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 eight of the eighteen 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 cells types including human intestinal epithelium, erythrocytes, neutrophils, and lymphocytes (Burchard and Bilke, 1992; Burchard *et al.*, 1992a; Burchard *et al.*, 1992b; Guerrant *et al.*, 1981; Ravdin and Guerrant, 1981). Cytolysis occurs as a step-wise 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 targets cells are not killed by the amoebae. Target cell death occurs within 5 to 15 minutes 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 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

1 glycosylphosphatidylinositol (GPI) anchors (Cheng *et al.*, 2001; McCoy *et al.*,
2 1993). Gal/GalNAc lectin subunits do not share any significant protein identity
3 or similarity to any other known proteins, though Hgl and Igl have some very
4 limited regions of similarity with known classes of proteins that will be
5 discussed below.
6
7 *3.1.1 The heavy (Hgl) subunit*
8 Based on pulse-field gel electrophoresis there are five loci in the genome with
9 similarity to the Hgl subunit. However, the current genome assembly only
10 identifies two complete genes, one of which corresponds to Hgl2 (Tannich *et*
11 *al.*, 1991b). The predicted proteins encoded by these loci are 92% identical. In
12 initial assemblies there were three other sequences with high similarity to the
13 Hgl subunit that were pseudogenes. These pseudogenes may account for the
14 additional loci detected by pulse-field gel electrophoresis. The large size of
15 these genes means that assembly problems may also be affecting our
16 interpretation.
17
18 Hgl subunit sequences can be divided in to domains based on amino acid
19 content and distribution (Figure 3). The amino-terminal domain of ca. 200
20 amino acids consist of 3.2% cysteine and 2.1% tryptophan residues. The next
21 domain, also ca. 200 amino acids, is completely devoid of these two amino
22 acids. The C-terminal domain of ca. 930 amino acids is cysteine-rich,
23 comprising 10.8 % cysteine. The number and spacing of all predicted
24 tryptophan and cysteine residues are 100% conserved in the two complete
25 genes. Although a portion of the C-terminal domain can be said to contain
26 cysteine-rich pseudo- repeats, there is no clear repetitive nature to the protein
27 (Tannich *et al.*, 1991b). The Hgl subunit has a single transmembrane domain
28 and a highly conserved 41 amino acid cytoplasmic domain. In addition to these
29 two *hgl* genes, the genome contains a newly identified divergent member of the
30 Hgl gene family (XP_650534). This ORF shares 43% similarity with the two

other Hgl isoforms, and is predicted to encode a protein with an almost identical domain structure to that of Hgl described above.

3.1.2 The light (Lgl) subunit

The Lgl subunit is encoded by five 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 subunits form a complex (Cheng *et al.*, 1998; Cheng *et al.*, 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 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 signaling 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

Entamoeba 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 proteinase 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 proteinases possess interleukin-1β convertase activity suggests that these enzymes use a mechanism that is novel in microbial pathogenicity (Zhang *et al.*, 2000).

1

2 Thiol-dependent proteolytic activity in *E. histolytica* was first attributed to a
3 neutral sulphydryl proteinase (McLaughlin and Faubert, 1977) and later to a
4 cytotoxic proteinase (Lushbaugh *et al.*, 1984). Other terms that have been used
5 to describe closely related or identical enzymes are cathepsin B (Lushbaugh *et*
6 *al.*, 1985), neutral proteinase (Keene *et al.*, 1990), histolysin (Luaces and
7 Barrett, 1988) (later changed to histolysain; Luaces *et al.*, 1992)), and
8 amoebapain (Scholze *et al.*, 1992). *E. histolytica* cysteine endopeptidases were
9 found to be secreted (Leippe *et al.*, 1995) and localised in lysosome-like
10 vesicles or at the surface of the cell (Garcia-Rivera *et al.*, 1999; Jacobs *et al.*,
11 1998). Molecular cloning has revealed a large number of cysteine
12 endopeptidase genes in the *E. histolytica* genome (Bruchhaus *et al.*, 2003;
13 Garcia-Rivera *et al.*, 1999; Reed *et al.*, 1993; Tannich *et al.*, 1991c; Tannich *et*
14 *al.*, 1992). Interestingly, most of these genes are not expressed during in vitro
15 cultivation (Bruchhaus *et al.*, 2003). As our current knowledge of *E. histolytica*
16 biology and pathogenicity is mostly based on analysis of cultured cells, the
17 function of most of the cysteine endopeptidases and their precise role in *E.*
18 *histolytica* virulence is largely unknown.

