, **14**:R101.

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

215 Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou W, Corbeil J, Del Fabbro C, Docking T, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca N, Ganapathy G, Gibbs R, Gnerre S, Godzaridis É, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt J, Ho I, Howard J, Hunt M, Jackman S, Jaffe D, Jarvis E, Jiang H, Kazakov S, Kersey P, Kitzman J, Knight J, Koren S, Lam T, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, MacCallum I, MacManes M, Maillet N, Melnikov S, Naquin D, Ning Z, Otto T, Paten B, Paulo O, Phillippy A, Pina-Martins F, Place M, Przybylski D, Przybylski D, Qin X, Qu C, Ribeiro F, Richards S, Rokhsar D, Ruby J, Scalabrin S, Schatz M, Schwartz D, Sergushichev A, Sharpe T, Sharpe T, Shaw T, Shendure J, Shi Y, Simpson J, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira B, Wang J, Worley K, Yin S, Yiu S, Yuan J, Zhang G, Zhang H, Zhou S, Korf I: **Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species.** *Gigascience* 2013, **2**:10.

216 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 Res.* 2012, **22**:557–67.

217 Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL: **De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae.** *Genome Res.* 2009, **19**:294–305.

218 Brent MR, Guigó R: **Recent advances in gene structure prediction** *Curr. Opin. Struct. Biol.* 2004, **14**:264–272.

219 Sleator RD: **An overview of the current status of eukaryote gene prediction strategies** *Gene* 2010, **461**:1–4.

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

221 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs** *Bioinformatics* 2015, **31**:3210–3212.

222 Madden TL, Tatusov RL, Zhang J: **Applications of network BLAST server.** *Methods Enzymol.* 1996, **266**:131–41.

223 Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19**:215–25.

224 Finn RD, Clements J, Eddy SR: **HMMER web server: interactive sequence similarity searching.** *Nucleic Acids Res.* 2011, **39**:W29–37.

225 Waterhouse RM, Zdobnov EM, Kriventseva E V: **Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi.** *Genome Biol. Evol.* 2011, **3**:75–86.

226 Clark G, Diamond LS: **Methods for cultivation of luminal parasitic protists of clinical importance.** *Clin. Microbiol. Rev.* 2002, **15**:329–41.

227 Ali IKM, Zaki M, Clark CG, Karim I, Ali M, Zaki M, Clark G: **Use of PCR amplification of tRNA gene-linked short tandem repeats for genotyping Entamoeba histolytica.** *J. Clin. Microbiol.* 2005, **43**:5842–7.

228 Clark G: **Fast CTAB DNA Isolation method** *London Sch. Trop. Med.* 2015.

229 Rohland N, Reich D: **Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture.** *Genome Res.* 2012, **22**:939–46.

230 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM: **Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.** *PLoS One* 2014, **9**:e112963.

231 Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, Gan J, Li N, Hu X, Liu B, Yang B, Fan, W: **Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph.** *Brief. Funct. Genomics* 2012, **11**:25–37.

232 Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754–60.

233 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools** *Bioinformatics* 2009, **25**:2078–2079.

234 R Core Team: **R: A language and environment for statistical computing**. 2014.

235 Wickham H: **ggplot2: Elegant Graphics for Data Analysis**. 2009.

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

237 Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res.* 1997, **25**:955–64.

238 Sehgal D, Mittal V, Ramachandran S, Dhar SK, Bhattacharya A, Bhattacharya S: **Nucleotide sequence organisation and analysis of the nuclear ribosomal DNA circle of the protozoan parasite Entamoeba histolytica.** *Mol. Biochem. Parasitol.* 1994, **67**:205–14.

239 Boetzer M, Pirovano W: **SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information.** *BMC Bioinformatics* 2014, **15**:211.

240 Nagarajan N, Read TD, Pop M: **Scaffolding and validation of bacterial genome assemblies using optical restriction maps.** *Bioinformatics* 2008, **24**:1229–35.

241 Cao H, Hastie AR, Cao D, Lam ET, Sun Y, Huang H, Liu X, Lin L, Andrews W, Chan S, *et al*: **Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology.** *Gigascience* 2014, **3**:34.

242 Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok P-Y, Deal KR, Dvorak J, Hernandez, P Martis M, Galvez S, Leroy P, Guilhot N, Sakai H, Bernard A, Choulet F, Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Li Y, Zheng H, Luo R, Wu H, Zhu H, Soderlund C, Longden I, Mott R, Warren RL, Varabei D, Platt D, Huang X, Messina D: **Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex Aegilops tauschii Genome.** *PLoS One* 2013, **8**:e55864.

243 Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M Kwok P.: **Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly** *Nat. Biotechnol.* 2012, **30**:771–776.

244 BioNano: **Animal Cell Culture DNA Extraction** *BioNano Genomics User Forum* 2015.

245 BioNano: **Documentation - Irys Instrument**. *BioNano Genomics User Forum* 2015..

246 Marie-Nelly H, Marbouty M, Cournac A, Flot J-F, Liti G, Parodi DP, Syan S, Guillén N, Margeot A, Zimmer C, Koszul R: **High-quality genome (re)assembly using chromosomal contact data.** *Nat. Commun.* 2014, **5**:5695.

247 English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology.** *PLoS One* 2012, **7**:e47768.

248 Lee H, Gurtowski J, Yoo S, Marcus S, McCombie WR, Schatz M: **Error correction and assembly complexity of single molecule sequencing reads.** *bioRxiv* 2014.

249 Koren S, Phillippy AM: **One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly** *Curr. Opin. Microbiol.* 2015, **23**:110–120.

250 Ono Y, Asai K, Hamada M: **PBSIM: PacBio reads simulator—toward accurate genome assembly** *Bioinformatics* 2013, **29**:119–121.

251 Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al.: **Nanopore sequencing and assembly of a human genome with ultra-long reads** *Nat. Biotechnol.* 2018, **36**:338–345.

252 Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, Luo F, Xie Z: **MECAT: fast mapping, error correction, and de novo assembly for single-**

molecule sequencing reads *Nat. Methods* 2017, **14**:1072–1074.

253 Jayakumar V, Sakakibara Y: **Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data.** *Brief. Bioinform.* 2017, **147**:1-11.

254 Jackson AP, Allison HC, Barry JD, Field MC, Hertz-Fowler C, Berriman M: **A Cell-surface Phylome for African Trypanosomes** *PLoS Negl. Trop. Dis.* 2013, **7**:e2121.

255 De Lange T, Borst P: **Genomic environment of the expression-linked extra copies of genes for surface antigens of Trypanosoma brucei resembles the end of a chromosome.** *Nature* 1982, **299**:451–3.

256 Williams RO, Young JR, Majiwa PA: **Genomic environment of T. brucei VSG genes: presence of a minichromosome.** *Nature* 1982, **299**:417–21.

257 Van der Ploeg LH, Schwartz DC, Cantor CR, Borst P: **Antigenic variation in Trypanosoma brucei analyzed by electrophoretic separation of chromosome-sized DNA molecules.** *Cell* 1984, **37**:77–84.

258 Wickstead B, Ersfeld K, Gull K: **The Small Chromosomes of Trypanosoma brucei Involved in Antigenic Variation Are Constructed Around Repetitive Palindromes** *Genome Res.* 2004, **14**:1014–1024.

259 Horn D: **Antigenic variation in African trypanosomes.** *Mol. Biochem. Parasitol.* 2014, **195**:123–9.

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

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

262 Haubold B, Wiehe T: **Gene Prediction**. In *An Introduction to Computational Biology: An Evolutionary Approach*. 2006:117–140.

263 Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SMJ, Clamp M: **The Ensembl Automatic Gene Annotation System** *Genome Res.* 2004, **14**:942–950.

264 Steinbiss S, Silva-Franco F, Brunk B, Foth B, Hertz-Fowler C, Berriman M, Otto TD: ***Companion* : a web server for annotation and analysis of parasite genomes** *Nucleic Acids Res.* 2016, **44**:W29–W34.

265 Otto TD, Dillon GP, Degrave WS, Berriman M: **RATT: Rapid Annotation Transfer Tool.** *Nucleic Acids Res.* 2011, **39**:e57.

266 Korf I: **Gene finding in novel genomes** *BMC Bioinformatics* 2004, **5**:59.

267 Li LL, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res.* 2003, **13**:2178–89.

268 Laslett D, Canback B: **ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences** *Nucleic Acids Res.* 2004, **32**:11–16.

269 Nawrocki EP, Eddy SR: **Infernal 1.1: 100-fold faster RNA homology searches** *Bioinformatics* 2013, **29**:2933–2935.

270 Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, Floden EW, Gardner PP, Jones TA, Tate J, Finn, RD: **Rfam 12.0: updates to the RNA families database.** *Nucleic Acids Res.* 2015, **43**:D130–D137.

271 Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, Parkhill J: **ACT: the Artemis Comparison Tool.** *Bioinformatics* 2005, **21**:3422–3.

272 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**:464–9.

273 Supek F, Bošnjak M, Škunca N, Šmuc T: **REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms** *PLoS One* 2011, **6**:e21800.

274 Smit A, Hubley R, Green P: **RepeatMasker**. Open-4.0. 2013-2015 <http://www.repeatmasker.org>

275 Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* 2010, **26**:841–842.

276 Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol. Cell* 2010, **38**:576–89.

277 Gouy M, Guindon S, Gascuel O: **SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building** *Mol. Biol. Evol.* 2010, **27**:221–224.

278 Criscuolo A, Gribaldo S: **BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments** *BMC Evol. Biol.* 2010, **10**:210.

279 Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution** *Bioinformatics* 2011, **27**:1164–1165.

280 Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0** *Syst. Biol.* 2010, **59**:307–321.

281 Kumar S, Stecher G, Tamura K: **MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets** *Mol. Biol. Evol.* 2016, **33**:1870–1874.

282 Koskiniemi S, Sun S, Berg OG, Andersson DI: **Selection-Driven Gene Loss in Bacteria** *PLoS Genet.* 2012, **8**:e1002787.

283 Alano P, Roca L, Smith D, Read D, Carter R, Day K: **Plasmodium falciparum: Parasites Defective in Early Stages of Gametocytogenesis** *Exp. Parasitol.* 1995, **81**:227–235.

284 Tibúrcio M, Dixon MWA, Looker O, Younis SY, Tilley L, Alano P: **Specific expression and export of the Plasmodium falciparum Gametocyte EXported Protein-5 marks the gametocyte ring stage.** *Malar. J.* 2015, **14**:334.

285 Lee SC, Kim IG, Marekov LN, O'Keefe EJ, Parry DA, Steinert PM: **The structure of human trichohyalin. Potential multiple roles as a functional EF-hand-like calcium-binding protein, a cornified cell envelope precursor, and an intermediate filament-associated (cross-linking) protein.** *J. Biol. Chem.* 1993, **268**:12164–76.

286 Moreno SN, Silva J, Vercesi AE, Docampo R: **Cytosolic-free calcium elevation in Trypanosoma cruzi is required for cell invasion.** *J. Exp. Med.* 1994, **180**:1535–40.

