high quality draft sequences will be produced for both in the near future. It is hoped that the *E. dispar* sequence will prove useful in identifying genomic differences linked to disease causation while that of *E. invadens* will be used to study patterns of gene expression during encystation. Small-scale genome surveys have been performed for two other species: *E. moshkovskii,* which is primarily a free-living species although it occasionally infects humans, and *E. terrapinae*, a reptilian commensal species (http://www.sanger.ac.uk/Projects/Comp_Entamoeba/).

2. GENOME STRUCTURE

2.1 The *E. histolytica* Genome Sequencing, Assembly and Annotation Process

The first choice to be made in the genome project was perhaps the easiest - the identity of the strain to be used for sequencing. A significant majority of the existing sequence data prior to the genome project was derived from one strain: HM-1:IMSS. This culture was established in 1967 from a rectal biopsy of a Mexican man with amoebic dysentery and axenised shortly thereafter. It has been used widely for virulence, immunology, cell biology and biochemistry in addition to genetic studies. In an attempt to minimise the effects of long-term culture cryopreserved cells that had been frozen in the early 1970s were revived and this uncloned culture used to generate the DNA for sequencing.

Before undertaking a genome scale analysis it is important to understand the quality and provenance of the underlying data. The *E. histolytica* genome was sequenced by whole genome shotgun approach with each center generating roughly half of the reads. Several different DNA libraries containing inserts of different sizes were produced using DNA that had been randomly sheared and sequences were obtained from both ends of each cloned fragment. The Phusion assembler (Mullikin and Ning, 2003) was used to assemble the 450,000 short reads into larger contigs (contiguous sequences), resulting in 1819 genome fragments that were approximately 12X deep, which means that each base has

been sequenced 12 times, on average. While the genome shotgun sequence provides high coverage of each base it is inevitable that there will be misassemblies and sequencing errors in the final consensus particularly towards each end of the contigs. Another problem with draft sequence is that it contains gaps, and while most of these will be small and will mostly contain repetitive non-coding "junk" sequence, some of the gaps will probably contain genes. This makes it impossible to be absolutely certain of the absence of particular genes in *E. histolytica* and, in some cases, the presence or absence of particular biological pathways. Due to the high repeat content and low GC content (24.1%) of the *E. histolytica* genome, closure of the remaining gaps is likely to be a lengthy process. Therefore it was decided to undertake and publish an analysis of the genome draft following assembly of the shotgun reads.

Annotation of the protein coding regions of the genome was initially carried out using two genefinders (GlimmerHMM (Majoros *et al.*, 2004) and Phat (Cawley *et al.*, 2001)) previously used successfully on another low G+C genome, that of *P. falciparum*. The software was re-trained specifically for analysis of the *E. histolytica* genome. The training process involved preparing a set of 600 manually edited genes to be used as models with the subsequent genefinding then being carried out on all of the assembled contigs to generate a 'complete' gene set. Predicted gene functions were generated automatically by homology searches using public protein and protein-domain databases, with subsequent refinement of identifications being carried out by manual inspection. For particular genes and gene families of special interest, members of the *Entamoeba* scientific community were involved throughout this process as expert curators with each individual assisting in the analysis and annotation of their genes of interest. Therefore although the manual curation of the genome has not been systematic, those areas of biology that are of primary interest to the *Entamoeba* community have been annotated most thoroughly. The publication of the genome by Loftus *et al* therefore represents a "first draft" of the complete genome sequence and the level of annotation is similar to the

initial publications of other genomes such as *Drosophila* (Adams *et al.*, 2000; Myers *et al.*, 2000) and human (Lander *et al.*, 2001).

2.2 Karyotype and Chromosome Structure.

The current *E. histolytica* genome assembly is approximately 23.7 million base pairs (Mbp) in size (Table 1). This figure is not likely to be a very accurate measure. In part this is due to misassembly of repetitive regions, which will cause the genome to appear smaller, and in part because of the possibility of aneuploidy in some regions of the genome, which would cause them to appear more than once in the assembly. Overall, however, this size is not inconsistent with data from pulse-field gels (Willhoeft and Tannich, 1999) and kinetic experiments (Gelderman *et al.*, 1971a,b) making the *E. histolytica* genome comparable in size (24 Mbp) to that of *Plasmodium falciparum* (23 Mbp) (Gardner *et al.*, 2002), *Trypanosoma brucei* (26 Mbp) (Berriman *et al.*, 2005), and the free living amoeba *Dictyostelium discoideum* (34 Mbp) (Eichinger *et al.*, 2005).