19

20 Homology searches using conserved active site regions revealed that the *E.*
21 *histolytica* genome contains at least 44 genes coding for cysteine
22 endopeptidases. Of these, the largest group is structurally related to the C1
23 papain superfamily (Table 4), whereas a few others are more similar to family
24 C2 (calpain-like cysteine proteinases), C19 (ubiquitinyl hydrolase), C54
25 (autophagin), and C65 (otubain), respectively (Table 5).

26

27 Phylogenetic analyses of the 36 C1-family members revealed that they
28 represent 3 distinct clades (A, B, C), consisting of 12, 11 and 13 members,
29 respectively. Clade A and B members correspond to the two previously
30 described subfamilies of *E. histolytica* cysteine proteinases, designated EhCP-A
31 and EhCP-B (Bruchhaus *et al.*, 2003). In contrast, clade C represents a new

1 group of *E. histolytica* cysteine endopeptidases that has not been described
2 before. EhCP-A and EhCP-B subfamily members are classical pre-pro enzymes
3 with an overall cathepsin L-like structure (Barrett 1998) as indicated by the
4 presence of an ERFNIN motif in the pro region of at least 21 of the 23 EhCP-A
5 and EhCP-B enzymes (Figure 4). Interestingly, biochemical studies with
6 purified EhCP-A indicated a cathepsin B-like substrate specificity (Scholze and
7 Schulte, 1988). This is likely due to the substitution of an alanine residue by
8 acidic or charged amino acids in the postulated S2 pocket, corresponding to
9 residue 205 of the papain sequence (Barrett 1998). As reported previously
10 (Bruchhaus *et al.*, 2003), the EhCP-A and EhCP-B subfamilies differ in the
11 length of the pro regions as well as of the catalytic domains, and have distinct
12 sequence motifs in the N-terminal regions of the mature enzymes (DWR vs.
13 PCNC). Moreover, none of the EhCP-A subfamily but 10 of the 11 EhCP-B
14 sequences contain hydrophobic stretches near or at the C-terminus, some of
15 which are predicted to constitute transmembrane helices (TMH) or GPI-
16 attachment moieties. This finding is consistent with previous reports on surface
17 localisation of *E. histolytica* cysteine proteinases but, so far, studies on the
18 cellular localisation of the various EhCP -B molecules have not been reported.
19
20 In contrast to the EhCP-A and EhCP-B subfamilies, primary structure
21 prediction indicates that EhCP-C members are not pre-proenzymes, as they
22 lack hydrophobic signal sequences as well as identifiable pro regions. Instead,
23 they contain a hydrophobic region located 11 to 28 amino acids from the N-
24 terminus, which is predicted to form a TMH (Figure 4). Therefore, this new
25 group of molecules appears to be membrane associated via a signal anchor. All
26 EhCP-C enzymes have a conserved motif of the sequence H/I(X)$_6$L/ICP in the
27 C-terminal half but they differ substantially in their pI, with values ranging
28 from 4.6 to 8.8. As there is no example of a structurally related cysteine
29 endopeptidase corresponding to the EhCP-C subfamily in other organisms, the
30 specific functions of this group of molecules remain completely unknown.
31

In addition to the large number of C1 superfamily members, the *E. histolytica* genome contains 2 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 remodeling of the cytoskeleton and membranes, signal transduction pathways and apoptosis.

Another 4 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* are found a family of small proteins, amoebapores, that are cytolytic towards human host cells,