287 Chuenkova M V., Furnari FB, Cavenee WK, Pereira MA: **Trypanosoma cruzi trans-sialidase: A potent and specific survival factor for human Schwann cells by means of phosphatidylinositol 3-kinase/Akt signaling** *Proc. Natl. Acad. Sci.* 2001, **98**:9936–9941.

288 Teixeira AAR, de Vasconcelos V de CS, Colli W, Alves MJM, Giordano RJ: **Trypanosoma cruzi Binds to Cytokeratin through Conserved Peptide Motifs Found in the Laminin-G-Like Domain of the gp85/Trans-sialidase Proteins** *PLoS Negl. Trop. Dis.* 2015, **9**:e0004099.

289 Alvarez AH, Martinez-Cadena G, Silva ME, Saavedra E, Avila EE: **Entamoeba histolytica: ADP-ribosylation of secreted glyceraldehyde-3-phosphate dehydrogenase** *Exp. Parasitol.* 2007, **117**:349–356.

290 Yao C, Donelson JE, Wilson ME: **The major surface protease (MSP or GP63) of Leishmania sp. Biosynthesis, regulation of expression, and function** *Mol. Biochem. Parasitol.* 2003, **132**:1–16.

291 Penuliar GM, Nakada-Tsukui K, Nozaki T: **Phenotypic and transcriptional profiling in Entamoeba histolytica reveal costs to fitness and adaptive responses associated with metronidazole resistance.** *Front. Microbiol.* 2015, **6**:354.

292 Gray Y: **It takes two transposons to tango:transposable-element-mediated chromosomal rearrangements**. *Trends Genet.* 2000, **16**:461–468.

293 Capy P, Gasperi G, Biémont C, Bazin C: **Stress and transposable elements: co-evolution or useful parasites?** *Heredity.* 2000, **85**:101–106.

294 McDonald J: **Transposable elements: possible catalysts of organismic evolution**. *Trends Ecol. Evol.* 1995, **10**:123–126.

295 Durand P, Oelofse A, Coetzet T: **An analysis of mobile genetic elements in three Plasmodium species and their potential impact on the nucleotide composition of the P. falciparum genome**. *BMC Genomics* 2006, **7**:Online publication.

296 Janoušek V, Laukaitis C, Yanchukov A, Karn R: **The Role of Retrotransposons in Gene Family Expansions in the Human and Mouse Genomes**. *Genome Biol. Evol.* 2016, **8**:2632–2650.

297 Janoušek V, Karn R, Laukaitis C: **The role of retrotransposons in gene family expansions: insights from the mouse Abp gene family**. *BMC Evol. Biol.* 2013, **13**:Online publication.

298 Kaźmierczuk A, Kiliańska ZM: **The pleiotropic activity of heat-shock proteins.** *Postepy Hig. Med. Dosw.* 2009, **63**:502–21.

299 Shilova VY, Garbuz DG, Myasyankina EN, Chen B, Evgen'ev MB, Feder ME, Zatsepina OG: **Remarkable site specificity of local transposition into the Hsp70 promoter of Drosophila melanogaster.** *Genetics* 2006, **173**:809–20.

300 Zatsepina OG, Velikodvorskaia V V, Molodtsov VB, Garbuz D, Lerman DN, Bettencourt BR, Feder ME, Evgenev MB: **A Drosophila**

melanogaster strain from sub-equatorial Africa has exceptional thermotolerance but decreased Hsp70 expression. *J. Exp. Biol.* 2001, **204**:1869–81.

301 Vembar SS, Seetin M, Lambert C, Nattestad M, Schatz MC, Baybayan P, Scherf A, Smith ML: **Complete telomere-to-telomere de novo assembly of the Plasmodium falciparum genome through long-read (>11 kb), single molecule, real-time sequencing** *DNA Res.* 2016, **23**:339-351.

302 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.** *Mol. Biol. Evol.* 2008, **25**:187–98.

303 Frydrychová R, Grossmann P, Trubac P, Vítková M, Marec F: **Phylogenetic distribution of TTAGG telomeric repeats in insects** *Genome* 2004, **47**:163–178.

304 Biessmann H, Carter SB, Mason JM: **Chromosome ends in Drosophila without telomeric DNA sequences.** *Proc. Natl. Acad. Sci. U. S. A.* 1990, **87**:1758–61.

305 Michel B, Lizardi PM, Alagon A, Zurita M: **Identification and analysis of the start site of ribosomal RNA transcription of Entamoeba histolytica**. *Mol. Biochem. Parasitol.* 1995, **73**:19–30.

306 Bhattacharya S, Bhattacharya A, Diamond LS: **Comparison of repeated DNA from strains of Entamoeba histolytica and other Entamoeba** *Mol. Biochem. Parasitol.* 1988, **27**:257–262.

307 Huber M, Koller B, Gitler C, Mirelman D, Revel M, Rozenblatt S, Garfinkel L: **Entamoeba histolytica ribosomal RNA genes are carried on palindromic circular DNA molecules.** *Mol. Biochem. Parasitol.* 1989, **32**:285–96.

308 Burch DJ, Li E, Reed S, Jackson TF, Stanley SL: **Isolation of a strain-specific Entamoeba histolytica cDNA clone.** *J. Clin. Microbiol.* 1991, **29**:696–701.

309 Clark CG, Diamond LS: **Entamoeba histolytica: A Method for Isolate Identification** *Exp. Parasitol.* 1993, **77**:450–455.

310 Santos HLC, Bandea R, Farnandes Martins LA, de Macedo HW, Peralta RHS, Ndubuisi M, da Silva A: **Differential identification of Entamoeba spp. based on the analysis of 18S rRNA**. *Parasitol. Res.* 2010, **106**:883–888.

311 Haghighi A, Kobayashi S, Takeuchi T, Thammapalerd N, Nozaki T: **Geographic diversity among genotypes of Entamoeba histolytica field isolates.** *J. Clin. Microbiol.* 2003, **41**:3748–56.

312 Haghighi A, Kobayashi S, Takeuchi T, Masuda G, Nozaki T: **Remarkable genetic polymorphism among Entamoeba histolytica isolates from a limited geographic area.** *J. Clin. Microbiol.* 2002, **40**:4081–90.

313 de la Vega H, Specht CA, Semino CE, Robbins PW, Eichinger D, Caplivski D, Ghosh S, Samuelson J: **Cloning and expression of chitinases of Entamoebae.** *Mol. Biochem. Parasitol.* 1997, **85**:139–47.

314 Beck DL, Tanyuksel M, Mackey AJ, Haque R, Trapaidze N, Pearson WR, Loftus B, Petri WA: **Entamoeba histolytica: sequence conservation of the Gal/GalNAc lectin from clinical isolates.** *Exp. Parasitol.* 2002, **101**:157–63.

315 Köhler S, Tannich E: **A family of transcripts (K2) of Entamoeba histolytica contains polymorphic repetitive regions with highly conserved elements.** *Mol. Biochem. Parasitol.* 1993, **59**:49–58.

316 Stanley SL, Becker A, Kunz-Jenkins C, Foster L, Li E, Li E: **Cloning and expression of a membrane antigen of Entamoeba histolytica possessing multiple tandem repeats.** *Proc. Natl. Acad. Sci. U. S. A.* 1990, **87**:4976–80.

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

318 Simonishvili S, Tsanava S, Sanadze K, Chlikadze R, Miskalishvili A, Lomkatsi N, Imnadze P, Petri WA, Trapaidze N: **Entamoeba histolytica: The serine-rich gene polymorphism-based genetic variability of clinical isolates from Georgia** *Exp. Parasitol.* 2005, **110**:313–317.

319 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** *Exp. Parasitol.* 2008, **118**:354–361.

320 Karim I, Ali M, Zaki M, Clark G: **Use of PCR Amplification of tRNA Gene-Linked Short Tandem Repeats for Genotyping Entamoeba histolytica**. *J. Clin. Microbiol.* 2005, **43**:5842–5847.

321 Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res.* 1999, **27**:573–80.

322 Vincze T, Posfai J, Roberts RJ: **NEBcutter: A program to cleave DNA with restriction enzymes.** *Nucleic Acids Res.* 2003, **31**:3688–91.

323 Klose RJ, Bird AP: **Genomic DNA methylation: the mark and its mediators** *Trends Biochem. Sci.* 2006, **31**:89–97.

324 Yunis JJ, Yasmineh WG: **Heterochromatin, Satellite DNA, and Cell Function** *Science.* 1971, **174**:1200–1209.

325 Nath J, Hussain G, Singha B, Paul J, Ghosh SK: **Burden of major diarrheagenic protozoan parasitic co-infection among amoebic dysentery cases from North East India: a case report** *Parasitology* 2015, **142**:1318–1325.

326 Nath J, Ghosh SK, Singha B, Paul J: **Molecular Epidemiology of Amoebiasis: A Cross-Sectional Study among North East Indian Population** *PLoS Negl. Trop. Dis.* 2015, **9**:e0004225.

327 Paul J, Srivastava S, Bhattacharya S: **Molecular methods for diagnosis of Entamoeba histolytica in a clinical setting: an overview.** *Exp. Parasitol.* 2007, **116**:35–43.

328 Gupta AK, Panigrahi SK, Bhattacharya A, Bhattacharya S: **Self-circularizing 5′-ETS RNAs accumulate along with unprocessed pre ribosomal RNAs in growth-stressed Entamoeba histolytica** *Sci. Rep.* 2012, **2**:303.

329 Gupta AK, Bhattacharya S: **Ribosomal RNA Genes and Their Regulation in Entamoeba histolytica**. In Amebiasis. Edited by Nozaki T, Bhattacharya A. Springer; 2015:119–135.

330 Holliday R: **Epigenetics: A Historical Overview** *Epigenetics* 2006, **1**:76–80.

331 Kouzarides T: **Chromatin Modifications and Their Function** *Cell* 2007, **128**:693–705.

332 Martin C, Zhang Y: **The diverse functions of histone lysine methylation** *Nat. Rev. Mol. Cell Biol.* 2005, **6**:838–849.

333 Suzuki MM, Bird A: **DNA methylation landscapes: provocative insights from epigenomics** *Nat. Rev. Genet.* 2008, **9**:465–476.

334 Hendrich B, Tweedie S: **The methyl-CpG binding domain and the evolving role of DNA methylation in animals** *Trends Genet.* 2003, **19**:269–277.

335 Jeltsch A: **Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases.** *Chembiochem* 2002, **3**:274–93.

336 Field LM, Lyko F, Mandrioli M, Prantera G: **DNA methylation in insects** *Insect Mol. Biol.* 2004, **13**:109–115.

337 Doerfler W: **Epigenetic consequences of foreign DNA insertions: de novo methylation and global alterations of methylation patterns in recipient genomes** *Rev. Med. Virol.* 2011, **21**:336–346.

338 Jones PA: **Functions of DNA methylation: islands, start sites, gene bodies and beyond** *Nat. Rev. Genet.* 2012, **13**:484–492.

