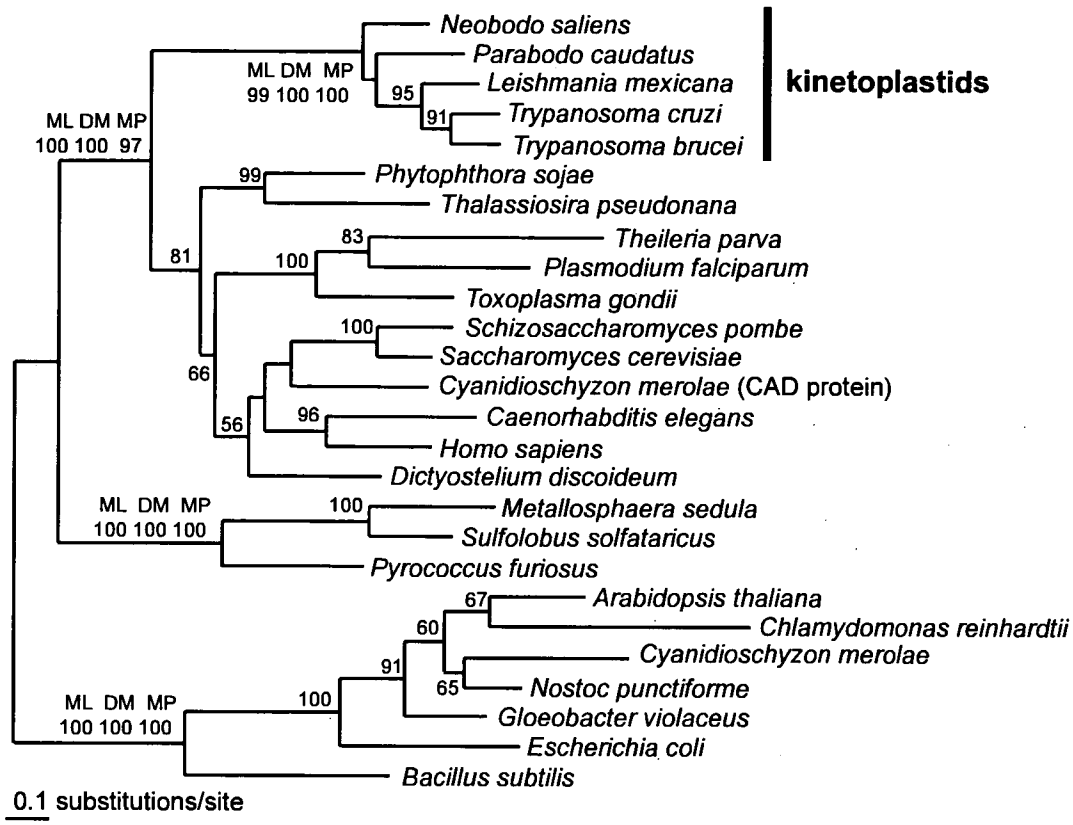
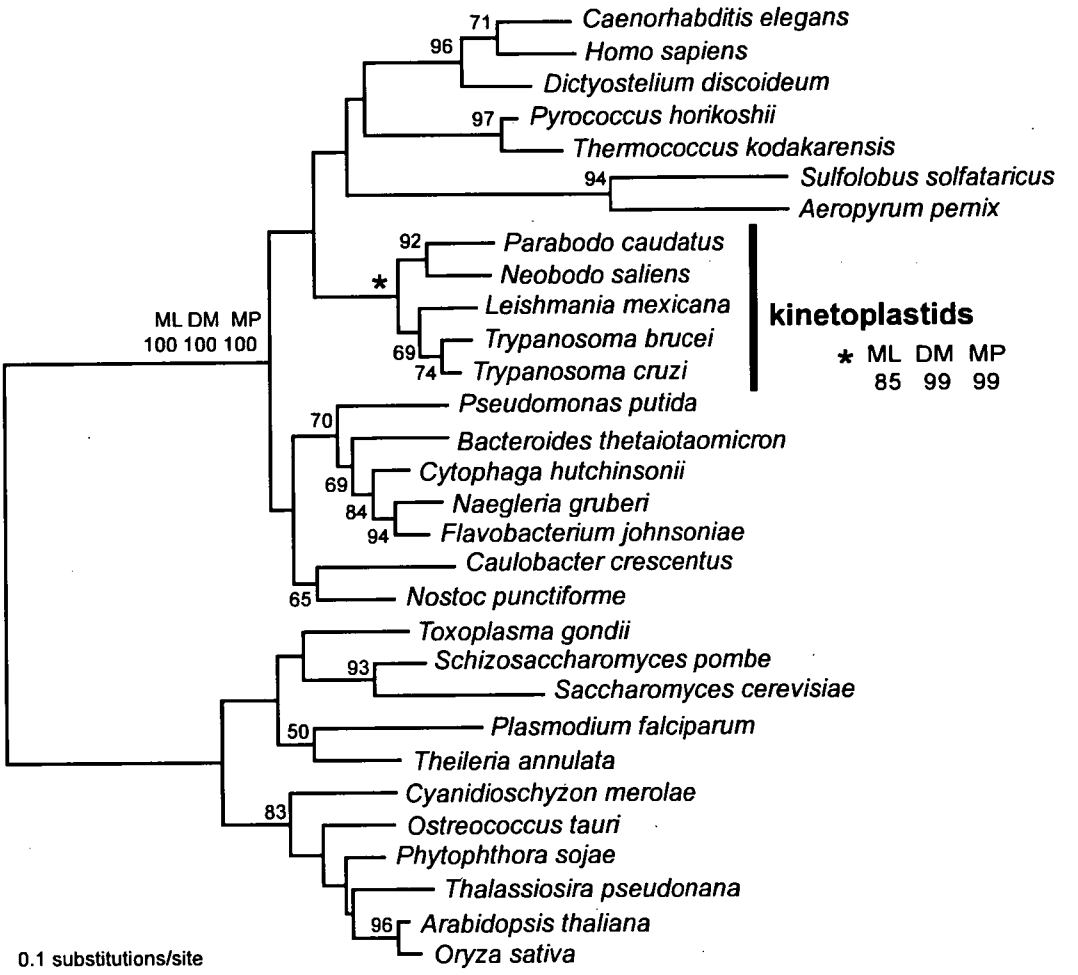
Makiuchi *et al.* Supplementary Fig. S2



