

The *Mastigamoeba balamuthi* Genome and the Nature of the Free-Living Ancestor of *Entamoeba*

Vojtěch Žárský,¹ Vladimír Klimeš,² Jan Pačes,³ Čestmír Vlček,³ Miluše Hradilová,³ Vladimír Beneš,⁴ Eva Nývltová,¹ Ivan Hrdý,¹ Jan Pyrih,¹ Jan Mach,¹ Lael Barlow,⁵ Courtney W. Stairs,^{6,7} Laura Eme,⁸ Neil Hall,^{9,10} Marek Eliáš,² Joel B. Dacks,^{5,11,12} Andrew Roger,⁶ and Jan Tachezy *,¹

¹Department of Parasitology, Faculty of Science, Charles University, BIOCEV, Vestec, Czech Republic

²Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czech Republic

³Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

⁴European Molecular Biology Laboratory (EMBL), Genomics Core Facility, Heidelberg, Germany

⁵Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

⁶Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada

⁷Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

⁸Diversity, Ecology and Evolution of Microbes (DEEM), Unité Ecologie Systématique Evolution Université Paris-Saclay, Orsay, France

⁹The Earlham Institute, Norwich Research Park, Norwich, United Kingdom

¹⁰School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom

¹¹Division of Infectious Diseases, Department of Medicine, University of Alberta, Edmonton, AB, Canada

¹²Institute of Parasitology, Biology Centre, CAS, v.v.i., Ceske Budejovice, Czech Republic

*Corresponding author: E-mail: tachezy@natur.cuni.cz.

Associate editor: Miriam Barlow

Abstract

The transition of free-living organisms to parasitic organisms is a mysterious process that occurs in all major eukaryotic lineages. Parasites display seemingly unique features associated with their pathogenicity; however, it is important to distinguish ancestral preconditions to parasitism from truly new parasite-specific functions. Here, we sequenced the genome and transcriptome of anaerobic free-living *Mastigamoeba balamuthi* and performed phylogenomic analysis of four related members of the Archamoebae, including *Entamoeba histolytica*, an important intestinal pathogen of humans. We aimed to trace gene histories throughout the adaptation of the aerobic ancestor of Archamoebae to anaerobiosis and throughout the transition from a free-living to a parasitic lifestyle. These events were associated with massive gene losses that, in parasitic lineages, resulted in a reduction in structural features, complete losses of some metabolic pathways, and a reduction in metabolic complexity. By reconstructing the features of the common ancestor of Archamoebae, we estimated preconditions for the evolution of parasitism in this lineage. The ancestor could apparently form chitinous cysts, possessed proteolytic enzyme machinery, compartmentalized the sulfate activation pathway in mitochondrion-related organelles, and possessed the components for anaerobic energy metabolism. After the split of *Entamoebidae*, this lineage gained genes encoding surface membrane proteins that are involved in host–parasite interactions. In contrast, gene gains identified in the *M. balamuthi* lineage were predominantly associated with polysaccharide catabolic processes. A phylogenetic analysis of acquired genes suggested an essential role of lateral gene transfer in parasite evolution (*Entamoeba*) and in adaptation to anaerobic aquatic sediments (*Mastigamoeba*).

Key words: evolution of parasitism, lateral gene transfer, pathway complexity, Archamoebae, chitinous cysts, *Mastigamoeba*.

Introduction

Parasitic organisms evolved multiple times in all major eukaryotic lineages (Blaxter and Koutsovoulos 2015), but evolutionary steps leading to their transition from free-living ancestors generally remain elusive. Parasites are organisms that employ various strategies to obtain nutritional benefits

for growth and reproduction at the expense of the host. These strategies are reflected by specific features of the parasites that resulted mainly from the character of the free-living ancestor and from the character of the host niche to which the evolving parasite needed to adapt. The best-known parasitic features are those involved in pathogenicity

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

and virulence. Although we often tend to view such features as innovations that evolved specifically in relation to parasitism, a growing list of examples suggests that some of them may be shared by nonparasitic relatives of the parasitic group concerned, and have thus evolved even before the transition to parasitism. Hence, some traits exhibited by free-living organisms may have served as preconditions facilitating the emergence of parasitism in the given lineage or were exapted by the parasitic descendants to foster their newly evolved lifestyle (Ginger and Field 2016; Janouskovec and Keeling 2016). On the other hand, other traits of the free-living ancestors and the genes that underpin them became dispensable once the organism became an obligate parasite that lives in a relatively narrow and nutrient-rich environment (Jackson et al. 2016; Janouskovec and Keeling 2016). This tendency is particularly apparent in intracellular parasites such as microsporidia and apicomplexans, in which reductions in metabolic pathways evolved together with dramatic reductions in the genome (Heinz et al. 2012; Woo et al. 2015; Mathur et al. 2019). However, other parasites, such as parasitic kinetoplastids and ciliates, seem to have maintained most of their canonical physiological functions, and reduction primarily manifests as losses of functionally redundant paralogs (Coyne et al. 2011; Jackson et al. 2016). In addition to gene loss, parasite genomes are shaped by innovations, that is, gains of truly novel parasite-specific functions. Such innovations include expansions in certain gene families and paralog specialization, the appearance of novel genes of unclear origin, and the acquisition of genes by lateral gene transfer (LGT) (Loftus et al. 2005; Carlton et al. 2007; Nývltová et al. 2015; Jackson et al. 2016; Janouskovec and Keeling 2016).

Investigations of gene histories to distinguish evolutionary preconditions from specific parasitic features are essential for understanding parasite evolution. Such studies require comparative analyses of genome sequences of parasitic species and their close free-living relatives. However, although genomes of most important parasites have been sequenced, the genomes of their free-living neighbors are rarely studied, rendering comparative genomic studies impossible. Moreover, those free-living lineages that have been sampled primarily include organisms thriving in aerobic environments, for example, free-living kinetoplastids related to trypanosomatids (Jackson et al. 2016), chrompodellids (Janouskovec et al. 2015) related to apicomplexans (Woo et al. 2015; Mathur et al. 2019), and free-living ciliates related to *Ichthyophthirius multifiliis* (Coyne et al. 2011). Therefore, there exists an unsampled diversity of anaerobic free-living organisms that could provide novel comparative insights into the genome evolution of their parasitic relatives.

Several important human parasites, such as *Entamoeba histolytica*, *Trichomonas vaginalis*, and *Giardia intestinalis*, are anaerobes living under oxygen-limited conditions. *Entamoeba histolytica* is a leading microbial parasitic cause of death and morbidity, with 41,000–74,000 deaths annually (Lozano et al. 2012). *Entamoeba* belongs to the Archamoebae group, which includes not only parasitic species—entamoebids, but also free-living pelobionts such as *Mastigamoeba balamuthi*

(Baptiste et al. 2002). All members of Archamoebae are anaerobes; they have no conventional mitochondria with cristae, and none of the species possess typical Golgi dictyosomes (Walker et al. 2001). Instead, Golgi-like vesicles scattered within the cells have been observed (Barlow et al. 2018). Entamoebids lack flagella, whereas a single flagellum exists in most pelobionts. The flagellum, when present, has a standard axoneme except for the lack of outer dynein arms, and the basal body is often associated with a cone of microtubules covering the nucleus (Walker et al. 2001; Pánek et al. 2016). Phylogenetic analyses revealed that Archamoebae clustered with the aerobic taxa Eumycetozoa and Variosea in the Conosa group (also called Evosea) and that the last common Conosa ancestor (LCCA) was clearly an aerobic free-living protist (Pánek et al. 2016; Kang et al. 2017). Therefore, the Conosa group provides an interesting opportunity to trace the history of parasitism from the aerobic LCCA via the last common anaerobic ancestor of Archamoebae (LCAA) to parasitic *Entamoeba*. In this sense, *M. balamuthi* represents a key species for the reconstruction of LCAA features.

Mastigamoeba balamuthi, originally described as *Phreatamoeba balamuthi*, was isolated from a public well in Gambia (Chavez et al. 1986). It exhibits multinucleated amoeboid stages, uninuclear flagellates, and cysts. Adaptation to anaerobiosis is reflected by the transformation of the mitochondrion into a hydrogenosome that contains typical enzymes of anaerobic energy metabolism (Nývltová et al. 2015), and the presence of an anaerobic peroxisome (Le et al. 2020). There are several unusual features that *M. balamuthi* shares with *E. histolytica* and that were most likely present in the Archamoebae stem lineage or in the LCAA itself. These include the loss of the mitochondrial iron–sulfur cluster (ISC) assembly system, which was replaced by an unrelated nitrogen fixation (NIF) system of ϵ -proteobacterial origin (Gill et al. 2007; Nývltová et al. 2013), and acquisition of components of the sulfate activation pathway that operate in *M. balamuthi* hydrogenosomes and *E. histolytica* mitosomes, the most reduced form of mitochondria (Mi-ichi et al. 2009; Nývltová et al. 2015). On the other hand, *M. balamuthi* exhibits features that are related to its free-living lifestyle and thus absent from *E. histolytica*. One interesting example is the acquisition of enzymes of Clostridiales origin that allow for the production of bactericidal *p*-cresol, presumably to suppress the growth of competitors in the microbial community in which *M. balamuthi* lives (Nývltová et al. 2017). However, the differences between parasitic and free-living Archamoebae have yet to be investigated systematically.

Here, we describe a draft genome assembly of *M. balamuthi* complemented by transcriptome sequencing, and report on analyses of the data aimed at the following key tasks: 1) to characterize the changes associated with the transition of the Archamoebae stem lineage to the anaerobic lifestyle; 2) to define features of free-living ancestors of *Entamoeba* that may have facilitated evolution of parasitism in the *Entamoeba* lineage; and 3) to identify innovations that appeared separately in free-living and parasitic Archamoebae as adaptations to two fundamentally different anaerobic environments.

Table 1. Comparison of Genome Statistics of *Mastigamoeba balamuthi* and Selected *Entamoeba* Species.

| Statistics | <i>Mastigamoeba balamuthi</i> | <i>Entamoeba histolytica</i> ^a | <i>Entamoeba invadens</i> ^a | <i>Entamoeba moshkovskii</i> ^a |
|----------------------------|-------------------------------|---|--|---|
| Genome size (Mb) | 57.2 | 23.7 | 40.8 | 25.2 |
| GC content (%) | 60.7 | 24.3 | 29.9 | 26.5 |
| % Coding DNA | 55.0 | 50.1 | 38.0 | 59.0 |
| Protein cod. genes | 16,287 | 9,938 | 12,007 | 12,449 |
| Av. protein size (aa) | 479.7 | 389 | 431.3 | 399.1 |
| Av. intergenic dist. (kb) | 1.3 | 1.2 | 2.1 | 0.8 |
| Av. intron size (bp) | 137 | 74 | 104 | 89 |
| Av. number of introns/gene | 3.4 | 1.27 | 1.5 | 1.3 |

NOTE.—Aa, amino acids; bp, basepairs; kb, kilobasepairs; Mb, million basepairs.

^aData according to Wilson et al. (2019).

Results and Discussion

The Genome of *Mastigamoeba balamuthi* Encodes up to Twice as Many Genes as Those of *Entamoeba* Species

The genome of axenically grown *M. balamuthi* was sequenced using the 454 GS FLX Titanium (Roche), MiSeq (Illumina), and PacBio RS (Pacific Biosciences) platforms and assembled into 1,926 scaffolds (N50 442.5 kb) at approximately 41× coverage, amounting to 57.27 Mb of DNA sequence (table 1). RNA-seq was used to sequence the *M. balamuthi* transcriptome, which aided in the prediction of 14,840 protein-coding genes. The completeness of the genome assembly was estimated using the Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment (Simão et al. 2015), which identified 82.8% conserved eukaryotic orthologs. The high level of completeness of the assembly was further evidenced by 98.6% of the transcriptome sequences being mapped to the genome using GMAP (Wu and Watanabe 2005). The characteristics of the *M. balamuthi* genome are predictably different from those of *Entamoeba* species (table 1). The *M. balamuthi* genome is larger and codes for 30–96% more genes than the genomes of various *Entamoeba* species. *Mastigamoeba balamuthi* protein-coding genes are richer in introns (3.4 introns/gene), with an average intron size of 137 bp. The genome has a high GC content (60.7%), which is at least twice as high as the GC content of different *Entamoeba* species (fig. 1, table 1, and supplementary table S1, Supplementary Material online). An unusual property of the *Entamoeba* genome is the very high number of tRNA genes (4,500) that form unique linear arrays (Clark et al. 2006). It has been speculated that these arrays may function as telomeres, as classical telomeres are absent in *Entamoeba* (Clark et al. 2006). We found a considerably lower number of tRNA genes (219) in the *M. balamuthi* genome, and they were randomly scattered within the genome rather than forming arrays (supplementary table S2, Supplementary Material online). *Mastigamoeba balamuthi* possesses a full set of genes for major spliceosome RNAs, whereas those specific for the minor (U12) spliceosome are absent, as are genes for minor spliceosome-specific proteins. The loss of the minor spliceosome, and by extension the corresponding intron

type, is shared with *Entamoeba* and contrasts with the situation in some other amoebozoans of the Conosa group, such as *Physarum polycephalum* (López et al. 2008). Further details on the general features of the *M. balamuthi* genome and its structural RNA genes are provided in supplementary Results and Discussion and table S3, Supplementary Material online.

A Survey of the *Mastigamoeba* Genome Illuminates the Character of the LCAA

To trace gene losses and gains that led to diversification of free-living *M. balamuthi* and parasitic *Entamoeba* species, we reconstructed gene histories within Archamoebae and their relatives. First, we collected conserved orthologous groups (OGs) in the predicted proteomes of eight amoebozoans (*M. balamuthi*, *E. histolytica*, *Entamoeba moshkovskii*, *Entamoeba invadens*, *Dictyostelium discoideum*, *Acytostelium subglobosum*, *Protostelium aurantium*, and *Acanthamoeba castellanii*) and four opisthokonts (*Homo sapiens*, *Amphimedon queenslandica*, *Saccharomyces cerevisiae*, and *Batrachochytrium dendrobatidis*) by searching the EggNOG database of OGs using a profile hidden Markov model (HMM)-based search. Based on the evolutionary relationships between these species and assuming vertical inheritance of OGs, we then reconstructed the evolutionary history of each of the OGs: 1) we predicted that a given OG was gained in the last common ancestor of species in which this OG was detected, and 2) we predicted loss of the OG in descending lineages in which the given OG was not detected (fig. 2 and supplementary table S4, Supplementary Material online).