The current assembly does not represent complete chromosomes. Analysis of pulse-field gels predicts 14 chromosomes ranging in sizes from 0.3 to 2.2 Mb and possibly a ploidy of four (Willhoeft and Tannich, 1999). There is no current information regarding the size and nature of the centromeres and there are no contigs that appear to contain likely centromeric regions based on comparisons with other organisms. A search for signature telomeric repeats within the data indicates that these are either not present in the genome, not present in our contigs, or are diverged enough to be unidentifiable. However, there is circumstantial evidence that the chromosome ends may contain arrays of tRNA genes (see 2.4 below).

2.3 Ribosomal RNA Genes

The organisation of the structural RNA genes in *E. histolytica* is unusual with the rRNA genes carried exclusively on 24 kb circular episomes (Bhattacharya

et al., 1998) that have two transcription units in an inverted repeat. These episomes are believed to make up about 20% of the total cellular DNA; indeed, roughly 15% of all of the sequencing reads generated in the genome project were derived from this molecule with the exception of certain libraries where attempts were made to exclude it. There are thought to be numerous other circular DNA molecules of varying sizes present with unknown functions (Dhar *et al.*, 1995; Lioutas *et al.*, 1995) but unfortunately they have not yet been identified in the genome shotgun sequence data. The exact reasons for this are unknown but the small size of the DNA may have prevented proper shearing during the library construction process. These molecules represent an intriguing unsolved aspect of the *E. histolytica* genome.

2.4 tRNA Genes

Perhaps the most unusual structural feature identified in the *E. histolytica* genome is the unprecedented number and organisation of its tRNA genes (Clark *et al.*, 2006a). Over 10% of the sequence reads contained tRNA genes and these are (with a few exceptions) organised in linear arrays. The array organisation of the tRNAs was immediately obvious in some cases from the presence of more than one repeat unit in individual sequence reads and in other cases from their presence in both reads from the two ends of the same clone. However because of the near complete identity of the array units they were impossible to assemble by the software used and therefore the size of the arrays cannot be estimated accurately.

By manual assembly of tRNA gene-containing reads, 25 distinct arrays with unit sizes ranging from under 500 bp to over 1750 bp were identified (Clark *et al.*, 2006a). The arrayed genes are predicted to be functional because of the 42 acceptor types found in arrays none has been found elsewhere in the genome. These array units encoded between one and five tRNAs and a few tRNA genes are found in more than one unit. Three arrays also encode the 5S RNA and one encodes what is thought to be a small nuclear RNA. Experimental quantitative

hybridisations suggest a copy number of between about 70 and 250 for various array units. In total it is estimated that there are about 4500 tRNA genes in the genome. The frequency of a particular tRNA isoacceptor appears to be independent of the codon usage in *E histolytica* protein-coding genes.

Between the genes in the array units are complex, non-coding, short tandem repeats ranging in size from 5 to over 36 bp. Some variation in short tandem repeat number is observed between copies of the same array unit but this variation is usually minor and not visible when inter-tRNA PCR amplification is performed. However, these regions often exhibit substantial variation when different isolates of *E. histolytica* are compared and this is the basis of a recently described genotyping method for this organism (Ali *et al.*, 2005).

There is indirect evidence to suggest that the tRNA arrays are present at the ends of chromosomes. Although allelic *E. histolytica* chromosomes often differ substantially in size in pulse-field gels, a central protein-encoding region appears to be conserved as DNA digested with rare cutting enzymes gives only a single band in Southern blots when most protein-coding genes are used as probes. In contrast, when some tRNA arrays are used as probes on such blots, the same number of bands is seen in digested and undigested DNA. It is therefore tempting to conclude that the tRNA genes are at the ends of the chromosomes and to speculate that these repeat units may perform a structural role. In *D. discoideum* it is thought that rDNA may function as a telomere in some cases (Eichinger *et al.*, 2005) and the tRNA arrays in *E. histolytica* may perform a similar role.

The chromosomal regions flanking the tRNA arrays are generally devoid of protein coding genes but often contain incomplete transposable elements (see next section) and other repetitive sequences (Clark *et al.*, 2006a). This is also consistent with a telomeric location.

2.5 LINEs

The *E. histolytica* genome is littered with transposable elements. There are two major types autonomous LINEs (Long Interspersed Elements) of which there are three subtypes (EhLINE 1, 2 and 3) and there are two types of SINEs (Short Interspersed Elements) (Eh SINE1 and 2) (Table 2a). The classification of these elements and their organisation has been reviewed recently (Bakre *et al.*, 2005). Phylogenetic analysis of the EhLINEs places them in the R4 clade of non- Long Terminal Repeat (LTR) elements, a mixed clade of elements that includes members from nematodes, insects, and vertebrates (Van Dellen *et al.*, 2002a). Analysis of the *E. histolytica* genome shows no evidence for the presence of LTR retrotransposons and very few DNA transposons (of the *Mutator* family) (Pritham *et al.*, 2005).