Makiuchi *et al.* Supplementary Fig. S3



CHAPTER 2

Structure and Content of the *Entamoeba histolytica* Genome

C. G. Clark,[*] U. C. M. Alsmark,[†] M. Tazreiter,[‡] Y. Saito-Nakano,[§] V. Ali,[¶] S. Marion,[‖,1] C. Weber,[‖] C. Mukherjee,[#] I. Bruchhaus,[**] E. Tannich,[**] M. Leippe,[††] T. Sicheritz-Ponten,[‡‡] P. G. Foster,[§§] J. Samuelson,[¶¶] C. J. Noël,[†] R. P. Hirt,[†] T. M. Embley,[†] C. A. Gilchrist,[|||] B. J. Mann,[|||] U. Singh,[##] J. P. Ackers,[*] S. Bhattacharya,[a] A. Bhattacharya,[b] A. Lohia,[#] N. Guillén,[‖] M. Duchêne,[‡] T. Nozaki,[¶] and N. Hall[c,2]

[*] Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
[†] Division of Biology, Newcastle University, Newcastle NE1 7RU, UK
[‡] Department of Specific Prophylaxis and Tropical Medicine, Center for Physiology and Pathophysiology, Medical University of Vienna, A-1090 Vienna, Austria
[§] Department of Parasitology, National Institute of Infectious Diseases, Tokyo, Japan
[¶] Department of Parasitology, Gunma University Graduate School of Medicine, Maebashi, Japan
[‖] Institut Pasteur, Unité Biologie Cellulaire du Parasitisme and INSERM U786, F-75015 Paris, France
[#] Department of Biochemistry, Bose Institute, Kolkata 700054, India
[**] Bernhard Nocht Institute for Tropical Medicine, D-20359 Hamburg, Germany
[††] Zoologisches Institut der Universität Kiel, D-24098 Kiel, Germany
[‡‡] Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, DK-2800 Lyngby, Denmark
[§§] Department of Zoology, Natural History Museum, London, SW7 5BD, UK
[¶¶] Department of Molecular and Cell Biology, Boston University Goldman School of Dental Medicine, Boston, Massachusetts 02118
[|||] Department of Medicine, Division of Infectious Diseases, University of Virginia Health Sciences Center, Charlottesville, Virginia 22908
[##] Departments of Internal Medicine, Microbiology, and Immunology, Stanford University School of Medicine, Stanford, California 94305
[a] School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110067, India
[b] School of Life Sciences and Information Technology, Jawaharlal Nehru University, New Delhi 110067, India
[c] The Institute for Genomic Research, Rockville, Maryland 20850
[1] Present address: Cell Biology and Biophysics Program, European Molecular Biology Laboratory, 69117 Heidelberg, Germany
[2] Present address: School of Biological Sciences, University of Liverpool, Liverpool L69 7ZB, United Kingdom

Advances in Parasitology, Volume 65
ISSN 0065-308X, DOI: 10.1016/S0065-308X(07)65002-7

Contents

Abstract The intestinal parasite *Entamoeba histolytica* is one of the first protists for which a draft genome sequence has been published. Although the genome is still incomplete, it is unlikely that many genes are missing from the list of those already identified. In this chapter we summarise the features of the genome as they are currently understood and provide previously unpublished analyses of many of the genes.

1. INTRODUCTION

Entamoeba histolytica is one of the most widespread and clinically important parasites, causing both serious intestinal (amoebic colitis) and extra-intestinal (amoebic liver abscess) diseases throughout the world. A recent World Health Organization estimate (WHO, 1998) places *E. histolytica* second after *Plasmodium falciparum* as causing the most deaths annually (70,000) among protistan parasites.

Recently a draft of the complete genome of *E. histolytica* was published (Loftus *et al.*, 2005) making it one of the first protist genomes to be sequenced. The *E. histolytica* genome project was initiated in 2000 with funding from the Wellcome Trust and the National Institute of Allergy and Infectious Diseases to the Wellcome Trust Sanger Institute and The Institute for Genomic Research (TIGR) in the UK and the USA, respectively. The publication describing the draft sequence concentrated on the expanded gene families, metabolism and the role of horizontal gene transfer in the evolution of *E. histolytica*. In this chapter we summarise

the structure and content of the *E. histolytica* genome in comparison to other sequenced parasitic eukaryotes, provide a description of the current assembly and annotation, place the inferred gene content in the context of what is known about the biology of the organism and discuss plans for completing the *E. histolytica* genome project and extending genome sequencing to other species of *Entamoeba*.

The fact that the genome sequence is still a draft has several important consequences. The first is that a few genes may be missing from the sequence data we have at present, although the number is likely to be small. For example, at least one gene (amoebapore B) is not present in the genome data despite it having been cloned, sequenced and the protein extensively characterised well before the start of the genome project. The second consequence is that the assembly contains a number of large duplicated regions that may be assembly artefacts, meaning that the number of gene copies is overestimated in several cases. These problems cannot as yet be resolved but should be eventually as more data becomes available. Nevertheless, it is important to remember these issues when reading the rest of this chapter.