In the Archamoebae lineage, a substantial loss of OGs (1,532) was associated with the first major evolutionary transition: the adoption of the anaerobic/microaerophilic lifestyle. This is supported by statistical evaluation of OG terms using the Ontologizer tool (Bauer et al. 2008), which revealed losses of OG terms for “mitochondrion” with the highest support (5.1e-34), followed by the term “oxidoreductase activity” (5.1e-17). That the mitochondrion of the LCAA was similar to the hydrogenosome of *M. balamuthi*, which lacks an organellar genome and most components of typical mitochondrial metabolism, including the pyruvate dehydrogenase complex (PDC), the TCA cycle, FoF1 ATPase, and all respiratory chain complexes except for complex II (succinate dehydrogenase). These were replaced by a simple anaerobic energy

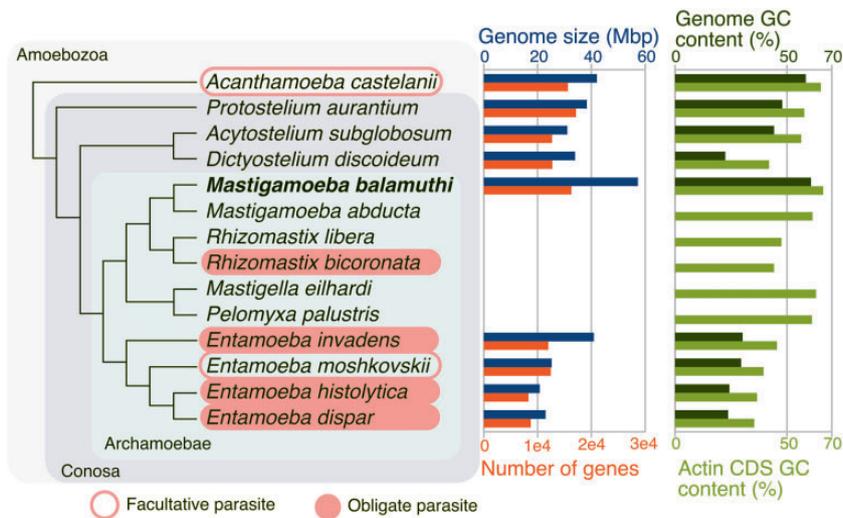


Fig. 1. Comparison of genome sizes, numbers of genes, and GC contents in selected members of Amoebozoa. Because genome sequences are available only for *Entamoeba* spp. and *Mastigamoeba balamuthi*, the GC content was estimated from actin gene sequences to allow for a broader comparison across the Archamoebae group.

metabolism based on substrate-level phosphorylation (Gill et al. 2007; Nývltová et al. 2015).

As the presence of mitochondrial cristae is correlated with oxidative phosphorylation (Davies et al. 2011), we predict that the mitochondria of the LCAA did not have cristae. This is supported by the lack of all components of the mitochondrial contact site and cristae organizing system (MICOS) in both *M. balamuthi* and *Entamoeba*. Concomitant with oxidative phosphorylation loss is the loss of electrochemical potential at the inner mitochondrial membrane, which is crucial for the import of proteins into aerobic mitochondria. This is reflected by reduction and modifications of the protein import machinery (Pyrihová et al. 2018; Makki et al. 2019). In *E. histolytica*, only the Tom40 component of the outer mitochondrial translocase (TOM) is known to exist together with the taxon-specific subunit Tom60 (Makiuchi et al. 2013) and the β -barrel insertion protein Sam50 (Dolezal et al. 2010), whereas no component of the inner membrane translocase (TIM) has been found so far. We expected that, similarly to the hydrogenosome of *T. vaginalis*, the *M. balamuthi* hydrogenosome may have retained more standard translocation machinery; however, only Tom40 and Sam50 were identified. We did not detect any components of TIM, which are likely too divergent in all Archamoebae to be recognized by current bioinformatic tools. However, we identified a standard mitochondrial protein-processing peptidase (MPP) that is required for the maturation of proteins imported into the mitochondrial matrix and is absent in *E. histolytica*. Unexpectedly, we also identified two paralogs of the GTPase MIRO (m51a1_g9410, and m51a1_g12175). Although MIRO is widely distributed in eukaryotes, it is absent in anaerobic protists with hydrogenosomes and mitosomes, including *Entamoeba* (Vlahou et al. 2011). To date, the stramenopile *Blastocystis* is the only known anaerobic organism with MIRO (Gentekaki et al. 2017), and its anaerobic mitochondria are generally much less modified than the hydrogenosome of *M. balamuthi*. MIRO is implicated in

numerous functions, such as interaction with the endoplasmic reticulum (ER) via the ER–mitochondria encounter structure (ERMES), and peroxisomal movement (Eberhardt et al. 2020). We did not identify ERMES components in the *M. balamuthi* genome and MIRO is absent from the hydrogenosomal proteome (Le et al. 2020). However, one of the paralogs (m51a1_g9410) is among the proteins detected in the proteome of *M. balamuthi* peroxisomes (Le et al. 2020), suggesting that its function might be related to these organelles instead of the hydrogenosomes.

Changes in physiological functions upon adaptation to anaerobiosis are further documented by correlation between the number of lost OGs and the corresponding COG categories (fig. 2). In addition to the reduction in pathways of energy metabolism, this analysis also pointed to a strong reduction in the COG category “lipid transport and metabolism” (fig. 2), which includes both biosynthetic and catabolic pathways. The data suggest the complete loss of fatty acid (FA) β -oxidation components, including acyl-CoA dehydrogenase, which in mitochondria is coupled with the respiratory chain to maintain redox balance and contributes to ATP synthesis. Alternatively, β -oxidation may be localized in peroxisomes where FA degradation is initiated by oxygen-dependent acyl-CoA oxidase. However, genes encoding all oxygen-dependent enzymes in peroxisomes, such as acyl-CoA oxidase and also catalase, were possibly lost no later than in the LCAA, as they are missing from the anaerobic peroxisomes of *M. balamuthi* (Le et al. 2020). The FA biosynthesis pathway is predicted to have been lost prior to the emergence of the LCAA, as both *E. histolytica* and *M. balamuthi* do not encode enzymes dedicated to de novo FA synthesis, including multifunctional FA synthase (type I FAS), which are all present in *D. discoideum*. However, similarly to *E. histolytica* (Castellanos-Castro et al. 2020), *Mastigamoeba* encodes enzymes of FA elongation (supplementary fig. S1, Supplementary Material online), which need a source of malonyl-CoA. Neither *Entamoeba* nor *Mastigamoeba* have the conventional

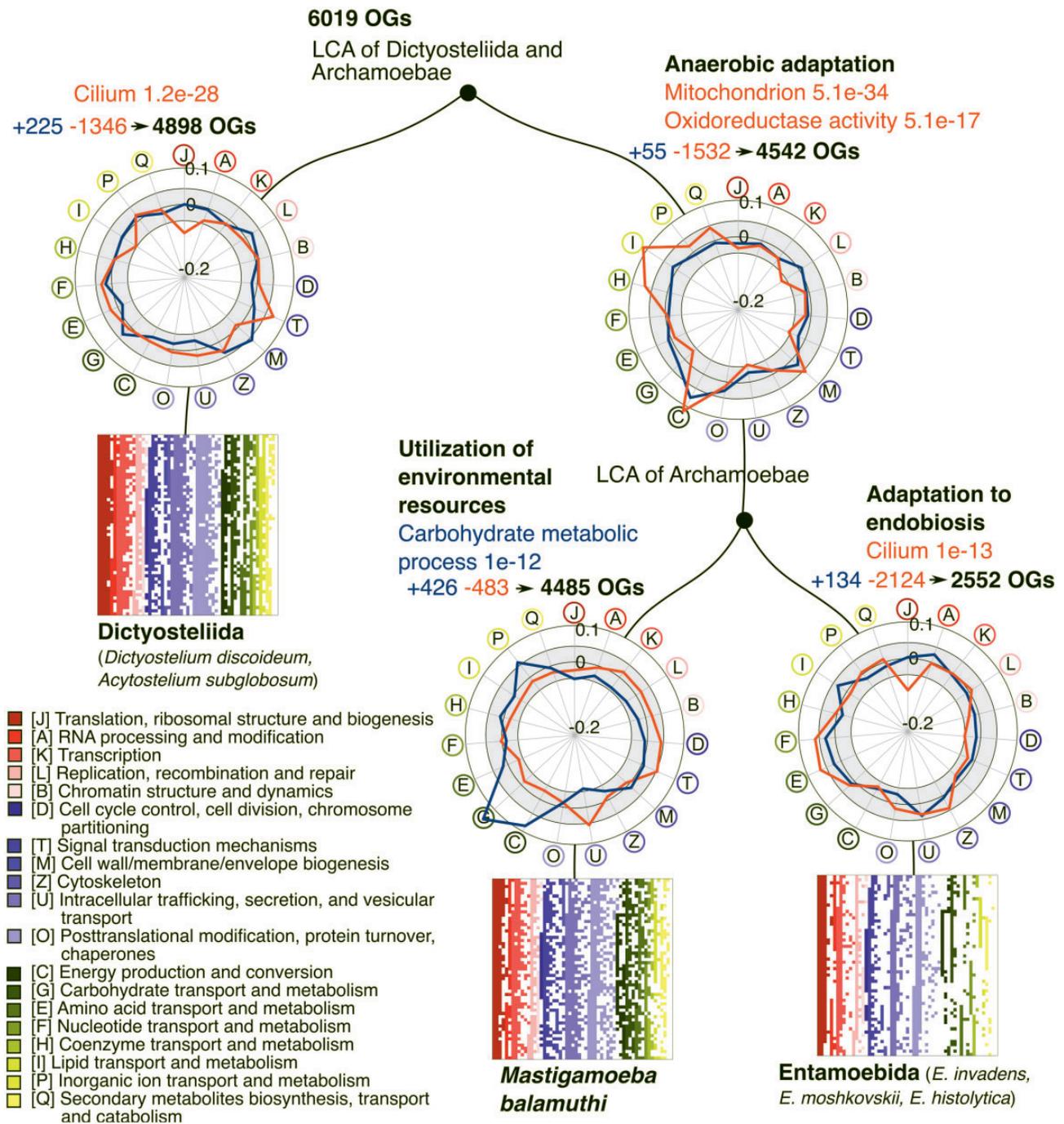


Fig. 2. Losses and gains of orthologous groups in dictyostelids and Archamoebae. Predicted proteins of amoebozoans and selected representatives of opisthokonts were assigned to the orthologous groups (OGs) of the EggNOG database (Powell et al. 2014) using profile-HMM homology searches (HMMER). Gains and losses of OGs were reconstructed for each node of the organismal phylogeny with the assumption of vertical inheritance of these OGs. For each branch these parameters are shown: 1) number of gains (in blue) and losses (in red) of OGs that were predicted to occur at that branch, and the resulting predicted number of OGs (in black). 2) Specific terms of the Gene Ontology enriched in the gained (blue) and lost (red) groups with high statistical significance using the Ontologizer tool (Bauer et al. 2008). Statistical significance is shown as the P value next to the term. 3) Net graphs showing the negative logarithms of the P values describing the statistical significance of the associations between the COG database functional categories (Tatusov et al. 2003), and gains (blue line) and losses (red line). The negative values represent an inverse proportion, that is, a negative association between the categories. At the terminal branches color-coded presence/absence plots of OGs with an assigned COG functional category are shown that demonstrate the overall OG reduction in *Entamoeba*. The description of the COG functional categories is shown in the bottom left corner.

enzymes for malonyl-CoA synthesis. However, it was previously demonstrated that *E. histolytica* possesses an alternative protein that functions as transcarboxylase catalyzing the synthesis of malonyl-CoA from acetyl-CoA and oxaloacetate (Clark et al. 2007; Barbosa-Cabrera et al. 2012), and *M. balamuthi* encodes an ortholog of this protein (supplementary fig. S1, Supplementary Material online). Therefore, similar to *Entamoeba*, *M. balamuthi* needs to acquire fatty acids from food; however, it has the capacity to elongate those fatty acids to very long-chain fatty acids that may not be available in external sources.

As with other anaerobic protists and intracellular parasites with reduced mitochondria (Kořený et al. 2013), *M. balamuthi* has lost haem synthesis. A possible reason is that at least one biosynthetic step in mitochondria requires oxygen that is limited or absent in anaerobic mitochondria/hydrogenosomes. Nevertheless, anaerobes do commonly contain several haem-binding proteins (Kořený et al. 2013; Pyrih et al. 2014). In particular, a soluble cytochrome b_5 of unknown function is present in most anaerobes including *M. balamuthi*. In addition, *M. balamuthi* possesses two other proteins fused with the cytochrome b_5 domain, cytochrome P450, a haem-binding subunit of the hydrogenosomal succinate dehydrogenase, a putative ferric reductase, and proteins with Per-ARNT-Sim (PAS) domains (supplementary table S5, Supplementary Material online). Therefore, like other anaerobes, *M. balamuthi* seems to be dependent on the acquisition of external haem (Pyrih et al. 2014). The possible external sources can include decomposed plant debris and phagocytosed microbes. Interestingly, no haem proteins have so far been identified in *E. histolytica* (Kořený et al. 2013), although it secretes two proteins that could function as haemophores to scavenge haem from environment (Cruz-Castañeda et al. 2011). However, acquired haem may not serve as a prosthetic group; rather, it represents a source of iron (Cruz-Castañeda et al. 2011).

As reported earlier, *M. balamuthi* lost mitochondrial ISC assembly, which was replaced by the NIF system (Gill et al. 2007). This system was most likely acquired via LGT from anaerobic ϵ -proteobacteria living in the same environment (Gill et al. 2007). There are two copies of the NIF system in *M. balamuthi*, one operating in the hydrogenosome and the second in the cytosol (Nylvtova et al. 2013). Interestingly, both *M. balamuthi* and *Entamoeba* possess a complete standard pathway mediating cytosolic FeS cluster assembly (CIA) (supplementary table S5, Supplementary Material online). In organisms with the ISC machinery, the CIA pathway is dependent on an FeS intermediate formed in mitochondria (Pandey et al. 2019). Since the dual localization of the NIF system and the absence of the ISC are inferred characteristics of the LCAA, we can expect that the NIF machinery interacts directly with CIA scaffold proteins such as Nbp35. In support of this hypothesis, the available data suggest that in *M. balamuthi*, Nbp35 has dual localization similar to the NIF system. Specifically, we identified three Nbp35 paralogs, one of which possesses a hydrogenosomal targeting

sequence, and whose presence in hydrogenosomes was supported by mass spectrometry (Le et al. 2020).

The most prominent evolutionary change in signaling pathways of the Archamoebae group concerns eukaryotic histidine kinases that serve as primary sensor proteins. We found that *M. balamuthi*, like *E. histolytica*, has no member of the eukaryotic families of histidine kinases, although they occur in other amoebozoans, including representatives of Discosea and Eumycetozoa (Kabbara et al. 2019). Hence, their absence from Archamoebae is most likely due to loss that occurred before the LCAA. However, this was not necessarily associated with adaptation to anaerobiosis or parasitism, as the absence of histidine kinases was noticed in several eukaryotic groups such as Metazoa, Apicomplexa, and Euglenozoa (Kabbara et al. 2019). Interestingly, we identified three FlxL-like histidine kinases with a generic structure comprising a PAS domain, a histidine kinase domain, and a receiver domain. Phylogenetic analysis revealed that these histidine kinases cluster with high statistical support within a group of bacterial homologs, which indicates acquisition from bacteria via LGT (supplementary fig. S2, Supplementary Material online). The closest bacterial homologs of these *M. balamuthi* histidine kinases serve as regulators that sense molecular oxygen through haem moiety to suppress the expression of oxygen-sensitive pathways (Gilles-Gonzalez et al. 1991). We can speculate that such a function would be beneficial for *M. balamuthi* as part of its defensive mechanisms against oxygen stress.