1 display potent antibacterial activity, and cause ion channel formation in
2 artificial membranes (for a review see Leippe (1997)). Three amoebapore
3 isoforms have been isolated and biochemically characterised, and their primary
4 structure has been elucidated by molecular cloning of the genes encoding their
5 precursors (Leippe *et al.*, 1991; Leippe *et al.*, 1992; Leippe *et al.*, 1994b).
6 These membrane-permeabilising polypeptides are discharged by *E. histolytica*
7 into bacteria-containing phagosomes to combat growth of engulfed
8 microorganisms (Andrä *et al.*, 2003). Because of their potent cytolytic activity
9 against human cells in vitro (Berninghausen and Leippe, 1997; Leippe *et al.*,
10 1994a), amoebapores have been viewed as a crucial element of the machinery
11 use by the parasite to kill host cells. Trophozoites of *E. histolytica* lacking the
12 major isoform amoebapore A, whether through antisense inhibition of
13 translation (Bracha *et al.*, 1999) or epigenetic silencing of the gene (Bracha *et*
14 *al.*, 2003), became avirulent demonstrating that this protein plays a key role in
15 pathogenesis. Relatives of these protistan polypeptides are found in granules of
16 porcine and human cytotoxic lymphocytes where they are termed NK-lysin and
17 granulysin, respectively. All of these polypeptides are 70-80 amino acids in
18 length and are characterised by a compact alpha-helical, disulphide-bonded
19 structure known as the saposin-like fold. The structures of the amoebic and
20 mammalian polypeptides have been solved and compared (Anderson *et al.*,
21 2003; Hecht *et al.*, 2004; Leippe *et al.*, 2005; Liepinsh *et al.*, 1997). The
22 biological activities have also been measured in parallel (Bruhn *et al.*, 2003;
23 Gutsmann *et al.*, 2003) to evaluate the similarities and differences of these
24 effector molecules from organisms whose evolutionary paths diverged very
25 early. As they are active against both prokaryotic andeukaryotic target cells,
26 they may be viewed as broad-spectrum effector molecules.
27
28 In the genome of *E. histolytica*, 16 genes coding for putative saposin-like
29 proteins (SAPLIPs) were identified. All of these genes are transcribed by cells
30 growing in axenic culture (Winkelmann *et al.*, 2006). Like amoebapores, the
31 predicted proteins all contain one C-terminal SAPLIP domain and (with one

exception) a putative signal peptide (Table 6). 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 four 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 phagocytising cell.

3.4 Antioxidants

Entamoeba 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 (Figure 5).

1
2 Among the enzymes in the first line of oxidative defence are superoxide
3 dismutases (SODs), which are metalloproteins that use copper/zinc (Cu/Zn),
4 manganese (Mn) or iron (Fe) as metal cofactors. SODs catalyse the dismutation
5 of superoxide radical anions to form H_2O_2 and O_2 (Fridovich, 1995). Analysis
6 of the *E. histolytica* genome revealed only a single gene coding for a FeSOD
7 and no sequences encoding MnSOD or Cu/ZnSOD. This reflects the situation
8 found in most protistan parasites and is consistent with biochemical studies
9 previously performed on *E. histolytica* lysates (Tannich *et al.*, 1991a).
10
11 *Entamoeba histolytica* lacks the tripeptide glutathione (Fahey *et al.*, 1984),
12 which constitutes the major low molecular weight thiol found in almost all
13 aerobic cells (Sies, 1999). Instead, *E. histolytica* uses cysteine as its principal
14 low molecular weight thiol (Ariyanayagam and Fairlamb, 1999; Fahey *et al.*,
15 1984; Nozaki *et al.*, 1999). As expected, coding sequences for enzymes that use
16 glutathione as a cofactor, such as glutathione-S-transferase, glutathione-
17 dependent peroxidase, glutathione reductase or glutaredoxin, are all absent from
18 the *E. histolytica* genome. In addition, genes encoding catalases and
19 peroxidases are also missing, as previously suggested (Sykes and Band, 1977;
20 Weinbach and Diamond, 1974).
21
22 Other genes were identified that code for proteins involved in detoxification of
23 H_2O_2, including one with homology to rubrerythrin. Rubrerythrin is a non-
24 haeme iron protein thought to be able to reduce H_2O_2 as part of an oxidative
25 stress protection system (Weinberg *et al.*, 2004). So far, the nature of its redox
26 partner is unknown in *E. histolytica* and it remains to be determined whether
27 protection against oxidative stress is indeed its main function. Another group of
28 H_2O_2-detoxifying proteins identified in *E. histolytica* are peroxiredoxins.
29 Peroxiredoxins are known from a wide variety of organisms. They are able to
30 reduce H_2O_2 as well as peroxynitrite with the use of electrons provided by
31 thiols. In addition to involvement in the detoxification of reactive oxygen