As the number of genes in *E. histolytica* runs into several thousands, it is not possible to discuss all of them. However, we have generated a number of tables that identify many genes and link them to their entries in GenBank using the relevant protein identifier. Only a few tables are included in the text of this chapter, but the others are available online as supplementary material, http://pathema.tigr.org/pathema/entamoeba_resources.shtml. The *E. histolytica* genome project data are being 'curated' at the J. Craig Venter Institute (JCVI, formerly TIGR), and it is on that site that the most current version of the assembled genome will be found. The 'Pathema' database will hold the data and the annotation (http://pathema.tigr.org/). The gene tables are also linked to the appropriate entry in the Pathema database, and the links will be maintained as the genome structure is refined over time.

Reference is made throughout the text to other species of *Entamoeba* where data are available. *Entamoeba dispar* is the sister species to *E. histolytica* and infects humans without causing symptoms. *Entamoeba invadens* is a reptilian parasite that causes invasive disease, primarily in snakes and lizards, and is widely used as a model for *E. histolytica* in the study of encystation, although the two species are not very closely related (Clark *et al.*, 2006b). Genome projects for both these species are under way at TIGR, and it is anticipated that 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: *Entamoeba moshkovskii*, which is primarily a free-living species although it occasionally infects humans,

and *Entamoeba 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 centre 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 ~12× 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 ~23.7 million basepairs (Mbp) in size (Table 2.1). 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 *P. 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 Mbp and possibly a ploidy of 4 (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

TABLE 2.1 Genome summary statistics for selected single celled organisms with sequenced genomes

Statistic[a]	*Entamoeba histolytica*	*Plasmodium falciparum*	*Dictyostelium discoideum*	*Saccharomyces cerevisiae*	*Encephalitozoon cuniculi*
Genome Size (Mbp)	23.7	22.8	33.8	12.5	2.5
G + C content (%)	24.1	19.4	22.5	38	45.5
Gene number	9938	5268	12,500	5538	1997
Av. gene size (bp)	1167	2534	1756	1428	1077
% coding DNA	49.2	52.6	ND	70.5	ND
Av. protein size (aa)	389	761	518	475	359
Av. intergenic dist. (kb)	0.8	1.7	0.8	0.6	0.1
Gene density (kb per gene)	1.9	4.3	2.5	2.2 kb	1.1
% Genes with introns	25.2	54	69	5	<1
Av. intron size (bp)	102.1	179	146	ND	–
Av. number of introns/gene	1.5	2.6	1.9	1	1

[a] Abbreviations: Mbp: million basepairs; kb: kilobasepairs; bp: basepairs; aa: amino acids, ND: not determined.

that the chromosome ends may contain arrays of transfer RNA (tRNA) genes (see Section 2.4).

2.3. Ribosomal RNA genes

The organisation of the structural RNA genes in *E. histolytica* is unusual with the ribosomal RNA (rRNA) genes carried exclusively on 24 kilobasepair (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 encode 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 gene 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 polymerase chain reaction (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 of 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 2.2a). 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

TABLE 2.2 Summary properties of the repeated DNAs

(a) References for data[a]

Repeat type	Size in kb	Estimated copy no. from genome sequence (Ref no.)	Estimated copy no. per haploid genome from hybridisation (Ref no.)	Transcript size in kb (Ref no.)
EhLINE1	4.8	142 (1) 409; 49 full-length (2)	140 (3)	No full-length transcript (4)
EhLINE2	4.72	79 (1) 290; 56 full-length (2)	Not Determined	Not Determined
EhLINE3	4.81	12 (1) 52; 3,full-length (2)	Not Determined	Not Determined
EhSINE1	0.5–0.6	219 (1) 272; 81 full-length (2) 214; >90 full-length (3)	500	0.7 (6)
EhSINE2	0.65	120 (1) 117; 62 full-length (2) 122; ~50 full-length (3)	Not Determined	0.75 (7)
EhSINE3	0.58	1 (1,2)	Not Determined	Not detected (3)
Tr	0.7	1 per rDNA episome (5)	Not Determined	0.7 (5)
BspA-like	0.96	77 (8)	190 (3)	Not detected (3)[b]
Ehssp1	0.9–1.1	Not Determined	306 (9)	1.5 (9)