Adaptation of *M. balamuthi* to an Anaerobic Free-Living Lifestyle

Mastigamoeba balamuthi lives in low-oxygen organic matter-rich freshwater sediments. The plant debris that is freely present in such a niche is decomposed mainly by saprophytic fungi and bacteria (Artzi, Bayer, and Morais 2017; Cheng et al. 2018). Interestingly, analysis of gene histories in *M. balamuthi* inferred the presence of a large set of genes that were gained to optimize *M. balamuthi* metabolism within this specific niche (supplementary table S6, Supplementary Material online). The OG term analysis revealed the highest statistical support ($1e-12$) for gains of genes involved in “carbohydrate metabolic processes” (fig. 2). Specifically, *M. balamuthi* gained genes encoding enzymes for plant cell wall degradation, including enzymes for the degradation of cellulose, pectins, and pentose-based hemicelluloses (xylans and arabinans) (supplementary fig. S3, Supplementary Material online). The cellulosytic enzymes include endoglucanase, exoglucanase, and cellobiase, which convert cellulose to glucose. The presence of xylose and arabinose isomerases suggests that *M. balamuthi* is capable of catabolizing pentoses via isomerization reactions typical for bacteria (Fang et al. 2018) (supplementary fig. S3, Supplementary Material online). An interesting enzyme of the xylose degradation is bifunctional xylulose 5-phosphate/fructose 6-phosphate phosphoketolase (Xfp), which converts D-xylulose 5-phosphate to D-glyceraldehyde and acetyl phosphate (acetyl-P) and thus links xylose

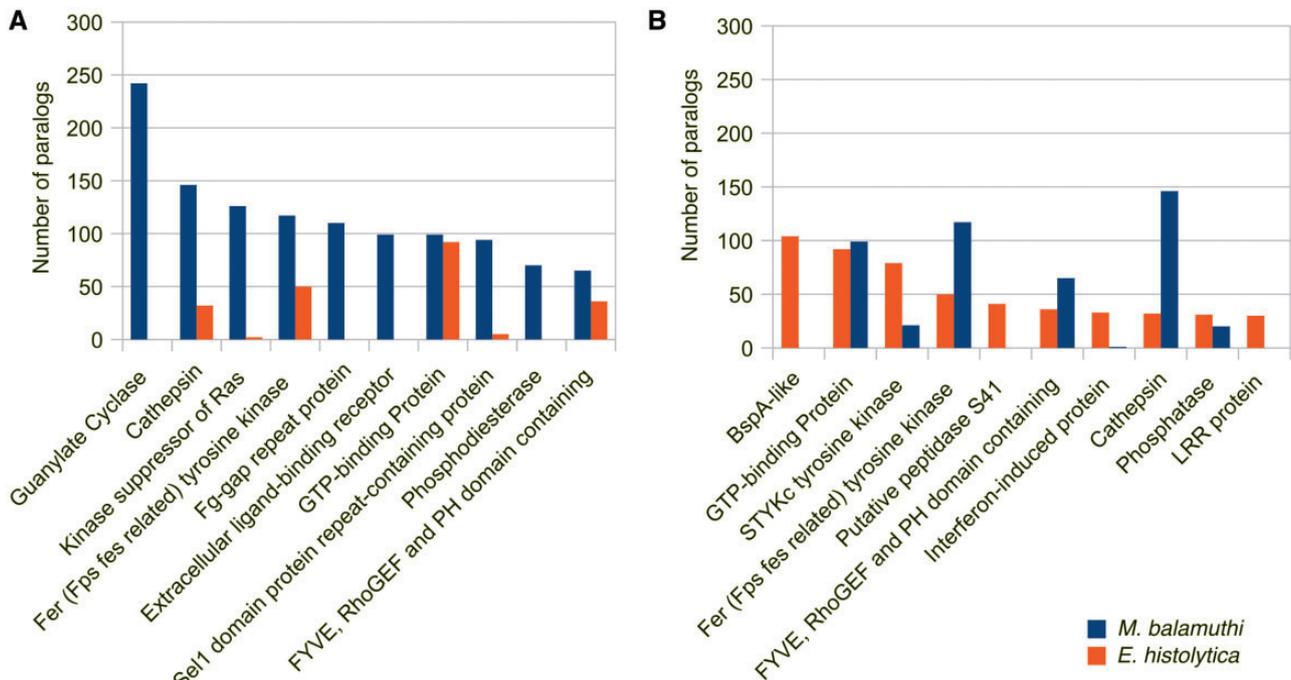


Fig. 3. The ten largest EggNOG orthologous groups of *Mastigamoeba balamuthi* (A) and *Entamoeba histolytica* (B). Note that some of the most abundant orthologous groups of *M. balamuthi* (blue bars) are completely missing in *E. histolytica* (red bars) and vice versa.

metabolism with glycolysis. Alternatively, it metabolizes fructose 6-phosphate to D-erythrose 4-phosphate and acetyl-P. Acetyl-P is then a substrate for acetate kinase (Ack), which produces acetate and pyrophosphate. This so-called pentose phosphoketolase pathway, previously reported in only some bacteria and fungi (Glenn et al. 2014), is also predicted to operate in *M. balamuthi* (supplementary fig. S3, Supplementary Material online). Interestingly, Ack was characterized in *E. histolytica* (Reeves and Guthrie 1975), but enzymes synthesizing acetyl-P, such as Xfp, are absent; thus, the physiological role of Ack in *E. histolytica* is unknown (Pineda et al. 2016).

When considering the largest OGs of *M. balamuthi* and *E. histolytica*, we observed an expansion of proteins involved in environmental sensing, interactions, and signaling (fig. 3). Both organisms have expanded sets of nontransmembrane tyrosine kinases of Fer family, small GTPases, their exchange factors, and cathepsins. *Mastigamoeba balamuthi* has a unique expansion of guanylate cyclases and cGMP-dependent phosphodiesterases, indicating intricate cGMP-dependent signaling.

Genome Streamlining in the *Entamoeba* Lineage

The transition to endobiosis/parasitism in the *Entamoeba* lineage was the second major evolutionary step in Archamoebae evolution, and was accompanied by a massive loss of OGs (2,552). These losses correspond either to the complete absence of certain physiological or structural features, or to reduced complexity, mainly in metabolic pathways. Here, we discuss some prominent or interesting

examples of reductive evolution in *Entamoeba* illuminated by our analyses.

Loss of Flagellum

Structurally, the most prominent change in the *Entamoeba* lineage is the loss of the flagellum and the associated microtubular cone that subtends the nucleus and was suggested as a synapomorphy for the Conosa group (Cavalier-Smith 1998). This is supported by the significantly reduced OG term for “cilium” ($P=1e-13$) in *Entamoeba* compared with *M. balamuthi*. Our search for flagellar components in the *M. balamuthi* genome revealed a complex set of proteins associated with the centriole, basal body, BBSome composed of Bardet–Biedl syndrome proteins (Nachury et al. 2007), and axoneme (supplementary fig. S4 and table S7, Supplementary Material online). Interestingly, *M. balamuthi* possesses an almost complete set of BBSome proteins, including an ortholog of BBS6, which has so far been reported only from metazoans and is absent in flagellated fungi. We did not detect any outer dynein arm components, in keeping with an analysis of an earlier genome draft (Kollmar 2016) and the apparent absence of outer dynein arms in the *M. balamuthi* axoneme as observed by transmission electron microscopy (Pánek et al. 2016). In the *Entamoeba* lineage, all components associated with the biogenesis and function of the flagellum were completely lost (supplementary fig. S4, Supplementary Material online). The loss of the flagellum is most likely associated with the character of the environment. *Mastigamoeba balamuthi* lives primarily as an aflagellated slowly crawling trophozoite, but upon specific environmental changes, it can enter a swimming unflagellated nondividing stage

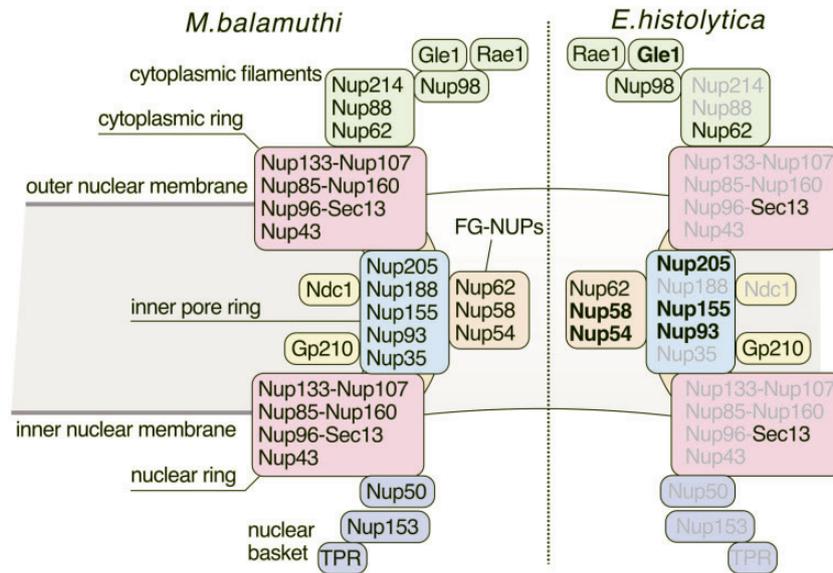


Fig. 4. Comparison of nuclear pore complexes in *Mastigamoeba balamuthi* and *Entamoeba histolytica*. Components of cytoplasmic filaments are in green boxes, Nups of cytoplasmic, and inner ring complexes are in pink boxes, Nups of the nuclear ring complex are in blue, FG-Nups of the central channel are in orange, the transmembrane Nups are in yellow boxes, and nuclear basket components are in violet. NPC subunits of *E. histolytica* identified in this work are in red. TPR, translocated promoter region; Ndc1, nuclear division 1; Gp201, glycoprotein 210; Rae1, mRNA export factor 1; Gle1, glycine-leucine-phenylalanine-glycine lethal. Topology of the components is according to Beck and Hurt (2017).

(Chavez et al. 1986), which presumably facilitates migration of the organism to a new spot. However, intestinal Archamoebae live in a narrowly defined stable environment that they can only exit as cysts, making the flagellated form dispensable.

The Nuclear Pore Complex

M. balamuthi trophozoites in axenic culture possess one to tens of pyriform nuclei with typical pores arranged in rows between cone microtubules (Chavez et al. 1986). The nuclear pore complex (NPC) typically consists of approximately 30 nucleoporins (Nups) that form cytoplasmic fibrils; the inner pore ring, including scaffold subunits and phenylalanine-glycine repeat-nucleoporins (FG-Nups) of the central channel; nuclear rings; and the peripheral basket (Beck and Hurt 2017). We identified a near-complete set of Nups (23 proteins) in the *M. balamuthi* genome (fig. 4 and supplementary table S8, Supplementary Material online). These include the transmembrane protein Ndc1, the scaffold subunits Nup205 and Nup188, and the basket subunit Nup153, which were previously suggested to be lost in Amoebozoa (Neumann et al. 2010). The finding of Ndc1 in Amoebozoa supports the previous prediction that Ndc1 together with Gp210 and Nup35 represent an ancestral system anchoring the central core within the nuclear membrane (Neumann et al. 2010). In contrast, most Nups seem to be lost in *E. histolytica*, in which only five have been reported before. In particular, no scaffold protein was identified, and only one candidate FG-Nup of the three conserved FG-Nup proteins in eukaryotes was previously predicted (Neumann et al. 2010). Therefore, we used profile HMMs built on *M. balamuthi* Nups to search for homologs of these proteins in *E. histolytica*. We discovered six more Nups, including three scaffold protein components

of the inner pore ring and a complete set of FG-Nups in *E. histolytica* (fig. 4). However, most components of the cytoplasmic and nuclear rings, and all basket components seem to be absent. Reductions in the numbers of Nups are often observed in parasitic species, for example, the microsporidian *Encephalitozoon cuniculi*, the metamonads *G. intestinalis* and *T. vaginalis*, and the apicomplexan *Theileria parva* (Mans et al. 2004; Neumann et al. 2010). However, the causal relationship, if any, between NPC and adaptation to parasitism remains to be elucidated. We also cannot rule out that Nups in these parasitic species are too divergent to be detected by the standard in silico approaches. Continued exploration of diverse, free-living protists will undoubtedly aid in our understanding of NPC in eukaryotes.