All copies of EhLINEs examined encode non-conservative amino acid changes, frame shifts, and/or stop codons and no copy with a continuous open reading frame (ORF) has yet been found. This suggests that the majority of these elements are inactive. However, a large number of EhLINE1 copies do contain long ORFs without mutations in the conserved protein motifs of the RT and EN domains, suggesting that inactivity is quite recent. ESTs corresponding to EhLINEs have been found suggesting that transcription of these elements still occurs. Although most R4 elements insert in a site-specific manner, EhLINEs do not show strict site-specificity and are widely dispersed in the genome. They are quite frequently found close to protein-coding genes and inserted near T-rich stretches (Bakre *et al.*, 2005).

All three EhLINE subtypes are of approximately equal size ranging from 4715 to 4811 bp in length. Individual members within an EhLINE family typically share >85% identity, while between families they are <60% identical. By aligning the available sequences, each EhLINE can be interpreted to encode a single predicted ORF that spans almost the entire element (EhLINE1, 1589 aa; EhLINE2, 1567 aa; EhLINE3, 1587 aa). However, a precise 5bp duplication at

nucleotide position 1442 in about 80% of the copies of EhLINE1 creates a stop codon, dividing the single ORF in two. Similarly in 92% of EhLINE2 copies, the single ORF contains a precise deletion of two nucleotides at position 1272, resulting in two ORFs. Very few intact copies of EhLINE3 are found. The location of the stop codon leading to two ORFs appears to be conserved since in both EhLINE1 and EhLINE2 the size of ORF1 is about half that of ORF2 (Bakre *et al.*, 2005). Among the identifiable domains in the predicted proteins are reverse transcriptase (RT) and a restriction enzyme-like endonuclease (EN). The putative 5' and 3' untranslated regions are very short (3-44 bp).

EhLINEs 1 and 2 appear to be capable of mobilising partner SINEs (see next section) for which abundant transcripts have been detected in *E. histolytica*. Putative LINE/SINE partners can be assigned on the basis of conserved sequences at the 3'-ends of certain pairs, which otherwise showed no sequence similarity. The relevance of this assignment for the EhLINE1/SINE1 pair has recently been demonstrated (Mandal *et al.*, 2004).

2.6 SINEs

The two EhSINEs are clearly related to the EhLINEs as they have a conserved 3' sequence. They are nonautonomous, non-LTR retrotransposons (nonautonomous SINEs). The genetic elements encoding the abundant polyadenylated but untranslatable transcripts found in *E. histolytica* cDNA libraries (initially designated IE elements (Cruz-Reyes and Ackers, 1992; Cruz-Reyes *et al.*, 1995) or *ehapt2* (Willhoeft *et al.*, 2002)) have now been designated EhSINE1 (VanDellen *et al.*, 2002a; Willhoeft *et al.*, 2002). BLAST searching with representative examples of the first 44 EhSINE1s detected has identified 90 full-length (= 99% complete) copies and at least a further 120 partial (= 50% of full length) copies in the genome. Length variation is observed among EhSINE1s and is largely due to variable numbers of internal 26-27 bp repeats (Ackers, unpublished). The majority contain two internal repeats and cluster closely around 546 bp in length.

A second *E. histolytica* SINE (EhSINE2) has recently been described (Van Dellen *et al.*, 2002a; Willhoeft *et al.*, 2002). Examination of the four published sequences again suggests the presence of variable numbers of short (20 bp) imperfect repeats. BLAST searching identified a total of 47 full-length (= 99%) and at least 60 partial copies in the genome. The 3'-end of EhSINE2 shows high similarity (76%) to the 3' end of EhLINE2.

A polyadenylated transcript designated UEE1 found commonly in cDNA libraries from *E. dispar* (Sharma *et al.*, 1999) is also a non-LTR retrotransposon. A single copy of a UEE1-like element has been identified in the *E. histolytica* genome and is here designated EhSINE3. There is no significant sequence identity between EhSINE3 and EhLINE3 but the 3' end of EhSINE3 is very similar to that of EhLINE1.

Analysis of an *E. histolytica* EST library identified over 500 significant hits to both EhSINE1 and EhSINE2. No convincing transcript from EhSINE3 could be identified although the nearly identical *E. dispar* UEE elements (EdSINE1; Shire and Ackers, submitted) are abundantly transcribed.

A very abundant polyadenylated transcript, *ehapt1*, was described by Willhoeft *et al.* (1999) in a cDNA library. However, only a small number of partial matches could be found in the current *E. histolytica* assembly and only 10-20 strong hits in the much larger *E. histolytica* EST library now available. *ehapt1* does not appear to be a SINE element and its nature is currently unclear. The lack of matches in the genome suggests either that it is encoded in regions missing from the current assembly or that it contains numerous introns.