339 Fournier A, Sasai N, Nakao M, Defossez P-A: **The role of methyl-binding proteins in chromatin organization and epigenome maintenance** *Brief. Funct. Genomics* 2012, **11**:251–264.

340 Wang Y, Jorda M, Jones P, Maleszka R, Robertson H, Mizzen C, Peinado M, Robinson G: **Functional CpG Methylation System in a Social Insect** *Science.* 2006, **314**:645–647.

341 Kucharski R, Maleszka J, Foret S, Maleszka R: **Nutritional control of reproductive status in honeybees via DNA methylation.** *Science.* 2008, **319**:1827–30.

342 Maleszka R: **Epigenetic integration of environmental and genomic signals in honey bees: the critical interplay of nutritional, brain and reproductive networks.** *Epigenetics* 2008, **3**:188–92.

343 Moczek AP, Snell-Rood EC: **The basis of bee-ing different: the role of gene silencing in plasticity** *Evol. Dev.* 2008, **10**:511–513.

344 Leonhardt H, Cardoso MC: **DNA methylation, nuclear structure, gene expression and cancer.** *J. Cell. Biochem. Suppl.* 2000, **35**:78–83.

345 Leonhardt H, Rahn H-P, Cardoso MC: **Functional Links between Nuclear Structure, Gene Expression, DNA Replication, and Methylation** *Crit. Rev. Eukaryot. Gene Expr.* 1999, **9**:345–351.

346 Jurkowski TP, Jeltsch A: **On the Evolutionary Origin of Eukaryotic DNA Methyltransferases and Dnmt2** *PLoS One* 2011, **6**:e28104.

347 Jeltsch A, Ehrenhofer-Murray A, Jurkowski TP, Lyko F, Reuter G, Ankri S, Nellen W, Schaefer M, Helm M: **Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation** *RNA Biol.* 2017, **14**:1108–1123.

348 Hermann A, Schmitt S, Jeltsch A: **The Human Dnmt2 Has Residual DNA-(Cytosine-C5) Methyltransferase Activity** *J. Biol. Chem.* 2003, **278**:31717–31721.

349 Katoh M, Curk T, Xu Q, Zupan B, Kuspa A, Shaulsky G: **Developmentally regulated DNA methylation in Dictyostelium discoideum.** *Eukaryot. Cell* 2006, **5**:18–25.

350 Phalke S, Nickel O, Walluscheck D, Hortig F, Onorati MC, Reuter G: **Retrotransposon silencing and telomere integrity in somatic cells of Drosophila depends on the cytosine-5 methyltransferase DNMT2** *Nat. Genet.* 2009, **41**:696–702.

351 Goll MG, Kirpekar F, Maggert KA, Yoder JA, Hsieh C-L, Zhang X, Golic KG, Jacobsen SE, Bestor TH: **Methylation of tRNAAsp by the DNA Methyltransferase Homolog Dnmt2** *Science.* 2006, **311**:395–398.

352 Darst RP, Pardo CE, Ai L, Brown KD, Kladde MP: **Bisulfite Sequencing of DNA** *Curr. Protoc. Mol. Biol.* 2010, **91**:7.9.1–7.9.17.

353 Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing.** *Nat. Methods* 2010, **7**:461–5.

354 Banerjee S, Fisher O, Lohia A, Ankri S: **Entamoeba histolytica DNA methyltransferase (Ehmeth) is a nuclear matrix protein that binds EhMRS2, a DNA that includes a scaffold/matrix attachment region (S/MAR)** *Mol. Biochem. Parasitol.* 2005, **139**:91–97.

355 Bernes S, Siman-Tov R, Ankri S: **Epigenetic and classical activation of Entamoeba histolytica heat shock protein 100 (EHsp100) expression** *FEBS Lett.* 2005, **579**:6395–6402.

356 Ali I, Ehrenkaufer GM, Hackney JA, Singh U: **Growth of the protozoan parasite Entamoeba histolytica in 5-azacytidine has limited effects on parasite gene expression.** *BMC Genomics* 2007, **8**:Online publication.

357 Fisher O, Siman-Tov R, Ankri S: **Characterization of cytosine methylated regions and 5-cytosine DNA methyltransferase (Ehmeth) in the protozoan parasite Entamoeba histolytica** *Nucleic Acids Res.* 2004, **32**:287–297.

358 Jones PA, Takai D: **The role of DNA methylation in mammalian epigenetics.** *Science* 2001, **293**:1068–70.

359 Geyer KK, Rodríguez López CM, Chalmers IW, Munshi SE, Truscott M, Heald J, Wilkinson MJ, Hoffmann KF: **Cytosine methylation regulates oviposition in the pathogenic blood fluke Schistosoma mansoni** *Nat. Commun.* 2011, **2**:424.

360 Tovy A, Ankri S: **Epigenetics in the unicellular parasite Entamoeba histolytica.** *Future Microbiol.* 2010, **5**:1875–84.

361 Cortés A, Crowley VM, Vaquero A, Voss TS: **A View on the Role of Epigenetics in the Biology of Malaria Parasites** *PLoS Pathog.* 2012, **8**:e1002943.

362 Gómez-Díaz E, Jordà M, Peinado MA, Rivero A: **Epigenetics of Host–Pathogen Interactions: The Road Ahead and the Road Behind** *PLoS Pathog.* 2012, **8**:e1003007.

363 Martin M: **CUTADAPT removes adapter sequences from high-throughput sequencing reads**. *EMBnet.journal* 2011, **17**.

364 Joshi NA, Fass JN: **Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.** *Github* 2011.

365 Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications** *Bioinformatics* 2011, **27**:1571–1572.

366 Broad Institute: **Picard Toolkit** *Github* 2018.

367 Bao W, Kojima KK, Kohany O: **Repbase Update, a database of repetitive elements in eukaryotic genomes** *Mob. DNA* 2015, **6**:11.

368 Jurka J: **Repeats in genomic DNA: mining and meaning.** *Curr. Opin. Struct. Biol.* 1998, **8**:333–7.

369 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner** *Bioinformatics* 2013, **29**:15–21.

370 Anders S, Pyl PT, Huber W: **HTSeq - A Python framework to work with high-throughput sequencing data** *Bioinformatics* 2014, **31**:166-9.

371 Tran H, Porter J, Sun M, Xie H, Zhang L: **Objective and Comprehensive Evaluation of Bisulfite Short Read Mapping Tools** *Adv. Bioinformatics* 2014, **2014**:1–11.

372 Raine A, Liljedahl U, Nordlund J: **Data quality of whole genome bisulfite sequencing on Illumina platforms.** *PLoS One* 2018, **13**:e0195972.

373 Hackney JA, Ehrenkaufer GM, Singh U: **Identification of putative transcriptional regulatory networks in Entamoeba histolytica using Bayesian inference.** *Nucleic Acids Res.* 2007, **35**:2141–52.

374 Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar HA, Thomson JA, Ren B, Ecker JR: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462**:315–322.

375 Beeler SM, Wong GT, Zheng JM, Bush EC, Remnant EJ, Oldroyd BP, Drewell RA: **Whole-genome DNA methylation profile of the jewel wasp (Nasonia vitripennis).** *G3.* 2014, **4**:383–8.

376 Drewell RA, Bush EC, Remnant EJ, Wong GT, Beeler SM, Stringham JL, Lim J, Oldroyd BP: **The dynamic DNA methylation cycle from egg to sperm in the honey bee Apis mellifera.** *Development* 2014, **141**:2702–11.

377 Bird A, Taggart M, Cell BS. **Methylated and unmethylated DNA compartments in the sea urchin genome** *Cell* 1979, **17**:889–901.

378 Tweedie S, Charlton J, Clark V, Bird A: **Methylation of genomes and genes at the invertebrate-vertebrate boundary.** *Mol. Cell. Biol.* 1997, **17**:1469–75.

379 Jeltsch A: **Phylogeny of methylomes.** *Science.* 2010, **328**:837–8.

380 Zemach A, McDaniel IE, Silva P, Zilberman D: **Genome-wide evolutionary analysis of eukaryotic DNA methylation.** *Science* 2010, **328**:916–9.

381 Que X, Reed SL: **Cysteine proteinases and the pathogenesis of amebiasis.** *Clin. Microbiol. Rev.* 2000, **13**:196–206.

382 Wion D, Casadesús J: **N6-methyl-adenine: an epigenetic signal for DNA–protein interactions** *Nat. Rev. Microbiol.* 2006, **4**:183–192.

383 Ratel D, Ravanat J-L, Berger F, Wion D: **N6-methyladenine: the other methylated base of DNA.** *Bioessays* 2006, **28**:309–15.

384 Lewinska A, Adamczyk-Grochala J, Kwasniewicz E, Wnuk M: **Downregulation of methyltransferase Dnmt2 results in condition-dependent telomere shortening and senescence or apoptosis in**

mouse fibroblasts *J. Cell. Physiol.* 2017, **232**:3714–3726.

385 Krauss V, Reuter G: **DNA Methylation in Drosophila—A Critical Evaluation** In *Progress in Molecular Biology and Translational Science*. Academic Press; 2011:177–191.

386 Lavi T, Isakov E, Harony H, Fisher O, Siman-Tov R, Ankri S: **Sensing DNA methylation in the protozoan parasite Entamoeba histolytica.** *Mol. Microbiol.* 2006, **62**:1373–86.

387 Loman N: **Thar she blows! Ultra long read method for nanopore sequencing · Loman Labs** *Loman Lab Blog* 2017 (lab.loman.net).

388 Abdel-Latif A, Osman G: **Comparison of three genomic DNA extraction methods to obtain high DNA quality from maize** *Plant Methods* 2017, **13**:1.

389 Workman R, Fedak R, Kilburn D, Hao S, Liu K, Timp W: **High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation Sequencing.** *Nat. Protoc. Exch.* 2018:Online publication.

390 Aitcheson N, Talbot S, Shapiro J, Hughes K, Adkin C, Butt T, Sheader K, Rudenko G: **VSG switching in Trypanosoma brucei: antigenic variation analysed using RNAi in the absence of immune selection.** *Mol. Microbiol.* 2005, **57**:1608–22.

391 del Portillo HA, Fernandez-Becerra C, Bowman S, Oliver K, Preuss M, Sanchez CP, Schneider NK, Villalobos JM, Rajandream MA, Harris D, et al.: **A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax.** *Nature* 2001, **410**:839–42.

392 Ersfeld K: **Fiber-FISH: Fluorescence &lt;I&gt; In Situ&lt;/I&gt; Hybridization on Stretched DNA** In *Parasite Genomics Protocols*. . Humana Press; 1994:395–402.

