The Genome of *Naegleria gruberi* Illuminates Early Eukaryotic Versatility

Lillian K. Fritz-Laylin[1*], Simon E. Prochnik[2*], Michael L. Ginger[3], Joel Dacks[4,5], Meredith L. Carpenter[1], Mark C. Field[5], Alan Kuo[2], Alex Paredez[1], Jarrod Chapman[2], Jonathan Pham[6], Shengqiang Shu[2], Rochak Neupane[7], Michael Cipriano[6], Joel Mancuso[8], Hank Tu[2,9], Asaf Salamov[2], Erika Lindquist[2], Harris Shapiro[2], Susan Lucas[2], Igor V. Grigoriev[2], W. Zacheus Cande[1], Chandler Fulton[10], Daniel S. Rokhsar[1,2‡], Scott C. Dawson[6‡]

[1] Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA.

[2] U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA.

[3] School of Health and Medicine, Division of Biomedical and Life Sciences, Lancaster University, Lancaster, LA1 4YQ, UK

[4] Department of Cell Biology, University of Alberta Edmonton, Alberta, Canada

[5] The Molteno Building,Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QT, UK

[6] Department of Microbiology, University of California, Davis, CA 95616, USA

[7] Center for Integrative Genomics, 545 Life Sciences Addition, University of California, Berkeley, Berkeley CA 84720, USA

[8] Gatan Inc., 5794 W. Las Positas Blvd., Pleasanton, CA 94588, USA

[9] Current address: Life Technologies, 850 Lincoln Center Drive, Foster City, CA 94404, USA

[10] Department of Biology, Brandeis University, Waltham MA, 02454-9110, USA

* These authors contributed equally to this work

‡ To whom correspondence should be addressed. dsrokhsar@gmail.com, (925) 296-5852 (D.S.R.); scdawson@ucdavis.edu, (530) 752-3633 (S.C.D.).

Summary

Genome sequences of diverse free-living protists are essential for understanding eukaryotic evolution, molecular and cell biology. The free-living amoeboflagellate *Naegleria gruberi* belongs to a varied and ubiquitous protist clade (Heterolobosea) that diverged from other eukaryotic lineages over a billion years ago. Analysis of the 15,727 protein-coding genes encoded by *Naegleria*'s 41 Mb nuclear genome indicates a capacity for both aerobic respiration and anaerobic metabolism with concomitant hydrogen production, with fundamental implications for the evolution of organelle metabolism. The *Naegleria* genome facilitates substantially broader phylogenomic comparisons of free-living eukaryotes than previously possible, allowing us to identify thousands of genes likely present in the pan-eukaryotic ancestor, with 40% likely eukaryotic inventions. Moreover, we construct a comprehensive catalog of amoeboid motility genes. The *Naegleria* genome, analysed in the context of other protists, reveals a remarkably complex ancestral eukaryote with a rich repertoire of cytoskeletal, sexual, signalling, and metabolic modules.

Introduction

Eukaryotes emerged and diversified at least a billion years ago (Brinkmann and Philippe, 2007), radiating into new niches by taking advantage of their metabolic, cytoskeletal, and compartmental complexity. Descendants of half a dozen deeply divergent, major eukaryotic clades survive, including diverse protists along with the more familiar plants, animals and fungi. These contemporary species retain some ancestral eukaryotic features along with novelties specific to their particular lineages. Here we report the genome sequence of *Naegleria gruberi*, the first from a free-living member of a major eukaryotic group which includes the pathogenic trypanosomatids. With the addition of *Naegleria*, five out of the six major eukaryotic clades now have genome sequence from free-living organisms. This is crucial as the genomes of obligate parasites are thought to be derived by gene loss and high sequence divergence (Carlton et al., 2007; Morrison et al., 2007), and are therefore not necessarily informative about the eukaryotic ancestor. Comparing the gene sets of diverse eukaryotes reveals thousands of genes present early in eukaryotic evolution, and also provides a new understanding of *Naegleria*'s remarkable versatility.

Naegleria gruberi is a free-living heterotrophic protist commonly found in both aerobic and microaerobic environments in freshwater and in moist soils around the world (De Jonckheere, 2002; Fulton, 1970, 1993). Its predominant form is a 15µm amoeba that can reproduce every 1.6 hr. when eating bacteria. Yet *Naegleria* is best known for its remarkably quick (<1.5 hr.) differentiation from amoebae to transitory streamlined flagellates with two anterior 9+2 flagella (Fig. 1) (Fulton, 1993). This change includes *de novo* assembly of an entire cytoplasmic microtubule cytoskeleton, including canonical basal bodies (Fig. 1) (Fulton, 1993). *Naegleria* also forms resting cysts, which excyst to produce amoebae (Fulton, 1970). Amoebae divide with neither nuclear envelope breakdown nor centrioles (Fulton, 1993).

Naegleria belongs to Heterolobosea, a major eukaryotic lineage that, together with the distantly related Euglenozoa (which include parasitic trypanosomes) and Jakobid flagellates, comprise the ancient and ecologically diverse clade termed “JEH” for Jakobids, Euglenozoa, Heterolobosea (Fig. 2) (Rodriguez-Ezpeleta et al., 2007). Within Heterolobosea, the genus *Naegleria* encompasses as much evolutionary diversity as the tetrapods (based on rDNA divergence (Fulton, 1993)) and includes the “brain-eating amoeba” *N. fowleri* which, although usually free-living in warm freshwater, is also an opportunistic pathogen that can cause fatal meningoencephalitis in humans (Visvesvara et al., 2007).

Although the position of the root of the eukaryotic tree remains controversial, three major hypotheses have emerged (Fig. 2 and Text S1) (Ciccarelli et al., 2006; Hampl et al., 2009;

Stechmann and Cavalier-Smith, 2002). In each hypothesis, *Naegleria* represents a critical taxon for comparative studies, alternately by being the first sequenced amoeboid bikont (Fig. 2, Root A), by allowing analysis of free-living descendants of an early common ancestor (Fig. 2, Root B), or by allowing analysis of free-living decendants of every major eukaryotic group via uniting JEH and POD into the Excavates (Fig. 2, Root C).

By parsimony, features shared between *Naegleria* and another major eukaryotic group likely existed in their common ancestor. These features would have been present early in eukaryotic evolution (i.e. before the divergence of the major eukaryotic groups that share those features (Fig. 2)), and perhaps in the ancestor of all eukaryotes. For example, *Naegleria* and humans (members of opisthokonts) diverged early (Fig 2 inset, green highlighting) so their common features were likely present by this time. (Lateral gene transfer (LGT) between eukaryotes may be the source of some shared genes, yet it is infrequent (Keeling and Palmer, 2008)).

