The genome of the protist parasite

Entamoeba histolytica

Brendan Loftus 1, Iain Anderson 1, Rob Davies 1, U. Cecilia M. Alsmark 2, John Samuelson 2, Paolo Amedeo 3, Paola Roncaglia 3, Matt Berriman 4, Robert P. Hirt 5, Barbara J. Mann 6, Tomo Nozaki 6, Bernard Suh 6, Mihai Pop 7, Michael Duchene 6, John Ackers 8, Egbert Tannich 9, Matthias Leippe 10, Margit Hofer 11, Iris Bruchhaus 12, Ute Willhoft 13, AloK Bhattacharya 14, Tracey Chillingworth 15, Carol Gilchrist 15, Sharon Moule 6, Karen Mungall 6, Doug Ormond 6, Rob Squares 6, Sally Whitehead 6, Michael A. Quail 6, Ester Rabbittsch 16, Halina Norbertzak 7, Claire Price 7, Zheng Wang 8, Nancy Guillon 12, Carol Gilchrist 8, Suzanne E. Stroup 12, Sudha Bhattacharya 17, Anuradha Lohia 18, Peter G. Foster 14, Thomas Schieritz-Ponten 15, Christian Weber 14, Upinder Singh 13, Chandrama Mukherjee 14, Najib M. El-Sayed 8, Alok Bhattacharya 19, William A. Petri Jr. 1, C. Graham Clark 1, T. Martin Embley 8, Bar t Barrell 12, Claire M. Fraser 8 & Neil Hall 11

1 TIGR, 9712 Medical Center Drive, Rockville, Maryland 20850, USA
2 The Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
3 School of Biological Sciences, University of Newcastle, King George VI Building, Newcastle upon Tyne NE1 7RU, UK
4 Department of Molecular and Cell Biology, Boston University Goldman School of Dental Medicine, 715 Albany Street, Boston, Massachusetts 02118, USA
5 Departments of Internal Medicine & Microbiology, University of Virginia, Charlottesville, Virginia 22908, USA
6 Department of Parasitology, National Institute of Infectious Diseases, 1-23-1 Toyama, Shinjuku-ku, Tokyo 162-8640, Japan
7 Division of Specific Prophylaxis and Tropical Medicine, Center for Physiology and Pathophysiology, Medical University of Vienna, Krankenspitze-Gasse 15, A-1095 Vienna, Austria
8 Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
9 Department of Molecular Parasitology, Bernhard Nocht Institute for Tropical Medicine, Bernhard Nocht Str. 74, 20359 Hamburg, Germany
10 Zoological Institute, University of Kiel, Olshausenstr. 40, 24098 Kiel, Germany
11 School of Environmental Sciences, Jawaharlal Nehru University, New Delhi 110067, India
12 Unite de Biologie Cellulaire du Parasitisme, INSERM U389, Institut Pasteur 28, rue du Dr Roux 75724, Paris Cedex 15, France
13 Department of Biochemistry, Bose Institute, P/11 CIT Scheme VIIM, Kolkata 700054, India
14 Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK
15 Center for Biological Sequence Analysis, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark
16 Departments of Internal Medicine, Microbiology, and Immunology, Stanford University School of Medicine, Stanford, California 94305-5107, USA
17 Present address: TIGR, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

Entamoeba histolytica is an intestinal parasite and the causative agent of amoebiasis, which is a significant source of morbidity and mortality in developing countries. Here we present the genome of E. histolytica, which reveals a variety of metabolic adaptations shared with other amitochondrial protist pathogens: Giardia lamblia and Trichomonas vaginalis. These adaptations include reduction or elimination of most mitochondrial metabolic pathways and the use of oxidative stress enzymes generally associated with anaerobic prokaryotes. Phylogenetic analysis identifies evidence for lateral gene transfer of bacterial genes into the E. histolytica genome, the effects of which centre on expanding aspects of E. histolytica’s metabolic repertoire. The presence of these genes and the potential for novel metabolic pathways in E. histolytica may allow for the development of new chemotherapeutic agents. The genome encodes a large number of novel receptor kinases and contains expansions of a variety of gene families, including those associated with virulence. Additional genome features include an abundance of tandemly repeated transfer-RNA-containing arrays, which may have a structural function in the genome. Analysis of the genome provides new insights into the workings and genome evolution of a major human pathogen.