1 species peroxiredoxins seem to play a role in other processes such as signalling
2 and differentiation (Hofmann *et al.*, 2002; Rhee *et al.*, 2005; Wood *et al.*,
3 2003a,b). All peroxiredoxins contain a conserved cysteine residue that
4 undergoes a cycle of peroxide-dependent oxidation and thiol-dependent
5 reduction during the reaction. The whole protein family can be divided into
6 three classes based on the number and position of active site Cys residues (2-
7 Cys, atypical 2-Cys, and 1-Cys peroxiredoxins; Wood *et al.*, 2003a,b). In *E.*
8 *histolytica* five different genes coding for peroxiredoxins were identified (Prx1-
9 5). They all belong to the 2-Cys peroxiredoxin family. Four of them (Prx1-4)
10 share 98% sequence identity and have an unusual N-terminal Cys-rich repeat
11 (KECCKKECQEKECQEKECCC) of unknown function. In contrast, the fifth
12 peroxiredoxin (Prx5) lacks the cysteine-rich N-terminal extension and shares
13 only 30% identity with Prx1-4. Biochemical studies have shown that *E.*
14 *histolytica* peroxiredoxins are able to detoxify H_2O_2 and cumene hydroperoxide
15 (Bruchhaus *et al.*, 1997; Poole *et al.*, 1997). Moreover, up-regulation of
16 peroxiredoxin and FeSOD was associated with metronidazole resistance in
17 cultured *E. histolytica* trophozoites (Samarawickrema *et al.*, 1997; Wassmann
18 *et al.*, 1999).
19
20 Reactions catalysed by peroxiredoxins are dependent on the presence of
21 physiological thiols like thioredoxin (Rhee *et al.*, 2005; Wood *et al.*, 2003b).
22 Thioredoxins are small proteins involved in thiol-redox processes (Holmgren,
23 2000). They contain two redox-active site cysteine residues of the motif CXXC
24 (Watson *et al.*, 2004). Five genes coding for classical cytoplasmic thioredoxins
25 were identified in the *E. histolytica* genome (Trx1-5). These thioredoxins have
26 a length of 103-114 amino acids and share 25 – 47 % sequence identity. Trx1-3
27 have identical active site motifs of the sequence WCGPC, whereas the active
28 sites of Trx4 and Trx5 have the sequences SCPSC and WCKDC, respectively.
29 In addition, another five thioredoxin-related proteins were identified (Trx6-10).
30 All have a signal sequence of 15 to 19 amino acid residues and the active site
31 motif WCGHC, which is also known from the active site of protein disulphide

1 isomerases. However, in contrast to the latter group of enzymes, the *E.*
2 *histolytica* thioredoxin-related molecules contain only one rather than two
3 active-site motifs and only two of the proteins have an ER membrane retention
4 signal (Freedman *et al.*, 2002). Thus it remains to be determined whether the
5 thioredoxin-related molecules of *E. histolytica* do constitute protein disulphide
6 isomerases or whether they undertake other functions within the cell.
7
8 Thioredoxins are kept in the reduced state by the enzyme thioredoxin reductase
9 which catalyses the reduction of oxidised thioredoxin by NADPH using FAD
10 and its redox-active disulphide (Nakamura, 2005). Two different genes with
11 homology to thioredoxin reductases have been previously described from *E.*
12 *histolytica* (thioredoxin reductase (TrxR) and NADPH:flavin oxidoreductase
13 (p34)). They share about 87% sequence identity and both contain the two
14 conserved sequence motifs forming the FAD and NAD(P)H binding domains.
15 p34 was shown to catalyse the NADPH-dependent reduction of oxygen to H_2O_2
16 as well as of disulphides like DTNB and cystine (Bruchhaus *et al.*, 1998; Lo
17 and Reeves, 1980). Therefore, in addition to disulphide reductase activity the
18 enzyme has H_2O_2-forming NADPH oxidase activity. It was also shown that p34
19 can transfer reducing equivalents to peroxiredoxin, converting the protein from
20 its non-active, oxidised form back into its active, reduced form (Bruchhaus *et*
21 *al.*, 1997). However, it is unlikely that peroxiredoxin is directly reduced by p34
22 *in vivo*. It is more likely that *E. histolytica* contains a classical thioredoxin
23 redox system consisting of thioredoxin reductase, thioredoxin and
24 peroxiredoxin (Poole *et al.*, 1997).
25
26 In addition to genes coding for proteins with homology to thioredoxin
27 reductase, four other gene families were identified that encode various
28 flavoproteins. One of these families includes 4 members that have between
29 53% and 61% sequence identity to A-type flavoproteins
30 (flavorubredoxin/flavodiiron). A-type flavoproteins belong to a large family of
31 enzymes that are widespread among anaerobic and facultatively anaerobic