What was the core eukaryotic gene repertoire and how did it arise and diversify? To date, eukaryotic genome sequencing has focused on opisthokonts and multicellular plants, as well as obligate parasitic protists (which tend to be genomically streamlined), although a number of free living protists have been sequenced (e.g., *Dictyostelium* (Eichinger et al., 2005), *Thalassiosira* (Armbrust et al., 2004), *Tetrahymena* (Eisen et al., 2006), *Paramecium* (Aury et al., 2006), *Chlamydomonas* (Merchant et al., 2007)). Several of these free-living protists are descendants of additional symbiosis events, so gene transfer from organellar to nuclear genomes may obscure gene ancestry. Previous phylogenomic comparisons of eukaryotes have been limited to species from two or three major groups (centered on opisthokonts and plants) (Hartman and Fedorov, 2002; Tatusov et al., 2003). Our genomic analysis includes all six major eukaryotic groups with genome sequences (circled 'G's in Fig. 2): opisthokonts, amoebozoa, plants, chromalveolates, JEH (now including free-living *Naegleria*) and POD (in which all sequenced species are obligate parasites). Analyses of individual genome sequences have tended to focus on known genes and protein domains in single taxa. Our analysis identifies both known and unknown eukaryotic gene families, begins to map out previously unexplored areas of eukaryotic biology, and highlights gene loss in every major lineage. Furthermore, we substantially extend the idea that early eukaryotes possessed extensive trafficking, cytoskeletal, sexual, metabolic, signaling, and regulatory modules (Dacks and Field, 2007; Eichinger et al., 2005; Merchant et al., 2007). We also generate a catalog of genes specifically associated with amoeboid motility, and identify an unusual capacity for both aerobic and anaerobic metabolism. Most importantly, the degree to which diverse gene families are shared among diverse major groups reveals an unexpectedly complex and versatile ancestral eukaryote.

Results and Discussion

Naegleria *genome sequence and gene set*

We assembled the 41 million base pair *N. gruberi* genome from ~8-fold redundant coverage of random paired-end shotgun sequence using genomic DNA prepared from an axenic, asexual culture of the NEG-M strain (ATCC 30224) (Fulton, 1974) (Table 1). *Naegleria* has at least twelve chromosomes (Fig. S1A), and only 5.1% repetitive sequence (Supplemental Experimental Procedures and Table S1). The genome is a mosaic of heterozygous and homozygous regions (Fig. S1A). Heterozygous regions showing two distinct haplotypes are found across 71% of the assembly, with a mean single nucleotide polymorphism frequency of 0.58%. The geometric distribution of variation in these polymorphic regions (Fig S1D) is consistent with the two haplotypes being randomly sampled from an interbreeding population (Nordborg, 2003). This implies a history of sexual recombination, despite recent clonal propagation in the laboratory. The remaining 29% of the genome comprises segments of up to hundreds of kilobases with little or no polymorphism. Assuming these homozygous regions are identical by descent, they could plausibly have arisen by gene conversion and/or inbreeding. Superimposed on the probable sexual history suggested by the geometric distribution of polymorphic variation, a genome duplication occurred in culture (Fulton, 1970, 1974), making NEG-M formally tetraploid.

In addition to its nuclear genome, NEG-M has ~4,000 copies of a sequenced extrachromosomal plasmid that encodes rDNA (Clark and Cross, 1987; Maruyama and Nozaki, 2007), and a 50 kb mitochondrial genome (GenBank AF288092).

We predicted 15,727 protein coding genes spanning 57.8% of the genome by combining *ab initio* and homology-based methods with 32,811 EST sequences (Tables S2 and S3). The assembly accounts for over 99% of the ESTs, affirming its near completeness. Nearly two-thirds (10,095) of the predicted genes are supported by EST, homology, and/or Pfam evidence. The remaining 5,632 genes may be novel, diverged, poorly-predicted or have low expression.

At least 191 *Naegleria* genes (1%) have homology to bacterial and/or archaeal, but not eukaryotic genes, making them candidates for LGT (or loss in other eukaryotic lineages). The number of potential LGT events is not unusual for free-living or parasitic protists (Armbrust et al., 2004; Berriman et al., 2005; Eichinger et al., 2005; Morrison et al., 2007). Phylogenetic analysis placed 45 of the *Naegleria* sequences in a prokaryotic clade with good bootstrap support, consistent with LGT from prokaryotes (yet coming from multiple phyla (Table S4)). Although most LGT candidate genes have unknown function, several have predicted metabolic function (including a class of formate nitrate transporter) (Table S4).

Cellular hallmarks of eukaryotes

Naegleria has many of the key features that distinguish eukaryotic cells from Bacteria and Archaea (Text S2). These features include complete actin and microtubule cytoskeletons (Tables S5 and S6 and Fig. S4), extensive meiotic, DNA replication, and transcriptional machinery (Tables S7 - S10 and see below), calcium/calmodulin mediated regulation (Table S11), transcription factors (Iyer et al., 2008), endosymbiotic organelles (mitochondria), and organelles of the membrane trafficking system (although it lacks visible Golgi, *Naegleria* contains the required genes (Dacks et al., 2003), Table S12). Additionally, *Naegleria* contains thousands more spliceosomal introns than parasitic JEH species such as *Trypanosoma brucei* (Table 1), which is consistent with other reports of parasitic JEH and POD taxa losing introns (Archibald et al., 2002; Slamovits and Keeling, 2006a). *Naegleria*'s introns include those in precisely orthologous positions in species from other eukaryotic groups (Text S2). The coding potential of the *Naegleria* genome clearly supports the early origin of all these eukaryotic hallmarks.

The sexuality of some protists, including *N. gruberi*, remains enigmatic. While many protists appear asexual, recent studies have indicated that most meiosis-specific genes were already present in the last common ancestor of all eukaryotes (Ramesh et al., 2005). These genes are present in *Naegleria* as well (Table S7). Strain NEG-M, and its parent NEG, have been maintained in the laboratory since 1967 without observing any sign of sex. However, NEG-M's heterozygosity suggests that *N. gruberi* NEG is the product of a mating. NEG is one of a cluster of independent globally-distributed isolates with consistent heterozygosity for electrophoretic variants of several enzymes (Robinson et al., 1992), a pattern which suggests asexual propagation of a widespread "natural clone" rather than frequent sexual recombination (Tibayrenc et al., 1990). The heterozygosity found in *Naegleria* is typical of a sexual organism, with perhaps infrequent matings. Additionally, identification of the core RNAi machinery indicates that *Naegleria* may use this mechanism (Table S13). Perhaps these results will encourage the discovery of conditions that induce sexuality or RNAi in *N. gruberi*, and thus bring genetic analysis to this organism.

