2.12. Translation-related proteins

Two of the predicted tRNAs (Ile^{TAT} and Tyr) need to be spliced because of 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 synthesising 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 β subunit of protein kinase CKII and to stimulate phosphorylation of DNA topoisomerase II α by CKII (Lee *et al.*, 1997b). If this gene is truly absent from *E. histolytica*, then 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).

In eukaryotic translation, elongation factor (EF)-1 is activated upon GTP binding and forms a ternary complex with aminoacyl tRNAs and ribosomes. EF-1 β and δ subunits work as GDP-GTP exchange factors to cycle EF-1 α between two forms while EF-1 γ 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 δ, a protein involved in exchanging GDP with GTP. This is also absent from *S. cerevisiae* and *P. falciparum*. It is likely that EF-1 β 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 that ensure progeny cells receive an exact copy of the parental genetic material (Hartwell and Weinert, 1989). Unlike most eukaryotes, *E. 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) in axenic cultures 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 yeasts 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), cell division cycle 6 (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 (minichromosome maintenance) 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 may function 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 cyclin-dependent kinase (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 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 yeasts. 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 and 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 (Wostmann *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 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 nine 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 eight 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 2× 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 α-amanitin-resistant (Lioutas and Tannich, 1995). The F homology block of the RNA polymerase II largest subunit (RPB1) has been identified as the putative α-amanitin binding site. This block is highly divergent in the α-amanitin resistant *T. 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 yeasts, 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 of subunit nine 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 transcription factor 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.*, 1997, 2002). 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 the 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 6 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. (A detailed analysis of *E. histolytica* transcription factors can be found at http://www.transcriptionfactor.org).

TAFs 5, 6, 10 and 12 are also components of the histone acetyltransferase (HAT) complexes in other organisms as are SPT6 and SPT16 (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 understood, histone acetylation complexes are known to be active in *E. histolytica* (Ramakrishnan *et al.*, 2004). Other potential members of chromatin remodelling complexes of *E. histolytica* include the TBP interacting helicase (RVB1 and 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* spliceosome components U1, U2, U4/6, U5 and the Prp19 complex have all been identified. In fact homologues of 10 of the 14 'core' small nuclear ribonucleoproteins (snRNPs), 2 of the U1-specific snRNPs, 7 of the 10 U2-specific snRNPs, 5 of the 6 U5-specific snRNPs, 3 of the U4/6 specific snRNPs and 4 of the 9 subunits of the Prp19 complex have been found. Indeed *E. histolytica* has homologues of ~80% of the *S. cerevisiae* splicing machinery (Jurica and Moore, 2003).

Like *Giardia 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; Vanacova *et al.*, 2003). Identification of homologues of the Ceg1 RNA guanylyltransferase—an enzyme that 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.

mRNAs in *E. histolytica* are polyadenylated, and the polyadenylation signal is found within the short 3′ untranslated region (Bruchhaus *et al.*, 1993; Li *et al.*, 2001). However, only 8 of the 18 yeast cleavage and polyadenylation specificity factor (CPSF) subunits are identifiable in *E. histolytica*.

3. VIRULENCE FACTORS

3.1. Gal/GalNAc lectin

One of the hallmarks of *E. histolytica* pathogenicity is contact-dependent killing of host cells. *E. histolytica* is capable of killing a variety of cell types, including human intestinal epithelium, erythrocytes, neutrophils and lymphocytes (Burchard and Bilke, 1992; Burchard *et al.*, 1992a,b; Guerrant *et al.*, 1981; Ravdin and Guerrant, 1981). Cytolysis occurs as a stepwise process that begins with adherence to target cells via galactose/*N*-acetyl D-galactosamine-inhibitable (Gal/GalNAc) lectin (Petri *et al.*, 1987; Ravdin and Guerrant, 1982). Adherence via the Gal/GalNAc lectin is a requirement for cell killing because in the presence of galactose or GalNAc target cells are not killed by the amoebae. Target cell death occurs within 5–15 min and is often followed by phagocytosis. Inhibition of the Gal/GalNAc lectin with galactose or specific antibody also blocks phagocytosis (Bailey *et al.*, 1990). Resistance to lysis by the complement system is also mediated in part by the Gal/GalNAc lectin. The lectin contains a CD59-like domain that likely helps protect the trophozoites from complement; CD59 is a surface antigen of many blood cells known to have this property (Braga *et al.*, 1992).