2.7 Other Repeats

The *E. histolytica* genome contains a number of other repetitive elements whose functions are not always clear. There are over 75 genes encoding

leucine-rich tandem repeats (LRR) of the type found in BspA-like proteins of the *Treponema pallidum* LRR (TpLRR) subfamily, which has a consensus sequence of LxxIxIxxVxxIgxxAFxxCxx (Davis *et al.*, 2006). These proteins generally have a surface location and may be involved in cell-cell interaction. Genes encoding such proteins are found mainly in Bacteria and some Archaea; so far they have been identified in only one other eukaryote, *Trichomonas vaginalis* (Hirt *et al.*, 2002). An extensive description of the BspA-like proteins *of E. histolytica* has recently been published (Davis *et al.*, 2006) and one of them has been shown to be surface exposed (Davis *et al.*, 2006).

E. histolytica stress sensitive protein (Ehssp) 1 is a dispersed, polymorphic and multicopy gene family (Satish *et al.*, 2003) and is present in ca. 300 copies per haploid genome as determined by hybridisation (Table 2a). The average Ehssp1 ORF is 1 kb in length with a centrally-located acidic-basic region (ABR) that is highly polymorphic. Unlike other such domains no clear repetitive motifs are present. The protein has, on average, 21% acidic (aspartate and glutamate) and 17% basic (arginine and lysine) amino acids, most of which are located in the ABR. The ABR varies in size from 5 to 104 amino acids among the various copies. No size polymorphism is seen outside the central ABR domain. The genes have an unusually long 5' untranslated region (UTR; 280 nucleotides). Only one or a few copies of the gene are transcribed during normal growth, but many are turned on under stress conditions. Homologues of this gene are present in *E. dispar*, but there is very little size polymorphism in the *E. dispar* gene family.

Eukaryotic genomes usually contain numerous microsatellite loci with repeat sizes of 2-3 basepairs. With the exception of di- and tri-nucleotides made up entirely of A+T such sequences are rare in the *E. histolytica* genome. In contrast, two dispersed repeated sequences of unknown function occur far more frequently than would be expected at random. Family 16 has a 42 base consensus sequence and occurs approximately 38 times in the genome while

family 17 has a 27 base consensus sequence and occurs 35 time in the genome (Table 2b). The significance of these sequences remains to be determined.

2.8 Gene Number

The current assembly predicts that the genome contains around 10,000 genes, almost twice as many as seen in *P. falciparum* (Gardner *et al.*, 2002) or *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996) but closer to that of the free living protist *Dictyostelium discoideum* (ca. 12,500; Eichinger *et al.*, 2005). It should be remembered that this number will change as the assembly improves, and is likely to decrease somewhat. Nevertheless, the comparatively large gene number when compared to some other parasitic organisms reflects both the relative complexity of *E. histolytica* and the presence of large gene families, despite the loss of certain genes as a consequence of parasitism. Gene loss and gain can both represent an adaptive response to life in the human host. Gene loss is most evident in the reconstruction of metabolic pathways of *E. histolytica*, which show a consistent pattern of loss of synthetic capacity as a consequence of life in an environment rich in complex nutrient sources. Similarly, analyses of expanded gene families with identifiable functions indicate that many are directly associated with the ability to sense and adapt to the environment within the human host and the ability to ingest and assimilate the nutrients present. One consequence of these gene family expansions being linked to phagocytosis of bacteria and other cells may be an association between many of these gene families and pathogenicity.

2.9 Gene Structure

Most *E. histolytica* genes comprise only a single exon; however as many as 25% may be spliced and 6% contain two or more introns. Therefore mRNA splicing is far less common than in the related protist *D. discoideum* or the malaria parasite *P. falciparum*. The genome contains all of the essential machinery for splicing (section 2.14) and a comparison of intron positions suggests that *D. discoideum* and *E. histolytica* have both lost introns since their

shared common ancestor with *P. falciparum,* although many more have been lost in the *E. histolytica* lineage. A good example of this intron loss is the vacuolar ATP synthase subunit D gene (Figure 1). This protein is highly conserved but the number of introns in each gene varies. *P. falciparum* has 5 introns, *D. discoideum* has two and *E. histolytica* has one. The positions of three of the five *P. falciparum* introns are conserved in one of the other species which suggests that these three (at least) were present in the common ancestor and that intron loss has led to the lower number seen in *E. histolytica* today. This loss is consistent with reverse transcriptase mediated 3' intron loss (Roy and Gilbert, 2005) as the 5' -most introns are retained. It would appear that this process has been more active in the *E. histolytica* and *D. discoideum* lineages than in *P. falciparum*, possibly because *Plasmodium* lacks a reverse transcriptase.