Metabolic flexibility

Like many microbial eukaryotes, *Naegleria* oxidises glucose, various amino acids, and fatty acids via the Krebs cycle and a branched mitochondrial respiratory chain using oxygen as a terminal electron acceptor (Fig. S2; Table S14; Text S3). However, *Naegleria*'s genome also encodes features of an elaborate and sophisticated anaerobic metabolism (Fig. 3; Fig. S2; Text S3) including i) substrate-level phosphorylation reactions of the type commonly found in microaerophilic eukaryotes, such as *Entamoeba*, *Giardia*, and *Trichomonas* (Hug et al., 2009; Sanchez et al., 2000; Slamovits and Keeling, 2006b; van Grinsven et al., 2008); ii) an ability to use fumarate as an electron sink; and iii) genes encoding an Fe-hydrogenase and its associated maturation system. *Naegleria*'s anerobic and aerobic metabolism parallels the recently discovered metabolic flexibility of another soil/pond dweller, the free-living alga *Chlamydomonas* (Fig. S2) (Atteia et al., 2006; Mus et al., 2007). These protists likely use their metabolic flexibility to take advantage of the intermittent hypoxia common to muddy environments (Mus et al., 2007).

Naegleria's branched mitochondrial respiratory chain (Fig. S2, Table S14) suggests the organism is capable of oxidative phosphorylation. Many complex I subunits (NADH:ubiquinone oxidoreductase) are encoded by the mitochondrial genome (GenBank accession NC_002573), but electrons can also be transferred to ubiquinone by two alternative NADH isoforms, succinate dehydrogenase (complex II), and electron transferring flavoprotein (Fig. S2). Two terminal oxidases (cytochrome *c* oxidase and alternative oxidase) catalyse the reduction of oxygen to water.

Surprisingly, we predict that *Naegleria*'s Fe-hydrogenase and three associated maturases contain N-terminal mitochondrial transit peptides (Table S15), suggesting *Naegleria* is capable of mitochondrial hydrogen production. Fe-hydrogenases are oxygen-sensitive enzymes, strongly suggesting that *Naegleria* only produces hydrogen anaerobically. Whereas organisms with authenticated organellar Fe-hydrogenases have an accompanying maturation system (e.g. *Trichomonas vaginalis* (Putz et al., 2006) and *Chlamydomonas reinhardtii* (Posewitz et al., 2004)), organisms with cytosolic Fe-hydrogenase (e.g. *Entamoeba histolytica* and *Giardia lamblia*) do not (Putz et al., 2006). Therefore, the prediction of an Fe-hydrogenase maturation system in *Naegleria* provides further evidence that the hydrogenase is organellar (discussed further in Text S3). We know of no other mitochondrion combining such a complete a repetoire of genes for both classic aerobic respiration with predicted anaerobic hydrogen production.

Diverse lineages of anaerobic eukaryotes possess mitochondrion-derived organelles (Embley, 2006). These organelles may have additional anaerobic metabolic capabilities and are typically, relative to traditional mitochondria, missing proteins involved in oxidative phosphorylation. The recent discovery of several additional anaerobic mitochondrial-derived organelles indicates that there is a continuum of gene loss, from the mitochondria-like organelles of *Blastocystis* and *Nyctotherus* (where cytochrome-dependent respiration, and perhaps ATP synthase, appear to have been lost, but mitochondrial complex I and complex II are retained (Boxma et al., 2005; Stechmann et al., 2008)) to mitosomes that contain only a handful of proteins (Maralikova et al., 2009). *Naegleria*'s metabolically-flexible mitochondrion (with both a complete traditional mitochondrial repetoire, and an Fe-hydrogenase and maturation machinery) thus resides at the far end of this continuum of mitochondrial functions.

Although it is clear that mitochondria-derived organelles have, in many cases, secondarily lost aerobic functionality, it is difficult to ascertain whether their anaerobic functions are ancestral or adaptive. For example, although *Naegleria* and chytrid fungi Fe-hydrogenases are monophyletic, eukaryotic Fe-hydrogenases are not (Fig. S5, and (Hug et al., 2009)). This suggests organellar Fe-hydrogenases were transferred laterally into diverse anaerobic lineages . This notion is further supported by the paucity of Fe-hydrogenases in extant alpha-proteobacteria, the bacteria that gave rise to the protomitochondrion (Hug et al., 2009). On the other hand, the conservation in all eukaryotes of an Fe-hydrogenase-related protein (Nar1 in yeast (Balk et al., 2004)) strongly suggests cytosolic Fe-hydrogenases existed early in eukaryotic biology.

Although lateral gene transfer is a likely source of some organellar iron hydrogengases (e.g. ciliate Fe-hydrogenases (Boxma et al., 2007)), other organellar Fe-hydrogenases could have arisen via retargetting of an ancestrally cytosolic Fe-hydrogenase. If the first eukaryotes lived in environments with dramatic fluctuations in oxygen tension, such retargeting would aid mitochondrial redox homeostasis.

Although *Naegleria*'s energy metabolism is flexible, the organism lacks several biosynthetic pathways found in most free-living eukaryotes and some parasitic taxa (Table S16; Text S3). This fits with *Naegleria*'s nutritional requirements (including auxotrophy for methionine, purine, heme and 19 other components that define an axenic medium (Fulton et al., 1984)) and reflects the importance of *Naegleria*'s microbial predation for obtaining these nutrients. However, the lack of cytoplasmic (Type I) fatty acid biosynthesis genes in *Naegleria* and *Dictyostelium* is particularly surprising, as both amoebae can grow without exogenous lipids (Franke and Kessin, 1977; Fulton et al., 1984). Both amoebae do contain multiple fatty acid elongases indicative of Type III fatty acid synthesis, suggesting that the Type III pathway substitutes for the missing Type I pathway in *Naegleria*. This also implies a wider phylogenetic distribution of a pathway previously limited to trypanosomes (Lee et al., 2007; Lee et al., 2006).

Conserved amoeboid and flagellar motility genes in the eukaryotic ancestor

Flagellar motility is found in every major eukaryotic group (Fig. 2), and is undoubtedly an ancestral feature (Cavalier-Smith, 2002). As actin-based amoeboid locomotion is found in many diverse eukaryotic lineages, this form of motility likely arose early in eukaryotic evolution, perhaps even in the eukaryotic ancestor (depending on the position of the eukayotic root, Fig. 2) (Cavalier-Smith, 2002; Fulton, 1970). By searching for genes present only in organisms that possess each type of locomotion (e.g. genes found in organisms with flagella and missing from organisms without flagella) we identified sets of genes enriched in functions specific to flagellar motility (Flagellar-Motility associated genes (FMs)) or amoeboid motility (Amoeboid-Motility associated genes (AMs)) (Fig. 4). These phylogenetic profiles (Li et al., 2004) exclude genes that are used both for motility and other processes (e.g. alpha tubulin, which is used in flagella, but also mitotic spindles), and will also include some false positives. *Naegleria*'s repertoire of 173 FMs is consistent with its typical eukaryotic flagellar structure (Dingle and Fulton, 1966) (Fig. 1). FMs also include proteins required for basal body assembly, flagellar beating, intraflagellar transport and 36 novel flagella-associated genes (Table S17).