The Gal/GalNAc lectin is a membrane complex that includes heavy (Hgl) 170 kilodalton (kDa), and light (Lgl) 30–35 kDa subunits linked by disulphide bonds, and a non-covalently associated intermediate (Igl) 150 kDa subunit (Cheng *et al.*, 2001; Petri *et al.*, 1989). The structure and function of the Gal/GalNAc lectin has recently been reviewed (Petri *et al.*, 2002). The heavy subunit is a type1 transmembrane protein while the light and intermediate subunits have glycosylphosphatidylinositol (GPI) anchors (Cheng *et al.*, 2001; McCoy *et al.*, 1993). Gal/GalNAc lectin subunits do not share any significant protein identity or similarity to any other known proteins, though Hgl and Igl have some very limited regions of similarity with known classes of proteins that will be discussed below.

3.1.1. The heavy (Hgl) subunit

On the basis of pulse-field gel electrophoresis there are five loci in the genome with similarity to the Hgl subunit. However, the current genome assembly only identifies two complete genes, one of which corresponds to Hgl2 (Tannich *et al.*, 1991b). The predicted proteins encoded by these loci are 92% identical. In initial assemblies there were three other sequences with high similarity to the Hgl subunit that were pseudogenes. These pseudogenes may account for the additional loci detected by pulse-field gel electrophoresis. The large size of these genes means that assembly problems may also be affecting our interpretation.

Hgl subunit sequences can be divided into domains based on amino acid content and distribution (Fig. 2.3). The amino-terminal domain of ~200 amino acids consists of 3.2% cysteine and 2.1% tryptophan residues. The next domain, also ~200 amino acids, is completely devoid of these 2 amino acids. The C-terminal domain of ~930 amino acids is cysteine-rich, comprising 10.8% cysteine. The number and spacing of all predicted tryptophan and cysteine residues are 100% conserved in the 2 complete genes. Although a portion of the C-terminal domain can be said to contain cysteine-rich pseudo-repeats, there is no clear repetitive structure to the protein (Tannich *et al.*, 1991b). The Hgl subunit has a single transmembrane domain and a highly conserved 41 amino acid cytoplasmic domain. In addition to these two *hgl* genes, the genome contains a newly identified divergent member of the Hgl gene family (XP_650534). This ORF shares 43% similarity with the 2 other Hgl isoforms, and is predicted to encode a protein with an almost identical domain structure to that of Hgl described earlier.

3.1.2. The light (Lgl) subunit

The Lgl subunit is encoded by 5 genes (*lgl1–5*) that share 74–85% amino acid identity. A sequence corresponding to Lgl2 is missing from the current genome assembly. The light subunits range from 270 to 294 amino acids in length. Each isoform has a 12 amino acid signal peptide, 5 conserved cysteine residues and a GPI-anchor addition site. Lgl1 has two potential glycosylation sites. Lgl2 has one of these sites, Lgl3 has one different site, and Lgl4 and Lgl5 have none.

3.1.3. The intermediate (Igl) subunit

The Igl subunit was first identified by a monoclonal antibody that blocked amoebic adherence to and cytotoxicity for mammalian cells (Cheng *et al.*, 1998). Co-purification of the Hgl, Lgl and Igl suggests that these three

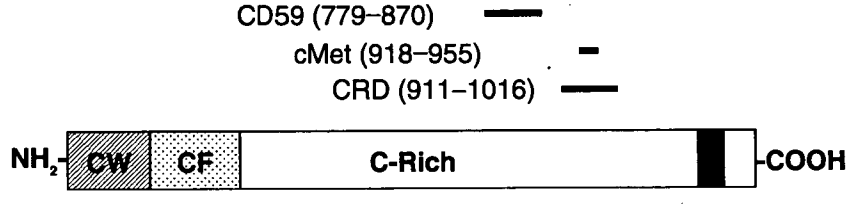

FIGURE 2.3 Domain diagram of the Hgl subunit of the Gal/GalNAc lectin. CW-Cysteine-Tryptophan region; CF-Cysteine free region; C-Rich- Cysteine rich region. The black vertical box near the carboxy terminus of the protein represents the single transmembrane domain. The horizontal black bars above the diagram indicate the location of a carbohydrate recognition domain (CRD), the region with similarity to the hepatic growth factor receptor, c-Met, and the region that has similarity to the CD59, the membrane inhibitor of the complement membrane attack complex. The numbers in parentheses indicate the location of these regions in the Hgl1 isoform (Mann *et al.*, 1991), where the methionine of the immature protein is residue 1.

subunits form a complex (Cheng *et al.*, 1998, 2001). The Igl subunit also has galactose-binding activity (Cheng *et al.*, 1998) and can serve as protective antigen in vaccine trials (Cheng and Tachibana, 2001). There are two loci that encode Igl subunits (Cheng *et al.*, 2001) and the predicted amino acid sequences are 81% identical. The Igl subunit, like the Hgl subunit, does not have any recognisable carbohydrate-binding domain.