(b) Consensus sequences of Family 16 and 17 repeats[c]

Family	Sequence
Family 16	GTAATGAATATAYAACTAAGAATTTCATT TAAAATGRATATG
Family 17	CAACAAATAAATRGKTTCAATAAAATA

[a] (1): Van Dellen *et al.* (2002a); (2): Bakre *et al.* (2005); (3): This analysis; (4): A. A. Bakre and S. Bhattacharya (unpublished data); (5): Burch *et al.* (1991); (6): Cruz-Reyes *et al.* (1995); (7): Shire and Ackers (2007); (8): Davis *et al.* (2006); (9): Satish *et al.* (2003).

[b] Although no transcript was detected, the protein has been demonstrated on the cell surface and in Western blots using antibodies (Davis *et al.*, 2006).

[c] Standard abbreviation for degenerate sequence positions are used: R = purine, Y = pyrimidine, K = G or T.

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 reverse transcriptase (RT) and restriction enzyme-like endonuclease (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 5-bp duplication at nucleotide position 1442 in about 80% of the copies of EhLINE1 creates a stop codon, dividing the single ORF into 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 RT and 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 interspersed elements (Cruz-Reyes and Ackers, 1992; Cruz-Reyes *et al.*, 1995) or *ehapt2* (Willhoeft *et al.*, 2002)] have now been designated EhSINE1 (Van Dellen *et al.*, 2002a; Willhoeft *et al.*, 2002). BLAST searching of databases 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 (J. P. Ackers, unpublished data). The majority contain 2 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 4 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, 2007) 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 that either it is encoded in regions missing from the current assembly or 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 mainly found 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 ~300 copies per haploid genome as determined by hybridisation (Table 2.2a). 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 two to three 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 ~38 times in the genome while family 17 has a 27 base consensus sequence and occurs 35 time in the genome (Table 2.2b). 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 *D. discoideum* (~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. Both gene loss and gain can 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 2 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 (see 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 (Fig. 2.1). This protein is highly conserved but the number of introns in each gene varies. *P. falciparum* has five 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

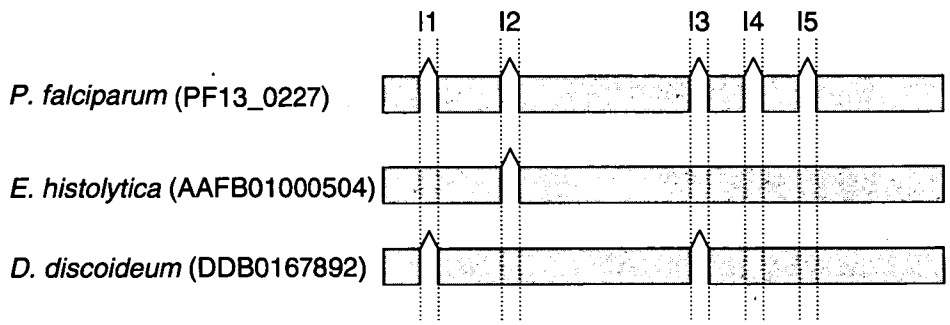


FIGURE 2.1 Positions of introns in the vacuolar ATPase subunit D gene in *P. falciparum, D. discoideum* and *E. histolytica.*

P. falciparum respectively. In fact the protein length distribution is most similar to that of the microsporidian *Encephalitozoon cuniculi* (Fig. 2.2) which has a very compact genome of 3 Mbp and <2000 genes. Direct