Here we present a catalog of proteins specifically associated with amoeboid motility. The actin cytoskeleton enables amoeboid motility and diverse cellular processes including cytokinesis, endocytosis, and maintenance of cell morphology and polarity. We identified 63 gene families (AMs) found only in organisms with cells capable of amoeboid locomotion (Table S18). By definition, the AM list does not include proteins which also play a role in non-motile functions such as actin, Arp2/3 (which nucleates actin filaments) or other general actin cytoskeletal components, since these genes are found across eukaryotes regardless of their capacity for amoeboid locomotion. Nineteen AMs have unknown function, but are strongly implicated in actin-based motility (Table S18).

The AMs include several genes thought to keep pseudopod actin filaments densely packed, highly branched, and properly positioned. For example, the Arp2/3 activator WASH (AM5) is proposed to activate actin filament formation in pseudopodia (Linardopoulou et al., 2007). The actin binding protein twinfilin (AM4) affects the relative sizes of functionally distinct pseudopodial subcompartments (Iwasa and Mullins, 2007). Filamin (AM3) stabilizes the three-dimensional actin networks necessary for amoeboid locomotion (Flanagan et al., 2001). Drebrin/ABP1 (AM2) aids in membrane attachment of actin filaments during endocytosis in yeast (Toret and Drubin, 2006), and could also function in cell migration (Peitsch et al., 2006; Song et al., 2008). The inclusion of both twinfilin and drebrin/ABP1 in the AMs argues that the actin patches formed during yeast endocytosis could have evolutionary origins in amoeboid motility.

Our analysis also suggests a role for the lipid sphingomyelin in amoeboid motility. AMs include a sphingomyelin-synthase-related protein (AM16) and Saposin-B-like proteins (AM17) that activate sphingomyelinase. (Sphingomyelinase is not in the AM set because it is found in the non-amoeboid *Paramecium* (Fig. 4).) As sphingomyelin is enriched in the pseudopodia of human amoeboid cells(Jandak et al., 1990), we suggest it (or perhaps a family of related ceramides) may contribute to motility via structural differentiation of the membrane, or as a second messenger in signaling pathways.

Signaling complexity

The genome encodes an extensive array of signaling machinery that likely orchestrates *Naegleria's* complex behavior. This repertoire includes entire pathways not found in parasitic protists (Fig. 5), as well as at least 265 predicted protein kinases, 32 protein phosphatases (Table S11), and 182 monomeric Ras-like GTPases. For example, *Naegleria* has thirty putative hybrid histidine kinases and six response receiver domain-proteins whereas *T. brucei*, *Giardia*, and *Entamoeba* have none (Berriman et al., 2005; Loftus et al., 2005; Morrison et al., 2007). *Naegleria* also contains extensive G-protein coupled receptor (GPCR) pathways missing from *Giardia* and *T. brucei* (Text S4).

Many organisms sense their environment via membrane-bound adenylate/guanylate cyclases. *Naegleria* contains at least 108 cyclases—almost twice that found in the human genome (Fig. S3), although the reason for this abundance remains puzzling. Nearly half contain PAS signal-sensing domains and four are paired with NIT domains that are used by bacteria to sense nitrate and nitrite concentrations (Shu et al., 2003). Four cyclases also have BLUF domains, a domain combination used by *Euglena* for photoresponsive behavior (Ntefidou et al., 2003). *Naegleria* might have subtle photoresponsive behavior, or use BLUF domains for redox sensing.

Inferring the protein complement of the eukaryotic ancestor

What genes were present in the common ancestor of all eukaryotes? Prior inventories of ancestral eukaryotic genes have been based on two or three eukaryotic groups (Hartman and Fedorov, 2002; Tatusov et al., 2003). This limited sampling, and the limited availability of free-living protist genome sequences, may have significantly underestimated the protein complement of the eukaryotic common ancestor. We used 17 genomes from all six major groups, and constructed 4,133 ancient eukaryotic gene families, requiring: i) a minimum of one *Naegleria* protein and two orthologs, and ii) one ortholog from another major eukaryotic group. These ancient gene families are conceptually similar to KOGs (euKaryotic clusters of Orthologous Groups), which were based on genes shared between several opisthokonts (Fig. 2) and *Arabidopsis* (Tatusov et al., 2003).

By including proteins from species in more diverse groups (i.e., in addition to plants and opisthokonts) as well as *Naegleria*, we added 1,292 ancient eukaryotic gene families to the KOG analysis. 481 of these additional ancient families also lack Pfam domains. This implies that these families encode deeply conserved, but as yet undetermined, biological activities. Further, these 481 ancient families are broadly conserved, with 45% present in at least five of the six major eukaryotic groups (Table S19).

As the number of major eukaryotic groups represented in an ancient protein family increases, we become more confident that the gene was present in the eukaryotic ancestor. The majority (92%) of the 4,133 ancient gene families are present in at least three eukaryotic groups, and nearly half (1,983) of the ancient gene families are present in all five major eukaryotic groups that include a genome sequence from a free-living species (Fig. 2). This estimate of the core eukaryotic gene repertoire is conservative, as it does not include ancestral genes lost from *Naegleria*, or genes whose sequence evolution prevents us from detecting homology.

Although pronounced gene loss from parasitic lineages has been well described (Berriman et al.; Morrison et al., 2007), loss of gene families from entire major eukaryotic groups has not been investigated on a genome-wide scale. Compared to the JEH group, other major lineages have lost 16 to 59% of the 4,133 ancient gene families, with substantially more losses observed in parasitic lineages (Table S20). Losses also likely occurred in the JEH lineage, as 1,139 KOGs are not found in JEH. Being the closest sequenced free-living organism to the parasitic trypanosomes, the genome of *Naegleria* provides new insight into the evolution of major pathogens such as *Trypanosoma brucei*, which has lost 2,424 ancient eukaryotic families (Table S20). Because all sequenced organisms (including *Naegleria*) have lost genes, sequencing more genomes, (particularly those of free-living species from groups where only parasitic taxa have been sequenced, e.g., POD), will likely reveal additional ancient gene families.