3.1.4. Conservation of Gal/GalNAc lectin subunits in other species of *Entamoeba*

There are clearly identifiable orthologues of the Hgl and Lgl subunits among the limited sequences of *E. dispar*, *E. invadens*, *E. moshkovskii* and *E. terrapinae* available at present (Dodson *et al.*, 1997; Pillai *et al.*, 1997; Wang *et al.*, 2003). Because these genomes are incomplete it is possible that as yet unidentified family members will show greater similarity to the *E. histolytica* sequences. Nevertheless, the Lgl subunit is quite conserved among the five *Entamoeba* species. For instance, the *E. terrapinae* gene is 56% identical and 62% similar to *E. histolytica* Lgl1 over a span of 201 amino acids. The Hgl subunits are more diverse. The *E. dispar* Hgl orthologue is highly similar to the *E. histolytica* subunit (86%) but the other species show more diversity, including the region that corresponds to the carbohydrate recognition domain (CRD). However, the number and positions of the cysteine residues are highly conserved, as is the sequence of the cytoplasmic domain, showing only a few changes. It is difficult to put precise numbers to these similarities because the complete sequences of Hgl subunits from the other species are not present in the database. The character of the conservation of the Hgl subunits suggests that the ligand specificity is different for the Hgl subunits of each species but the signalling functions of the cytoplasmic domains are similar, if not perhaps identical. Only *E. dispar* has an identifiable Igl subunit. The other three species clearly have paralogues of the CXXC repeat family to which Igl belongs, but their similarity to Igl is mostly restricted to the CXXC and CXC repeat motifs.

3.2. Cysteine endopeptidases

E. histolytica is characterised by its extraordinary capacity to invade and destroy human tissues. The main lytic activity has been attributed to cysteine endopeptidases. This class of enzymes, which is found in all organisms, plays a major role in the pathogenicity of *E. histolytica* as demonstrated in a large number of *in vitro* and *in vivo* studies (Ankri *et al.*, 1999; Gadasi and Kessler, 1983; Keene *et al.*, 1990; Li et *al.*, 1995; Luaces and Barrett, 1988; Lushbaugh *et al.*, 1985; Reed *et al.*, 1989; Schulte and Scholze, 1989; Stanley *et al.*, 1995). Most striking are results from laboratory animal infections showing that *E. histolytica* trophozoites

with reduced cysteine protease activity are greatly impaired in their ability to induce amoebic disease (Ankri *et al.*, 1999; Stanley *et al.*, 1995). In addition, the discovery that *E. histolytica* cysteine proteases possess interleukin-1β convertase activity suggests that these enzymes use a mechanism that is novel in microbial pathogenicity (Zhang *et al.*, 2000).

Thiol-dependent proteolytic activity in *E. histolytica* was first attributed to a neutral sulphydryl protease (McLaughlin and Faubert, 1977) and later to a cytotoxic protease (Lushbaugh *et al.*, 1984). Other terms that have been used to describe closely related or identical enzymes are cathepsin B (Lushbaugh *et al.*, 1985), neutral proteinase (Keene *et al.*, 1990), histolysin (Luaces and Barrett, 1988) (later changed to histolysain; Luaces *et al.*, 1992) and amoebapain (Scholze *et al.*, 1992). *E. histolytica* cysteine endopeptidases were found to be secreted (Leippe *et al.*, 1995) and localised in lysosome-like vesicles or at the surface of the cell (Garcia-Rivera *et al.*, 1999; Jacobs *et al.*, 1998). Molecular cloning has revealed a large number of cysteine endopeptidase genes in the *E. histolytica* genome (Bruchhaus *et al.*, 2003; Garcia-Rivera *et al.*, 1999; Reed *et al.*, 1993; Tannich *et al.*, 1991c., 1992). Interestingly, most of these genes are not expressed during *in vitro* cultivation (Bruchhaus *et al.*, 2003). As our current knowledge of *E. histolytica* biology and pathogenicity is mostly based on analysis of cultured cells, the function of most of the cysteine endopeptidases and their precise role in *E. histolytica* virulence is largely unknown.