2.10 Gene Size

Genes in *E. histolytica* are surprisingly short, not only due to the loss of introns but also in the predicted lengths of the proteins they code for. On average the predicted length of a protein in *E. histolytica* is 389 amino acids (aa) which is 129 aa and 372 aa shorter than in *D. discoideum* and *P. falciparum* respectively. In fact the protein length distribution is most similar to that of the microsporidian *Encephalitozoon cuniculi* (Figure 2) which has a very compact genome of 3Mb and less than 2000 genes. Direct comparison of orthologous genes between *E. histolytica* and its closest sequenced relative *D. discoideum* demonstrates this phenomenon quite well, with the majority of *E. histolytica* proteins being shorter than the *D. discoideum* counterpart (Hall, unpublished). Protein length is normally very well conserved among eukaryotes so the reason for protein shortening is unclear. It has been postulated that in bacteria reduced protein lengths reflects a reduced capacity for signaling (Zhang, 2000). This would not seem to be the case here as the number of genes identified as having a role in signaling suggests quite the opposite. An alternative theory is that as *E.*

histolytica has reduced organelles it is possible that its proteins contain fewer or simpler targeting signals.

2.11 Protein Domain Content

The most common protein family (Pfam) domains of *E. histolytica* are shown in Table 3. The domains that are unusually common in *E. histolytica* reflect some of the more unusual aspects of the biology of this protist. For example, the Rab and Rho families that are involved in signaling and vesicle trafficking are among the most common domains in *E. histolytica* while in other species they are not often among the top 50 families. This could well be due to the fact that *E. histolytica* has a 'predatory' life style and these domains are intimately involved in environmental sensing, endocytosis and delivery of lysosomes to the phagosome. There are also a number of domains involved in actin dynamics and cytoskeletal rearrangement that are not common in non-phagocytic species, such as the gelsolin and SH3 domains. Myb domains are the most common transcription regulatory domains in *E. histolytica*; this domain is also common in plants where the proteins regulate many plant-specific pathways (Ito, 2005). An important finding from an initial analysis was the presence of unusual multidomain proteins, including five proteins containing both RhoGEF and Arf-GAP domains, suggesting a mechanism for direct communication between the regulators of vesicle budding and cytoskeletal rearrangement. Over 80 receptor kinases were identified (section 7.2.2), each containing a kinase domain and a C rich extracellular domain. These kinases fall into distinct classes depending on the presence of CXC or CXXC repeats. There are also domains that are common in most other sequenced genomes but rare or missing from *E. histolytica*. For example, most mitochondrial carrier domain proteins are not needed in *E. histolytica* as it lacks a normal mitochondrion (section 8).

2.12 Translation-Related Proteins

Two of the predicted tRNAs (Ile^{TAT} and Tyr) need to be spliced due to the presence of an intron. tRNA introns are distinct in structure from those in protein-coding genes and require a distinct splicing machinery. The expected enzymes required for this splicing are present as are a number of tRNA modification enzymes (including those for sythesising queuine and pseudouridine) and rRNA methylases that act on specific bases in their respective RNA molecules. The expected panel of tRNA synthetases necessary for aminoacylating the tRNAs is also present, with one or two gene copies for each type.

The majority of ribosomal protein genes are well-conserved in *E. histolytica* and only the gene for large subunit protein L41 could not be identified. The missing protein is only 25 amino acids in length, 17 of which are arginines or lysines, which would make it difficult to identify in this A+T-rich genome, but it is highly conserved, having been reported from Archaea to mammals. However, it also appears to be dispensable, as *S. cerevisiae* can grow relatively normally after deletion of both its copies (Yu and Warner, 2001). Nevertheless, deletion of L41 in *S. cerevisiae* reduces the level of 80S ribosomes, suggesting that it is involved in ribosomal subunit association, reduces peptidyl transferase activity, and increases translocation (Dresios *et al.*, 2003). In addition, L41 has been shown to interact with the beta subunit of protein kinase CKII and to stimulate phosphorylation of DNA topoisomerase II alpha by CKII (Lee *et al.*, 1997b). If this gene is truly absent from *E. histolytica* it may have important consequences for the cell.

No genes for mitochondrial ribosomal proteins were found. Their absence is not surprising since *E. histolytica* lacks typical mitochondria (see section 8 below).