Origin of eukaryotic genes

Which of these ancient gene families are shared with archea and/or bacteria, and which are specific to eukayotes? To investigate the origin of ancient eukaryotic gene families, we compared each of the 4,133 families to prokaryotic (archaeal and bacterial) protein sequences. Approximately 57% (2,361) have clearly recognizable homologs in prokaryotes, and therefore arose before the emergence of eukaryotes (and possibly were transferred to eukaryotes from the mitochondrial genome) (“ancient”; Fig. 6A). Conversely, 40% (1,421) appear to be novel to the eukaryotic lineage, with no detectable homology in prokaryotic genomes (“novel”, Table S21). A similar analysis that required presence in the parasite *Giardia* found only 347 Eukaryotic Signature Proteins (Hartman and Fedorov, 2002). The 1,421 novel eukaryotic genes emerged in recognizably modern form early in eukaryotic history, if not on the eukaryotic stem, and likely encode much of what is needed to be a eukaryote. The novel protein set is most enriched in functions relating to intracellular trafficking, signal transduction and ubiquitin-based protein degradation, and to a lesser extent, cytoskeletal and RNA-processing genes (Fig. 6B). About 40% of protein families in the eukaryotic lineage are novel compared to prokaryotes. In contrast, only about 20% of protein families in metazoa are novel relative to other eukaryotes (Fig. 6A) (Putnam et al., 2007). The larger fraction of eukaryotic novelties (compared to metazoan novelties) may reflect the magnitude of change accompanying the transition to early eukaryotes, whether eukaryotes arose from bacteria/archaeal ancestors or another ancestral life form (Hartman and Fedorov, 2002; Kurland et al., 2006).

In addition to *de novo* inventions, 232 eukaryotic proteins arose by evolutionary tinkering such as domain addition. The proteins in 140 families (Table S22) share a domain with the prokaryotic homolog, but have gained a novel eukaryotic-specific domain (“additions”). An example is the addition of a eukaryotic poly-A binding domain to a

RNA-recognition motif that is also present in prokaryotes (Mangus et al., 2003). An additional 92 families (Table S23) are eukaryotic fusions of domains found in separate polypeptides in prokaryotes ("fusions"), including a previously described example of archeal DNA ligase that combined with a BRCT domain in eukaryotes (Bork et al., 1997).

Concluding discussion

Evolutionary biologist George Gaylord Simpson presciently claimed that "All the essential problems of living organism[s] are already solved in the one-celled ... protozoan and these are only elaborated in man" (Simpson, 1949). Simpson's intuition runs counter to the long-held view that a great gulf separates "simple" or "lower" unicellular protists from "higher" multicellular organisms. By comparing eukaryotic genomes across a greater evolutionary span than previously possible (Fig. 2), the genome of *Naegleria* reveals unexpectedly rich versatility in early eukaryotic ancestors, and well as highlighting losses in parasites. *Naegleria*'s numerous introns, complex DNA and RNA metabolism, flexible metabolic and signaling capabilities, and capacity for both amoeboid and flagellar motility provide direct genomic evidence for the early evolution of molecular hallmarks of so-called "complex" eukaryotes. These extensive capabilities were required by the long-extinct common ancestor, and are still needed for *Naegleria*'s versatility as a free-living, predatory cell, able to assume radically distinct phenotypes and to live in diverse environments. In Simpson's sense, it was a giant step to an amoeba, yet a small step to man.

Experimental Procedures

See Supplemental Experimental Procedures for further details for all procedures.

Genome sequencing, assembly, annotation

We sequenced genomic DNA from an axenic culture of *Naegleria gruberi* strain NEG-M (ATCC 30224) grown from a frozen stock. The draft *N. gruberi* assembly was generated from paired-end whole genome shotgun sequence at 8× coverage using v. 2.9 of the assembler JAZZ. 15,727 gene models were predicted by combining EST, homology and *ab initio* data and annotated using the JGI annotation pipeline.

Curation of genes associated with cellular functions

Naegleria homologs of proteins involved in cellular processes were identified by BLAST and PFAM searches using published proteins as queries.

Determining lateral gene transfer

We added homologs to *Naegleria* proteins that have homology to prokaryotes but not eukaryotes and built phylogenetic trees to assess the evolutionary origin of these proteins.

Construction of protein families

To create protein families, we BLASTed (Altschul et al., 1990) each of the 15,727 protein sequences in *Naegleria* to all protein sequences in a wide range of eukaryotes and a cyanobacterium, then generated ortholog pairs (mutual best BLAST hits with E-value <

1E-10) consisting of one *Naegleria* protein and a protein from another organism. Paralogs from a given organism were added whenever a paralog's p-dist (defined as 1 - the fraction of identical amino acids in the two proteins' alignment) from the putative ortholog in the same organism was less than a certain fraction (0.5 for comparisions between two eukaryotes and 0.1 for *Naegleria* and the cyanobacterium) of the p-dist between the two orthologs in the pair. Lastly, all sets of two orthologs plus paralogs were merged if they contained the same *Naegleria* protein. We created 5,107 families of homologous proteins, plus 8 families restricted to *Naegleria* and the cyanobacterium *Prochlorococcus*.

Inferring the protein complement of the eukaryotic ancestor

We identified a subset of 4,133 ancient eukaryotic gene families that contain a minimum of one *Naegleria* protein and two orthologs, and that at least one of the orthologs be from another major eukaryotic group.

To predict protein function where possible, we assigned majority rule KOG annotations (Tatusov et al., 2003) to each family in two steps. First, each protein in the family was searched against the KOG sequence database (Tatusov et al., 2003) with RPS-BLAST (Altschul et al., 1990) and the best hit with E-value < 1E-5 was retained. (This slightly relaxed E-value was chosen because *Naegleria*'s protein sequences are divergent and the value had worked well compared to more stringent cutoffs for assigning PFAMs.) Second, if the commonest KOG annotation in a protein family was in at least half the proteins in a family, that KOG was assigned to the family.

While it is possible that an ancestral eukaryotic protein could be present in more than one eukaryotic group due to inter-eukaryotic lateral gene transfer, this process is rare (Keeling and Palmer, 2008). In addition 92% of the 4,133 ancient eukaryotic gene families are present in at least three major eukaryotic groups making lateral gene transfer unparsimonious in most scenarios.