Homology searches using conserved active site regions revealed that the *E. histolytica* genome contains at least 44 genes coding for cysteine endopeptidases. Of these, the largest group is structurally related to the C1 papain superfamily (Table 2.4), whereas a few others are more similar to family C2 (calpain-like cysteine proteases), C19 (ubiquitinyl hydrolase), C54 (autophagin) and C65 (otubain), respectively (Table 2.5).

Phylogenetic analyses of the 36 C1-family members revealed that they represent three distinct clades (A, B and C), consisting of 12, 11 and 13 members, respectively. Clades A and B members correspond to the two previously described subfamilies of *E. histolytica* cysteine proteases, designated EhCP-A and EhCP-B (Bruchhaus *et al.*, 2003). In contrast, clade C represents a new group of *E. histolytica* cysteine endopeptidases that has not been described before. EhCP-A and EhCP-B-subfamily members are classical pre-pro enzymes with an overall cathepsin L-like structure (Barrett, 1998) as indicated by the presence of an ERFNIN motif in the pro region of at least 21 of the 23 EhCP-A and EhCP-B enzymes (Fig. 2.4). Interestingly, biochemical studies with purified EhCP-A indicated a cathepsin B-like substrate specificity (Scholze and Schulte, 1988). This is likely due to the substitution of an alanine residue by acidic or charged amino acids in the postulated S2 pocket, corresponding to residue 205 of the papain sequence (Barrett, 1998). As reported previously (Bruchhaus *et al.*, 2003), the EhCP-A and EhCP-B subfamilies differ in the length of the pro regions

TABLE 2.4 Family C1-like cysteine endopeptidases of *E. histolytica*

Protein name	Previous designation	Accession No.	Protein length (aa) Total (pre, pro, mature)	Active site residues	Conserved motifs	Remarks
EhCP-A1	EhCP1	XP_650156	315 (13,80,222)	QCHN	ERFNIN, DWR	
EhCP-A2	EhCP2	XP_650642	315 (13,80,222)	QCHN	ERFNIN, DWR	
EhCP-A3	EhCP3	XP_653254	308 (13,79,216)	QCHN	ERFNIN, DWR	
EhCP-A4	EhCP4	XP_656602	311 (20,73,218)	QCHN	ERFNIN, DWR	
EhCP-A5	EhCP5	XP_650937	318 (20,72,225)	QCHN	ERFNIN, DWR, RGD	Degenerate in *E. dispar*
EhCP-A6	EhCP6	XP_657364	320 (17,79,224)	QCHN	ERFNIN, DWR	
EhCP-A7	EhCP8	XP_648996	315 (13,80,222)	QCHN	ERFNIN, DWR	
EhCP-A8	EhCP9	XP_657446	317 (15,82,220)	QCHN	ERFNIN, DWR	
EhCP-A9	EhCP10	XP_655675	297 (17,90,190)	QCHN	ERFNIN, DWR	
EhCP-A10	EhCP17	XP_651147	420 (18,148,254)	QCHN	ERFNIN, DWR	
EhCP-A11	EhCP19	XP_651690	324 (17,79,228)	QC IN[a]	ERFNIN, DWR	
EhCP-A12	New	XP_653823	317 (14,83,220)	(d)	ERFNIN, DWR	

(*continued*)

TABLE 2.4 *(continued)*

Protein name	Previous designation	Accession No.	Protein length (aa) Total (pre, pro, mature)	Active site residues	Conserved motifs	Remarks
EhCP-B1	EhCP7	XP_651581	426 (15,106,305)	QCHN	ERFNIN, PCNC	Hydrophobic C-terminus
EhCP-B2	EhCP11	AAO03568	431 (15,106,310)	QCHS[a]	ERFNIN, PCNC	GPI cleavage site
EhCP-B3	EhCP12	XP_656747	474 (16,107,351)	QCHN	ERFNIN, PCNC	TMH:444–466 aa
EhCP-B4	EhCP13	XP_648501	379 (16,105,258)	QCHN	ERFNIN, PCNC	TMH or GPI cleavage site
EhCP-B5	EhCP14	XP_652671	434 (12,108,314)	QCHN	ERFNIN, PCNC	GPI cleavage site
EhCP-B6	EhCP15	XP_652465	300 (14,55,231)	QCHN	PCNC	Hydrophobic C-terminus
EhCP-B7	EhCP16	XP_650400	650 (18,144,488)	QCHN	ERFNIN, PCNC	Hydrophobic C-terminus, Cys-rich profile
EhCP-B8	EhCP18	XP_651049	473 (15,105,353)	QCHN	ERFNIN, PCNC, RGD	GPI cleavage site
EhCP-B9	EhCP112	XP_652993	446 (19,112,315)	QCHN	ERFNIN, PCNC, RGD	Hydrophobic C-terminus, Cys-rich profile
EhCP-B10	New	XP_648306	372 (b)	QCHN	ERFNIN, PCNC, RGD	Hydrophobic C-terminus
EhCP-B11	New	XP_648013	133 (b)	Q ? ? ?	PCNC	