Appendices

Appendix 2 – Chapter Two addition materials in thesis

Appendix 3 – Chapter Three additional materials in thesis

Appendix 4 – Chapter Four additional materials in thesis

Appendix 5 – Chapter Five additional materials in thesis

Appendix 2

S2.1. assemblyStats.pl

```
#! /usr/bin/perl

##USAGE: perl assemblyStats.pl [FASTA file]

use strict;
use warnings;
use List::Util qw(sum min max);
use Getopt::Long;
use File::Basename;

#Define basetype variables
my $A = 0;
my $T = 0;
my $G = 0;
my $C = 0;
my $N = 0;

# Define file variables
my $file=shift;
my $outFile = "";


my ($fileName, $filePath) = fileparse($file);
$outFile = $file . "_n50_stat";


#Open files

open(I, "<$file") or die "Can not open file: $file\n";
open(O, ">$outFile") or die "Can not open file: $outFile\n";


my @len = ();

my $prevFastaSeqId = "";
my $fastaSeqId = "";
my $fastaSeq = "";

while(my $line = <I>) {
      chomp $line;
      if($line =~ /^>/) {
            $prevFastaSeqId = $fastaSeqId;
            $fastaSeqId = $line;
            if($fastaSeq ne "") {
                  push(@len, length $fastaSeq);
                  baseCount($fastaSeq);
            }
            $fastaSeq = "";
      }
      else {
           $fastaSeq .= $line;
      }
}
if($fastaSeq ne "") {
      $prevFastaSeqId = $fastaSeqId;
      push(@len, length $fastaSeq);
      baseCount($fastaSeq);
}

my $totalReads = scalar @len;
my $bases = sum(@len);
my $minReadLen = min(@len);
my $maxReadLen = max(@len);
my $n25 = calcN50(\@len, 25);
my $n50 = calcN50(\@len, 50);
my $n75 = calcN50(\@len, 75);
my $n90 = calcN50(\@len, 90);
my $n95 = calcN50(\@len, 95);

printf O "%-25s %d\n" , "Total sequences", $totalReads;
printf O "%-25s %d\n" , "Total bases", $bases;
printf O "%-25s %d\n" , "Min sequence length", $minReadLen;
printf O "%-25s %d\n" , "Max sequence length", $maxReadLen;
printf O "%-25s %0.2f\n", "Average sequence length",
$avgReadLen;
printf O "%-25s %0.2f\n", "Median sequence length",
$medianLen;
printf O "%-25s %d\n", "N25 length", $n25;
printf O "%-25s %d\n", "N50 length", $n50;
printf O "%-25s %d\n", "N75 length", $n75;
printf O "%-25s %d\n", "N90 length", $n90;
printf O "%-25s %d\n", "N95 length", $n95;
printf O "%-25s %0.2f %s\n", "A", $A/$bases*100, "%";
printf O "%-25s %0.2f %s\n", "T", $T/$bases*100, "%";
printf O "%-25s %0.2f %s\n", "G", $G/$bases*100, "%";
printf O "%-25s %0.2f %s\n", "C", $C/$bases*100, "%";
printf O "%-25s %0.2f %s\n", "(A + T)s", ($A+$T)/$bases*100,
"%";
printf O "%-25s %0.2f %s\n", "(G + C)s", ($G+$C)/$bases*100,
"%";
printf O "%-25s %0.2f %s\n", "Ns", $N/$bases*100, "%";

print "N50 Statisitcs file: $outFile\n";

exit;

sub calcN50 {
      my @x = @{$_[0]};
      my $n = $_[1];
      @x=sort{$b<=>$a} @x;
      my $total = sum(@x);
      my ($count, $n50)=(0,0);
      for (my $j=0; $j<@x; $j++){
        $count+=$x[$j];
        if(($count>=$total*$n/100)){
            $n50=$x[$j];
            last;
```

```
		}
	}
	return $n50;
}


sub baseCount {
	my $seq = $_[0];
	my $tA += $seq =~ s/A/A/gi;
	my $tT += $seq =~ s/T/T/gi;
	my $tG += $seq =~ s/G/G/gi;
	my $tC += $seq =~ s/C/C/gi;
	$N += (length $seq) - $tA - $tT - $tG - $tC;
	$A += $tA;
	$T += $tT;
	$G += $tG;
	$C += $tC;
}
```

Appendix 3

Table S3.1. Non-default parameters used when running PhyML on gene families in close proximity to transposable elements. PhyML version 20120412 was used from command line.

Parameter	**Value**
Data Type	aa
Sequence Format	sequential
Number of bootstrapped datasets	1,000
Model Name	[Best model produced by ProtTest3]
Proportion of invariable sites	[P-inv value produced by ProtTest3]
Gamma distribution parameter	[alpha value produced by ProtTest3]
Tree topology search	SPR (Subtree pruning and regrafting)

Appendix 4

Table S4.1. Full list of tRNA array types in the *E. histolytica* genome

tRNA isoacceptor types are given with their anti-codon.

Array Name	Isoacceptor Types	GenBank Accession
[A^{AGC}]	Ala^{AGC}	BK005648
[ALL]	Ala^{CGC}, Leu^{TAA}, Leu^{CAA}	BK005649
[ASD]	Ala^{TGC}, Ser^{GCT}, Asp^{GTC}	BK005650
[G^{GCC}]	Gly^{GCC}	BK005662
[G^{TCC}]	Gly^{TCC}	BK005663
[H^{GTG}]	His^{GTG}	BK005654
[LS]	Leu^{CAG}, Ser^{CGA}	BK005667
[LT]	Leu^{AAG}, Thr^{AGT}	BK005666
[MR]	eMet^{CAT},Arg^{TCG}	BK005653
[NK1]	Asn^{GTT},Lys^{CTT}	BK005655
[NK2]	Asn^{GTT},Lys^{CTT}	BK005656
[P^{TGG}]	ProTGG	BK005669
[R^{TCT}]	Arg^{TCT}	BK005654
[R5]	Arg^{ACG}	BK005651
[RT]	Arg^{CCT}, Thr^{AGT}	BK005652
[SD]	Ser^{TGA}, Asp^{GTC}	BK005657
[SPPCK]	Ser^{AGA}, Pro^{AGG}, Pro^{CGG}, Cys^{GCA}, Lys^{TTT}	BK005659
[SQCK]	Ser^{AGA}, Gln^{CTG}, Cys^{GCA},Lys^{TTT}	BK005658
[TQ]	Thr^{CGT}, Gln^{TTG}	BK005660
[TX]	Thr^{TGT}	BK005670
[V5]	Val^{TAC}	BK005671
[VF]	Val^{GAC}, Phe^{GAA}	BK005668
[VME5]	Val^{CAC}, iMet^{CAT}, Glu^{CTC}	BK005672
[WI]	Trp^{CCA}, Ile^{AAT}	BK005665
[YE]	Tyr^{GTA}, Glu^{TTC}	BK005661

S4.2. Extracting codons from CDS sequences

```
cat input.fasta|\
awk  '/^>/  {printf("%s%s\n",(N==0?"":"\n"),$0);N++;next;}
{printf("%s",$0);}END{printf("\n");}' |\
sed -e $'/^[^>]/s/\([A-Z][A-Z][A-Z]\)/\\1\\\n/g'
```

Appendix 5

Table S5.1. UniProt accession numbers of DNA methyltransferases (Dnmt) in *Homo Sapiens* and *Arabidopsis thaliana*

Organism	Dnmt type	UniProt Accession Number
Homo Sapiens	Dnmt1	P26358
	Dnmt2	Q61C57
	Dnmt3a	Q946K1
	Dnmt3b	Q2PJ58
Arabidopsis thaliana	Dnmt1	P34881
	Dnmt2	F4JWT7
	Dnmt3a	Q9T0I1
	Dnmt3b	O23273

S5.2. Methylation_protocol.sh

```
#!/bin/bash


echo "Enter name and location of Reference genome folder,
followed by [ENTER]:"
read reference_genome_folder
echo "Enter name and location of Reference genome fasta file,
followed by [ENTER]:"
read reference_genome_fasta
echo "Enter name and location of one zipped fastq file (not
whole path make sure in same directory), followed by [ENTER]:"
read fastq1
echo "Enter name and location of one zipped fastq file (not
whole path make sure in same directory), followed by [ENTER]:"
read fastq2
echo "Enter name only of fastq1 (without .fastq.gz) folowed by
enter:"
read fastq3
echo "Enter study name, followed by [ENTER]:"
read name

echo "Number of reads R1:" >> $name"_Coverage_stats"
gunzip -c $fastq1 | grep "^+$" | wc -l >>
$name"_Coverage_stats"
echo "Number of reads R2:" >> $name"_Coverage_stats"
gunzip -c $fastq2 | grep "^+$" | wc -l >>
```

```
$name"_Coverage_stats"
/pub34/laura/bismark_v0.18.1/bismark_genome_preparation --
bowtie2 $reference_genome_folder

/pub34/laura/bismark_v0.18.1/bismark --bowtie2 --
non_directional $reference_genome_folder -1 $fastq1 -2 $fastq2
2> Bismark_stout

cp $fastq3"_bismark_bt2_pe.bam"
$name"_1.fastq_bismark_bt2_pe.bam"
samtools sort $name"_1.fastq_bismark_bt2_pe.bam"
$name"_fastq_bismark_sort"
samtools index $name"_fastq_bismark_sort.bam"

java -jar -Xmx10g /pub35/xliu/software/picard-tools-
1.85/MarkDuplicates.jar I= $name"_fastq_bismark_sort.bam" O=
$name"_remdups.bam" M=duplication.txt REMOVE_DUPLICATES=true
AS=true

samtools sort $name"_remdups.bam" $name"_remdups_sort"
samtools index $name"_remdups_sort.bam"

perl /pub34/laura/coverageStatsSplitByChr_v2.pl -i
$name"_remdups_sort.bam" > $name"_coverage"

awk '{sum=sum+$4} END {print "Average % coverage of reference
contigs=\t" sum/NR}' $name"_coverage" >>
$name"_Coverage_stats"
awk '{sum=sum+$5} END {print "Average depth of coverage of
reference contigs=\t" sum/NR}' $name"_coverage" >>
$name"_Coverage_stats"
echo "Number of mapped contigs:" >> $name"_Coverage_stats"
wc -l $name"_coverage" >> $name"_Coverage_stats"

echo "Number of mapped reads:" >> $name"_Coverage_stats"
samtools view $name"_1.fastq_bismark_bt2_pe.bam" | grep -v
"^@" | wc -l >> $name"_Coverage_stats"
echo "Number of mapped reads after duplicate removal:" >>
$name"_Coverage_stats"
samtools view $name"_remdups_sort.bam" | grep -v "^@" | wc -l
>> $name"_Coverage_stats"

/pub34/laura/bismark_v0.18.1/bismark_methylation_extractor -s
-comprehensive $name"_remdups_sort.bam"

perl
/pub34/laura/Watkins_collection/CS_bis/Map_direct_to_TGAC/Are_
SNP_reads_methylated.pl "CpG_context_"$name"_remdups_sort.txt"
"CHG_context_"$name"_remdups_sort.txt"
"CHH_context_"$name"_remdups_sort.txt" >
$name"_%_C_sites_meth.txt"
echo "Number of cytosines to analyze:" >>
$name"_Coverage_stats"
wc -l $name"_%_C_sites_meth.txt" >> $name"_Coverage_stats"

awk '($3 >= 10) { print $0}' $name"_%_C_sites_meth.txt" >
$name"_%_C_sites_meth_g10x.txt"
```

```
echo "Number of cytosines to analyze 10x:" >>
$name"_Coverage_stats"
wc -l $name"_%_C_sites_meth_g10x.txt" >>
$name"_Coverage_stats"
```