1 prokaryotes. In addition to bacteria, homologous genes are also found in the
2 genomes of the pathogenic amitochondriate protistan parasites *Trichomonas*
3 *vaginalis* and *Giardia intestinalis* (Andersson *et al.*, 2003; Sarti *et al.*, 2004).
4 The A-type flavoproteins are made up of two independent structural modules.
5 The N-terminal region forms a metallo-beta-lactamase-like domain, containing
6 a non-haeme di-iron site, whereas the C-terminal region is a flavodoxin-like
7 domain, containing one FMN moiety. These enzymes have significant nitric
8 oxide reductase activity (Gomes *et al.*, 2002; Sarti *et al.*, 2004). For *Escherichia*
9 *coli* it is known that the nitric oxide reductase (FIRd) receives electrons from a
10 NADH:oxidoreductase (FIRd-red). Consistent with that situation, the *E.*
11 *histolytica* genome contains a gene encoding an NADH oxidase with 25%
12 sequence identity to several bacterial FIRd-reds.
13
14 The three other *E. histolytica* gene families with homology to iron-sulphur
15 flavoproteins (families B-D) are characterised by the presence of a flavodoxin-
16 like domain forming a typical FMN binding site. Family B and family C consist
17 of three members each, which share sequence identity of 42% and 46%,
18 respectively. Family D consists of two members, which share only 33%
19 sequence identity. At present, the function of the various flavodoxin-like
20 molecules remains to be determined and fully deserves to be investigated,
21 particularly as to whether they do indeed have antioxidant capacity.
22
23 **4. METABOLISM**
24 Biochemical analysis of *E. histolytica* metabolism has a long history (Reeves,
25 1984), dating back to shortly after the development of culture media that
26 allowed the generation of substantial numbers of axenic cells. The genome
27 sequence has confirmed most of the predicted metabolic pathways shown
28 biochemically to be present or absent in *E. histolytica* in the past. As with most
29 parasites, secondary loss of biosynthetic pathways is a recurring theme.
30 However, a few surprises have also been uncovered. Every single enzyme
31 involved in metabolism cannot realistically be discussed in this review. In this

section the only the major energy generating and biosynthetic aspects of metabolism will be covered. Enzyme rames, EC numbers and accession numbers are given in the the supplementary table for this section.

4.1 Energy Metabolism

4.1.1 Glycolysis

E. histolytica lacks a functional tricarboxylic acid (TCA) cycle and oxidative phosphorylation. It is not able to convert organic substrates such as glucose into H_2O and CO_2, but has to rely on the energy generated by various types of substrate level phosphorylation (Reeves, 1984). Glycolysis is the major pathway of ATP generation, but in addition the genome project has identified a number of genes that could result in more ATP generation through the catabolism of amino acids. These enzymes will be described further below. As *E. histolytica* lacks compartmentalised energy generation, it has been classified as a type I amitochondriate protist (Martin and Müller , 1998) in contrast to the type II amitochondriate protists containing hydrogenosomes, such as *Trichomonas vaginalis*. Nevertheless, it does contain a mitochondrial remnant, the mitosome (see section 8).

In *E. histolytica*, glycolysis appears to be localised in the cytosol. This is in contrast to trypanosomes where a major part is carried out in the glycosomes (Parsons, 2004) and the pathway is regarded as a potential target for chemotherapy (Opperdoes and Michels, 2001). The kinetic properties of recombinant *E. histolytica* glycolysis enzymes have recently been studied by Saavedra *et al.* (2005). Their analysis suggested that fructose-1,6-bisphosphate aldolase, phosphoglycerate mutase, glyceraldehyde-3-phosphate dehydrogenase, and pyruvate phosphate dikinase might be regulating the glycolytic flux.