EhCP-C1	New	XP_654453	586 (c)	QCIN[a]	$HS(X)_6ICP$	TMH:12–34
EhCP-C2	New	XP_656632	567 (c)	QCHN	$HS(X)_6ICP$	TMH:27–49
EhCP-C3	New	XP_655128	572 (c)	QCHN	$HS(X)_6LCP$	TMH:17–39
EhCP-C4	New	XP_655800	502 (c)	QCHN	$LT(X)_6LCP$	
EhCP-C5	New	XP_654800	557 (c)	QCHN	$IS(X)_6ICP$	TMH:20–42
EhCP-C6	New	XP_651553	557 (c)	QCHD[a]	$HS(X)_6LCA$	TMH:14–36
EhCP-C7	New	XP_657273	595 (c)	QCHN	$IS(X)_6LCP$	TMH:19–41
EhCP-C8	New	XP_655479	627 (c)	QCHN	$IS(X)_6ICP$	TMH:29–51
EhCP-C9	New	XP_655011	518 (c)	(d)	$HS(X)_6ICP$	TMH:12–34
EhCP-C10	New	XP_654829	530 (c)	QCHN	$IS(X)_6ICP$	TMH:15–37
EhCP-C11	New	XP_648083	526 (c)	(d)	$HS(X)_6ICP$	TMH:20–42
EhCP-C12	New	XP_650829	473 (c)	(d)	$MS(X)_6LCG$	TMH:26–48 & 449–471
EhCP-C13	New	XP_656556	564 (c)	QCHN	$VS(X)_6RCG$	TMH:21–43

a: active sites that lack the canonical motif QCHN; b: incomplete sequence; c: cleavage sites to be determined; d: not conserved; ???: incomplete active site sequence.

TABLE 2.5 Family C2-, C19-, C54- and C65-like cysteine endopeptidases of *E. histolytica*

Name	Homology	Family	ProteinID	Protein length	Active site
EhCALP1	Calpain-like	C2	XP_649922	591 aa	Not conserved
EhCALP2	Calpain-like	C2	XP_657312	473 aa	QCHN
EhUBHY	Ubiquitin Hydrolase-like	C19	XP_657356	444 aa	NDTN
EhAUTO1	Autophagin-like	C54	XP_651386	325 aa	YCHS
EhAUTO2	Autophagin-like	C54	XP_653798	364 aa	YCHD
EhAUTO3	Autophagin-like	C54	XP_652043	364 aa	YCHD
EhAUTO4	Autophagin-like	C54	XP_656724	348 aa	YCHD
EhOTU	Otubain-like	C65	XP_654013	259 aa	DCH