In eukaryotic translation, elongation factor EF-1 is activated upon GTP binding and forms a ternary complex with aminoacyl tRNAs and ribosomes. EF-1 beta and delta subunits work as GDP-GTP exchange factors to cycle EF-1 alpha

between two forms while EF-1 gamma provides structural support for the formation of this multimeric complex. EF2 assists in the translocation of tRNAs on the mRNA by exactly one codon. *E. histolytica* has most of the expected factors except for EF-1 delta, a protein involved in exchanging GDP with GTP. This is also absent from *S. cerevisiae* and *P. falciparum*. It is likely that EF-1 beta carries out this activity. It is thought that the EF-1 complex can exist in two forms, EF-1-alpha/beta/gamma and EF-1-alpha/delta/gamma. In *E. histolytica*, probably only the former complex exists.

Eukaryotes typically have two polypeptide release factors, eRF1 and eRF3. Both of these factors have been found in *E. histolytica*.

2.13 Analysis of Cell Cycle Genes

Alternation of DNA duplication and chromosome segregation is a hallmark in the cell cycle of most eukaryotes. Carefully orchestrated processes coordinate an ensemble of cell cycle regulating 'checkpoint' proteins ensure that progeny cells receive an exact copy of the parental genetic material (Hartwell and Weinert, 1989). Unlike most eukaryotes, *Entamoeba histolytica* cells can reduplicate their genome several times before cell division occurs (Gangopadhyay *et al.*, 1997). Approximately 5-20% of the trophozoites (depending on the growth phase) of axenic culture are multi-nucleated. Additionally, DNA reduplication may occur without nuclear division so that single nuclei contain 1X -6X or more genome contents (Das and Lohia, 2002). Thus axenically cultured *E. histolytica* trophozoites display heterogeneity in their genome content suggesting that eukaryotic cell cycle checkpoints are either absent or altered in this organism. Around 200 genes have been identified in yeast that play a direct role in cell cycle progression.

2.13.1 DNA replication initiation and DNA duplication

The DNA replication licensing system is one of the crucial mechanisms that ensures the alternation of S-phase with mitosis in most cells (Tye, 1999).

Initiation of DNA replication involves binding of the replicative helicases to DNA replication origins in late mitosis. Loading of the replicative helicase Mcm2-7 proteins is preceded by formation of the pre-replicative complex (pre-RC) and its subsequent activation. Formation of pre-RC requires the ordered assembly of the origin recognition complex (ORC), Cdc6, Cdt1 and the Mcm2–7 proteins. The pre-RC is activated by the protein kinase Cdc7p and its regulatory subunit Dbf4 (Masai and Arai, 2002). Other factors that regulate the transition from pre-RC to replication initiation are Mcm10p, Cdc45p, TopBP1, RecQL4 and the GINS complex (Gregan *et al.*, 2003; Machida *et al.*, 2005; Merchant *et al.*, 1997; Wohlschlegel *et al.*, 2002). Two other Mcm proteins – Mcm8 and Mcm9 - have been identified in metazoan systems and are believed to be part of the replicative helicase (Maiorano *et al.*, 2006). Replication origin licensing is inactivated during S-phase but Mcm2–9p mayfunction as a helicase that unwinds DNA ahead of the replication fork during S-phase (Maiorano *et al.*, 2006). Once S-phase has begun, the formation of new pre-RC is kept in check by high CDK activity and by the activity of the protein geminin Bell and Dutta, 2002).

A detailed analysis of the *E. histolytica* genome shows that homologues of several proteins required for DNA replication initiation are absent. These include ORC (Origin Recognition Complex) 2-6, Cdt1, geminin, Cdc7/Dbf4 and Mcm10. A single gene encoding a homologue of the archaeal and human Cdc6/Orc1p (Capaldi and Berger, 2004) was identified. This suggests that DNA replication initiation in *E.histolytica* is likely similar to archaeal replication initiation where a single Cdc6p/ORC1p replaces the hetero-hexameric ORC complex (Kelman and Kelman, 2004). Several proteins described from metazoa, such as Cdt1, geminin, Mcm8 and Mcm9, have not been found in yeast. Surprisingly, Mcm8 and Mcm9 were identified in the *E.histolytica* genome.

Of the four known checkpoint genes that regulate DNA replication in *S. cerevisiae* only Mec1 and Mrc1 have homologues in *E. histolytica*. *E. histolytica* homologues of several proteins involved in G1-S transitions are absent, such as Sic1, Chk1. The S-phase checkpoint genes p21, p27, p53 and retinoblastoma (RB) required for transition from G1 to S-phase in humans were absent in *E. histolytica*. Chk1 and Chk2 genes encode kinases that act downstream from the ATM and ATR kinases (intra-S phase checkpoint genes). The Chk1 homologue is absent but a Chk2 homologue has been identified in *E. histolytica* and partially characterised (Iwashita *et al.*, 2005).