S5.3. Are_SNP_reads_methylated.pl Perl script provided by Dr Laura Gardiner (IBM Research, Daresbury, UK).

```
#!/usr/bin/perl
use strict;
my $line;
my @temp;
my $ref_contig;
my %hash3;
my $counter=0;
my $line2;
my @temp2;
my %hash2;
my @array;
my $ratio;
use Bio::SeqIO;
my %hash1;
my @array2;
my @array3;
my %hash_CS;
my %hash_CHH;
my %hash_CHG;

##Usage ./Are_SNP_reads_methylated.pl
Bismark_CpG_methylation_file Bismark_CHH_methylation_file
Bismark_CHG_methylation_file Output1 > Output2

#open (INPUT, $ARGV[0]);
open(INPUT,"/pub9/laura/Indian_wheat_grant/CS_new/CS_new_SPLIT
_TREATED/CpG_context_CS_new_split_treated.fastq.gz_bismark_rem
dups_sort.txt") || die "cannot open file\n";
while($line=<INPUT>){
chomp $line;
my @array=split(/\t/,$line);
my $seq_read = $array[0];
my $pos=$array[3];
my $meth=$array[1];
my $contig=$array[2];
my $detail=($contig . ":" . $pos . ":" . "Z" . ":" . $meth);
    push @{$hash_CS{$seq_read}}, $detail;
    }
close INPUT;

#open (INPUT, $ARGV[1]);
open(INPUT,"/pub9/laura/Indian_wheat_grant/CS_new/CS_new_SPLIT
_TREATED/CHH_context_CS_new_split_treated.fastq.gz_bismark_rem
```

```
dups_sort.txt") || die "cannot open file\n";
while(my $line2=<INPUT>){
chomp $line2;
my @array2=split(/\t/,$line2);
my $seq_read2 = $array2[0];
my $meth2=$array2[1];
my $pos2=$array2[3];
my $contig2=$array2[2];
my  $detail2=($contig2  .  ":"  .  $pos2  .  ":"  .  "H"  .  ":"  .
$meth2);
        push @{$hash_CS{$seq_read2}}, $detail2;
        }
close INPUT;

#open (INPUT, $ARGV[2]);
open(INPUT,"/pub9/laura/Indian_wheat_grant/CS_new/CS_new_SPLIT
_TREATED/CHG_context_CS_new_split_treated.fastq.gz_bismark_rem
dups_sort.txt") || die "cannot open file\n";
while(my $line3=<INPUT>){
chomp $line3;
my @array3=split(/\t/,$line3);
my $seq_read3 = $array3[0];
my $meth3=$array3[1];
my $pos3=$array3[3];
my $contig3=$array3[2];
my  $detail3=($contig3  .  ":"  .  $pos3  .  ":"  .  "X"  .  ":"  .
$meth3);
        push @{$hash_CS{$seq_read3}}, $detail3;
        }
close INPUT;


print
"Probe\tPosition\tRef\tAlt\tVarfreq\tSeq_read_ID\tMapping_orie
ntation\tLeftmost_mapping_pos\tCigar\tRightmost_mapping_positi
on\tSNP_position_on_read\tSNP_allele_in_read\tSequencing_read\
tNumber_of_C's_hit_by_read\tListed_C's_hit_by_read(Probe:Pos:t
ype_of_C:meth(+)unmeth(-
))\tMeth_Cs_hit_by_read\tunmeth_Cs_hit_by_read\n";
    #\tMethylation_site_on_read(Probe:Position_on_probe:Type)
\n";
#open (INPUT, $ARGV[3]);
open(INPUT,"/pub9/laura/Indian_wheat_grant/CS_new/CS_new_SPLIT
_TREATED/CS_new_split_treated_reads_mapping_homoeologous.snp")
|| die;
while(my $liner=<INPUT>){
chomp $liner;
@temp=split(/\t/,$liner);
my $read_ID=$temp[3];
    if(exists($hash_CS{$read_ID})){
    my $value=$hash_CS{$read_ID};
    my @val=@$value;

    my $sites=@val;
    print "$liner\t$sites\t";
            my @metharray;
            my @unmetharray;
    foreach my $element(@val){
            if($element =~ /:\+/){
            push (@metharray, $element);
            print "$element ";
            }
            elsif($element =~ /:\-/){
            push(@unmetharray, $element);
            print "$element ";
            }
            }
            my $ll=@metharray;
            my $tt=@unmetharray;
            print "\t$ll\t$tt\t";
            print "\n";
            }
            else{
            print "$liner\t0\n";
            }
}
close INPUT;
```

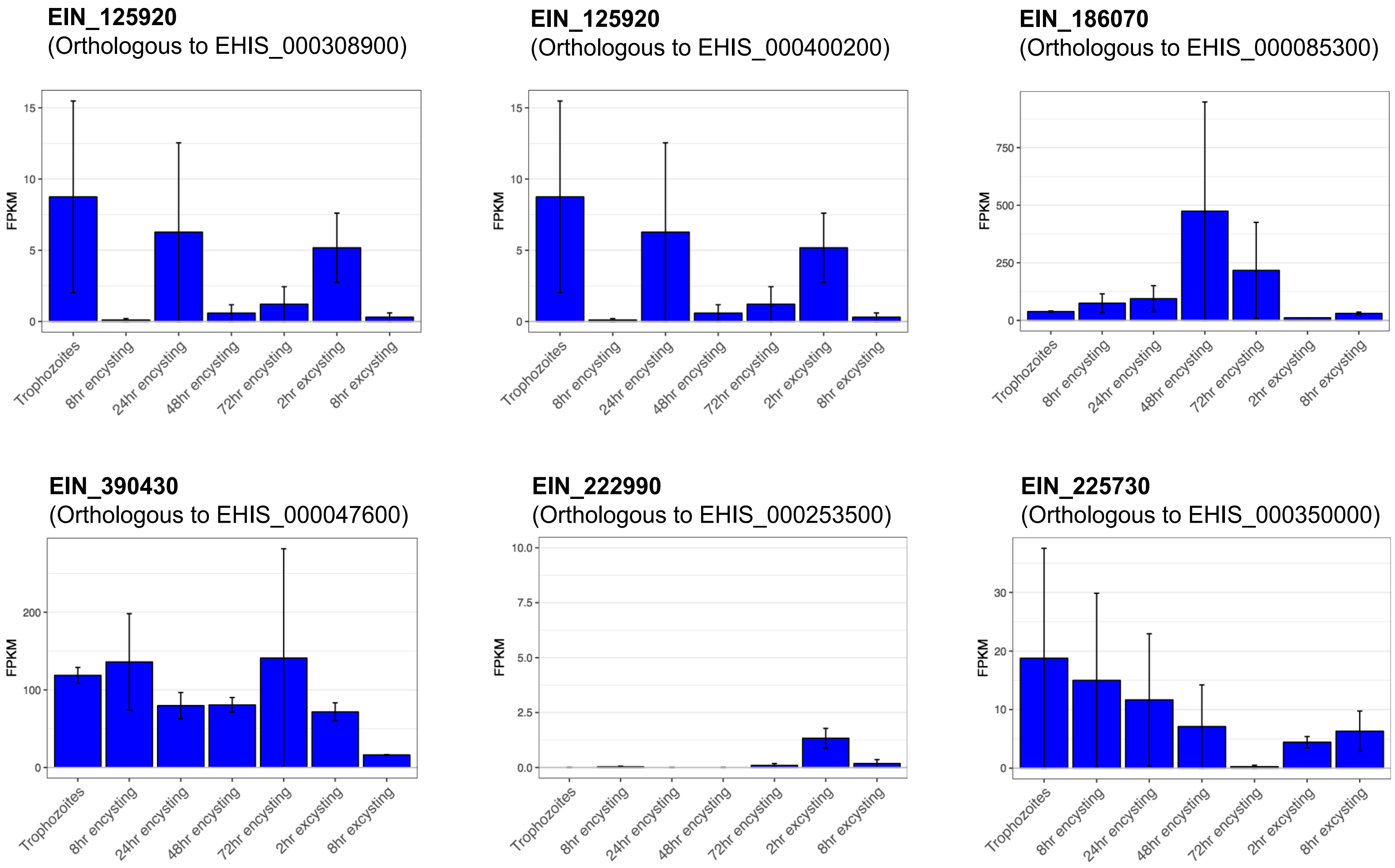


Figure S5.4. Expression of *Entamoeba invadens* -IP-1 genes whose orthologues are methylated in *Entamoeba histolytica* HM-1:IMSS. Methylated genes were identified in *E. histolytica* HM-1:IMSS using bisulphite sequencing. Orthologues to these genes were identified in *E. invadens* IP-1 and expression data collected from AmoebaDB. (Plot accessed taken AmoebaDB, June 2018).