Genome analysis was carried out on a 12.5-fold coverage genome assembly consisting of 23,751,783 base pairs (bp) distributed among 888 scaffolds. The 9,938 predicted genes average 1.17 kilobases (kb) in size and comprise 49% of the genome. One-quarter of E. histolytica genes are predicted to contain introns, with 6% of genes containing multiple introns. No homologues could be identified for a third of predicted proteins (31.8%) from the public databases (see Methods). E. histolytica chromosomes do not condense, and the uncertainty surrounding its ploidy and the extensive length variability observed between homologous chromosomes from different isolates makes the exact chromosome number difficult to determine. The chromosome size variation observed may be due to expansion and contraction of subtelomeric repeats, as in other protists, and it is tempting to speculate that in E. histolytica these regions consist of rRNA-containing arrays. Comprising almost 10% of the sequence reads, 25 types of long tandem array, each containing between one and five tRNA types per repeat unit, could be identified from the genome data. The full complement of tRNAs required for translation has been identified, and all but four of the tRNA genes are encoded exclusively in arrays. These unique tRNA gene arrays are thus predicted to be functional as well as potentially fulfilling a structural role in the genome. No association could be determined between codon usage and the relative copy numbers of their cognate tRNA species.

The metabolism of E. histolytica seems to have been shaped by secondary gene loss and lateral gene transfer (LGT), primarily from bacterial lineages (Fig. 1). E. histolytica is an obligate fermenter, using bacterial-like fermentation enzymes and lacking proteins of the tricarboxylic acid cycle and mitochondrial electron transport chain. An atrophic, mitochondrion-derived organelle has been identified in E. histolytica, and the genome data support the absence of a mitochondrial genome. Glucose is the main energy source; however, in place of the typical eukaryotic glucose transporters those of E. histolytica are related to the prokaryote glucose/ribose porter family, with the amino- and carboxy-terminal domains switched relative to their prokaryotic counterparts.

As a phagocytic resident of the human gut, E. histolytica has access to many bacterial and host-derived preformed organic compounds. Most pathways for amino acid biosynthesis have been eliminated, except those for serine and cysteine, which are probably retained for the production of cysteine, the major intracellular thiol. The high levels of cysteine in E. histolytica may compensate for the lack of glutathione and its associated enzymes, a major component of oxidative stress resistance in many organisms. E. histolytica lacks de novo purine, pyrimidine and thymidylate synthesis and must rely on salvage pathways, similar to G. lamblia and T. vaginalis. In addition, E. histolytica appears to lack ribonucleotide reductase, a characteristic that it shares with G. lamblia. E. histolytica is unable to synthesize fatty acids but retains the ability to synthesize a variety of phospholipids. The absence of identifiable pathways for the synthesis of isoprenoids and the sphingolipid head group aminoethyolphosphonate suggest the existence of novel pathways. These pathways, once characterized, might represent attractive drug targets. Two unusual enzymes of fatty acid elongation are shared between E. histolytica and G. lamblia, including a predicted acetyl-CoA carboxylase with two carboxyltransferase domains. We propose that this enzyme removes a carboxyl group from oxaloacetate and transfers it to acetyl-CoA to form malonyl-CoA and pyruvate. E. histolytica also has five members of a fatty acid elongase family, previously identified only in plants, green algae and G. lamblia. Polate is
a cofactor essential for thymidylate synthesis and methionine recycling, and genome analysis reveals a complete lack of genes coding for known folate-dependent enzymes and folate transporters. Folate is also required for organelle protein synthesis in mitochondria and chloroplasts, and loss of the mitochondrial genome may have paved the way for the loss of these folate-dependent functions.

LGT is an important force in the evolution of prokaryotes but significantly less is known about its importance in eukaryotic evolution\(^1\). We conducted a phylogenetic screen of the Entamoeba genome for cases of relatively recent prokaryote to eukaryote LGT (see Methods), and for 96 genes we believe that this is the simplest explanation for the tree topologies obtained (see Supplementary Information). These genes are embedded among typically eukaryotic genes on Entamoeba scaffolds and do not seem to represent contaminating prokaryotic sequences. Most (58%) of the LGT genes encode a variety of metabolic enzymes, whereas most of the remaining genes (41%) encode proteins of unknown function (Supplementary Fig. 1). The major impact is in the area of carbohydrate and amino acid metabolism, where they have increased the range of substrates available for energy generation including tryptophanase and aspartase, which contribute to the use of amino acids. Several glycosidases and sugar kinases appear to have been acquired through LGT and would probably enable Entamoeba to use sugars other than glucose; for example, fructose and galactose. There is a strong bias in the data for a major donor being in the Cytophaga–Flavobacterium–Bacteroides (CFB) group of the phylum Bacteroidetes; however, this should be interpreted with caution, as current sampling of prokaryotic genomes is still relatively incomplete. It is clear that among the 96 genes, some result in significant enhancements to Entamoeba metabolism, thus contributing to its biology to a greater extent than indicated by the numbers alone.