2.13.2 Chromosome segregation and cell division

A large number of genes are known to regulate different events during the transition from G2-Mitosis - spindle formation checkpoint, chromosome segregation, mitosis, exit from mitosis, and cytokinesis - in *S.cerevisiae*. Many of the proteins required by yeast for kinetochore formation have no obvious homologues in *E. histolytica* suggesting that amoeba kinetochores may have an altered composition and structure. Proteins of the Anaphase Promoting Complex (APC) regulate transition from metaphase to anaphase. With the exception of APC11, none of the APC proteins could be identified in *E.histolytica*. In contrast two genes encoding CDC20 homologues, which are known to activate the APC complex, were identified in *E.histolytica* along with ubiquitin and related proteins (Wöstmann *et al.*, 1992), indicating that although most APC subunit homologues were absent the pathway of proteasomal degradation for regulation of cell cycle proteins may still be functional in *E.histolytica*. Effectors of the apoptotic pathway and meiosis were also largely absent.

2.13.3 CDKs and cyclins

The CDC28 gene encodes the single cyclin dependant kinase (CDK) in *S. cerevisiae* and regulates cell cycle progression by binding to different cyclins at the G1/S or G2/M boundaries (Reed, 1992; Surana *et al.*, 1991; Wittenberg *et*

al., 1990). Similarly, *Schizosaccharomyces pombe* also encodes a single CDK (*cdc2*) (Simanis and Nurse, 1986) . Mammals and plants can encode multiple CDKs and an equally large number of cyclins (Morgan, 1995; Vandepoele *et al.*, 2002). Association of different CDKs with specific cyclins regulates the cell cycle in different developmental stages as well as in specific tissues. CDKs belong to the serine/threonine family of kinases with a conserved PSTAIRE domain where cyclins are believed to bind (Jeffrey *et al.*, 1995; Morgan, 1996) although some mammalian and plant CDKs have been shown to have divergent PSTAIRE motifs. This heterogeneity may or may not affect cyclin binding (Poon *et al.*, 1997). The *E. histolytica* genome encodes at least 9 different CDKs among which not even one has the conserved PSTAIRE motif. The closest homologue of the CDC28/cdc2 gene, which shows only conservative substitutions in the PSTAIRE motif (PVSTVRE), was cloned previously (Lohia and Samuelson, 1993). The remaining 8 CDK homologues exhibit even greater divergence in this motif. Eleven putative cyclin homologues with a high degree of divergence have been found. Identifying their CDK/cyclin partner along with their roles in the cell cycle is a major task that lies ahead. Some of the CDKs may not function by associating with their functional cyclin partners but may play a role in regulating global gene expression, either by activation from non-cyclin proteins or by other mechanisms (Nebreda, 2006).

E. histolytica presents a novel situation where the eukaryotic paradigm of a strictly alternating S-phase and mitosis is absent. Discrete G1, S and G2 populations of cells are not routinely found in axenic cultures. Instead cells in S-phase show greater than 2x genome contents, suggesting that the G2 phase is extremely short and irregular. This observation together with the absence of a large number of checkpoint genes suggests that regulation of genome partitioning and cell division in *E. histolytica* may be additionally dependant on extracellular signals. *E. histolytica* must however contain regulatory mechanisms to ensure that its genome is maintained and transmitted with precision even in the absence of the expected checkpoint controls. The

discovery of these mechanisms will be crucial to our understanding of how the *E. histolytica* cell divides.

2.14 Transcription

RNA polymerase II transcription in *E. histolytica* is known to be []-amanitin-resistant (Lioutas and Tannich, 1995). The F homology block of the RNA polymerase II large subunit has been identified as the putative []-amanitin binding site. This block is highly divergent in the []-amanitin resistant *Trichomonas vaginalis* RNA polymerase II (Quon *et al.*, 1996). The *E. histolytica* RPB1 homologue also diverges from the consensus in this region but, interestingly, it is also quite dissimilar to the *T. vaginalis* sequence.

The heptapeptide repeat (TSPTSPS) common to other eukaryotic RNA polymerase II large subunit C terminal domains (CTD) is not present in the *E. histolytica* protein. Indeed, the *E. histolytica* CTD is not similar to any other RNA polymerase II domain in the current database. However, the CTD of the *E. histolytica* enzyme does remain proline/serine-rich (these amino acids constitute 40% of the CTD sequence). The *E. histolytica* CTD also retains the potential to be highly phosphorylated: of the 24 serines, 6 threonines and 3 tyrosines within the CTD, 9 serines, 3 threonines and 1 tyrosine are predicted to be within potential phosphorylation sites. It is therefore possible that, despite its divergence, modification of the CTD by kinases and phosphatases could modulate protein-protein interactions as is postulated to occur in other RNA polymerases (Yeo *et al.*, 2003). In yeast, phosphorylation of the CTD regulates association with the mediator protein (Davis *et al.*, 2002; Kang *et al.*, 2001; Kornberg, 2001). The yeast mediator protein complex consists of 20 subunits. However, perhaps due to the divergence of the CTD, only two of these proteins have been identified in *E. histolytica* (Med7 and Med10). Homologues of the Spt4 and Spt5 elongation factors, also thought to interact with the CTD, have been identified.