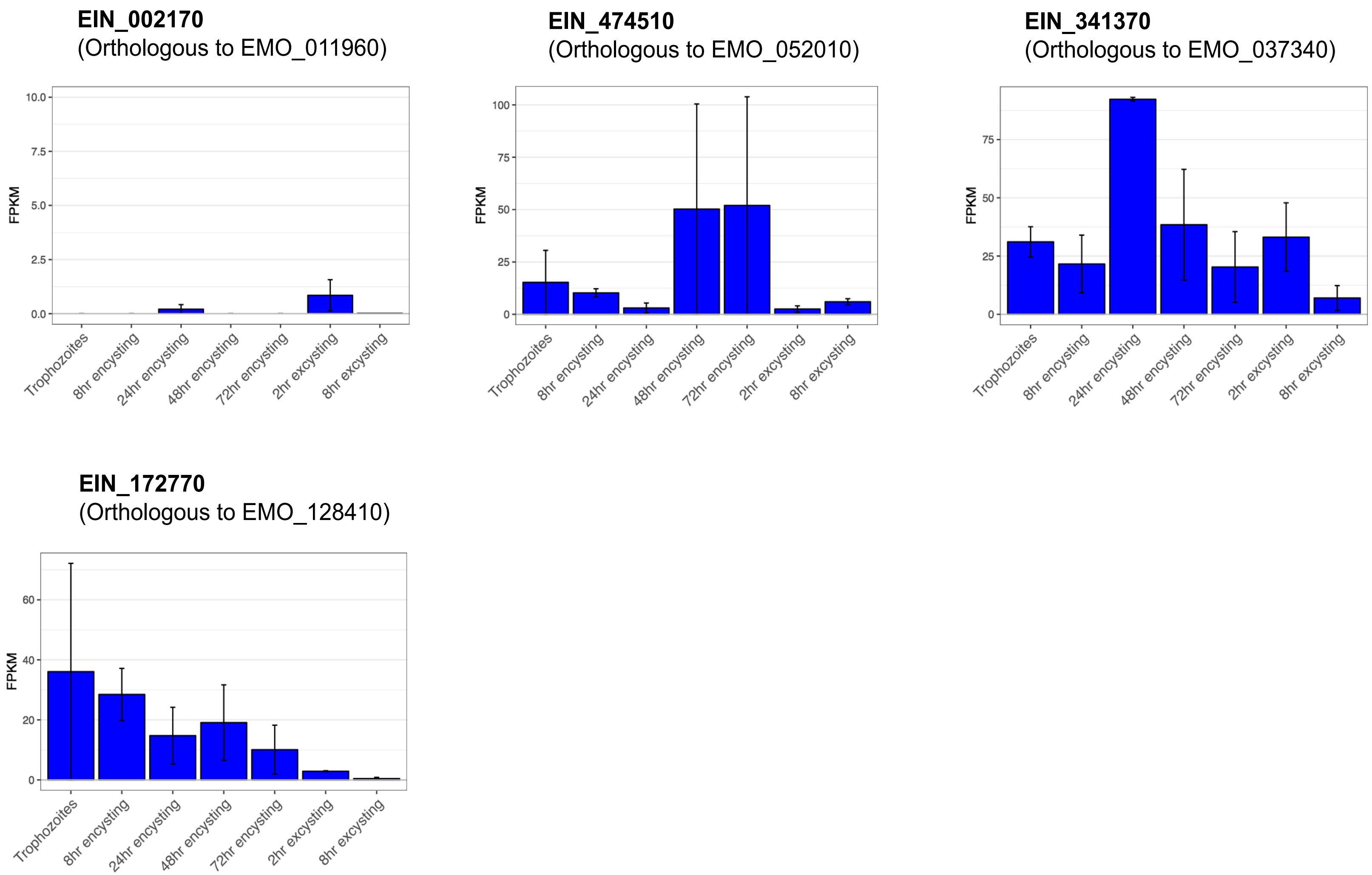


Figure S5.5. Expression of *Entamoeba invadens* -IP-1 genes whose orthologues are methylated in *Entamoeba moshkovskii* Laredo. Methylated genes were identified in *E. moshkovskii* Laredo using bisulphite sequencing. Orthologues to these genes were identified in *E. invadens* IP-1 and expression data collected from AmoebaDB. (Plots taken from AmoebaDB, June 2018).

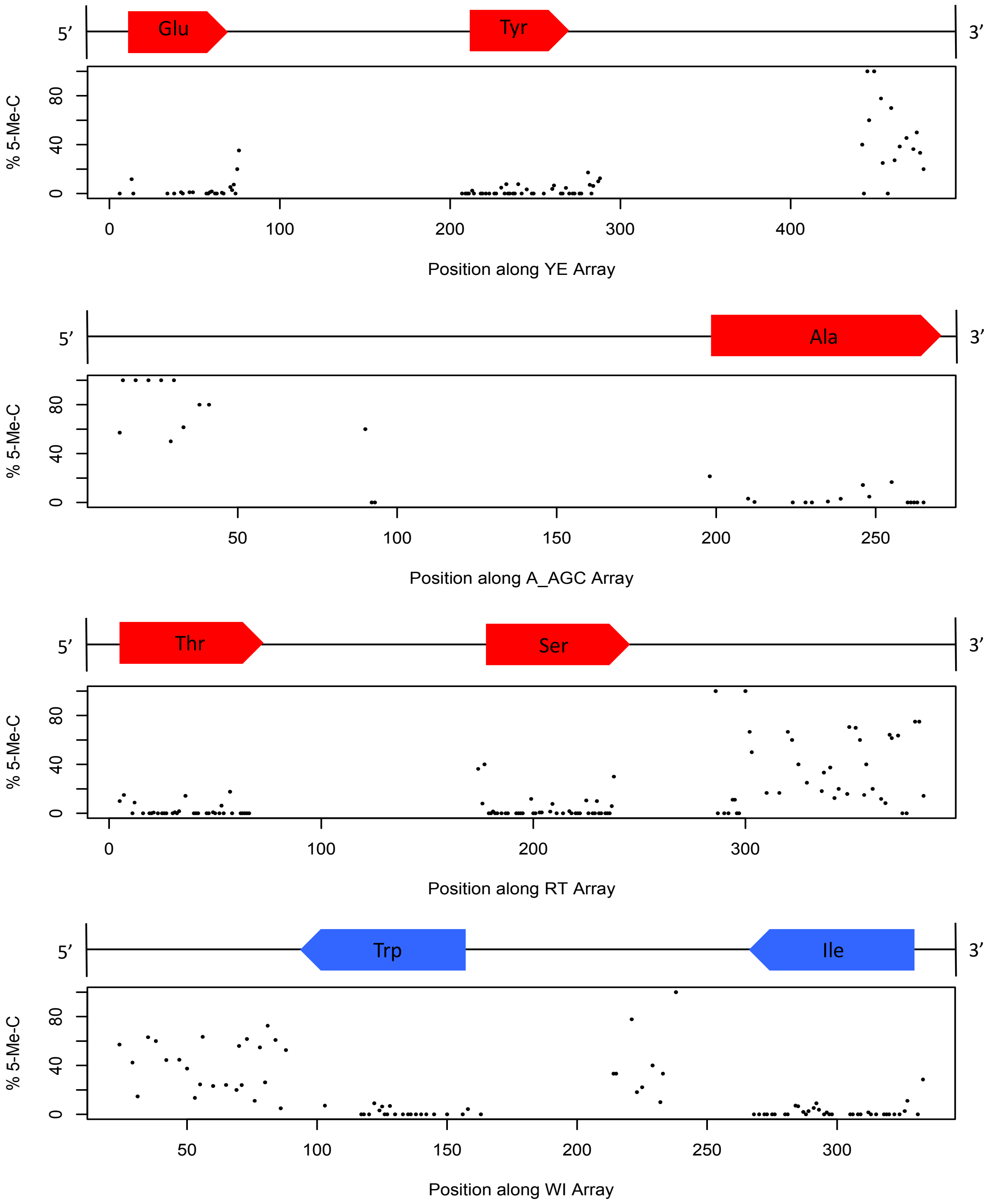


Figure S5.6. Methylation of *Entamoeba moshkovskii* Laredo tRNA array units YE, A^{AGC}, RT and WI. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 5X coverage.

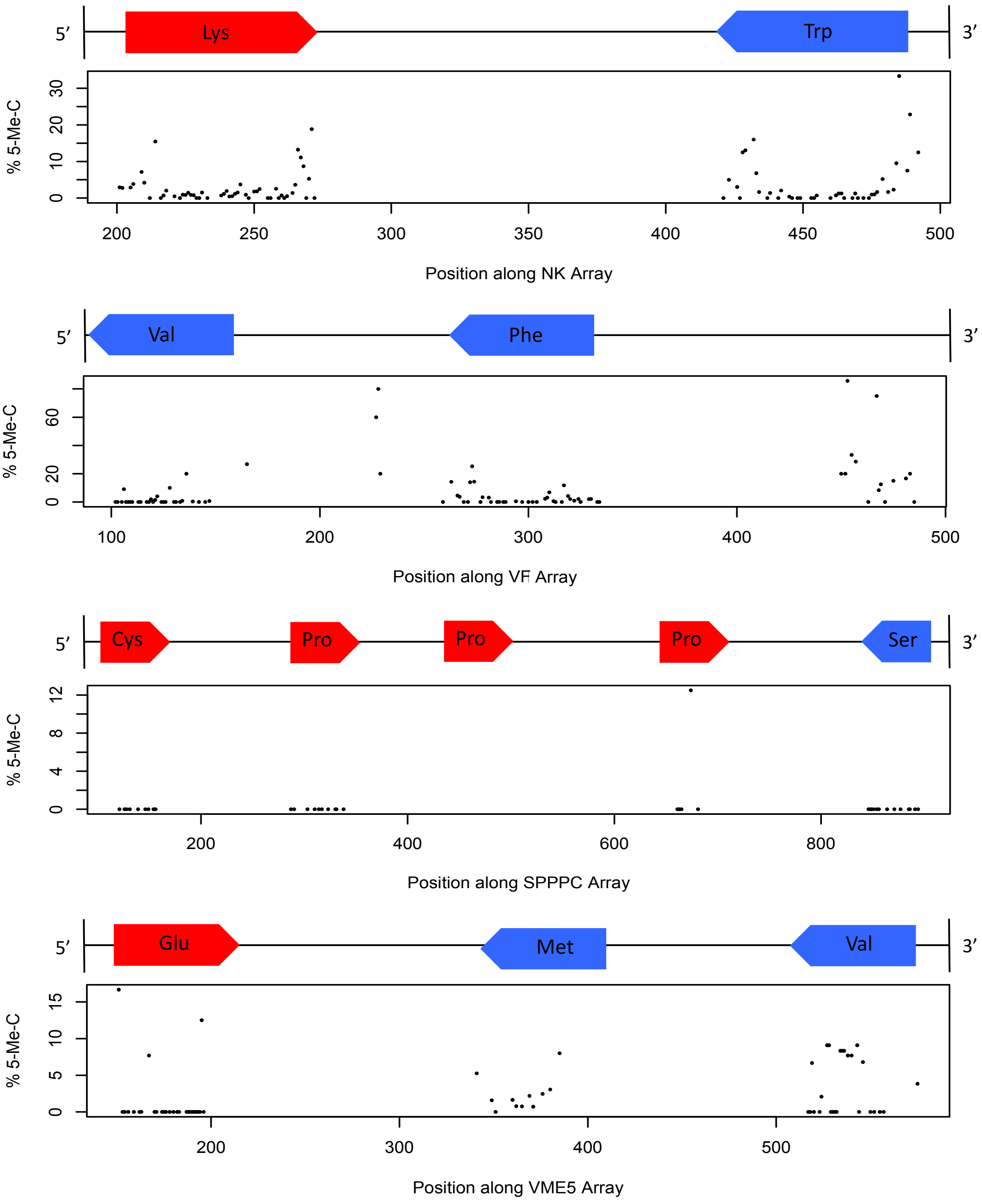


Figure S5.7. Methylation of *Entamoeba moshkovskii* Laredo tRNA array units NK, VF, SPPPC and VME5. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 5X coverage.