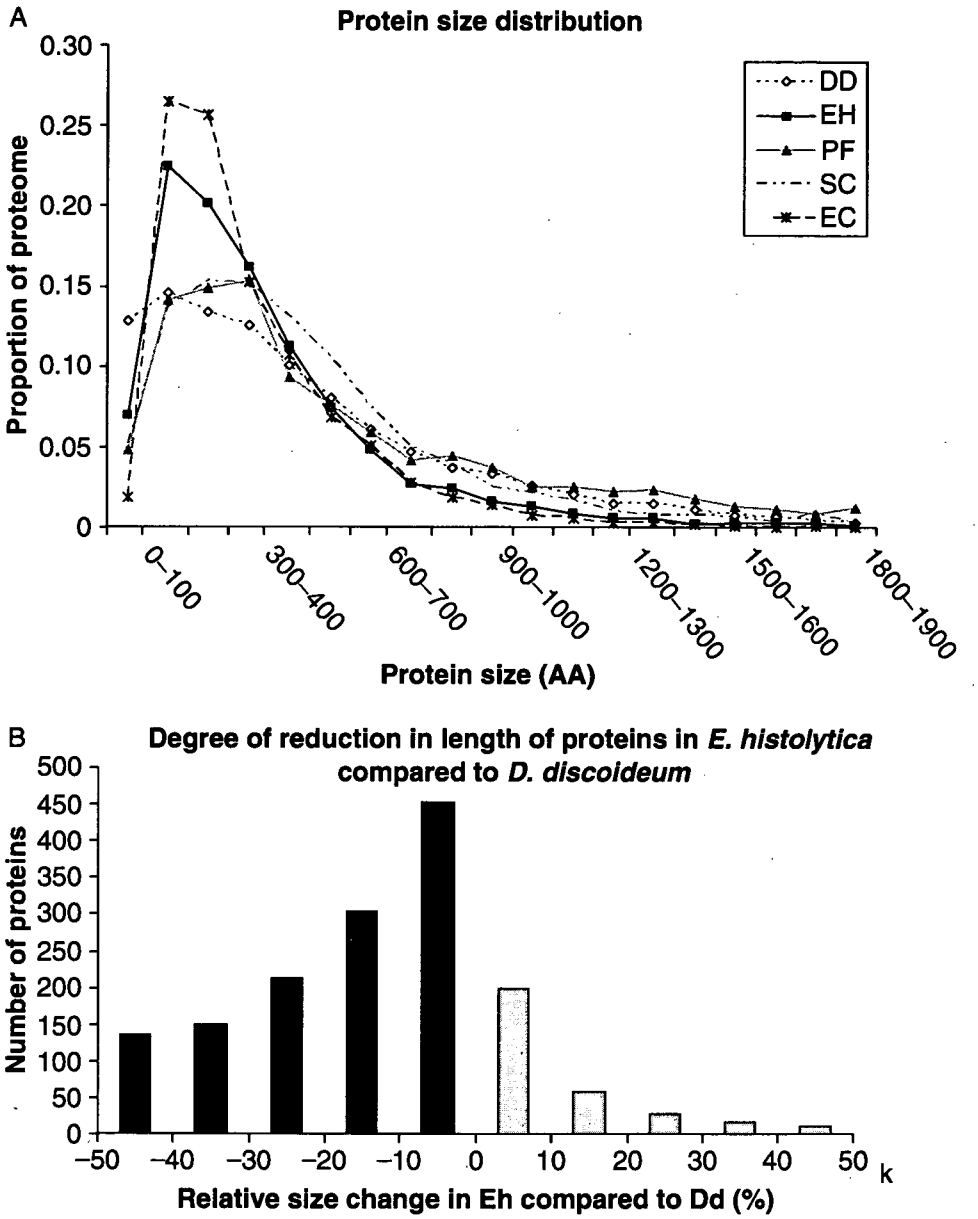


FIGURE 2.2 Comparison of protein sizes in *E. histolytica* and *D. discoideum.* (A) The graph shows the distribution of predicted amino acid length across sequenced genomes from single celled eukaryotes: *D. discoideum* (DD) *Encephalitozoon cuniculi* (EC), *P. falciparum* (PF), *E. histolytica* (EH), and *S. cerevisiae* (SC). *Entamoeba histolytica* and *E. cuniculi* have a distribution that is skewed towards smaller proteins relative to the other species. (B) The histogram displays the degree of size change of genes in *E. histolytica* relative to *D. discoideum* when comparing orthologous genes identified by reciprocal best BLAST hits. The black bars show genes that are smaller in *E. histolytica* where as the gray bars are smaller in *D. discoideum.*