4.1.1 (a) Hexokinases

Glucose taken up by *E. histolytica* is phosphorylated by two hexokinase (EC 2.7.1.1) isoenzymes (Hxk1 and Hxk2). The two *E. dispar* isoenzymes are shifted towards a slightly more basic pI, which is the basis of the classical biochemical method for distinguishing *E. histolytica* from *E. dispar* by starch gel electrophoresis (Farri *et al.*, 1980). The pI differences among the two *E. histolytica* isoforms (Ortner *et al.*, 1995) and between the two species (Ortner *et al.*, 1997b) are the result of genetic differences that lead to different amino acid sequences and charge differences. Hxk1 phosphorylates glucose and mannose, while Hxk2 phosphorylates mainly glucose and is much less active with mannose as a substrate (Kroschewski *et al.*, 2000).

4.1.1 (b) Glucose-6-phosphate isomerase

Glucose 6-phosphate is converted to fructose 6-phosphate by glucose-6-phosphate isomerase (EC 5.3.1.9). The genome has 2 genes for this enzyme, which code for proteins that differ only by a single insertion/deletion of 7 amino acid residues. Glucose-6-phosphate isomerase is another of the enzymes for the classical differentiation of *Entamoeba* zymodemes by starch gel electrophoresis (Sargeaunt, 1987).

4.1.1 (c) Phosphofructokinases

The main phosphofructokinase activity in *E. histolytica* is PPi-dependent (EC 2.7.1.90; Reeves *et al.*, 1976). There is a single gene (Deng *et al.*, 1998) encoding this 60 kDa enzyme. The gene is a candidate for lateral transfer from bacteria (Loftus *et al.*, 2005) (see section 10). The enzyme is expressed at a tenfold higher level and displays about tenfold higher activity than a second phosphofructokinase of 48 kDa (XP_653373) (Chi *et al.*, 2001). The substrate specificity of the smaller enzyme is disputed. Whereas Bruchhaus *et al.* (1996) reported that this minor enzyme also used PPi as phosphate donor, Chi *et al.* (2001) found only an ATP-dependent activity. The 48 kDa and 60 kDa enzymes are highly divergent with less than 20% sequence identity. Interestingly, the specificity of the 60 kDa phosphofructokinase can be changed

from PPi to ATP by mutation of a single amino acid residue (Chi and Kemp, 2000). The authors concluded that ATP rather than PPi was the primordial high energy compound. In the genome, there are two additional genes encoding isoforms of the 48 kDa enzyme, which have not been studied on the protein level.

4.1.1 (d) Fructose-1,6-bisphosphate aldolase

Fructose 1,6-bisphosphate is cleaved to glyceraldehyde 3-phosphate and dihydroxyacetone 3-phosphate by fructose-1,6-bisphosphate aldolase (EC 4.1.2.13). The enzyme, a Class II aldolase (Marsh and Lebherz, 1992) has been cloned (XP_650373) and exhibits strong sequence similarity to eubacterial aldolases (Sanchez *et al.*, 2002). A second gene (XP_655966) encodes a protein differing from the first by a single deletion of 28 amino acids flanked by short divergent stretches. These bacterial-type aldolases are also found in *Trichomonas vaginalis*, *Giardia intestinalis* and other protists (Sanchez *et al.*, 2002). *E. histolytica* has no gene coding for a Class I aldolase like those found in animals, which might make aldolase an interesting target for chemotherapy.

4.1.1 (e) Triose-phosphate isomerase

Triose-phosphate isomerase (EC 5.3.1.1) converts dihydroxyacetone 3-phosphate into glyceraldehyde 3-phosphate. The gene was previously cloned (Landa *et al.*, 1997), and is highly similar to the annotated gene product. This dimer-forming enzyme represents the first *E. histolytica* protein for which the structure has been solved by X-ray crystallography (Rodriguez-Romero *et al.*, 2002).