The origin of eukaryotic genes

To ask whether each of the 4,133 ancient eukaryotic protein families (see above) had been inherited from prokaryotes (i.e. from Archaea/Bacteria), or were eukaryotic inventions, or some combination of these two scenarios, we first constructed a “centroid” sequence for each of ancient protein family, defined as the hypothetical protein sequence that maximizes the sum of BLAST alignment scores between the centroid and the protein sequences in the family. Thus, each centroid sequence acts as a proxy for the ancestral protein sequence. We next made a set of all prokaryotic (taxonomy ID = 2 (Bacteria) or 2157 (Archaea)) proteins in the UniRef90 protein database (Benson et al., 2009) and searched these proteins for homology (E-value < 1E-6) to each centroid sequence. If the centroid sequence had no hit to a prokaryotic protein it was classified as eukaryotic-specific (Fig 6A, “novel”). We found 1,421 such “novel” protein families.

In the following classification steps, we compared Pfam domain annotations in the eukaryotic centroid and prokaryotic sequences. We classified protein families as “ancient” if the centroid and the best hitting prokaryotic protein met any of the following criteria: i) neither sequence has a Pfam (Finn et al., 2008) domain; ii) the two sequences have the same combination of pairwise domains; iii) the two sequences have another simple pattern of domain gain/loss that does not imply novelty in the eukaryotic lineage. This class of ancient proteins has 2,361 protein families. The remaining protein families showed some degree of innovation in eukaryotes relative to their prokaryotic homologs. The first class had no homolog in prokaryotic genomes (1,421 "novel" families, Table S21). The second class had extra eukaryotic-specific domain(s) (140 "addition" families, Table S22). The third class had been formed by the fusion in eukaryotes of multiple ubiquitous domains into a single polypeptide (92 "fusion" families, Table S23). Some proteins showed domain innovations in both the second and third classes, in which case the commonest type of innovation was chosen. Ties were left unclassified and joined the remaining 119 families with more complex evolutionary patterns. These proteins showed for example evidence of evolutionary splitting of multi-domain prokaryotic polypeptides into different proteins in eukaryotes, conceptually the opposite of the "fusion" category.

Majority-rule KOGs were assigned as described above (Fig. 6B).

Generation of Flagellar Motility-associated proteins (FMs)

Genes associated with flagellar function have been identified by phylogenetic profiling (Avidor-Reiss et al., 2004; Li et al., 2004; Merchant et al., 2007). We generated a list of proteins associated with flagellar function by searching the *Naegleria* protein families (see above) for those that contain proteins from organisms with flagella (*Naegleria*, *Chlamydomonas*, and human) and none from organisms lacking flagella (*Dictyostelium*, *Neurospora*, *Arabidopsis* and *Prochlorococcus*). This analysis resulted in 182 *Naegleria* proteins in 173 families (Table S17), which we named FMs (Flagellar Motility associated proteins).

Generation of Amoeboid Motility-associated proteins (AMs)

We used phylogenetic profiling (see above) to generate a catalog of proteins associated with amoeboid motility. We searched the *Naegleria* protein families (see above) for those that contain proteins from organisms that undergo amoeboid movement [*Naegleria*, human, and at least one Amoebozoan (*Dictyostelium* or *Entamoeba*)], but not in organisms that have no amoeboid movement [*Prochlorococcus*, Arabidopsis, *Physcomitrella*, Diatom, *Paramecium*, Trypanosome, *Giardia*, *Chlamydomonas*] (Table S18).

Acknowledgements

We thank Woodrow Fischer, Matt Welch, Dyche Mullins, David Drubin and Anosha Siripala for discussions; Jeremy Thorner, Nicole King, Jason Stajich, Elaine Lai and Stephen Remillard for comments on the manuscript; Zoe Assaf for manuscript editing. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. Additional support was provided by a Royal Society University Research

Fellowship to MLG, a Tien Scholars Fellowship in Environmental Sciences and Biodiversity to LKFL, a Wellcome Trust and Parke-Davis Fellowship, as well as start-up funding from the U. of Alberta, to JBD, and funding by Wellcome Trust for MCF.

Accession numbers: The genome assembly, predicted gene models and annotations are being deposited at DDBJ/EMBL/GenBank under accession number ACER00000000.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Archibald, J.M., O'Kelly, C.J., and Doolittle, W.F. (2002). The chaperonin genes of jakobid and jakobid-like flagellates: implications for eukaryotic evolution. Mol Biol Evol *19*, 422-431.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M.*, et al.* (2004). The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science *306*, 79-86.

Atteia, A., van Lis, R., Gelius-Dietrich, G., Adrait, A., Garin, J., Joyard, J., Rolland, N., and Martin, W. (2006). Pyruvate formate-lyase and a novel route of eukaryotic ATP synthesis in Chlamydomonas mitochondria. J Biol Chem *281*, 9909-9918.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Segurens, B., Daubin, V., Anthouard, V., Aiach, N.*, et al.* (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature *444*, 171-178.

Avidor-Reiss, T., Maer, A.M., Koundakjian, E., Polyanovsky, A., Keil, T., Subramaniam, S., and Zuker, C.S. (2004). Decoding cilia function: defining specialized genes required for compartmentalized cilia biogenesis. Cell *117*, 527-539.

Balk, J., Pierik, A.J., Netz, D.J., Muhlenhoff, U., and Lill, R. (2004). The hydrogenase-like Nar1p is essential for maturation of cytosolic and nuclear iron-sulphur proteins. EMBO J *23*, 2105-2115.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2009). GenBank. Nucleic Acids Res *37*, D26-31.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D.C., Lennard, N.J., Caler, E., Hamlin, N.E., Haas, B.*, et al.* (2005). The genome of the African trypanosome Trypanosoma brucei. Science *309*, 416-422.

Bork, P., Hofmann, K., Bucher, P., Neuwald, A.F., Altschul, S.F., and Koonin, E.V. (1997). A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. FASEB J *11*, 68-76.

Boxma, B., de Graaf, R.M., van der Staay, G.W., van Alen, T.A., Ricard, G., Gabaldon, T., van Hoek, A.H., Moon-van der Staay, S.Y., Koopman, W.J., van Hellemond, J.J.*, et al.* (2005). An anaerobic mitochondrion that produces hydrogen. Nature *434*, 74-79.

Boxma, B., Ricard, G., van Hoek, A.H., Severing, E., Moon-van der Staay, S.Y., van der Staay, G.W., van Alen, T.A., de Graaf, R.M., Cremers, G., Kwantes, M.*, et al.* (2007). The [FeFe] hydrogenase of Nyctotherus ovalis has a chimeric origin. BMC Evol Biol *7*, 230.

Brinkmann, H., and Philippe, H. (2007). The diversity of eukaryotes and the root of the eukaryotic tree. Adv Exp Med Biol *607*, 20-37.

Burki, F., Shalchian-Tabrizi, K., and Pawlowski, J. (2008). Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. Biol Lett *4*, 366.