**Figure 1** Predicted metabolism of *E. histolytica* based on analysis of the genome sequence data. Arrows indicate enzyme reactions. Glycolysis and fermentation are the major energy generation pathways. Green arrows represent enzymes encoded by genes that are among the 96 candidates for LGT into the *E. histolytica* genome. Broken arrows indicate enzymes for which no gene could be identified using searches of the genome data, although the activity is likely to be present. The yellow arrow points to the source of electrons for activation of metronidazole, the major drug for treatment of amoebic liver abscess. DK, dihydroxyacetone kinase; GlcNdc, N-acetylglucosamine; G6P, glucose-6-phosphate; 1-3β-D-Glu, 1-3β-D-glucan; 

\(^1\) E. histolytica feed on bacteria in the lumen of the colon and lyse host epithelial cells after invasion of the intestinal wall\(^2\). A number of amoebic virulence determinants have been characterized, including a multi-subunit GalGalNAc lectin involved in adhesion to host cells, cysteine proteases that degrade host extracellular matrix, and pore-forming peptides (amoebapores) capable of lysing target cells\(^3\). Analysis of the genome reveals redundancy in the genes encoding these virulence factors. Thirty homologues of the intermediate subunit and one homologue of the heavy subunit of the GalGalNAc lectin were identified. Ten new cysteine proteinases with predicted N-terminal transmembrane anchors, which might allow them to be localized on the amoeba cell surface, were identified. In addition to three new amoebapores a homologue of haemolysin III was identified, suggesting that, in addition to amoebapores, haemolysins may have a role in host cell lysis.

Vesicle trafficking has a role in *E. histolytica* pathogenesis through phagocytosis and the delivery of secreted hydrolytic enzymes and amoebapores to the cell surface\(^4\). *E. histolytica* lacks morphologically identifiable rough endoplasmic reticulum and the Golgi apparatus\(^5\) but encodes the basic elements of the vesicle transport machinery common to other eukaryotic cells, with the coat complexes COPI, COPII, clathrin and retromer all being present. Rab and Arf protein family expansions reflect the increased complexity and number of vesicle fusion and recycling steps that have been associated with phagocytosis and pinocytosis in amoebae\(^6\). The cytoskeleton has a number of important roles in parasite motility, contact-dependant killing and phagocytosis of host intestinal epithelial cells\(^7\). This is reflected in expansions of Rho GTPases and their regulators RhoGAPs and RhoGEFs, which control a number of processes involving the actin cytoskeleton. Five proteins with a unique domain architecture containing both RhoGEF and ArfGAP domains were identified, suggesting a mechanism for direct
seven-transmembrane receptors. PTEN, phosphatase and tensin homologue; TyrP, tyrosine phosphatase; 7TM receptors, phosphatidylinositol-3,4,5-trisphosphate; PKC, protein kinase C; PLC, phospholipase C; protein kinase superfamily were identified. These include tyrosine kinases representing members of all seven families of the eukaryotic E. histolytica encounters (Fig. 2). Almost 270 putative order to sense and interact with the different environments it cytoskeletal rearrangement.