4.1.1 (f) Glyceraldehyde-3-phosphate dehydrogenase

Glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) oxidises and phosphorylates glyceraldehyde 3-phosphate to 1,3-bisphosphoglycerate in two coupled reactions using NAD^+ as cofactor (Reeves, 1984). The genome project revealed five putative genes, three of which encode the identical protein

1 sequence of 36.0 kDa and a predicted pI of 7.04. The fourth gene product,
2 XP_648981, differs from these three only by a 13 amino acid deletion, while
3 XP_650370 is a clearly distinct 34.8 kDa isoform with a lower predicted pI of
4 5.80. Interestingly, the isoforms XP_650356 and XP_650370 of different pI are
5 encoded within the same contig.
6
7 4.1.1 (g) Phosphoglycerate kinase
8 Phosphoglycerate kinase has an unusual substrate (Reeves and South, 1974),
9 transferring the high energy phosphate group from 3-phosphoglyceroyl
10 phosphate to GDP leading to the formation of GTP (EC 2.7.2.10). There is one
11 candidate gene encoding a 45 kDa protein.
12
13 4.1.1 (h) Phosphoglycerate mutase
14 Phosphoglycerate mutase (Reeves, 1984) isomerises 3-phosphoglycerate to 2-
15 phosphoglycerate (EC 5.4.2.1). Five divergent putative genes for this enzyme
16 are found in the genome. Two gene products of 62 kDa were classified as 2,3-
17 bisphosphoglycerate-independent phosphoglycerate mutases (XP_649031 and
18 XP_654182); they differ only at their C-termini and display significant
19 similarity to bacterial phosphoglycerate mutases. The three other genes are very
20 divergent. XP_651808 was identified as a candidate for lateral gene transfer
21 (Loftus *et al.*, 2005) (see section 10). The remaining two gene products
22 XP_649053 and XP_657284 are related to genes found in both prokaryotes and
23 eukaryotes.
24
25 4.1.1 (i) Enolase (2-phosphoglycerate dehydratase)
26 Enolase (EC 4.2.1.11) converts 2-phosphoglycerate to phosphoenolpyruvate.
27 The gene has been cloned (Beanan and Bailey, 1995) and the protein
28 characterised (Hidalgo *et al.*, 1997) previously. The 47 kDa gene product is a
29 typical eukaryotic enolase (XP_649161). A carboxy-terminally truncated
30 incomplete ORF is also found.
31

1 4.1.1 (j) Pyruvate, orthophosphate dikinase and pyruvate kinase
2 In *E. histolytica*, both activities forming ATP and pyruvate from
3 phosphoenolpyruvate have been found. The exergonic pyruvate kinase reaction
4 uses ADP (Saavedra *et al.*, 2004), and the pyruvate, orthophosphate dikinase
5 uses AMP and PPi in a slightly endergonic reaction (Varela-Gomez *et al.*,
6 2004). The dikinase activity is found in C4 plants where it is involved in
7 phosphoenolpyruvate generation for gluconeogenesis. In *E. histolytica* it was
8 discovered long before the pyruvate kinase (Reeves, 1968).
9
10 The cloning of pyruvate, orthophosphate dikinase (EC 2.7.9.1) was reported by
11 two groups. The published sequences (Bruchhaus and Tannich, 1993; Saavedra
12 Lira *et al.*, 1992) are highly similar or identical to XP_657332 and XP_654666.
13 In addition there are two shorter related open reading frames.
14
15 In the genome 3 putative pyruvate kinase genes (EC 2.7.1.40) have been
16 identified. The 3 are identical except for an amino-terminal deletion in
17 XP_648240 and an internal deletion in XP_653635.
18
19 4.1.1 (k) Pyruvate:ferredoxin oxidoreductase (PFOR) and ferredoxin
20 PFOR (EC 1.2.7.1) is an enzyme of major importance to *E. histolytica*, as the
21 parasite lacks NAD^+-dependent pyruvate dehydrogenase and pyruvate
22 decarboxylase (Reeves, 1984). No evidence for the latter two genes was found
23 in the genome, confirming the biochemical results. PFOR oxidatively
24 decarboxylates pyruvate to acetyl-CoA. The electrons are transferred to
25 ferredoxin which, in its reduced form, can activate and reduce metronidazole,
26 the major anti-amoebic drug (Müller , 1986). The activated form of
27 metronidazole can potentially react with a number of biomolecules and is able
28 to cleave the parasite DNA. In human cells, metronidazole is not activated and
29 is much less toxic. In *T. vaginalis*, down-regulation of PFOR is one mechanism
30 of producing metronidazole resistance (Kulda, 1999); however PFOR
31 expression appears unaltered in partially resistant *E. histolytica*