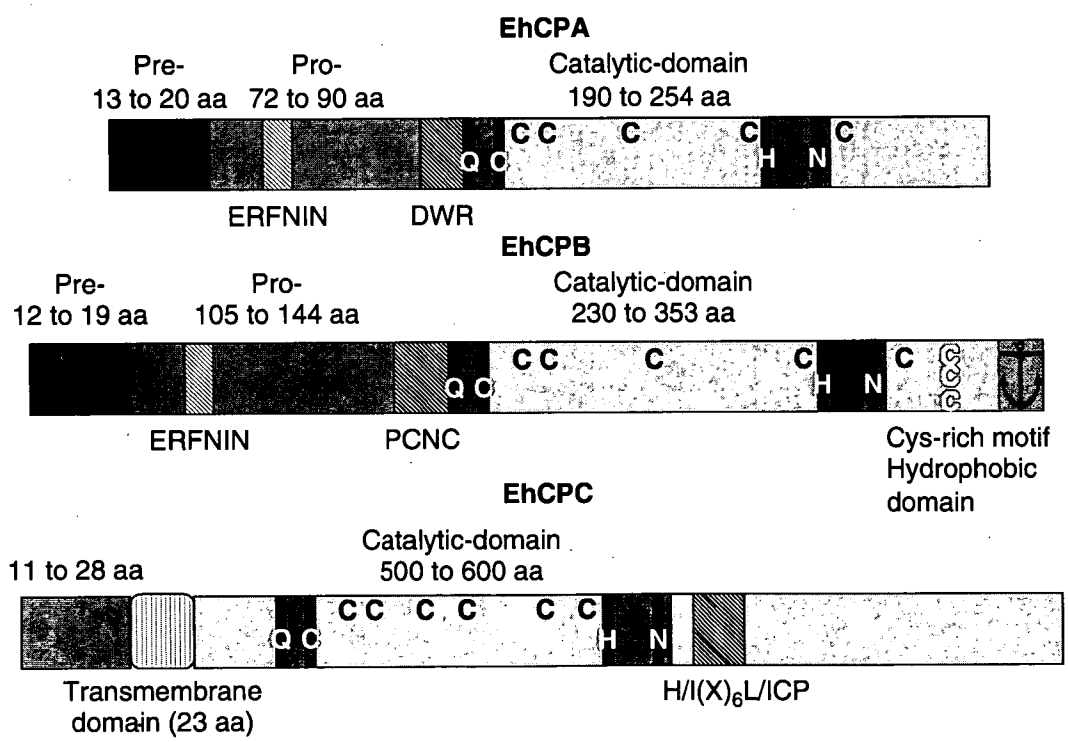

FIGURE 2.4 Structural domains of the three different types of family C1-like cysteine endopeptidases EhCP-A, EhCP-B and EhCP-C. Shown are the location and length of domains specific for each the three types as well as the conserved active site and cysteine residue.

as well as of the catalytic domains, and have distinct sequence motifs in the N-terminal regions of the mature enzymes (DWR vs. PCNC). Moreover, none of the EhCP-A subfamily but 10 of the 11 EhCP-B sequences contain hydrophobic stretches near or at the C-terminus, some of which are predicted to constitute transmembrane helices (TMH) or GPI-attachment moieties. This finding is consistent with previous reports on surface localisation of *E. histolytica* cysteine proteases but, so far, studies on the cellular localisation of the various EhCP-B molecules have not been reported.

In contrast to the EhCP-A and EhCP-B subfamilies, primary structure prediction indicates that EhCP-C members are not pre-pro enzymes, as they lack hydrophobic signal sequences as well as identifiable pro regions. Instead, they contain a hydrophobic region located 11–28 amino acids from the N-terminus, which is predicted to form a TMH (Fig. 2.4). Therefore, this new group of molecules appears to be membrane associated via a signal anchor. All EhCP-C enzymes have a conserved motif of the sequence H/I(X)$_6$L/ICP in the C-terminal half but they differ substantially in their pI, with values ranging from 4.6 to 8.8. As there is no example of a structurally related cysteine endopeptidase corresponding to the EhCP-C subfamily in other organisms, the specific functions of this group of molecules remain completely unknown.

In addition to the large number of C1 superfamily members, the *E. histolytica* genome contains two genes encoding cysteine endopeptidases homologous to family C2 or calpain-like cysteine proteases (EhCALP1 and EhCALP2). Enzymes of this class contain several calcium-binding domains and have been shown to participate in a variety of cellular processes, including remodelling of the cytoskeleton and membranes, signal transduction pathways and apoptosis.

Another four genes were identified coding for enzymes with homology to the peptidase family C54, also termed autophagins (EhAUTO1–4). The process of autophagy has been studied in human and yeast cells (Kirisako *et al.*, 2000; Marino *et al.*, 2003). Autophagy is a mechanism for the degradation of intracellular proteins and the removal of damaged organelles. During this process the cellular components become enclosed in double membranes and are subsequently degraded by lysosomal peptidases. Autophagins seem to be important for cytoplasm-to-vacuole targeting.

Two other genes encoding putative cysteine endopeptidases of *E. histolytica* show homology to the C19 and C65 families. These two groups of enzymes are known to be involved in ubiquitin degradation. Family C19 are ubiquitinyl hydrolases described as having ubiquitin-specific peptidase activity in humans. C65 or otubains are a group of enzymes with isopeptidase activity, which releases ubiquitin from polyubiquitin.

In summary, the *Entamoeba* genome contains a considerable number of endopeptidase genes. Elucidation of the precise role of each of the various enzymes will be a major challenge but may help us to understand the mechanism(s) of virulence and other unique properties of this protistan parasite.