Carlton, J.M., Hirt, R.P., Silva, J.C., Delcher, A.L., Schatz, M., Zhao, Q., Wortman, J.R., Bidwell, S.L., Alsmark, U.C., Besteiro, S., *et al.* (2007). Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. Science *315*, 207-212.

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. Int J Syst Evol Microbiol *52*, 297-354.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283-1287.

Clark, C.G., and Cross, G.A. (1987). rRNA genes of Naegleria gruberi are carried exclusively on a 14- kilobase-pair plasmid. Mol Cell Biol *7*, 3027-3031.

Dacks, J.B., Davis, L.A.M., Sjogren, A.M., Andersson, J.O., Roger, A.J., and Doolittle, W.F. (2003). Evidence for Golgi bodies in proposed 'Golgi-lacking' lineages. Proc Biol Sci *270 Suppl 2*, S168-171.

Dacks, J.B., and Field, M.C. (2007). Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. J Cell Sci *120*, 2977-2985.

De Jonckheere, J.F. (2002). A century of research on the amoeboflagellate genus *Naegleria*. Acta Protozoologica *41*, 309-342.

Dingle, A.D., and Fulton, C. (1966). Development of the flagellar apparatus of *Naegleria.* J Cell Biol *31*, 43-54.

Eichinger, L., Pachebat, J.A., Glockner, G., Rajandream, M.A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., *et al.* (2005). The genome of the social amoeba Dictyostelium discoideum. Nature *435*, 43-57.

Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., *et al.* (2006). Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol *4*, e286.

Embley, T.M. (2006). Multiple secondary origins of the anaerobic lifestyle in eukaryotes. Philos Trans R Soc Lond B Biol Sci *361*, 1055-1067.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., *et al.* (2008). The Pfam protein families database. Nucleic Acids Res *36*, D281-288.

Flanagan, L.A., Chou, J., Falet, H., Neujahr, R., Hartwig, J.H., and Stossel, T.P. (2001). Filamin A, the Arp2/3 complex, and the morphology and function of cortical actin filaments in human melanoma cells. J Cell Biol *155*, 511-517.

Franke, J., and Kessin, R. (1977). A defined minimal medium for axenic strains of Dictyostelium discoideum. Proc Natl Acad Sci U S A *74*, 2157-2161.

Fulton, C. (1970). Amebo-flagellates as research partners: The laboratory biology of *Naegleria* and *Tetramitus*. Methods Cell Physiol *4*, 341-476.

Fulton, C. (1974). Axenic cultivation of *Naegleria gruberi*. Requirement for methionine. Exp Cell Res *88*, 365-370.

Fulton, C. (1993). *Naegleria* : A research partner for cell and developmental biology. Journal of Eukaryotic Microbiology *40*, 520-532.

Fulton, C., Webster, C., and Wu, J.S. (1984). Chemically defined media for cultivation of *Naegleria gruberi*. Proc Natl Acad Sci USA *81*, 2406-2410.

Hampl, V., Hug, L., Leigh, J.W., Dacks, J.B., Lang, B.F., Simpson, A.G., and Roger, A.J. (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A *106*, 3859-3864.

Hartman, H., and Fedorov, A. (2002). The origin of the eukaryotic cell: a genomic investigation. Proc Natl Acad Sci U S A *99*, 1420-1425.

Hug, L.A., Stechmann, A., and Roger, A.J. (2009). Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes. Mol Biol Evol.

Iwasa, J.H., and Mullins, R.D. (2007). Spatial and temporal relationships between actin-filament nucleation, capping, and disassembly. Curr Biol *17*, 395-406.

Iyer, L.M., Anantharaman, V., Wolf, M.Y., and Aravind, L. (2008). Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. Int J Parasitol *38*, 1-31.

Jandak, J., Li, X.L., Kessimian, N., and Steiner, M. (1990). Unequal distribution of membrane components between pseudopodia and cell bodies of platelets. Biochim Biophys Acta *1029*, 117-126.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet *9*, 605-618.

Kurland, C.G., Collins, L.J., and Penny, D. (2006). Genomics and the irreducible nature of eukaryote cells. Science *312*, 1011-1014.

Lee, S.H., Stephens, J.L., and Englund, P.T. (2007). A fatty-acid synthesis mechanism specialized for parasitism. Nat Rev Microbiol *5*, 287-297.

Lee, S.H., Stephens, J.L., Paul, K.S., and Englund, P.T. (2006). Fatty acid synthesis by elongases in trypanosomes. Cell *126*, 691-699.

Li, J.B., Gerdes, J.M., Haycraft, C.J., Fan, Y., Teslovich, T.M., May-Simera, H., Li, H., Blacque, O.E., Li, L., Leitch, C.C., *et al.* (2004). Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell *117*, 541-552.

Linardopoulou, E.V., Parghi, S.S., Friedman, C., Osborn, G.E., Parkhurst, S.M., and Trask, B.J. (2007). Human subtelomeric WASH genes encode a new subclass of the WASP family. PLoS Genet *3*, e237.

Loftus, B., Anderson, I., Davies, R., Alsmark, U.C., Samuelson, J., Amedeo, P., Roncaglia, P., Berriman, M., Hirt, R.P., Mann, B.J., *et al.* (2005). The genome of the protist parasite Entamoeba histolytica. Nature *433*, 865-868.

Mangus, D.A., Evans, M.C., and Jacobson, A. (2003). Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. Genome Biol *4*, 223.

Maralikova, B., Ali, V., Nakada-Tsukui, K., Nozaki, T., van der Giezen, M., Henze, K., and Tovar, J. (2009). Bacterial-type oxygen detoxification and iron-sulphur cluster assembly in amoebal relict mitochondria. Cell Microbiol.

Maruyama, S., and Nozaki, H. (2007). Sequence and intranuclear location of the extrachromosomal rDNA plasmid of the amoebo-flagellate Naegleria gruberi. J Eukaryot Microbiol *54*, 333-337.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Marechal-Drouard, L., *et al.* (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science *318*, 245-250.

Morrison, H.G., McArthur, A.G., Gillin, F.D., Aley, S.B., Adam, R.D., Olsen, G.J., Best, A.A., Cande, W.Z., Chen, F., Cipriano, M.J., *et al.* (2007). Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science *317*, 1921-1926.

Mus, F., Dubini, A., Seibert, M., Posewitz, M.C., and Grossman, A.R. (2007). Anaerobic acclimation in Chlamydomonas reinhardtii: anoxic gene expression, hydrogenase induction, and metabolic pathways. J Biol Chem *282*, 25475-25486.

Nordborg, M. (2003). Coalescent theory. In Handbook of statistical genetics, D.J. Balding, M. Bishop, and C. Cannings, eds. (Hoboken, NJ, Wiley).