1 (Samarawickrema *et al.*, 1997; Wassmann *et al.*, 1999). All eukaryotic PFOR
2 genes, including that of *E. histolytica*, appear to have been acquired during an
3 ancient lateral gene transfer event from bacteria (Horner *et al.*, 1999; Rotte *et*
4 *al.*, 2001). There are two putative PFORs in the *E. histolytica* genome
5 displaying minor sequence differences.
6
7 The genome contains 7 ferredoxin genes in total with 5 quite divergent
8 sequences. All are related to eubacterial and archaeal ferredoxins (Nixon *et al.*,
9 2002). The gene pairs XP_655183 / XP_655182 and XP_654311 / XP_652694
10 are identical. The other three gene products represent more divergent open
11 reading frames. The deduced proteins have similar molecular masses, between
12 6.1 kDa and 8.8 kDa, and different predicted isoelectric points between 4.2 and
13 8.6.
14
15 4.1.1 (l) Acetyl-CoA synthetase (acetate thiokinase)
16 The normal fate of acetyl-CoA in mitochondriate organisms is entry into the
17 tricarboxylic acid cycle. However this pathway is absent from *E. histolytica.*
18 Instead, the cleavage energy of the thioester bond of acetyl-CoA can be used to
19 generate one ATP molecule. One of the known acetyl-CoA synthetases
20 generates ATP from ADP and Pi (EC 6.2.1.13). Such an enzyme has been
21 characterised by Reeves *et al.* (1977) and cloned (Field *et al.*, 2000), and
22 reported to be a 77 kDa protein. The common acetyl-CoA synthetase activity
23 that produces ATP from AMP and PPi (EC 6.2.1.1) appears to be absent in *E.*
24 *histolytica*.
25
26 4.1.1 (m) Aldehyde and alcohol dehydrogenases
27 The *E. histolytica* genome encodes a complex system of alcohol and/or
28 aldehyde dehydrogenases. In total, there are 25 predicted genes, 3 of which are
29 on the list of lateral gene transfer (LGT) candidates.
30

1 Alcohol dehydrogenase ADH1 was the first alcohol dehydrogenase to be
2 characterised in *E. histolytica* (Reeves *et al.*, 1971) and is a NADPH-dependent
3 enzyme (EC 1.1.1.2). The gene was previously cloned (Kumar *et al.*, 1992); in
4 the genome three genes are almost identical to that sequence, while one
5 (XP_652772) has 67% identity.
6
7 Fermentation in *E. histolytica* uses the bifunctional alcohol dehydrogenase /
8 aldehyde dehydrogenase NADH-dependent enzyme ADH2, which belongs to
9 the ADHE family and has both alcohol dehydrogenase and aldehyde
10 dehydrogenase activities (Lo and Reeves, 1978). Under anaeobic conditions,
11 reduction of the acetyl-CoA generated by PFOR to ethanol is one way to
12 regenerate the NAD+ used by glyceraldehyde-3-phosphate dehydrogenase.
13 ADH2 first reduces acetyl-CoA to an enzyme-bound hemiacetal which is then
14 hydrolysed to acetaldehyde (EC 1.2.1.10) and further reduced to ethanol (EC
15 1.1.1.1). If the enzyme is also able to work in the reverse direction, *E.*
16 *histolytica* would be able to generate acetyl-CoA and energy from ethanol in
17 the presence of oxygen. This would explain older reports of ethanol stimulated
18 oxygen uptake in *E. histolytica* (Weinbach and Diamond, 1974). The enzyme is
19 closely related to AdhE from *E. coli* and other bacteria (Reid and Fewson,
20 1994) and there is strong support for its aquisition by LGT (Andersson *et al.*,
21 2006; Field *et al.*, 2000; Loftus *et al.*, 2005) (see section 10). Like its bacterial
22 homologue, ADH2 appears to form helical rods that sediment with membrane
23 fractions (Avila *et al.*, 2002). Two groups have previously cloned ADH2
24 (Bruchhaus and Tannich, 1994; Yang *et al.*, 1994), and in total the genome
25 contains 5 full-length ADH2 genes and one that is truncated. All share between
26 98% and 100% sequence identity.
27
28 In total, there are 11 alcohol dehydrogenase ADH3 genes in the genome, two of
29 which been reported previously (Kimura *et al.*, 1996; Rodriguez *et al.*, 1996).
30 The recombinant enzyme characterised by Rodriguez *et al.* (1996) was
31 NADPH-specific, like ADH1. There are five genes similar to these previously