Membrane Trafficking Complexes

A comparison of membrane trafficking complexes revealed that *M. balamuthi* retains most components found in the free-living *D. discoideum* and *A. castellanii*, whereas the encoded *E. histolytica* complement reflected both losses and expansions after the split from the *Mastigamoeba* lineage. We were able to identify only a single case where *M. balamuthi* shared a partial loss of a complex with *Entamoeba*. The TSET complex that was recently described to be involved in endocytosis from the cell surface is complete in *D. discoideum*, present in part in *M. balamuthi*, and entirely absent from *E. histolytica* (supplementary fig. S5 and table S9, Supplementary Material online). Notably, the *Entamoeba* complement seems to reflect remodeling rather than reduction, with losses and expansions in functionally interconnected machinery. The TRAPP III tethering complex, which acts in the late Golgi, is absent but there are increased paralogs of the COPI complex with which it interacts. In the

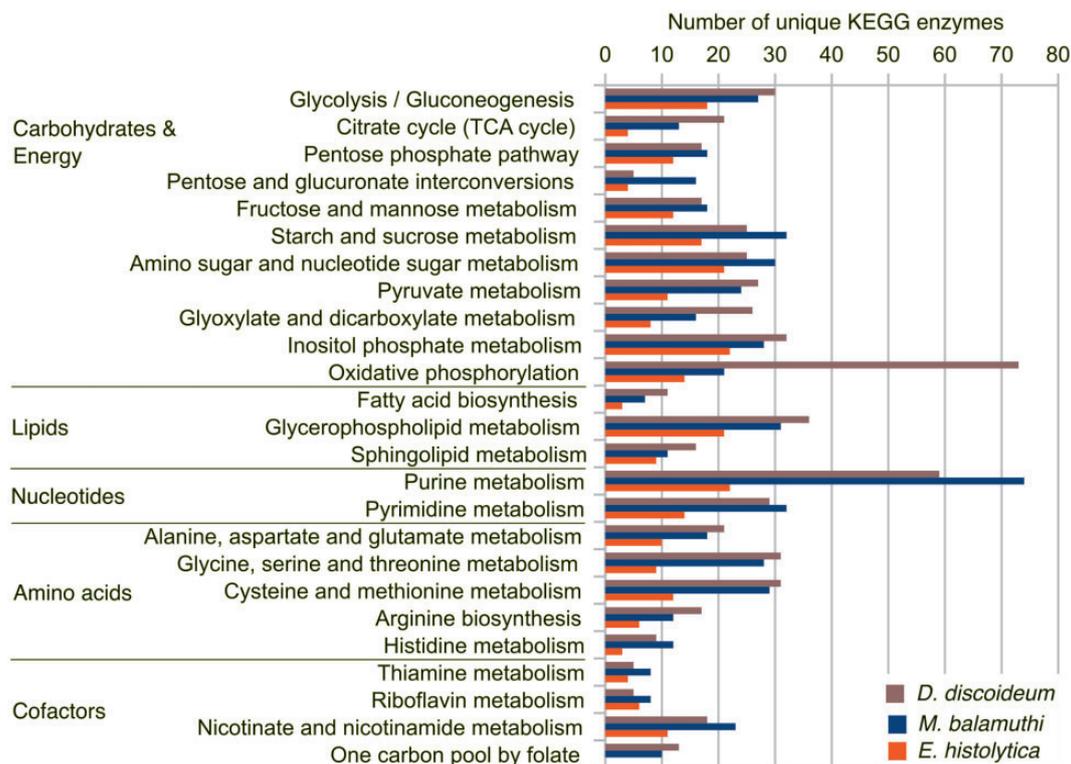


Fig. 5. Comparison of metabolic pathway complexity in *Dictyostelium discoideum*, *Mastigamoeba balamuthi*, and *Entamoeba histolytica* using KEGG pathway mapping. KEGG enzymes were detected in the genomes using KofamKOALA (Aramaki et al. 2020). Numbers of unique enzymes (not counting paralogs) of each KEGG category are shown.

endocytic machinery, there are expansions in the number of AP2 complex subunits, but loss of the TSET complex, both of which mediate internalization from the cell surface. There is a loss of HOPS and CORVET-specific subunits that act at early and late endosomes (and act in the process of phagosomal maturation) but an expansion in retromer subunits and retention of all five AP complexes, even the poorly conserved AP5 complex that acts at the late endosome. Our analysis and comparison with *M. balamuthi* serves to highlight sets of machinery in *Entamoeba* where lineage-specific modulation has taken place, and thus are important areas for molecular cell biological investigation to understand *Entamoeba*-specific biology.

Metabolic Pathways

To compare the complexities of metabolic pathways, we mapped enzyme-coding genes of *D. discoideum*, *M. balamuthi*, and *E. histolytica* to Kyoto Encyclopedia of Genes and Genome (KEGG) pathways. We mapped 802 unique enzymes encoded in the *D. discoideum* genome, 677 in *M. balamuthi*, and only 301 in *E. histolytica* (fig. 5). The reduction in the unique enzyme spectrum in *E. histolytica* was distributed across the majority of pathways tested. The most striking losses are apparent in nucleotide metabolism. Whereas *M. balamuthi* possesses pathways for de novo purine and pyrimidine synthesis (supplementary fig. S6, Supplementary Material online), these pathways are absent in *E. histolytica*, and the parasite entirely relies on nucleotide salvage pathways (Clark et al. 2007). It has been shown that

Entamoeba is able to import external nucleosides and secrete nucleases that hydrolyze RNA (Das et al. 1997; McGugan et al. 2007). Thus, the inability to synthesize nucleotides could be compensated by their acquisition from the nutrient-rich environment of the host. *Mastigamoeba balamuthi* also possesses a ribonucleotide reductase that was lost in *E. histolytica* but retained in *E. moshkovskii* and *E. invadens* (Loftus et al. 2005). Furthermore, *E. histolytica* has substantial reductions in amino acid synthesis pathways compared with *M. balamuthi* (supplementary fig. S6, Supplementary Material online) (Clark et al. 2007).

Next, we were interested to find the extent to which the reduction in biosynthesis of nucleotides and amino acids pathways observed in *E. histolytica* is similar or different in comparison to losses described in other related pairs of free-living and parasitic taxa, specifically the autotrophic *Chromera velia* versus the apicomplexan parasite *Toxoplasma gondii* (Woo et al. 2015) and free-living (*Bodo saltans*) versus parasitic (*Trypanosoma brucei*) kinetoplastids (Jackson et al. 2016). Therefore, we mapped enzyme-coding genes of the selected protists to KEGG pathways and filtered enzymes for the biosynthesis of amino acids, purines, and pyrimidines. As expected, the comparison of de novo synthesis of amino acids and nucleotides between selected free-living protists and their parasitic relatives showed a general tendency to reduce the biosynthetic capacity of the latter taxa (supplementary table 10, Supplementary Material online). The losses of amino acid synthesis included the loss of histidine, glycine, proline, and threonine synthesis in *E. histolytica*; arginine, histidine, aromatic acids, and branch

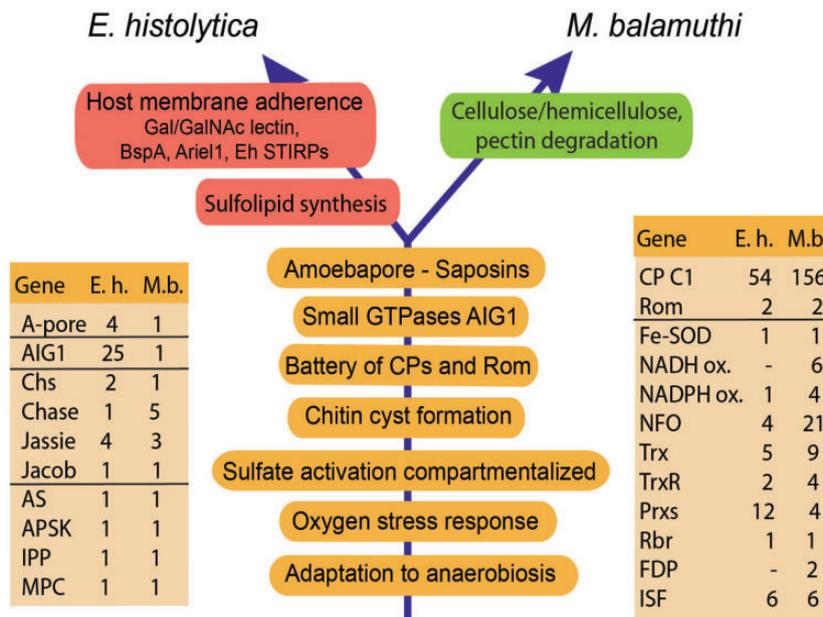


FIG. 6. Summary of features representing preconditions and adaptations for parasitic and free-living lifestyles in anaerobic niches. Predicted preconditions are in brown, *Entamoeba histolytica* specific features are in red, and *Mastigamoeba balamuthi* specific features are in green. A-pore, amoebapore/saplip; Chs, chitin synthase; Chase, chitinase; AS, ATP sulfurylase; APSK, adenosine phosphosulfate kinase; IPP, inorganic pyrophosphatase; MCF, mitochondrial carrier; SPLP, saposin-like protein; CL, cysteine protease cathepsin L; CF, cysteine protease cathepsin F; CP C1, cysteine protease C1 family; Rom, rhomboid protease; Fe-SOD, iron-dependent superoxide dismutase; NADH ox., NADH oxidase; NADPH ox., NADPH oxidase; NFO, NADPH: flavin oxidoreductase; Prxs, peroxiredoxins; Rbr, rubrerythrin; Trx, thioredoxin; TrxR, thioredoxin reductase; FDP, flavodiiron protein; ISF, FeS flavoprotein.

chain amino acids in *To. gondii*; and glycine formation via hydroxymethyltransferase, and cysteine synthesis from serine in *Tr. brucei*. None of the parasitic species synthesize purine, however pyrimidine de novo synthesis was lost only in *E. histolytica*. A comparison of the pathways is further discussed in supplementary Results and Discussion, [Supplementary Material](#) online. Importantly, the particular enzymes lost in the parasitic species are highly varied, reflecting significant differences in the synthetic capacities of the corresponding free-living ancestors (approximated by the free-living relatives of the parasitic taxa compared). Importantly, *Tr. brucei* revealed minimal losses of complete pathways when compared with *B. saltans* (Jackson et al. 2016). In this case, the reduction of most biosynthetic pathways occurred already in free-living bodonids without relation to the parasitism (Butenko et al. 2020). The possible reason is that bodonids, as a heterotroph preying on other organisms, may acquire essential nutrients from this source (Payne and Loomis 2006). Similarly, *E. histolytica* is highly active in phagocytosis of the host gut bacteria (Iyer et al. 2019). Thus, the access of *E. histolytica* to nutrients from bacterial sources may have contributed to the loss of dispensable biosynthetic pathways.

Preconditions to Parasitism

Comparison of the free-living *M. balamuthi* and its parasitic relatives provides an opportunity to define the features of the free-living ancestors of *Entamoeba* that may have facilitated the transition to parasitism in this lineage; that is, that may be

considered putative preconditions for parasitism (Janouskovec and Keeling 2016). We also identified features that were either specifically gained in the parasitic *Entamoeba* lineage or had specific properties associated with the free-living lifestyle of *M. balamuthi* (fig. 6). As we discuss below, the key putative preconditions for the evolution of parasitism in Archamoebae include the inferred ability of the LCAA to live under anaerobic conditions, the ability to form resistant cysts, proteolytic enzyme equipment, and compartmentalization of the sulfate activation pathway.

Adaptation to anaerobiosis is certainly advantageous for colonizing low-oxygen niches such as the human gut. Nevertheless, an efficient antioxidant system is still needed during periods of aerobic stress caused by the host immune response, and during the invasion of oxygenated tissues. *Mastigamoeba balamuthi* and *E. histolytica* share antioxidant system characteristics with anaerobic and microaerophilic eukaryotic protists (supplementary table S11, [Supplementary Material](#) online). Their antioxidant machinery relies on the thioredoxin-based system (thioredoxin, NADPH: flavin oxidoreductase, peroxiredoxin), Fe-superoxide dismutase, rubrerythrin, and flavodiiron proteins (fig. 6). Interestingly, *M. balamuthi* also possesses a homolog of the common bacterial osmotically inducible protein C (OsmC), which may serve as a peroxidase (Nývltová et al. 2016), and multiple homologs of hemerythrin may be involved in oxygen sensing (Xiong et al. 2000). However, both *M. balamuthi* and *E. histolytica*, and hence probably the LCAA before them, lack glutathione-based pathways and catalases (fig. 6).

Another crucial precondition for parasitism in Archamoebae is the ability to form resistant chitinous cysts. *Mastigamoeba balamuthi* forms cysts that allow survival under adverse conditions, such as oxygen exposure or desiccation stress. The cysts are periodic acid-Schiff positive (Chavez et al. 1986) and labeled with calcofluor white (fig. 7), which is used to detect chitin and other polysaccharides. This is reminiscent of the chitinous cysts that have been studied in *E. histolytica* and *E. invadens* (Chatterjee et al. 2009). The *Entamoeba* cyst wall is known to be composed of chitin and a unique set of lectins, including chitinase and Jacob and Jessie lectins with cysteine-rich chitin-binding domains (CBDs). In *M. balamuthi*, we found a complete pathway for chitin synthesis and degradation and Jessie and Jacob lectins, which were previously thought to be unique to *Entamoeba* (Van Dellen et al. 2002) (supplementary figs. S7–S11, Supplementary Material online). We identified a single *M. balamuthi* chitin synthase (chs) with a C-terminal transmembrane domain (TMD) that is in a monophyletic group with the membrane-anchored chs-1 of *E. invadens* (supplementary fig. S8, Supplementary Material online). There are five *M. balamuthi* chitinases, of which three possess an N-terminal CBD, whereas the other two chitinases have no CBD but possess a TMD domain at the C-terminus (supplementary fig. S9, Supplementary Material online). A common feature of Jessie lectins is the presence of an N-terminal CBD with eight conserved cysteine residues. *Entamoeba* Jessie-1 and Jessie-2 are short Jessie variants, whereas Jessie-3 contains an additional C-terminal specific domain downstream of the CBD. We found three orthologs of Jessie-3 in *M. balamuthi* based on a phylogenetic analysis of CBDs (supplementary fig. S10A, Supplementary Material online): a short variant (MbJessie-A), a long variant with a C-terminal domain of unknown function (MbJessie-B), and another long variant (MbJessie-C) with the Jessie-3 specific C-terminal domain (supplementary fig. S10A, Supplementary Material online). Homology searches revealed that this domain corresponds to a catalytic domain of a prokaryotic chitinase whose structure was recently determined (supplementary fig. S10B, Supplementary Material online) (Nishitani et al. 2018). The four proposed catalytic residues (Glu532, Asp566, Glu572, His693) are conserved in all Jessie-3 proteins and in MbJessie-C (supplementary fig. S10C, Supplementary Material online). *Entamoeba* Jacob lectins have a variable number (two to six) of CBDs (Frisardi et al. 2000). *Mastigamoeba balamuthi* possesses a single Jacob ortholog with two CBDs and a C-terminal TMD (supplementary fig. S11, Supplementary Material online). Although *Entamoeba* and *M. balamuthi* possess a similar set of lectins, the cyst wall structure may differ in detail. It has been proposed that during the initial stage of encystation, Jacob is bound to the surface of encysting *Entamoeba* by the membrane-anchored Gal/GalNAc lectin, then chitin is synthesized, and finally, the wall is solidified by the addition of Jessie (Chatterjee et al. 2009). However, the Gal/GalNAc lectin is absent in *M. balamuthi* (fig. 6). Notably, unlike the *Entamoeba* Jessie and Jacob proteins, MbJessie-C and *M. balamuthi* Jacob possess C-terminal TMDs, which might function as direct

anchors to the plasma membrane instead of Gal/GalNAc lectin. The chitin cyst wall is possibly degraded during excystation (Chavez et al. 1986) by chitinase activity (fig. 7). The use of indirect attachment of *Entamoeba* chitin-binding moieties through the Gal/GalNAc lectins may be a useful innovation that allows rapid hatching through perforation of the cyst wall instead of its degradation (Makioka et al. 2005). Another difference between *M. balamuthi* and its parasitic relatives is that the latter lacks homologs of chitin deacetylases previously characterized from *Entamoeba* (Das et al. 2006), suggesting that in contrast to *Entamoeba*, *M. balamuthi* lacks chitosan (deacetylated chitin) in the cyst cell wall. Altogether, these findings suggest that the free-living LCAA could likely form chitin-based cysts, and that this ability may preadapt the parasitic *Entamoeba* ancestors to survive passage through the host stomach and in the outer environment.