The RNA polymerase core is composed of 12 putative subunits in *S. cerevisiae* (Young, 1991), while *S. pombe* contains a subset of 10 of these proteins, lacking the equivalents of subunits 4 and 9 (Yasui *et al.*, 1998). In *E. histolytica* only 10 of the RNA polymerase subunits have been identified, identifiable homologues of subunits 4 and 12 being absent. While the homologue to subunit 9 was present it lacks the first of the two characteristic zinc binding motifs of this protein and the DPTLPR motif in the C terminal region. A similar sequence, DPTYPK, is however present and a homologue of the TFIIE large subunit Tfa1, which is proposed to interact with this region of the protein, has been identified (Hemming and Edwards, 2000; Van Mullem *et al.*, 2002). The conserved N terminal portion (residues 1-52) of Rpb9 is thought to interact with both Rpb1 and Rpb2 in *S. cerevisiae* (Hemming and Edwards, 2000) and homologues of these have been identified.

The core promoter of *E. histolytica* has an unusual tripartite structure consisting of the three conserved elements TATA, GAAC and INR (Purdy *et al.*, 1996; Singh and Rogers, 1998; Singh *et al.*, 2002; Singh *et al.*, 1997). Singh and Rogers (1998) have speculated that the GAAC motif may be the binding site of a second or alternative *E. histolytica* DNA binding protein in the preinitiation complex. It is therefore of interest that, in addition to the *E. histolytica* TATA-binding protein (TBP), two other proteins contain the TATA-binding motif (Hernandez *et al.*, 1997). TBP is a subunit of the TFIID general transcription factor (GTF) which in other organisms is required for the recognition of the core promoter. In light of the variation in the core promoter previously mentioned, and the divergence in proteins that bind to the core promoter in other parasitic protists, it is not surprising that only six of the 14 evolutionary conserved subunits of TFIID, TBP Associated Factors (TAFs) 1, 5, 6, 10, 12 and 13 were identified. Homologues of some of the global regulatory subunits of the Ccr4/Not complex, which interacts with TBP and TAFs 1 and 13, have also been identified.

TAFs 5, 6, 10 and 12 are also components of the histone acetyltransferase (HAT) complexes in other organisms as is SPT6 and 16 (Carrozza *et al.*, 2003). While all known components of the HAT complexes have by no means been identified or the role of the previously unknown bromodomain containing proteins encoded in the *E. histolytica* genome, histone acetylation complexes are known to be active in *E. histolytica* (Ramakrishnan *et al.*, 2004). Other potential members of chromatin remodeling complexes of *E. histolytica* include the TBP interacting helicase (RVB1 & 2) and the SNF2 subunit of the SWI/SNF complex.

Homologues of some of the other GTFs (TFII E, F and H) but not the large or small subunits were identified. In contrast to the difficulty identifying some of the GTFs, the *E. histolytica* spliceosomes components U1, U2, U4/6, U5 and the Prp19 complex have all been identified. In fact homologues of ten of the fourteen "core" snRNP proteins, two of the U1 specific snRNPs, seven of the ten U2 specific snRNPs, five of the six U5 specific snRNPs, three of the U4/6 specific snRNPs, and four of the nine subunits of the Prp19 complex have been found. In fact *E. histolytica* has homologues of approximately 80% of the *S. cerevisiae* splicing machinery (Jurica and Moore, 2003).

Like *G. intestinalis, E. histolytica* has short 5' untranslated regions on its mRNAs. However, unlike those of *G. intestinalis, E. histolytica* mRNA has been shown to be capped (Ramos *et al*, 1997; Vanacová *et al.*, 2003). Identification of homologues of the Ceg1 RNA guanylyltransferase - an enzyme which adds an unmethylated GpppRNA cap to new transcripts - and of Abd1 - which methylates the cap to form m7GpppRNA - gives new insight into the probable cap structure in *E. histolytica* (Hausmann *et al.*, 2001; Pillutla *et al.*, 1998). It has been proposed that the capping enzymes interact with the phosphorylated CTD of RNA polymerase (Schroeder *et al.*, 2000). The CTD of *E. histolytica* large subunit is, as discussed earlier, not well conserved but contains several probable phosphorylation sites.