Communication between the regulators of vesicle budding and cytoskeletal rearrangement. E. histolytica uses a complex mix of signal transduction systems in order to sense and interact with the different environments it encounters (Fig. 2). Almost 270 putative E. histolytica protein kinases representing members of all seven families of the eukaryotic protein kinase superfamily were identified. These include tyrosine kinases with SH2 domains, tyrosine kinase-like protein kinases and 90 putative receptor Ser/Thr kinases. These Ser/Thr kinases are uncommon in protists, appear to be absent from Dictyostelium and have previously been described only in plants, animals and Choano-flagellates. The E. histolytica receptor Ser/Thr kinases all contain an N-terminal signal peptide, a predicted extracellular domain and a single transmembrane helix followed by a cytosolic tyrosine kinase-like domain. The receptor kinases fall into three groups on the basis of differences in their predicted extracellular domains. The first group of 50 receptor kinase proteins contains CXXC-rich repeats similar to those found in the intermediate subunit (Igl) of the Gal/GalNAc lectin and G. lamblia variant-specific surface proteins. A second group of 32 proteins encodes cysteine-rich domains containing CXC repeats. The third group of eight receptor kinase-like proteins lacks cysteine-rich extracellular domains. Although no immediate downstream effectors to the amoebic receptor kinases could be identified, E. histolytica contains greater than 100 protein phosphatases, which dephosphorylate proteins. An unusual feature of some of the phosphatases is the presence of varying numbers of leucine-rich repeat (LRR) domains that are involved primarily in protein–protein interactions and have not previously been associated with phosphatases. The E. histolytica genome encodes numerous putative seven-transmembrane receptors and trimeric G proteins, which are probably involved in mediating autocrine stimulation of encystation. In contrast to autocrine stimulation of Dictyostelium sporulation, which uses secreted cyclic AMP, E. histolytica encystment is self-stimulated by secreted catecholamines. Finally, E. histolytica has numerous cytosolic proteins involved in signal transduction, including Ras-family proteins, seven-transmembrane receptors and trimeric G proteins.

Predicted signal transduction mechanisms of E. histolytica based on analysis of the genome sequence data. E. histolytica possesses three types of receptor serine/threonine kinases: one group has CXCC repeats in the extracellular domain; a second has CXC repeats, and a third has non-cysteine rich (NCR) repeats. E. histolytica has cytosolic tyrosine kinases (TyrK), but not receptor tyrosine kinases. Some serine/threonine phosphatases (S/Tp) have an attached LRR domain. CaBP, calcium-binding protein; DAG, diacylglycerol; G, G protein; GAP, GTPase-activating protein; GEF, guanine nucleotide exchange factor; IP3, inositol-1,4,5-trisphosphate; P(1,4,5)K, phosphatidylinositol-3,4,5-trisphosphate; PIP3, phosphatidylinositol-3,4,5-trisphosphate; PKC, protein kinase C; PLC, phospholipase C; protein kinase superfamily were identified. These include tyrosine kinases representing members of all seven families of the eukaryotic E. histolytica encounters (Fig. 2). Almost 270 putative order to sense and interact with the different environments it cytoskeletal rearrangement.

Communication between the regulators of vesicle budding and cytoskeletal rearrangement. E. histolytica uses a complex mix of signal transduction systems in order to sense and interact with the different environments it encounters (Fig. 2). Almost 270 putative E. histolytica protein kinases representing members of all seven families of the eukaryotic protein kinase superfamily were identified. These include tyrosine kinases with SH2 domains, tyrosine kinase-like protein kinases and 90 putative receptor Ser/Thr kinases. These Ser/Thr kinases are uncommon in protists, appear to be absent from Dictyostelium and have previously been described only in plants, animals and Choano-flagellates. The E. histolytica receptor Ser/Thr kinases all contain an N-terminal signal peptide, a predicted extracellular domain and a single transmembrane helix followed by a cytosolic tyrosine kinase-like domain. The receptor kinases fall into three groups on the basis of differences in their predicted extracellular domains. The first group of 50 receptor kinase proteins contains CXXC-rich repeats similar to those found in the intermediate subunit (Igl) of the Gal/GalNAc lectin and G. lamblia variant-specific surface proteins. A second group of 32 proteins encodes cysteine-rich domains containing CXC repeats. The third group of eight receptor kinase-like proteins lacks cysteine-rich extracellular domains. Although no immediate downstream effectors to the amoebic receptor kinases could be identified, E. histolytica contains greater than 100 protein phosphatases, which dephosphorylate proteins. An unusual feature of some of the phosphatases is the presence of varying numbers of leucine-rich repeat (LRR) domains that are involved primarily in protein–protein interactions and have not previously been associated with phosphatases. The E. histolytica genome encodes numerous putative seven-transmembrane receptors and trimeric G proteins, which are probably involved in mediating autocrine stimulation of encystation. In contrast to autocrine stimulation of Dictyostelium sporulation, which uses secreted cyclic AMP, E. histolytica encystment is self-stimulated by secreted catecholamines. Finally, E. histolytica has numerous cytosolic proteins involved in signal transduction, including Ras-family proteins, seven-transmembrane receptors and trimeric G proteins.