The LCAA most likely possessed the sulfate activation pathway in its anaerobic type of mitochondria (fig. 6) (Mi-ichi et al. 2009; Nývltová et al. 2015). This pathway produces 3'-phosphoadenosine-5'-phosphosulfate (PAPS), which is exported to the cytosol via a mitochondrial carrier family protein (EhMCP) in *E. histolytica* (Mi-ichi, Miyamoto, et al. 2015). The identification of an EhMCP homolog in *M. balamuthi* (supplementary fig. S12, Supplementary Material online) suggests that PAPS export from the hydrogenosome is also likely. In *E. histolytica*, PAPS is then utilized by sulfotransferases (SULTs) to form sulfolipids (SLs), including cholesteryl sulfate synthesized by the activity of SULT6 and involved in the regulation of cyst formation (Mi-ichi, Nozawa, et al. 2015). Catabolic pathways of sulfated molecules in *Entamoeba* include multiple sulfatases (SFs) (Mi-ichi et al. 2017). *Mastigamoeba balamuthi* was shown to have no capacity to synthesize sulfolipids and was suggested to lack homologs of the *E. histolytica* SULTs (Mi-ichi, Nozawa, et al. 2015). In fact, *M. balamuthi* does possess SULT genes, but its substrate specificity cannot be predicted, and all four genes constitute a clade specifically related to genes from various bacteria rather than to SULTs from *Entamoeba* species (supplementary fig. S13, Supplementary Material online). Hence, the *Entamoeba* and *M. balamuthi* lineages seem to have acquired SULTs via independent LGT events. SFs in the two Archamoebae lineages are likewise of a different origin, as those identified in *M. balamuthi* are related to eukaryotic enzymes (supplementary fig. S14, Supplementary Material online), whereas the *E. histolytica* counterparts are related to bacterial Zn-dependent alkyl SFs (Mi-ichi et al. 2017). It is plausible that the paralogous SULTs resulting from gene duplications in the *Entamoeba* lineage gained new diverse substrate specificities associated with adaptation to parasitism (Mi-ichi et al. 2017) and that this was complemented by the acquisition of different catabolic SFs to metabolize *Entamoeba*-specific sulfated molecules, whereas *M. balamuthi* retained the ancestral eukaryotic SFs.

Cysteine proteases (CPs) are important virulence factors of *E. histolytica* (Hellberg et al. 2001; Lidell et al. 2006). The most important are CPs of the C1 papain superfamily, including 36 members categorized into three clades (EhCP-A, EhCP-B, and

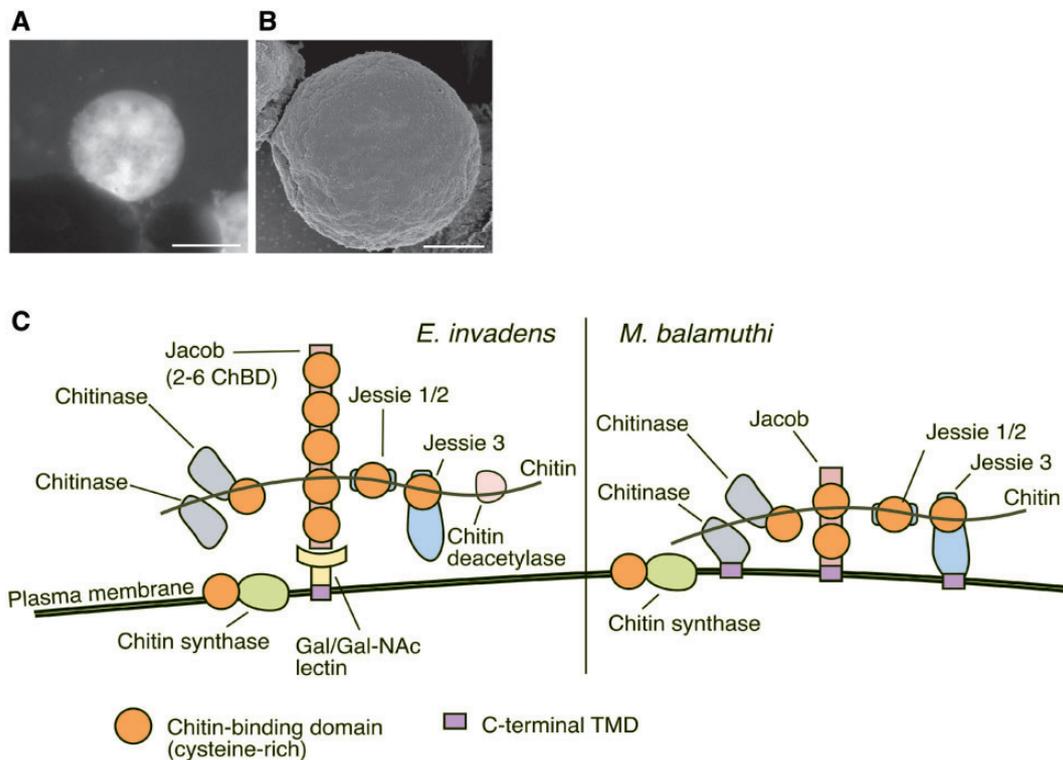


Fig. 7. Cysts of *Mastigamoeba balamuthi*. (A) Labeling of *M. balamuthi* cyst with calcofluor white stain. (B) Scanning electron microscopy of *M. balamuthi* cyst. (C) Comparison of cyst wall components in *M. balamuthi* and *Entamoeba invadens* and their predicted structure. The model for *E. invadens* is based on Chatterjee et al. (2009).

EhCP-C) (Bruchhaus et al. 2003; Clark et al. 2007). We found 54 C1 papain CPs in *E. histolytica* and a considerably larger repertoire in *M. balamuthi* (156 CPs) (fig. 6). The *M. balamuthi* CPs clustered into ten groups that also included EhCPs (supplementary fig. S15, Supplementary Material online). Two *M. balamuthi* CPs were closely related to EhCP-B, and five *M. balamuthi* CPs were related to EhCP-C. Importantly, although the overall CP repertoire was reduced in *E. histolytica*, the number of paralogs in each *E. histolytica* clade was expanded. For example, there are 11 *E. histolytica* EhCP-A paralogs and 13 EhCP-B paralogs (supplementary fig. S15, Supplementary Material online). In contrast, some of the *M. balamuthi* clades without *E. histolytica* representatives were highly expanded, such as the 84 paralogs of the CP group with m51a1_g146 as the representative gene. Therefore, we can predict that the LCAA was equipped with a broad spectrum of CPs, from which the *E. histolytica* CP complement associated with virulence evolved by reduction of the ancestral set and secondary expansion of the CP types that had been retained.

Entamoeba histolytica amoebapore is a pore-forming protein of the saposin-like protein (SAPLIP) family that is involved in the degradation of bacterial (Andrá et al. 2003) and possibly host cell membranes (Leippe et al. 1991; Bujanover et al. 2003). We found a gene of the SAPLIP family in the *M. balamuthi* genome that resembles the *E. histolytica* amoebapore gene (fig. 6 and supplementary table S11, Supplementary Material online). Thus, members of the SAPLIP family were likely present in the LCAA.

The features of *Entamoeba* that could be considered specific adaptations to parasitism and that are absent in free-living *M. balamuthi* include a set of membrane proteins that are involved in parasite–host cell interactions. These proteins include Gal/GalNAc lectin (Petri et al. 2002), a highly expanded family of BspA adhesins (111 paralogs in *E. histolytica*, 1,320 in *E. moshkovskii*), an Ariel1 surface antigen family protein (Willhoeft et al. 1999), a lysine- and glutamic acid-rich protein (KERP1) (Seigneur et al. 2005), and *E. histolytica* serine-, threonine-, and isoleucine-rich proteins (EhSTIRP) involved in adhesion and cytotoxicity (MacFarlane and Singh 2007). *Mastigamoeba balamuthi* also lacks GTPases of the AIG1 family (Nakada-Tsukui et al. 2018), which are involved in *E. histolytica* virulence via regulation of host cell adhesion (fig. 6). These features were most likely absent in the LCAA.

The Role of LGT

The comparison of OGs within the Conosa group revealed that Archamoebae and *M. balamuthi* in particular have gained a number of genes that provide functions specific to these organisms. Some of these genes have previously been studied and identified as putative acquisitions from prokaryotic lineages via LGT (Gill et al. 2007; Nylytova et al. 2013; Nýlytová et al. 2015). To systematically assess the overall contribution of LGT to the *M. balamuthi* gene repertoire, we conducted a large-scale analysis of gene phylogenies. Based on homology searches against the UniProt database, we selected conserved domains of *M. balamuthi* proteins and built

phylogenies for each of them. For the LGT evaluation, we introduced an LGT score (for details see [supplementary Materials](#) and [Methods](#) and [fig. S16, Supplementary Material](#) online) as a tool to identify cases when the following two conditions are met: 1) the gene transfer was directed from prokaryotes to eukaryotes, and 2) the gene tree topology was more plausibly interpreted as a result of LGT than as independent losses of an ancestral eukaryotic gene. Altogether, we calculated the LGT score for 5,254 genes (paralogs resulting from duplications in the *M. balamuthi* lineage, i.e., in-paralogs, counted as a single gene). Of these, 528 *M. balamuthi* genes (or groups of in-paralogs) have strong support for having been gained by LGT from prokaryotic sources (LGT score >0.75) ([fig. 8A](#) and [supplementary table S12, Supplementary Material](#) online).

To evaluate which functional gene categories were most frequently transferred from prokaryotes to *M. balamuthi*, we correlated the LGT score with the functional annotation of *M. balamuthi* proteins ([fig. 8D](#)). The most significant positive correlation between the LGT score and COG enrichment was observed for enzymes and small molecule transporters, namely for products of genes functioning in amino acid metabolism, carbohydrate transport and metabolism, and energy conversion (with Pearson correlation coefficients of 0.31, 0.25, and 0.23, respectively). Conversely, genes involved in genetic information processing, signal transduction, and intracellular trafficking correlated negatively with the LGT score, except for the COG category “cell wall/membrane/envelope biogenesis,” which includes genes for cyst wall components ([fig. 8D](#)).

The more than 500 LGT candidates in *M. balamuthi* represent one of the largest LGT sets reported in any protist species, although any comparison of LGTs frequencies inferred in various organisms must be done with caution, as various approaches have been used for LGT detection in different studies. The estimated numbers of LGTs that have contributed to the genetic toolkit of particular eukaryotic species vary immensely from zero to hundreds ([Keeling and Palmer 2008](#)). For example, only one and ten LGT events were inferred for the yeasts *Ashbya gossypii* and *S. cerevisiae*, respectively ([Hall et al. 2005](#)). Considerably more frequent LGTs have been observed in extremophiles and anaerobes, for whom LGT allows them to live in highly specialized environments ([Keeling and Palmer 2008](#); [Husnik and McCutcheon 2018](#)). Thus, 75 gene acquisitions were detected in the extremophilic red alga *Galdieria sulphuraria*, which lives in hot, toxic metal-rich, acidic environments ([Schönknecht et al. 2013](#)), whereas the anaerobic commensal *Blastocystis* sp. living in the human intestinal tract was found to contain up to 167 LGT candidates ([Eme et al. 2017](#)). The largest set of LGT candidates was estimated in anaerobic rumen-dwelling ciliates (~1,000 genes) ([Ricard et al. 2006](#); [Keeling and Palmer 2008](#)). The majority of LGTs in ciliates were associated with the ability to degrade plant cell wall-derived carbohydrates and anaerobic energy metabolism. In this context, the large set of LGT candidates in *M. balamuthi* is not surprising, as similarly to rumen ciliates, *M. balamuthi* is adapted for living under anaerobiosis and in an environment rich in metabolisable plant polymers. In contrast, the role of LGT in the adaptation of the *Entamoeba* lineage to the

endobiotic lifestyle appears to have relied less on acquiring genes from external sources, as indicated by the only 96 *E. histolytica* genes of prokaryotic origin identified in an early genome analysis ([Loftus et al. 2005](#)).

To infer possible prokaryotic donors for the 528 best LGT candidates in *M. balamuthi*, we calculated the so-called domain boundary frequency (DBF) value for each taxon within all 528 phylogenetic trees. The DBF value is used to evaluate prokaryotic taxa that are in the closest proximity to the boundary between eukaryotes and prokaryotes (see [supplementary Materials](#) and [Methods, Supplementary Material](#) online, for details, [supplementary table S12](#) and [fig. S16, Supplementary Material](#) online). Strong support for emergence from within a particular prokaryotic taxon was obtained for only 32 out of 528 LGT candidates ([supplementary table S12, Supplementary Material](#) online). The most common putative donor taxon was Actinobacteria (8 genes), followed by anaerobic bacteria of the Bacteroidetes group (6 genes) ([fig. 8B](#)). Among these well-supported LGTs, we identified two genes that encode enzymes catalyzing consecutive reactions in proline biosynthesis: glutamate 5-kinase (G5K) and glutamate-5-semialdehyde dehydrogenase (G5SD). In the *M. balamuthi* genome, these genes are organized in a head-to-head arrangement, and both show specific affinities to homologs from the same bacterial taxon in the class Verrucomicrobiae ([supplementary fig. S17, Supplementary Material](#) online). These bacteria also tend to have these two genes physically clustered in the genome, so it is likely that both were acquired by *M. balamuthi* from the same source *en bloc*. Neither of these genes is present in *Entamoeba*, which supports their possibly recent acquisition after the *M. balamuthi/Entamoeba* split. We attribute the small number of well-supported prokaryotic groups as LGT sources to several factors: generally low resolution of single gene phylogenies, insufficient taxon sampling or poor taxonomic placement of prokaryote metagenomic sequences, and possibly complicated nonvertical histories of many of these gene families in prokaryotes.