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 (N. Hall, unpublished data). 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 signalling (Zhang, 2000). This would not seem to be the case here as the number of genes identified as having a role in signalling 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 2.3. 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 signalling 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 Src-homology 3 (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 (Rho GTPase guanine nucleotide exchange factor) and Arf-GAP (ADP ribosylation factor GTPase activating protein) domains, suggesting a mechanism for direct communication between the regulators of vesicle budding and cytoskeletal rearrangement. Over 80 receptor kinases were identified (see 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).

TABLE 2.3 Number and ranking of Pfam domains across different genomes

		EH		EC		PF		SC		AT		CE		DD	
Domain name	Domain detail	#	Rank	#	Rank	#	Rank	#	Rank	#	Rank	#	Rank	#	Rank
WD40	WD domain, G-β repeat	249	1	139	1	287	2	414	1	1137	3	694	1	719	2
LRR_1	Leucine-rich repeat	131	2	40	2	55	12	43	17	3793	2	494	5	372	4
Pkinase	Protein kinase domain	95	3	27	5	78	8	116	2	839	4	405	8	225	7
HEAT	HEAT repeat	70	4	13	15	44	17	114	3	220	17	162	26	108	12
efhand	EF hand	58	5	7	28	80	7	29	25	422	8	213	20	153	9
RRM_1	RNA recognition motif	57	6	30	3	95	6	86	6	375	10	223	19	134	10
Ras	Ras family	46	7	9	22	13	44	25	28	78	68	66	76	126	11
TPR_1	Tetratricopeptide repeat	42	8	23	7	48	15	103	4	334	12	180	22	168	8
Ank	Ankyrin repeat	34	9	6	34	55	12	61	9	431	6	629	2	446	3
PUF	Pumilio-family RNA binding repeat	33	10	8	23	15	34	51	13	142	32	75	68	34	62
RhoGAP	RhoGAP domain	27	11	2	118	1	520	11	80	9	559	31	138	45	39
Myb_DNA-binding	Myb-like DNA-binding domain	22	12	15	12	10	62	21	34	424	7	30	141	55	26
RhoGEF	RhoGEF domain	22	12	1	230	0	1215	3	366	0	2581	34	130	47	37

(continued)

TABLE 2.3 *(continued)*

		EH		EC		PF		SC		AT		CE		DD	
Domain name	Domain detail	#	Rank	#	Rank	#	Rank	#	Rank	#	Rank	#	Rank	#	Rank
Helicase_C	Helicase conserved C-terminal domain	20	14	28	4	64	11	74	8	150	31	98	49	84	20
DEAD	DEAD/DEAH box helicase	20	14	22	9	49	14	59	10	103	50	76	67	48	35
PH	PH domain	19	16	1	230	5	123	25	28	22	255	77	63	94	16
Metallophos	Calcineurin-like phosphoesterase	19	16	6	34	16	32	21	34	66	83	78	62	31	67
Gelsolin	Gelsolin repeat	18	18	2	118	2	295	4	255	33	169	12	323	29	68
LIM	LIM domain	17	19	0	703	0	1,215	8	116	16	341	103	47	56	25
CH	Calponin homology (CH) domain	16	20	4	54	1	520	7	137	26	211	57	87	49	33
Filamin	Filamin/ABP280 repeat	16	20	0	703	1	520	0	1842	2	1450	55	91	10	203

Note: Columns labeled '#' give the total number of occurrences of a particular domain. Columns labeled 'Rank' give the ranking of the domain where the most common domain is ranked 1. The organisms shown are *E. histolytica* (EH), *Encephalitozoon cuniculi* (EC), *Plasmodium falciparum* (PF), *Arabidopsis thaliana* (AT), *Saccharomyces cerevisae* (SC), *Dictyostelium discoideum* (DD).