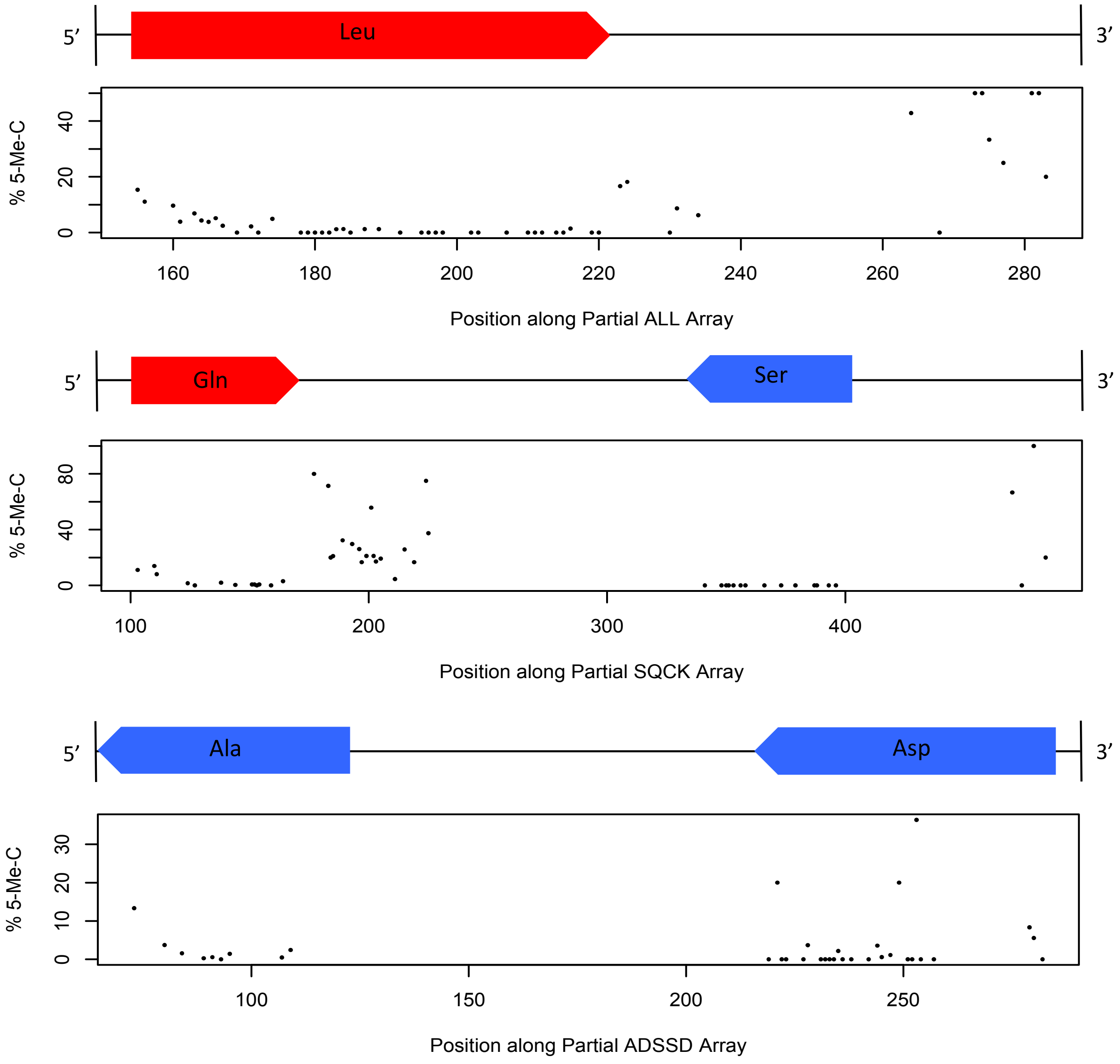


Figure S5.8. Methylation of *Entamoeba moshkovskii* Laredo tRNA array units ALL, SQCK and ADSSD. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 5X coverage.

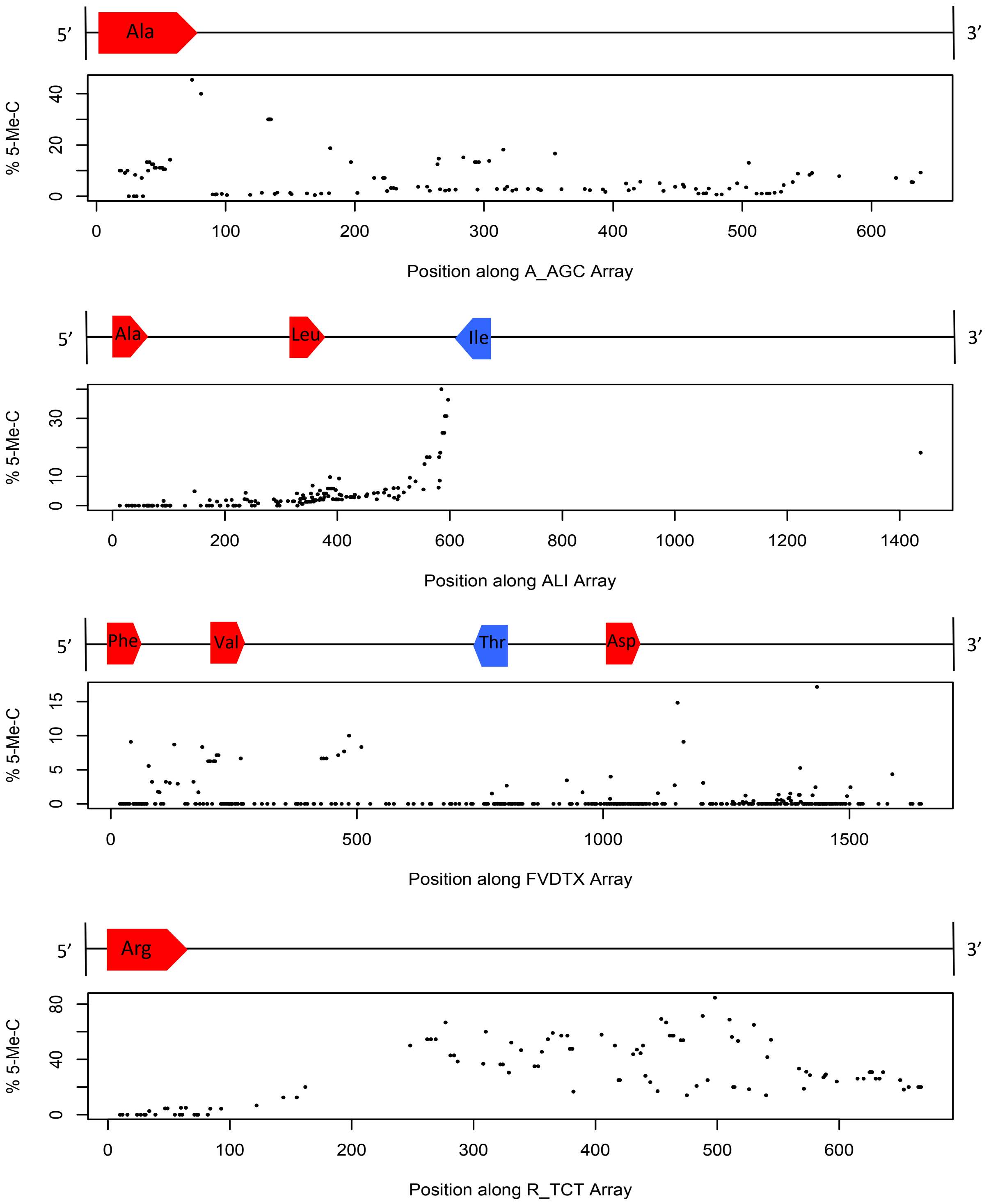


Figure S5.9. Methylation of *Entamoeba invadens* IP-1 tRNA array units A^{AGC}, ALI, FVDTX and R^{TCT}. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 10X coverage.

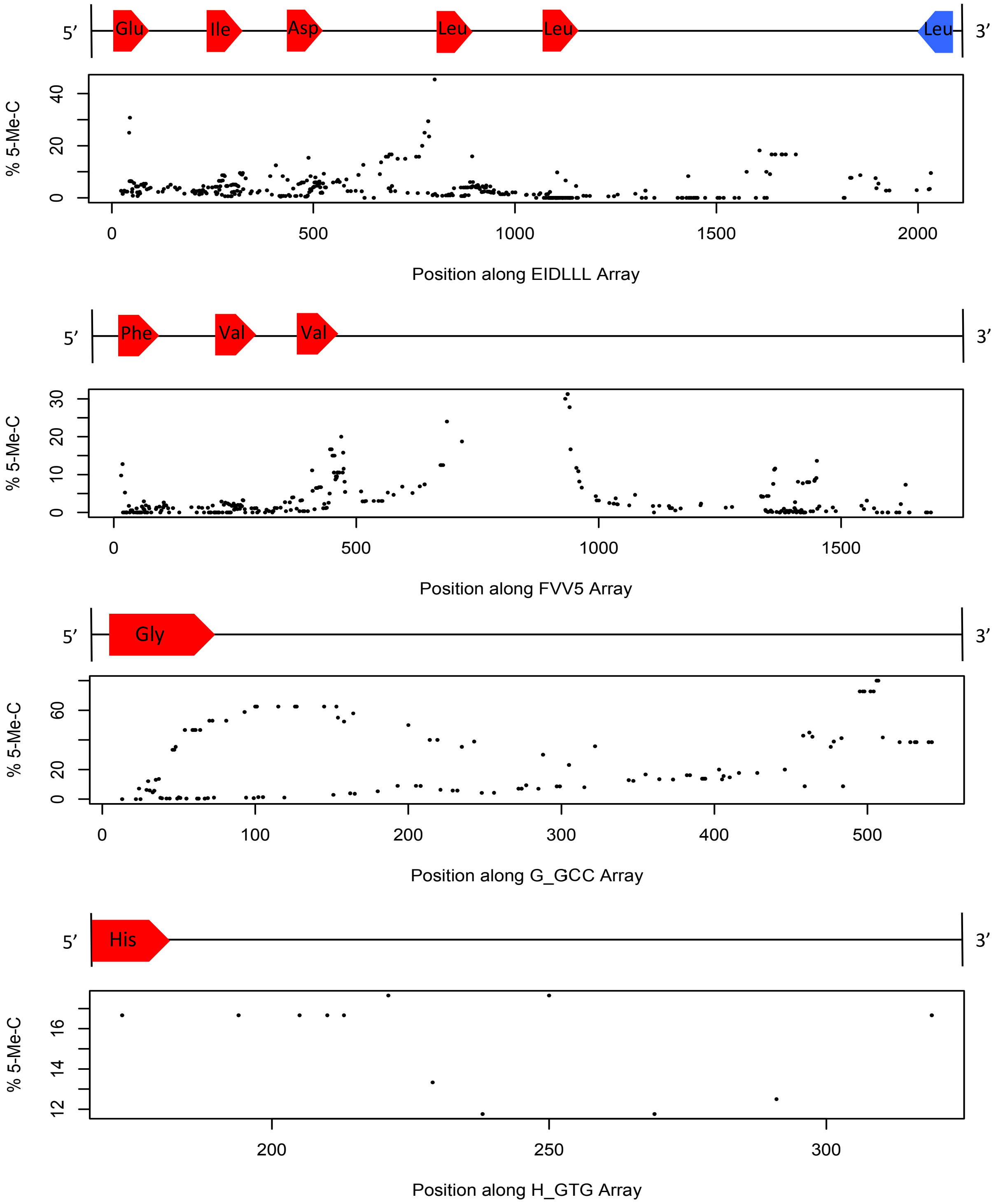


Figure S5.10. Methylation of *Entamoeba invadens* IP-1 tRNA array units EIDLLL, FVV5, G^{GCC} and H^{GTG}. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 10X coverage.