Finally, we investigated the extent to which *M. balamuthi* shares LGT candidates with other eukaryotes, particularly with *Entamoeba*. To do so, we searched for eukaryotic genes that form monophyletic groups with the *M. balamuthi* LGT candidates. This analysis revealed that 358 out of 528 LGT candidates were unique to *M. balamuthi* (i.e., resulting from an independent LGT from a prokaryotic source to the *M. balamuthi* lineage), whereas 170 LGT candidates were shared with various eukaryotes as a result of vertical inheritance from a deeper eukaryotic ancestor or intereukaryote LGT following the initial gain from prokaryotes ([fig. 8C](#)). Of them 34 had genes from the *Entamoeba* genus as the immediate sister group, which was the highest number among all different eukaryotic taxa with genes directly affiliated to the *M. balamuthi* LGT candidates. These 34 genes were thus presumably inherited from a common ancestor of *M. balamuthi* and *Entamoeba*, and 24 were unique to Archamoebae, indicating that they were gained from prokaryotes specifically in the Archamoebae stem lineage ([supplementary table S12, Supplementary Material](#) online). Previous analyses by different authors predicted 68–138 gene transfers from prokaryotic

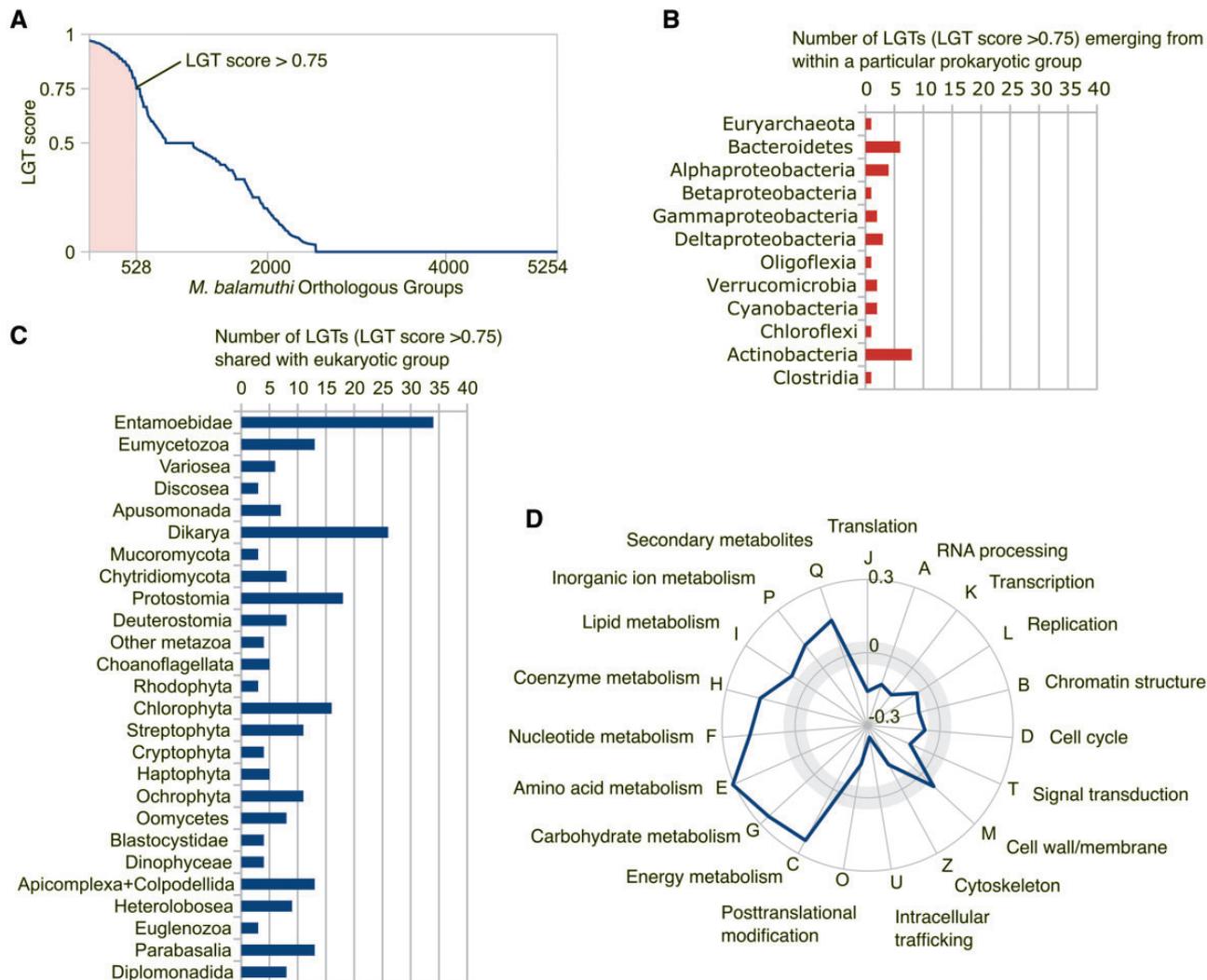


Fig. 8. Evaluation of LGT contribution to *Mastigamoeba balamuthi* genome. (A) Distribution of LGT scores among *M. balamuthi* orthologous groups. (B) Predicted prokaryotic sources of LGT (for LGT score >0.75). (C) A number of genes acquired by LGT in eukaryotic groups that are shared with *M. balamuthi* (LGT score >0.75). (D) Pearson correlation coefficients between the LGT score and COG functional categories. The gray area demarcates correlation coefficient values with low statistical significance (P value ≥ 0.01).

sources to the *E. histolytica* lineage (Loftus et al. 2005; Clark et al. 2007; Grant and Katz 2014). The 138 LGT candidates in *E. histolytica* identified by the most recent analysis (Grant and Katz 2014) include 14 of the 34 LGTs identified by us as shared by *M. balamuthi* and *Entamoeba* (supplementary table S12, Supplementary Material online). The 17 missing cases correspond to two genes missing in *E. histolytica* but present in other *Entamoeba* species, one gene that was not analyzed by Grant and Katz (2014), two genes that were interpreted by these authors as an LGT into the *Entamoeba* lineage from an unrelated eukaryotic donor, and 12 genes that were considered as unique for *Entamoeba* or whose phylogenetic placement was not resolved in previous study.

Hence, our analyses refine the picture of the LGT history in the *Entamoeba* lineage by expanding the list of acquired genes and distinguishing gains specific for *Entamoeba* from those preceding the radiation of extant Archamoebae. Encouragingly, the NIF system components, which were previously suggested to have been acquired by an archamoebal ancestor (Nyvltova et al. 2013), were recovered using our

approach. Furthermore, these ancestral acquisitions included several enzymes crucial for anaerobic metabolism, such as [Fe, Fe] hydrogenase, pyruvate-phosphate dikinase, pyruvate kinase, acetate kinase, and iron-sulfur flavoprotein; and also enzymes for amino acid metabolism including serine O-acetyltransferase of the de novo cysteine biosynthetic pathway, methionine gamma-lyase, and arginine decarboxylase. Interestingly, the latter enzyme mediates the survival of enteric bacteria during passage through acidic parts of the digestive tract, a need that is shared by *E. histolytica* (Iyer et al. 2003). Flavodiiron protein (A-type flavoprotein) is shared by *M. balamuthi* and *Entamoeba* and, interestingly, by the free-living anaerobic jakobid *Stygiella incarcerata*.

Dikaryan fungi were the group with the second highest number (26) of prokaryotic LGTs shared with *M. balamuthi*. These OGs included enzymes important for polysaccharide degradation, such as xylan 1,4- β -xylosidase, α -amylase, and cellulase (fig. 8D and supplementary table S12, Supplementary Material online). Based on the tree topologies, it seems likely that these genes were first gained by dikaryan fungi and subsequently

transferred to the *M. balamuthi* lineage, although we cannot exclude the possibility that the two eukaryote lineages acquired these enzymes independently from the same group of prokaryotic decomposers of organic matter.

Conclusions

Analysis of the *M. balamuthi* genome has enabled us to make inferences about the transition of the common ancestor of Eumycetozoa and Archamoebae to the common anaerobic ancestor of *M. balamuthi* and *Entamoeba* species (the LCAA) and to separate the evolutionary origins of parasitism in the *Entamoeba* lineage from earlier events of adaptation of free-living amoebae to low-oxygen organic matter-rich water sediments. Two events, adaptation to anaerobiosis and adaptation to parasitism, were both accompanied by massive gene loss, although the gene categories affected were quite different. During the transition to the anaerobic lifestyle, the loss concerned mainly oxygen-dependent pathways and functions related to mitochondria and peroxisomes. In the second step, gene loss in the parasitic *Entamoeba* lineage was significantly associated with amino acid and nucleotide metabolism, but the majority of losses were distributed across most OGs, resulting in an overall decrease in metabolic complexity. In contrast, there was no such dramatic gene loss in the *M. balamuthi* lineage.

On the other hand, gene loss in Archamoebae has been compensated for by numerous LGTs from prokaryotes that have conferred new metabolic capacities. In general, eukaryotes are expected to acquire genes from their cohabitants, and Archamoebae seem to comply with this notion. Members of the *Entamoeba* lineage interact with the rich and dense mucosal microbiota within the host, and the donors of the putative LGTs identified in *E. histolytica* are indeed enriched for the bacterial phyla Bacteroidetes and Firmicutes, which constitute a substantial portion of the mucosal bacteria (Alsmark et al. 2009). The same analysis revealed that donors of LGT candidates in the aerobic free-living *D. discoideum* are instead enriched for Proteobacteria, consistent with their prevalence in the environment inhabited by dictyostelids. However, this picture did not consider adaptation to anaerobiosis, which in Archamoebae most likely preceded parasitism, and consequently, most LGTs related to adaptation to an oxygen-poor niche may have happened under the conditions of a free-living lifestyle. Our analysis revealed that Bacteroidetes, as putative gene donors, are also enriched among LGT cases in free-living *M. balamuthi*. As these bacteria are present in the gut but are also widely distributed in water, sediments, and soil, we can predict that ancestral Archamoebae gained Bacteroidetes genes related to anaerobiosis from environmental bacteria and not from intestinal microbiota. After the split of the *M. balamuthi* and *Entamoeba* lineages, the former seems to have gained genes to increase its metabolic capacity as a decomposer, whereas the latter has acquired genes related to pathogenicity and virulence. Genes providing these specific properties are indeed of different origins. For example, *Entamoeba* BspA surface adhesins were suggested to have

been acquired from Bacteroidetes, whereas most transferred OGs of *M. balamuthi* are related to Actinobacteria. However, we cannot rule out the possibility that some of the gains that we have inferred as having occurred along the *M. balamuthi* lineage had already happened in or before the evolution of the free-living LCAA, but were followed by the loss of the respective genes in the *Entamoeba* lineage. Our study of gene histories of the Conosa group underlines the importance of comparative genomics for understanding how aerobic protists become anaerobes and how free-living anaerobes become parasites.

Materials and Methods

This section summarizes the most important methods used in this study. Further technical details are provided in the [supplementary Materials and Methods](#), [Supplementary Material](#) online.

Cell Culture

The *M. balamuthi* strain is a descendant of the culture isolated by R.S. Bray in Gambia and axenized by L.A. Chavez and colleagues (Chavez et al. 1986). The culture was kindly provided by M. Müller (Rockefeller University) and it is also available at the American Type Culture Collection (ATCC 30984). *Mastigamoeba balamuthi* was grown in PYGC medium (Chavez et al. 1986) at 24 °C in 50-ml tissue culture flasks. The cells were transferred to fresh medium weekly.

DNA Isolation and Sequencing

DNA from the axenically grown *M. balamuthi* culture was isolated using phenol–chloroform extraction.

Shotgun (500–800 bp) and paired-end (8–20 kb) fragment genomic libraries were prepared and sequenced on the 454 GS FLX platform with Titanium chemistry following the manufacturer's protocol (Roche Diagnostic, Rotkreuz, Switzerland) to generate 76.8 million reads. A TruSeq DNA library was analyzed using Illumina HiSeq 2000 (San Diego) at EMBL (Genomics Core Facility, Heidelberg, Germany) to generate 42.9 million paired-end reads. PacBio RS (EA Sequencing & Bioinformatics, Durham, NC) was used to generate 295,000 long reads (details are given in the [supplementary Materials and Methods](#), [Supplementary Material](#) online).

RNA Isolation and Sequencing

Total RNA was isolated using TRIzol (Total RNA Isolation Reagent, Gibco BRL) and purified from DNA with the RNeasy Minelute Cleanup Kit (Qiagen). cDNA was prepared using the Transcriptor High Fidelity cDNA Synthesis Kit (Roche). The transcriptome was sequenced with Illumina HiSeq 2000 (San Diego) at EMBL (Genomics Core Facility, Heidelberg, Germany) using the TruSeq RNA library (non-stranded) to generate 86.6 million paired-end reads. Transcriptome assembly was obtained using Trinity with the default parameters (Grabherr et al. 2011).

Genome Assembly and Annotation

The genome sequence was assembled using Newbler (Software Release: 2.6 20110517_1502) (Margulies et al.

2005), with further refinements described in detail in the [supplementary Materials](#) and Methods and [figure S18, Supplementary Material](#) online. Gene prediction was performed using Augustus 3.0.2 ([Stanke and Morgenstern 2005](#)) with a set of 298 manually curated training gene models and hints from the transcriptome mapped onto the genome assembly, which yielded 14,840 protein-coding genes. Manual curation was employed for genes of interest using the interface provided by the Online Resource for Community Annotation of Eukaryotes (ORCAE) webserver ([Sterck et al. 2012](#)). Gene annotation was performed using InterProScan ([Jones et al. 2014](#)) and Blast2GO ([Conesa et al. 2005](#)). Structural RNA genes were predicted and annotated using the Rfam database ([Griffiths-Jones et al. 2004](#)) ([supplementary table S3, Supplementary Material](#) online). The annotated genome sequence has been deposited in the online resource for community annotation of eukaryotes (ORCAE, <https://bioinformatics.psb.ugent.be/orcae/overview/Masba>).

Analysis of Orthologous Groups

We gathered predicted proteomes of selected amoebozoans (*A. castellanii*, *Protostelium aurantium*, *Acytostelium subglobosum*, *D. discoideum*, *Mastigamoeba balamuthi*, *E. invadens*, *E. moshkovskii*, and *E. histolytica*) and opisthokonts (*Homo sapiens*, *Amphimedon queenslandica*, *S. cerevisiae*, *Batrachochytrium dendrobatidis*) ([supplementary table S4, Supplementary Material](#) online). The protein sequences were searched against the profile HMMs of the EggNOG database version 4.5 using HMMER3 ([Eddy 2011](#)). Only EggNOG OGs with the best hit e-values $\leq 1e-10$ were considered. For each protein sequence, nonoverlapping hits with e-values $\leq 1e-5$ were selected. The gain of the orthologous group was assigned to the last common ancestor of all the species in which the orthologous group was detected. The losses were assigned to the nodes descending from the last common ancestor in which none of the descendants had the orthologous group considered. At each node of the tree of selected amoebozoan and opisthokont species, for each event type (gain or loss) and each COG category, we created a contingency table of orthologous groups categorizing them according to the presence of the given COG category in the EggNOG annotation, and according to whether the orthologous group was gained or lost, matching the given event type. The statistical significance of the association between those categorical variables was tested using the χ^2 test. Protein sequences of *E. histolytica*, *To. gondii*, *Chromera velia*, *Tr. brucei*, and *B. saltans* were downloaded from the UniProt database (<https://www.uniprot.org/>) and annotated using KofamKOALA ([Aramaki et al. 2020](#)).