3.3. Amoebapores and related proteins

In the lysosome-like granular vesicles of *E. histolytica* is found a family of small proteins, amoebapores, that are cytolytic towards human host cells, display potent antibacterial activity and cause ion channel formation in artificial membranes (for a review, see Leippe, 1997). Three amoebapore isoforms have been isolated and biochemically characterised, and their primary structure has been elucidated by molecular cloning of the genes encoding their precursors (Leippe *et al.*, 1991, 1992, 1994b). These membrane-permeabilising polypeptides are discharged by *E. histolytica* into bacteria-containing phagosomes to combat growth of engulfed microorganisms (Andrä *et al.*, 2003). Because of their potent cytolytic activity against human cells *in vitro* (Berninghausen and Leippe, 1997; Leippe *et al.*, 1994a), amoebapores have been viewed as a crucial element of the machinery use by the parasite to kill host cells. Trophozoites of *E. histolytica* lacking the major isoform amoebapore A, whether through antisense inhibition of translation (Bracha *et al.*, 1999) or epigenetic silencing of the gene (Bracha *et al.*, 2003), became avirulent demonstrating that this protein plays a key role in pathogenesis. Relatives of these protistan polypeptides are found in granules of porcine and human cytotoxic lymphocytes where they are termed NK-lysin and granulysin, respectively. All of these polypeptides are 70–80 amino acids in length and are characterised by a compact α-helical, disulphide-bonded structure known as the saposin-like fold. The structures of the amoebic and mammalian polypeptides have been solved and compared (Anderson *et al.*, 2003; Hecht *et al.*, 2004; Leippe *et al.*, 2005; Liepinsh *et al.*, 1997). The biological activities have also been measured in parallel (Bruhn *et al.*, 2003; Gutsmann *et al.*, 2003) to evaluate the similarities and differences of these effector molecules from organisms whose evolutionary paths diverged very early. As they are active against both prokaryotic and eukaryotic target cells, they may be viewed as broad-spectrum effector molecules.

In the genome of *E. histolytica*, 16 genes coding for putative saposin-like proteins (SAPLIPs) were identified. All of these genes are transcribed by cells growing in axenic culture (Winkelmann *et al.*, 2006). Like amoebapores, the predicted proteins all contain one C-terminal SAPLIP domain and (with one exception) a putative signal peptide (Table 2.6).

TABLE 2.6 Attributes of the identified SAPLIPs of *E. histolytica*

Name	Size, aa			SAPLIP domain		Identical to annotated protein	SAPLIP domain found[g]	Similar to (aa sequence identity, %)[f]	Homologous proteins (aa sequence identity, %)
	Entire	Signal peptide[a]	Proform/ Mature[c]	Position aa residues	Similarity Name, Acc. no.[e]				
Amoebapore A	98	21[b]	77	22–98	SAPOSIN B, IPR 008139	Amoebapore A precursor XP_653265		Amoebapore A AAA29111 (100%)	Disparpore A AAA18632 *E. dispar* (94%)
SAPLIP 1	92	15	77	16–92	SAPOSIN B, IPR 008139	Saposin-like protein XP_655836		Amoebapore A AAA29111 (64%)	Disparpore A AAA18632 *E. dispar* (68%)
Amoebapore B	96	19[b]	77	20–96	SAPOSIN B, IPR 008139	Amoebapore B precursor (EH-APP) Q24824		Amoebapore B CAA54226 (100%)	Disparpore B AAF04195 *E. dispar* (90%)
Amoebapore C	101	24[b]	77	25–101	SAPOSIN B, IPR 008139	Amoebapore C XP_656029		Amoebapore C CAA54225 (100%)	Disparpore C AAF04196 *E. dispar* (88%)
SAPLIP 2	153	15	138	71–153	SAPOSIN B, IPR 008139	Hypothetical protein XP_656037		—	—
SAPLIP 3	94	16	78	18–94	SAPOSIN B, IPR 008139	Hypothetical protein XP_656682		Amoebapore A AAA29111 (30%)	Invapore X AAP80381 *E. invadens* (67%)
SAPLIP 4	96	17	79	18–96	SAPOSIN B, IPR 008139	Hypothetical proteins XP_652159 and XP_652303		Amoebapore C CAA54225 (27%)	Disparpore C AAF04196 *E.dispar* (30%)
SAPLIP 5	1026	18	1008	946–1026	SAPOSIN B, IPR 008139	Chromosome partition protein XP_655789		—	—

(continued)