Figure 2 Predicted signal transduction mechanisms of E. histolytica based on analysis of the genome sequence data. E. histolytica possesses three types of receptor serine/threonine kinases: one group has CXCC repeats in the extracellular domain; a second has CXC repeats, and a third has non-cysteine rich (NCR) repeats. E. histolytica has cytosolic tyrosine kinases (TyrK), but not receptor tyrosine kinases. Some serine/threonine phosphatases (S/Tp) have an attached LRR domain. CaBP, calcium-binding protein; DAG, diacylglycerol; G, G protein; GAP, GTPase-activating protein; GEF, guanine nucleotide exchange factor; IP3, inositol-1,4,5-trisphosphate; P(1,4,5)K, phosphatidylinositol-3,4,5-trisphosphate; PIP3, phosphatidylinositol-3,4,5-trisphosphate; PKC, protein kinase C; PLC, phospholipase C; protein kinase superfamily were identified. These include tyrosine kinases representing members of all seven families of the eukaryotic E. histolytica encounters (Fig. 2). Almost 270 putative order to sense and interact with the different environments it cytoskeletal rearrangement.

Communication between the regulators of vesicle budding and cytoskeletal rearrangement. E. histolytica uses a complex mix of signal transduction systems in order to sense and interact with the different environments it encounters (Fig. 2). Almost 270 putative E. histolytica protein kinases representing members of all seven families of the eukaryotic protein kinase superfamily were identified. These include tyrosine kinases with SH2 domains, tyrosine kinase-like protein kinases and 90 putative receptor Ser/Thr kinases. These Ser/Thr kinases are uncommon in protists, appear to be absent from Dictyostelium and have previously been described only in plants, animals and Choano-flagellates. The E. histolytica receptor Ser/Thr kinases all contain an N-terminal signal peptide, a predicted extracellular domain and a single transmembrane helix followed by a cytosolic tyrosine kinase-like domain. The receptor kinases fall into three groups on the basis of differences in their predicted extracellular domains. The first group of 50 receptor kinase proteins contains CXXC-rich repeats similar to those found in the intermediate subunit (Igl) of the Gal/GalNAc lectin and G. lamblia variant-specific surface proteins. A second group of 32 proteins encodes cysteine-rich domains containing CXC repeats. The third group of eight receptor kinase-like proteins lacks cysteine-rich extracellular domains. Although no immediate downstream effectors to the amoebic receptor kinases could be identified, E. histolytica contains greater than 100 protein phosphatases, which dephosphorylate proteins. An unusual feature of some of the phosphatases is the presence of varying numbers of leucine-rich repeat (LRR) domains that are involved primarily in protein–protein interactions and have not previously been associated with phosphatases. The E. histolytica genome encodes numerous putative seven-transmembrane receptors and trimeric G proteins, which are probably involved in mediating autocrine stimulation of encystation. In contrast to autocrine stimulation of Dictyostelium sporulation, which uses secreted cyclic AMP, E. histolytica encystment is self-stimulated by secreted catecholamines. Finally, E. histolytica has numerous cytosolic proteins involved in signal transduction, including Ras-family proteins, seven-transmembrane receptors and trimeric G proteins.
tRNAs were detected using the tRNAscan-SE program with default parameters. The Combiner gene predictions were then manually curated. A set of analyses. Each candidate LGT was analysed by MrBayes using the WAG matrix, a gamma values for these 96 candidate LGT are provided as Supplementary Information.

On the basis of these analyses we identified 96 genes in which the tree topology is consistent with prokaryote to eukaryote LGT. Blast summary statistics, trees and support The specificity of cortical neuron connections creates columns of functionally similar neurons spanning from the pia to the white matter. Here we investigate whether there is an additional, finer level of specificity that creates subnetworks of excitatory neurons within functional columns. We tested for fine-scale specificity of connections to cortical layer 2/3 pyramidal neurons in rat visual cortex by using cross-correlation analyses of synaptic currents evoked by photostimulation. Recording simultaneously from adjacent layer 2/3 pyramidal cells, we find that when they are connected to each other (20% of all recorded pairs) they share common input from layers 4 and within layer 2/3. When adjacent layer 2/3 neurons are not connected to each other, they share very little (if any) common excitatory input from layers 4 and 2/3. In contrast, all layer 2/3 neurons share common excitatory input from the leading edge of motile cells. Mol. Biochem. Parasitol. 135, 185–190 (2004).