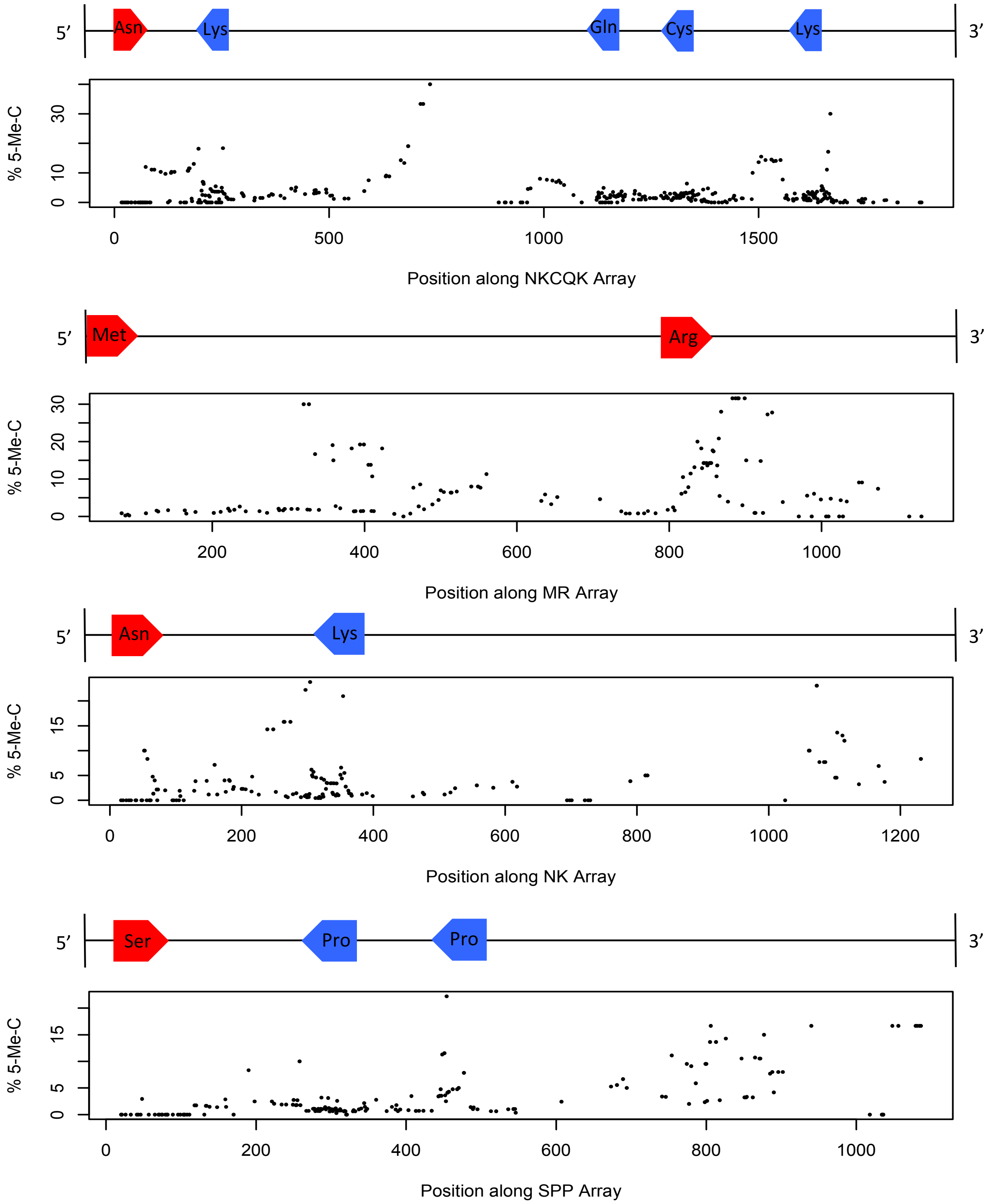


Figure S5.11. Methylation of *Entamoeba invadens* IP-1 tRNA array units NKCQK, MR, NK and SPP. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 10X coverage.

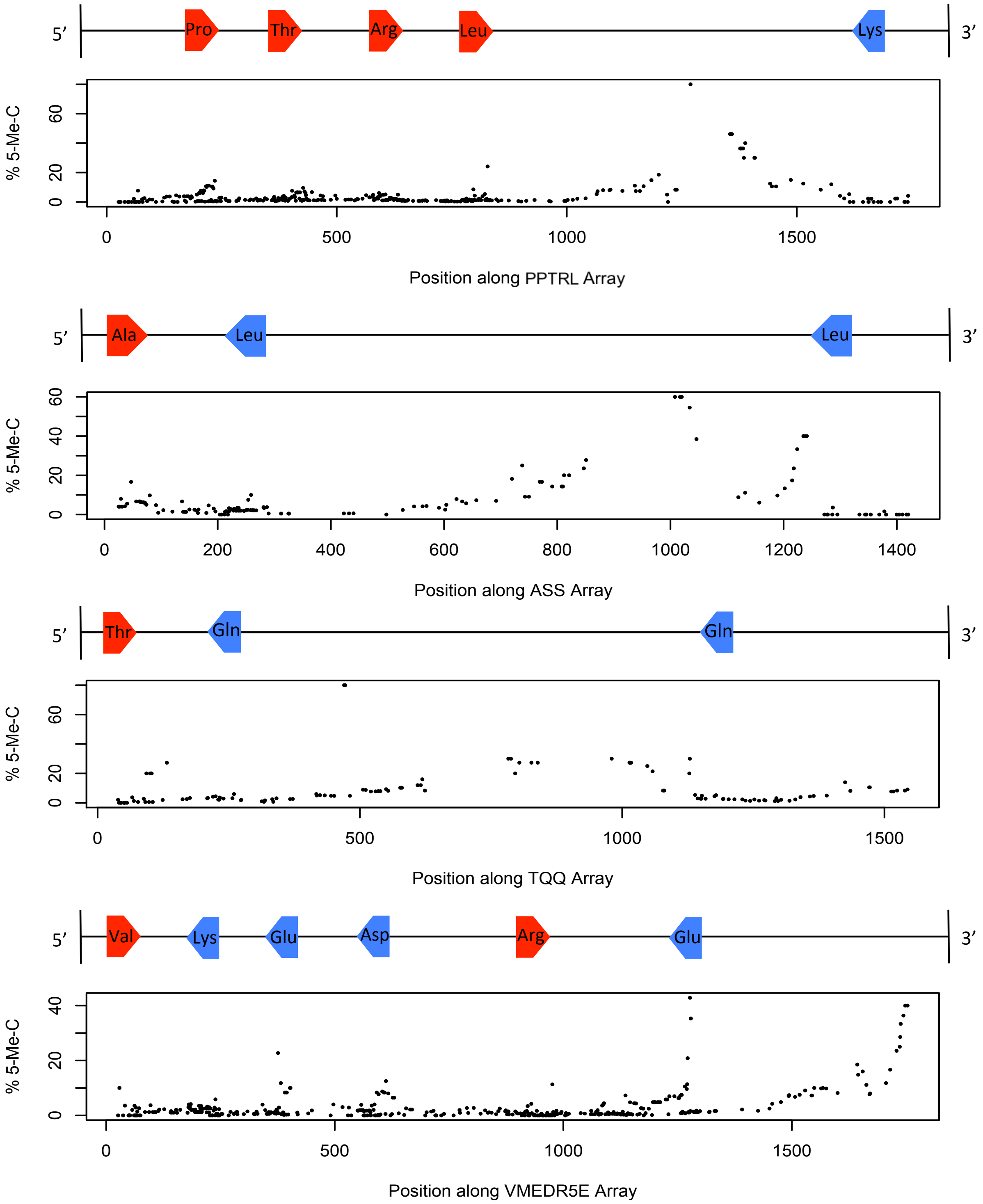


Figure S5.12. Methylation of *Entamoeba invadens* IP-1 tRNA array units PPTRL, ASS, TWW and VMEDR5E. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 10X coverage.

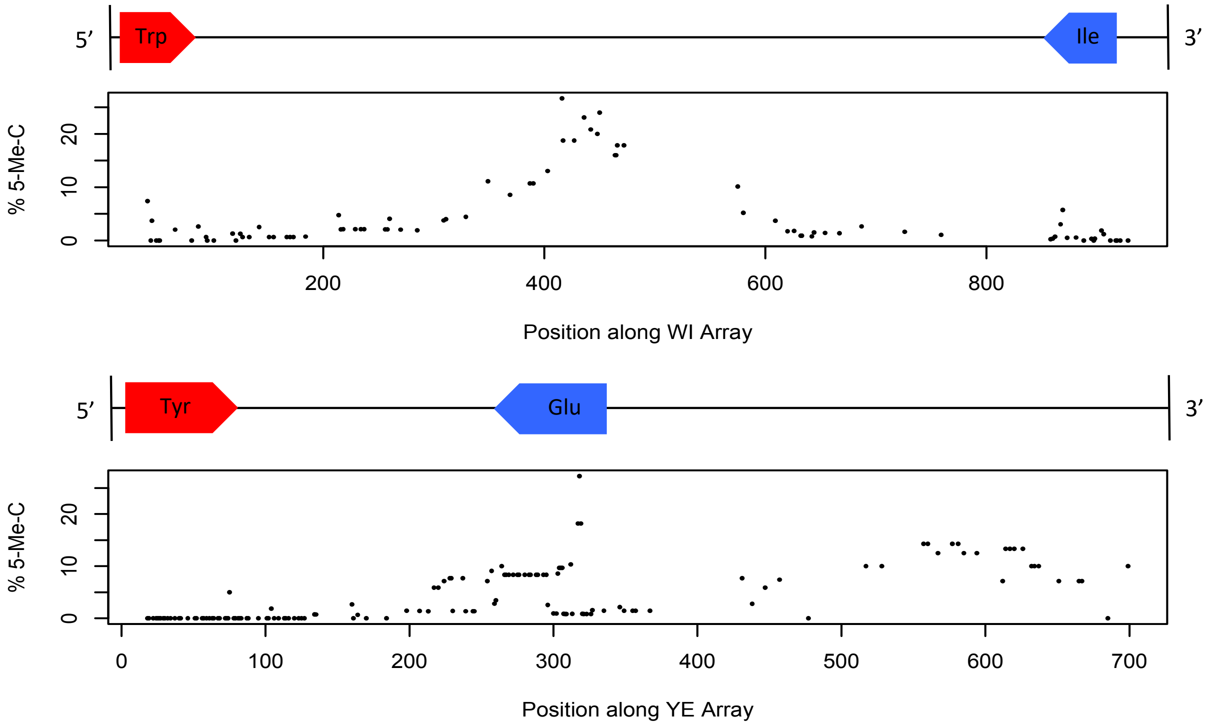


Figure S5.13. Methylation of *Entamoeba invadens* IP-1 tRNA array units WI and YE. Differential methylation of cytosine bases along the tRNA arrays was observed. Tracks show a singular tRNA array unit with the points demonstrating the position of cytosines and the percentage of methylated reads that mapped to the each position. All cytosines presented had at least 10X coverage.