Phylogenetic Analyses

Phylogenies were estimated for each predicted *M. balamuthi* protein. The respective protein sequences were used as queries to search the UniProt protein database using the DIAMOND protein aligner (version diamond/0.9.10) ([Buchfink et al. 2015](#)) in sensitive mode. Up to 10,000 best hits were used for further analysis, discarding those with e-values higher than $1e-5$. NCBI taxonomy was assigned to

homologous sequences (hits), and hit regions homologous to the query sequence were clustered at the 80% identity level using CD-HIT ([Fu et al. 2012](#)). The presence of hits from both eukaryotes and prokaryotes in a given 80% identity group was suspected to result from contamination. The source of contamination was determined based on the number of unique taxa in the 80% identity group and the contamination was removed. The sequences were assigned to one of the predefined major eukaryotic taxonomic groups (generally corresponding to an intermediate hierarchy level in the NCBI taxonomic database, see [supplementary table S13, Supplementary Material](#) online), and from each 80% identity group up to one sequence per each taxonomic group was chosen for further analysis to minimize redundancy. A multiple sequence alignment was constructed for each sequence group using MAFFT with the default settings ([Katoh and Standley 2013](#)) and trimmed with BMGE ([Criscuolo and Grimaldo 2010](#)) using the BLOSUM30 matrix with a block size of 1. Phylogenetic trees were inferred with IQ-TREE ([Nguyen et al. 2015](#)) with automatic model selection and 1,000 ultrafast bootstraps or 500 bootstraps. The selected models are summarized in [supplementary table S12, Supplementary Material](#) online. Nodes with statistical support below 90% were removed. Analysis and manipulation of the trees were implemented using the ETE package ([Huerta-Cepas et al. 2016](#)).

Analysis of LGT

Taxon Frequency

For the analysis of gene trees, we introduced a measure, computed for each node of a rooted tree, which we named “taxon frequency” (see [supplementary fig. S16C, Supplementary Material](#) online). It provides information about the possible taxonomic identity of the organism carrying an ancestral sequence that corresponds to the particular node of the gene tree. The taxon frequency is calculated iteratively from the tips of the tree to the root. Starting at the tips of the tree, the frequencies of taxa (considering taxa listed in [supplementary table S13, Supplementary Material](#) online) at the descendant tips of a node were calculated and assigned to that internal node. Then, taxon frequencies of the next “deeper” node are calculated as the average of the frequencies calculated for the two descendant nodes. This procedure is iterated until the root is reached.

T , taxon
 f_T , frequency of taxon T at a given node
 N_{children} , number of children at the node
 i , identifier of a child (1, 2, ..., N_{children})
 f_{Ti} , frequency of taxon T at child i

$$f_T = \frac{\sum_{i=1}^{N_{\text{children}}} f_{Ti}}{N_{\text{children}}}$$

Estimation of Eukaryotic Orthologs

Sequences of eukaryotic organisms in the gene trees (inferred as described in the Materials and Methods) were evaluated to

determine whether we could refute that they formed a clade of eukaryotic sequences together with the query *M. balamuthi* sequence (see [supplementary fig. S16D](#), [Supplementary Material](#) online). The tree was arbitrarily rooted by the *M. balamuthi* query sequence. Taxon frequencies at the level of superkingdoms (i.e., Bacteria, Archaea, and Eukaryota) were computed. For each eukaryotic sequence, the path to the bottom of the gene tree (rooted by the query sequence) was followed, and the taxonomic frequency of the “Eukaryota” taxon was recorded. Whenever the frequency of the “Eukaryota” taxon was below 0.5, it was an indication that the eukaryotic sequence in question may not form a clade of eukaryotic sequences with the query sequence. We computed such a measure of orthology as the minimal “Eukaryota” frequency along the path.

Orthology Score = Min (Taxon Frequency of “Eukaryota”).

For the subsequent analyses, eukaryotic sequences with orthology scores above 0.25 were considered to possibly form a eukaryotic clade with the query *M. balamuthi* sequence.

LGT Evaluation

We focused specifically on LGTs from prokaryotes to eukaryotes that occurred after the diversification of eukaryotes (excluding, for example, endosymbiotic gene transfers from the ancestors of mitochondria). The directionality of the LGT was evaluated by comparing the number of all unique prokaryotic taxa in the phylogenetic tree and the number of unique eukaryotic taxa that formed a eukaryotic clade with the *M. balamuthi* query. The taxa considered are listed in [supplementary table S13](#), [Supplementary Material](#) online. Each of the selected prokaryotic and eukaryotic taxa has a column in [supplementary table S11](#), [Supplementary Material](#) online. The measure of this directionality is called the “LGT prokaryotic coefficient” (LGTPC) and is computed accordingly:

N_{proktaxa} the number of prokaryotic taxa

N_{euktaxa} the number of eukaryotic taxa that formed a eukaryotic monophylum with the *M. balamuthi* query

$$\text{LGTPC} = \frac{N_{\text{proktaxa}}}{N_{\text{proktaxa}} + N_{\text{euktaxa}}}.$$

To distinguish genes that were not present in the common ancestor of eukaryotes, we evaluated the scenarios of both an early transfer (the gene was present in the common ancestor of eukaryotes) and a late transfer (the gene was acquired later, after the radiation of eukaryotes), and compared them using maximum parsimony, that is, favoring those that involve a lesser number of events. We call the measure of the late transfer scenario the LGT eukaryotic coefficient (LGTEC), and compute it accordingly:

$N_{\text{ancestral losses}}$ minimal number of losses if the gene was present in the common ancestor of eukaryotes

$N_{\text{new transfers}}$ minimal number of transfers between eukaryotic lineages if the gene was gained after the radiation of eukaryotes

$N_{\text{new losses}}$ minimal number of losses if the gene was gained after the radiation of eukaryotes

$$\text{LGTEC} = \frac{N_{\text{ancestralLosses}}}{N_{\text{ancestralLosses}} + N_{\text{newTransfers}} + N_{\text{newLosses}}}.$$

For the LGT score, we selected the lower of the values calculated for the two coefficients ([supplementary fig. S16](#), [Supplementary Material](#) online). Strong support for LGT was considered for a gene if both coefficients were ≥ 0.75 ([supplementary table S11](#), [Supplementary Material](#) online).

Domain Boundary Frequency

To assess possible sources of prokaryote-to-eukaryote LGTs based on the phylogenetic tree, we introduced a measure that we call domain boundary frequency (DBF). It is defined as the prokaryotic taxon frequencies at the root of the eukaryotic subtree, that is, at the boundary between eukaryotic and prokaryotic sequences (see [supplementary fig. S16A](#), [Supplementary Material](#) online). The most significant prokaryotic taxon was that with the highest DBF value. We classified the DBFs into three classes: 1) $0 < \text{DBF} \leq 0.25$: unclear origin, 2) $0.25 < \text{DBF} \leq 0.75$: possible sister relationship with the given taxon, and 3) $0.75 < \text{DBF} \leq 1$: the given taxon is a likely source of the LGT. Automated tools for the calculation of taxon frequency, LGTPC, LGTEC, and DBF are available at <https://github.com/vojtech-zarsky/vojta-tools/blob/master/LGTanalysis.py>.

Supplementary Material

[Supplementary data](#) are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank John Cawley for proofreading the manuscript. This work was supported by the Czech Science Foundation (P305/11/1061, 16-06123S), Centre for research of pathogenicity and virulence of parasites (CZ.02.1.01/0.0/0.0/16_019/0000759) provided by European Regional Development Fund, Microbial Communities in Biomedical and Environmental Areas, and Systems Biology funded from EU H2020 (No 810224), University Research Centre (UNCE, SCI/12) from Charles University, and the ELIXIR CZ research infrastructure project (Ministry of Education, Youth, and Sport Grant No: LM2018131) including access to computing and storage facilities. The genomic sequences have been deposited in the online resource for community annotation of eukaryotes (ORCAE, <https://bioinformatics.psb.ugent.be/orcae/overview/Masba>).

References

- Alsmark UC, Sicheritz-Ponten T, Foster PG, Hirt RP, Embley TM. 2009. Horizontal gene transfer in eukaryotic parasites: a case study of *Entamoeba histolytica* and *Trichomonas vaginalis*. *Methods Mol Biol.* 532:489–500.
- Andr  J, Herbst R, Leippe M. 2003. Amoebapores, archaic effector peptides of protozoan origin, are discharged into phagosomes and kill bacteria by permeabilizing their membranes. *Dev Comp Immunol.* 27(4):291–304.

- Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36(7):2251–2251.
- Artzi L, Bayer EA, Morais S. 2017. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat Rev Microbiol* 15(2):83–95.
- Baptiste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, Gordon P, Duruffé L, Gaasterland T, Lopez P, Müller M, et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci U S A* 99(3):1414–1419.
- Barbosa-Cabrera E, Salas-Casas A, Rojas-Hernández S, Jarillo-Luna A, Abarca-Rojano E, Rodríguez MA, Campos-Rodríguez R. 2012. Purification and cellular localization of the *Entamoeba histolytica* transcarboxylase. *Parasitol Res* 111(3):1401–1405.
- Barlow LD, Nývltová E, Aguilar M, Tachezy J, Dacks JB. 2018. A sophisticated, differentiated Golgi in the ancestor of eukaryotes. *BMC Biol* 16(1):27.
- Bauer S, Grossmann S, Vingron M, Robinson PN. 2008. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* 24(14):1650–1651.
- Beck M, Hurt E. 2017. The nuclear pore complex: understanding its function through structural insight. *Nat Rev Mol Cell Biol* 18(2):73–89.
- Blaxter M, Koutsovoulos G. 2015. The evolution of parasitism in Nematoda. *Parasitology* 142(S1):S26–S39.
- Bruchhaus I, Loftus BJ, Hall N, Tannich E. 2003. The intestinal protozoan parasite *Entamoeba histolytica* contains 20 cysteine protease genes, of which only a small subset is expressed during in vitro cultivation. *Eukaryot Cell* 2(3):501–509.
- Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12(1):59–60.
- Bujanover S, Katz U, Bracha R, Mirelman D. 2003. A virulence attenuated amoebapore-less mutant of *Entamoeba histolytica* and its interaction with host cells. *Int J Parasitol* 33(14):1655–1663.
- Butenko A, Opperdoes FR, Flegontova O, Hork A, Hampf V, Keeling P, Gawryluk RMR, Tikhonov D, Flegontov P, Lukeš J. 2020. Evolution of metabolic capabilities and molecular features of diplomonads, kinetoplastids, and euglenids. *BMC Biol* 18(1):23.
- Carlton JM, Hirt RP, Silva JC, Delcher AL, Schatz M, Zhao Q, Wortman JR, Bidwell SL, Alsmark UCM, Besteiro S, et al. 2007. Draft genome sequence of the sexually transmitted pathogen *Trichomonas vaginalis*. *Science* 315(5809):207–212.
- Castellanos-Castro S, Bolaños J, Orozco E. 2020. Lipids in *Entamoeba histolytica*: host-dependence and virulence factors. *Front Cell Infect Microbiol* 10:75.
- Cavalier-Smith T. 1998. A revised six-kingdom system of life. *Biol Rev Camb Rev* 73(3):203–266.
- Chatterjee A, Ghosh SK, Jang K, Bullitt E, Moore L, Robbins PW, Samuelson J. 2009. Evidence for a “wattle and daub” model of the cyst wall of entamoeba. *PLoS Pathog* 5(7):e1000498.
- Chavez LA, Balamuth W, Gong T. 1986. A light and electron microscopical study of a new, polymorphic free-living amoeba, *Phreatamoeba balamuthi* n. g., n. sp. *J Protozool* 33(3):397–404.
- Cheng Y, Shi Q, Sun R, Liang D, Li Y, Li Y, Jin W, Zhu W. 2018. The biotechnological potential of anaerobic fungi on fiber degradation and methane production. *World J Microbiol Biotechnol* 34(10):155.
- Clark CG, Ali IKM, Zaki M, Loftus BJ, Hall N. 2006. Unique organisation of tRNA genes in *Entamoeba histolytica*. *Mol Biochem Parasitol* 146(1):24–29.
- Clark CG, Alsmark UCM, Tazreiter M, Saito-Nakano Y, Ali V, Marion S, Weber C, Mukherjee C, Bruchhaus I, Tannich E, et al. 2007. Structure and content of the *Entamoeba histolytica* genome. *Adv Parasitol* 65:51–190.
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal annotation and visualization tool in functional genomics research. Application note. *Bioinformatics* 21(18):3674–3676.
- Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Bami D, Joardar VS, Johnson J, Radune D, Singh I, Badger JH, et al. 2011. Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. *Genome Biol* 12(10):R100.
- Crisuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol* 10(1):210.
- Cruz-Castañeda A, López-Casamichana M, Olivares-Trejo JJ. 2011. *Entamoeba histolytica* secretes two haem-binding proteins to scavenge haem. *Biochem J* 434(1):105–111.
- Das P, Das SR, Moorji A, Baer HP. 1997. Characterization of nucleoside uptake and transport in *Entamoeba histolytica*. *Parasitol Res* 83(4):364–369.
- Das S, Van Dellen K, Bulik D, Magnelli P, Cui J, Head J, Robbins PW, Samuelson J. 2006. The cyst wall of *Entamoeba invadens* contains chitosan (deacetylated chitin). *Mol Biochem Parasitol* 148(1):86–92.
- Davies KM, Strauss M, Daum B, Kief JH, Osiewacz HD, Rycovska A, Zickermann V, Kühlbrandt W. 2011. Macromolecular organization of ATP synthase and complex I in whole mitochondria. *Proc Natl Acad Sci U S A* 108(34):14121–14126.
- Dolezal P, Dagley MJ, Kono M, Wolyneć P, Likić VA, Foo JH, Sedinová M, Tachezy J, Bachmann A, Bruchhaus I, et al. 2010. The essentials of protein import in the degenerate mitochondrion of *Entamoeba histolytica*. *PLoS Pathog* 6(3):e1000812.
- Eberhardt EL, Ludlam AV, Tan Z, Cianfrocco MA. 2020. Miro: a molecular switch at the center of mitochondrial regulation. *Protein Sci* 29(6):1269–1284.
- Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol* 7(10):e1002195.
- Eme L, Gentekaki E, Curtis B, Archibald JM, Roger AJ. 2017. Lateral gene transfer in the adaptation of the anaerobic parasite *Blastocystis* to the gut. *Curr Biol* 27(6):807–820.
- Fang Z, Zhang W, Zhang T, Guang C, Mu W. 2018. Isomerases and epimerases for biotransformation of pentoses. *Appl Microbiol Biotechnol* 102(17):7283–7292.
- Frisardi M, Ghosh SK, Field J, Van Dellen K, Rogers R, Robbins P, Samuelson J. 2000. The most abundant glycoprotein of amebic cyst walls (Jacob) is a lectin with five Cys-rich, chitin-binding domains. *Infect Immun* 68(7):4217–4224.
- Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152.
- Gentekaki E, Curtis BA, Stairs CW, Klimeš V, Eliáš M, Salas-Leiva DE, Herman EK, Eme L, Arias MC, Henrissat B, et al. 2017. Extreme genome diversity in the hyper-prevalent parasitic eukaryote *Blastocystis*. *PLoS Biol* 15(9):e2003769.
- Gill EE, Diaz-Trivino S, Barbera MJ, Silberman JD, Stechmann A, Gaston D, Tamas I, Roger AJ. 2007. Novel mitochondrion-related organelles in the anaerobic amoeba *Mastigamoeba balamuthi*. *Mol Microbiol* 66(6):1306–1320.
- Gilles-Gonzalez MA, Ditta GS, Helinski DR. 1991. A haemoprotein with kinase activity encoded by the oxygen sensor of *Rhizobium meliloti*. *Nature* 350(6314):170–172.
- Ginger M, Field MC. 2016. Making the pathogen: evolution and adaptation in parasitic protists. *Mol Biochem Parasitol* 209(1–2):1–2.
- Glenn K, Ingram-Smith C, Smith KS. 2014. Biochemical and kinetic characterization of xylulose 5-phosphate/fructose 6-phosphate phosphoketolase 2 (Xfp2) from *Cryptococcus neoformans*. *Eukaryot Cell* 13(5):657–663.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644–652.
- Grant JR, Katz LA. 2014. Phylogenomic study indicates widespread lateral gene transfer in *Entamoeba* and suggests a past intimate relationship with parabasalids. *Genome Biol Evol* 6(9):2350–2360.

- Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2004. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33(Database issue):D121–D124.
- Heinz E, Williams TA, Nakjang S, Noël CJ, Swan DC, Goldberg AV, Harris SR, Weinmaier T, Markert S, Becher D, et al. 2012. The genome of the obligate intracellular parasite *Trachipleistophora hominis*: new insights into microsporidian genome dynamics and reductive evolution. *PLoS Pathog.* 8(10):e1002979.
- Hall C, Brachat S, Dietrich FS. 2005. Contribution of horizontal gene transfer to the evolution of *Saccharomyces cerevisiae*. *Eukaryot Cell.* 4(6):1102–1115.
- Hellberg A, Nickel R, Lotter H, Tannich E, Bruchhaus I. 2001. Overexpression of cysteine proteinase 2 in *Entamoeba histolytica* or *Entamoeba dispar* increases amoeba-induced monolayer destruction in vitro but does not augment amoebic liver abscess formation in gerbils. *Cell Microbiol.* 3(1):13–20.
- Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* 33(6):1635–1638.
- Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nat Rev Microbiol.* 16(2):67–79.
- Iyer R, Williams C, Miller C. 2003. Arginine-aggmatine antiporter in extreme acid resistance in *Escherichia coli*. *J Bacteriol.* 185(22):6556–6561.
- Iyer LR, Verma AK, Paul J, Bhattacharya A. 2019. Phagocytosis of Gut Bacteria by *Entamoeba histolytica*. *Front Cell Infect Microbiol.* 9:34.
- Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, Hartley C, Sanders M, Wastling JM, Dacks JB, et al. 2016. Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. *Curr Biol.* 26(2):161–172.
- Janouskovec J, Keeling PJ. 2016. Evolution: causality and the origin of parasitism. *Curr Biol.* 26(4):R174–R177.
- Janouskovec J, Tikhonenkov DV, Burki F, Howe AT, Kolisko M, Mylnikov AP, Keeling PJ. 2015. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc Natl Acad Sci U S A.* 112(33):10200–10207.
- Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236–1240.
- Kabbara S, Hérviaux A, Dugé de Bernonville T, Courdavault V, Clastre M, Gastebois A, Osman M, Hamze M, Cock JM, Schaap P, et al. 2019. Diversity and evolution of sensor histidine kinases in eukaryotes. *Genome Biol Evol.* 11(1):86–108.
- Kang S, Tice AK, Spiegel FW, Silberman JD, Pánek T, Čepička I, Kostka M, Kosakyan A, Alcántara DMC, Roger AJ, et al. 2017. Between a pod and a hard test: the deep evolution of amoebae. *Mol Biol Evol.* 34(9):2258–2270.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4):772–780.
- Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet.* 9(8):605–618.
- Kollmar M. 2016. Fine-tuning motile cilia and flagella: evolution of the dynein motor proteins from plants to humans at high resolution. *Mol Biol Evol.* 33(12):3249–3267.
- Kořený L, Oborník M, Lukeš J. 2013. Make It, Take It, or Leave It: heme Metabolism of Parasites. *PLoS Pathog.* 9(1):e1003088.
- Le T, Žárský V, Nývltová E, Rada P, Harant K, Vancová M, Verner Z, Hrdý I, Tachezy J. 2020. Anaerobic peroxisomes in *Mastigamoeba balamuthi*. *Proc Natl Acad Sci U S A.* 117(4):2065–2075.
- Leippe M, Ebel S, Schoenberger OL, Horstmann RD, Müller-Eberhard HJ. 1991. Pore-forming peptide of pathogenic *Entamoeba histolytica*. *Proc Natl Acad Sci U S A.* 88(17):7659–7663.
- Lidell ME, Moncada DM, Chadee K, Hansson GC. 2006. *Entamoeba histolytica* cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel. *Proc Natl Acad Sci U S A.* 103(24):9298–9303.
- Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, Roncaglia P, Berriman M, Hirt RP, Mann BJ, et al. 2005. The genome of the protist parasite *Entamoeba histolytica*. *Nature* 433(7028):865–868.
- López MD, Alm Rosenblad M, Samuelsson T. 2008. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res.* 36(9):3001–3010.
- Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, et al. 2012. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380(9859):2095–2128.
- MacFarlane RC, Singh U. 2007. Identification of an *Entamoeba histolytica* serine-, threonine-, and isoleucine-rich protein with roles in adhesion and cytotoxicity. *Eukaryot Cell.* 6(11):2139–2146.
- Makioka A, Kumagai M, Kobayashi S, Takeuchi T. 2005. *Entamoeba invadens*: cysteine protease inhibitors block excystation and metacystic development. *Exp Parasitol.* 109(1):27–32.
- Makiuchi T, Mi-ichi F, Nakada-Tsukui K, Nozaki T. 2013. Novel TPR-containing subunit of TOM complex functions as cytosolic receptor for *Entamoeba* mitochondrial transport. *Sci Rep.* 3(1):1129.
- Makki A, Rada P, Žárský V, Kereiče S, Kováčik L, Novotný M, Jores T, Rapaport D, Tachezy J. 2019. Triplet-pore structure of a highly divergent TOM complex of hydrogenosomes in *Trichomonas vaginalis*. *PLoS Biol.* 17(1):e3000098.
- Mans BJ, Anantharaman V, Aravind L, Koonin EV. 2004. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* 3(12):1625–1650.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376–380.
- Mathur V, Kolisko M, Hehenberger E, Irwin NAT, Leander BS, Kristmundsson Á, Freeman MA, Keeling PJ. 2019. Multiple independent origins of apicomplexan-like parasites. *Curr Biol.* 29(17):2936–2941.e5.
- McGugan GC, Joshi MB, Dwyer DM. 2007. Identification and biochemical characterization of unique secretory nucleases of the human enteric pathogen, *Entamoeba histolytica*. *J Biol Chem.* 282(43):31789–31802.
- Mi-ichi F, Miyamoto T, Takao S, Jeelani G, Hashimoto T, Hara H, Nozaki T, Yoshida H. 2015. *Entamoeba* mitochondria play an important role in encystation by association with cholesterol sulfate synthesis. *Proc Natl Acad Sci U S A.* 112(22):E2884–E2890.
- Mi-ichi F, Miyamoto T, Yoshida H. 2017. Uniqueness of *Entamoeba* sulfur metabolism: sulfolipid metabolism that plays pleiotropic roles in the parasitic life cycle. *Mol Microbiol.* 106(3):479–491.
- Mi-ichi F, Nozawa A, Yoshida H, Tozawa Y, Nozaki T. 2015. Evidence that the *Entamoeba histolytica* mitochondrial carrier family links mitochondrial and cytosolic pathways through exchange of 3= phosphoadenosine 5=phosphosulfate and ATP. *Eukaryot Cell.* 14(11):1144–1150.
- Mi-ichi F, Yousuf MA, Nakada-Tsukui K, Nozaki T. 2009. Mitosomes in *Entamoeba histolytica* contain a sulfate activation pathway. *Proc Natl Acad Sci U S A.* 106(51):21731–21736.
- Nachury MV, Loktev AV, Zhang Q, Westlake CJ, Peränen J, Merdes A, Slusarski DC, Scheller RH, Bazan JF, Sheffield VC, et al. 2007. A core complex of BBS proteins cooperates with the GTPase Rab8 to promote ciliary membrane biogenesis. *Cell* 129(6):1201–1213.
- Nakada-Tsukui K, Sekizuka T, Sato-Ebine E, Escueta-de Cadiz A, Ji D, Tomii K, Kuroda M, Nozaki T. 2018. AIG1 affects in vitro and in vivo virulence in clinical isolates of *Entamoeba histolytica*. *PLoS Pathog.* 14(3):e1006882.
- Neumann N, Lundin D, Poole AM. 2010. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. *PLoS One* 5(10):e13241.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1):268–274.

- Nishitani Y, Horiuchi A, Aslam M, Kanai T, Atomi H, Miki K. 2018. Crystal structures of an archaeal chitinase ChiD and its ligand complexes. *Glycobiology* 28:418–426.
- Nývltová E, Smutná T, Tachezy J, Hrdý I. 2016. OsmC and incomplete glycine decarboxylase complex mediate reductive detoxification of peroxides in hydrogenosomes of *Trichomonas vaginalis*. *Mol Biochem Parasitol*. 206(1-2):29–38.
- Nývltová E, Stairs CW, Hrdý I, Rídl J, Mach J, Pačes J, Roger AJ, Tachezy J. 2015. Lateral gene transfer and gene duplication played a key role in the evolution of *Mastigamoeba balamuthi* hydrogenosomes. *Mol Biol Evol*. 32(4):1039–1055.
- Nývltová E, Šut'ák R, Žárský V, Harant K, Hrdý I, Tachezy J. 2017. Lateral gene transfer of p-cresol- and indole-producing enzymes from environmental bacteria to *Mastigamoeba balamuthi*. *Environ Microbiol*. 19(3):1091–1102.
- Nývltová E, Sutak R, Harant K, Sedinova M, Hrdy I, Paces J, Vlcek C, Tachezy J. 2013. NIF-type iron-sulfur cluster assembly system is duplicated and distributed in the mitochondria and cytosol of *Mastigamoeba balamuthi*. *Proc Natl Acad Sci U S A*. 110(18):7371–7376.
- Pandey AK, Pain J, Dancis A, Pain D. 2019. Mitochondria export iron-sulfur and sulfur intermediates to the cytoplasm for iron-sulfur cluster assembly and tRNA thiolation in yeast. *J Biol Chem*. 294(24):9489–9502.
- Pánek T, Zadravilková E, Walker G, Brown MW, Gentekaki E, Hroudová M, Kang S, Roger AJ, Tice AK, Vlček Č, et al. Č. 2016. First multigene analysis of Archamoebae (Amoebozoa: Conosa) robustly reveals its phylogeny and shows that Entamoebidae represents a deep lineage of the group. *Mol Phylogenet Evol*. 98:41–51.
- Payne SH, Loomis WF. 2006. Retention and loss of amino acid biosynthetic pathways based on analysis of whole-genome sequences. *Eukaryot Cell*. 5(2):272–276.
- Petri WA, Haque R, Mann BJ. 2002. The bittersweet interface of parasite and host: lectin-carbohydrate interactions during human invasion by the parasite *Entamoeba histolytica*. *Annu Rev Microbiol*. 56(1):39–64.
- Pineda E, Vázquez C, Encalada R, Nozaki T, Sato E, Hanadate Y, Néquiz M, Olivos-García A, Moreno-Sánchez R, Saavedra E. 2016. Roles of acetyl-CoA synthetase (ADP-forming) and acetate kinase (PPi-forming) in ATP and PPi supply in *Entamoeba histolytica*. *Biochim Biophys Acta*. 1860(6):1163–1172.
- Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T, Rattei T, Creevey C, Kuhn M, et al. 2014. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucl Acids Res*. 42(D1):D231–D239.
- Pyrh J, Harant K, Martinová E, Sutak R, Lesuisse E, Hrdý I, Tachezy J. 2014. *Giardia intestinalis* Incorporates Heme into Cytosolic Cytochrome b5. *Eukaryot Cell* 13(2):231–239.
- Pyrhová E, Motycková A, Voleman L, Wandyszewska N, Fišer R, Seydlová G, Roger A, Kolísko M, Doležal P. 2018. A single Tim translocase in the mitosomes of *Giardia intestinalis* illustrates convergence of protein import machines in anaerobic eukaryotes. *Genome Biol Evol*. 10(10):2813–2822.
- Reeves RE, Guthrie JD. 1975. Acetate kinase (pyrophosphate). A fourth pyrophosphate-dependent kinase from *Entamoeba histolytica*. *Biochem Biophys Res Commun*. 66(4):1389–1395.
- Ricard G, McEwan NR, Dutilh BE, Jouany JP, Macheboeuf D, Mitsumori M, McIntosh FM, Michalowski T, Nagamine T, Nelson N, et al. 2006. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*. 7(1):22.
- Schönknecht G, Chen WH, Ternes CM, Barbier GG, Shrestha RP, Stanke M, Bräutigam A, Baker BJ, Banfield JF, Garavito RM, et al. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*. 339(6124):1207–1210.
- Seigneur M, Mounier J, Prevost M-C, Guillén N. 2005. A lysine- and glutamic acid-rich protein, KERP1, from *Entamoeba histolytica* binds to human enterocytes. *Cell Microbiol*. 7(4):569–579.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212.
- Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*. 33(Web Server):W465–W467.
- Sterck L, Billiau K, Abeel T, Rouzé P, Van De Peer Y. 2012. ORCAE: online resource for community annotation of eukaryotes. *Nat Methods*. 9(11):1041–1041.
- Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4(1):41.
- Van Dellen K, Ghosh SK, Robbins PW, Loftus B, Samuelson J. 2002. *Entamoeba histolytica* lectins contain unique 6-Cys or 8-Cys chitin-binding domains. *Infect Immun*. 70(6):3259–3263.
- Vlahou G, Eliás M, von Kleist-Retzow JC, Wiesner RJ, Rivero F. 2011. The Ras related GTPase Miro is not required for mitochondrial transport in *Dictyostelium discoideum*. *Eur J Cell Biol*. 90(4):342–355.
- Walker G, Simpson AGB, Edgcomb V, Sogin ML, Patterson DJ. 2001. Ultrastructural identities of *Mastigamoeba punctachora*, *Mastigamoeba simplex* and *Mastigella commutans* and assessment of hypotheses of relatedness of the pelobionts (Protista). *Eur J Protistol*. 37(1):25–49.
- Willhoeft U, Buß H, Tannich E. 1999. DNA sequences corresponding to the ariel gene family of *Entamoeba histolytica* are not present in *E. dispar*. *Parasitol Res*. 85(8–9):787–789.
- Wilson IW, Weedall GD, Lorenzi H, Howcroft T, Hon C-C, Deloger M, Guillén N, Paterson S, Clark CG, Hall N. 2019. Genetic diversity and gene family expansions in members of the genus *Entamoeba*. *Genome Biol Evol*. 11(3):688–705.
- Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michálek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A, et al. 2015. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *Elife* 4:1–41.
- Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859–1875.
- Xiong J, Kurtz DM, Ai J, Sanders-Loehr J. 2000. A hemerythrin-like domain in a bacterial chemotaxis protein. *Biochemistry* 39(17):5117–5125.