TABLE 2.6 *(continued)*

Name	Size, aa			SAPLIP domain		Identical to annotated protein	SAPLIP domain found[g]	Similar to (aa sequence identity, %)[f]	Homologous proteins (aa sequence identity, %)
	Entire	Signal peptide[a]	Proform/ Mature[c]	Position aa residues	Similarity Name, Acc. no.[e]				
SAPLIP 6	92	15	77	14–92	SAPOSIN B, IPR 008139	Hypothetical protein XP_655820		—	—
SAPLIP 7	926	17	909	855–926	SAPOSIN B, IPR 008139	Conserved hypothetical protein XP_656441		—	—
SAPLIP 8	980	15	965	902–980	SAPOSIN B, IPR 008139	Hypothetical protein XP_656913		—	—
SAPLIP 9	140	15	125	61–140	SAPOSIN B, IPR 008139	Hypothetical protein XP_650376		—	—
SAPLIP 10	657	16	641	577–657	SAPOSIN B, IPR 008139	—	Genomic survey sequence AZ687176	—	—
SAPLIP 11	693	17	676	615–693[d]	—	—	Genomic survey sequence AZ692153	—	—
SAPLIP 12	873	16	857	793–873	SAPOSIN B, IPR 008139	Hypothetical protein XP_652721		—	—

SAPLIP 13	1009	None predicted	1009	931–1005	SAPOSIN B, IPR 008139	Hypothetical protein XP_655089		—	—
SAPLIP 14	915	17	898	834–915	SAPOSIN B, IPR 008139	—	Genomic survey sequence AZ690015	—	—
SAPLIP 15	804	17	787	728–800	SAPOSIN B, IPR 008139	—	Genomic survey sequence BH132588	—	—
SAPLIP 16	921	15	906	842–921[d]	—	—	Genomic survey sequence AZ546519	—	—

Note: SAPLIPs were named according to the similarity of their SAPLIP domain to amoebapore A.
[a] By the programme SignalP and manually corrected if predicted cleavage site is within the SAPLIP domain.
[b] Verified by experimental data.
[c] With the exception of amoebapores it is not possible to decide whether proteins are further processed.
[d] Identified manually.
[e] Extracted from InterPro databases.
[f] If no similarity is reported, there is none outside the SAPLIP domain.
[g] Sequences only found in GSS section of GenBank with given identifier.

As a transmembrane domain is not apparent in these proteins, it may well be that they are secretory products stored in the cytoplasmic vesicles and act synergistically with the amoebapores. However, only 4 of them have a similar size to amoebapores, the others being considerably larger (up to 1009 residues). At present, it is not clear whether these larger gene products represent precursor molecules that are processed further. None of the novel SAPLIPs contain the conserved unique histidine residue at the C-terminus that is a key residue for the pore-forming activity of amoebapores (Andrä and Leippe, 1994; Hecht *et al.*, 2004; Leippe *et al.*, 2005). Indeed, it has recently been shown that recombinant SAPLIP3 has no pore-forming or bactericidal activity, although it does cause membrane fusion *in vitro* (Winkelmann *et al.*, 2006). This is in agreement with the experimental evidence for only three pore-forming entities being present in trophozoite extracts. Therefore, it is most likely that the three amoebapores are the sole pore-forming molecules of the parasite. However, the lipid-interacting activity present in all SAPLIP proteins (Munford *et al.*, 1995) and a function that helps to kill bacterial prey may well characterise all members of the amoebapore/SAPLIP superfamily of this voraciously phagocytic cell.

3.4. Antioxidants

E. histolytica trophozoites usually reside and multiply within the human gut, which constitutes an anaerobic or microaerophilic environment. However, during tissue invasion, the amoebae are exposed to an increased oxygen pressure and have to eliminate toxic metabolites such as reactive oxygen or nitrogen species (ROS/RNS) produced by activated phagocytes during the respiratory burst. *E. histolytica* lacks a conventional respiratory electron transport chain that terminates in the reduction of O_2 to H_2O. However, *E. histolytica* does respire and tolerates up to 5% oxygen in the gas phase (Band and Cirrito, 1979; Mehlotra, 1996; Weinbach and Diamond, 1974). Thus, *E. histolytica* trophozoites must use different antioxidant enzymes for the removal of ROS, RNS and oxygen (Fig. 2.5).

Among the enzymes in the first line of oxidative defence are superoxide dismutases (SODs), which are metalloproteins that use copper/zinc (Cu/ZnSOD), manganese (MnSOD) or iron (FeSOD) as metal cofactors. SODs catalyse the dismutation of superoxide radical anions to form H_2O_2 and O_2 (Fridovich, 1995). Analysis of the *E. histolytica* genome revealed only a single gene coding for a FeSOD and no sequences encoding MnSOD or Cu/ZnSOD. This reflects the situation found in most protistan parasites and is consistent with biochemical studies previously performed on *E. histolytica* lysates (Tannich *et al.*, 1991a).

E. histolytica lacks the tripeptide glutathione (Fahey *et al.*, 1984), which constitutes the major low molecular weight thiol found in almost all