Ntefidou, M., Iseki, M., Watanabe, M., Lebert, M., and Hader, D.P. (2003). Photoactivated adenylyl cyclase controls phototaxis in the flagellate Euglena gracilis. Plant Physiol *133*, 1517-1521.

Peitsch, W.K., Bulkescher, J., Spring, H., Hofmann, I., Goerdt, S., and Franke, W.W. (2006). Dynamics of the actin-binding protein drebrin in motile cells and definition of a juxtanuclear drebrin-enriched zone. Exp Cell Res *312*, 2605-2618.

Posewitz, M.C., King, P.W., Smolinski, S.L., Zhang, L., Seibert, M., and Ghirardi, M.L. (2004). Discovery of two novel radical S-adenosylmethionine proteins required for the assembly of an active [Fe] hydrogenase. J Biol Chem *279*, 25711-25720.

Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., *et al.* (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science *317*, 86-94.

Putz, S., Dolezal, P., Gelius-Dietrich, G., Bohacova, L., Tachezy, J., and Henze, K. (2006). Fe-hydrogenase maturases in the hydrogenosomes of Trichomonas vaginalis. Eukaryot Cell *5*, 579-586.

Ramesh, M.A., Malik, S.B., and Logsdon, J.M., Jr. (2005). A phylogenomic inventory of meiotic genes; evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. Curr Biol *15*, 185-191.

Robinson, B.S., Christy, P., Hayes, S.J., and Dobson, P.J. (1992). Discontinuous genetic variation among mesophilic *Naegleria* isolates: further evidence that *N. gruberi* is not a single species. Journal of Protozoology *39*, 702-712.

Rodriguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A.J., Gray, M.W., Philippe, H., and Lang, B.F. (2007). Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Curr Biol *17*, 1420-1425.

Sanchez, L.B., Galperin, M.Y., and Muller, M. (2000). Acetyl-CoA synthetase from the amitochondriate eukaryote Giardia lamblia belongs to the newly recognized superfamily of acyl-CoA synthetases (Nucleoside diphosphate-forming). J Biol Chem *275*, 5794-5803.

Shu, C.J., Ulrich, L.E., and Zhulin, I.B. (2003). The NIT domain: a predicted nitrate-responsive module in bacterial sensory receptors. Trends Biochem Sci *28*, 121-124.

Simpson, G.G. (1949). The Meaning of Evolution. A Study of the History of Life and of Its Significance for Man (New Haven, Yale University Press).

Slamovits, C.H., and Keeling, P.J. (2006a). A high density of ancient spliceosomal introns in oxymonad excavates. BMC Evol Biol *6*, 34.

Slamovits, C.H., and Keeling, P.J. (2006b). Pyruvate-phosphate dikinase of oxymonads and parabasalia and the evolution of pyrophosphate-dependent glycolysis in anaerobic eukaryotes. Eukaryot Cell *5*, 148-154.

Song, M., Kojima, N., Hanamura, K., Sekino, Y., Inoue, H.K., Mikuni, M., and Shirao, T. (2008). Expression of drebrin E in migrating neuroblasts in adult rat brain: coincidence between drebrin E disappearance from cell body and cessation of migration. Neuroscience *152*, 670-682.

Stechmann, A., and Cavalier-Smith, T. (2002). Rooting the eukaryote tree by using a derived gene fusion. Science *297*, 89-91.

Stechmann, A., Hamblin, K., Perez-Brocal, V., Gaston, D., Richmond, G.S., van der Giezen, M., Clark, C.G., and Roger, A.J. (2008). Organelles in Blastocystis that blur the distinction between mitochondria and hydrogenosomes. Curr Biol *18*, 580-585.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N.*, et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tibayrenc, M., Kjellberg, F., and Ayala, F.J. (1990). A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. Proc Natl Acad Sci U S A *87*, 2414-2418.

Toret, C.P., and Drubin, D.G. (2006). The budding yeast endocytic pathway. J Cell Sci *119*, 4585-4587.

van Grinsven, K.W., Rosnowsky, S., van Weelden, S.W., Putz, S., van der Giezen, M., Martin, W., van Hellemond, J.J., Tielens, A.G., and Henze, K. (2008). Acetate:succinate CoA-transferase in the hydrogenosomes of Trichomonas vaginalis: identification and characterization. J Biol Chem *283*, 1411-1418.

Visvesvara, G.S., Moura, H., and Schuster, F.L. (2007). Pathogenic and opportunistic free-living amoebae: Acanthamoeba spp., Balamuthia mandrillaris, Naegleria fowleri, and Sappinia diploidea. FEMS Immunol Med Microbiol *50*, 1-26.

Yoon, H.S., Grant, J., Tekle, Y.I., Wu, M., Chaon, B.C., Cole, J.C., Logsdon, J.M.J., Patterson, D.J., Bhattacharya, D., and Katz, L.A. (2008). Broadly sampled multigene trees of eukaryotes. BMC Evol Biol *8*, 14.

Figure Legends

Figure 1. Schematic of *Naegleria* amoeba and flagellate forms.

Naegleria amoebae move along a surface with a large blunt pseudopod. Changing direction (arrows) follows the eruption of a new, usually anterior, pseudopod. *Naegleria* maintains fluid balance using a contractile vacuole. The nucleus contains a large nucleolus. The cytoplasm has many mitochondria and food vacuoles which are excluded from pseudopods. Flagellates also contain canonical basal bodies and flagella (insets). Basal bodies are connected to the nuclear envelope via a single striated rootlet. See also Tables S7, S12, S13, and Text S2.

Figure 2. Consensus cladogram of selected eukaryotes.

Consensus cladogram of selected eukaryotes relevant to our comparative analyses, highlighting six major groups with widespread support in diverse molecular phylogenies (Burki et al., 2008; Rodriguez-Ezpeleta et al., 2007; Yoon et al., 2008). The dotted polytomy indicates uncertainty regarding the order of early branching events. Representative taxa are shown on the right, with glyphs indicating flagellar and/or actin-based amoeboid movement. Although commonly referred to as “amoeboid”, *Trichomonas* does not undergo amoeboid locomotion. The inset depicts three contending hypotheses for the root. Root A: early divergence of unikonts and bikonts (Stechmann and Cavalier-Smith, 2002). Root B: the largely parasitic POD lineage branching first, followed by JEH (including *Naegleria*) (Ciccarelli et al., 2006). Root C: POD and JEH uniting to form the "excavates" (Supplemental Data). The branches connecting *Naegleria* to humans are highlighted in green, with a black triangle indicating their last common ancestor. See also Text S1.

Figure 3. A model for anaerobic fermentation in *Naegleria*.

Likely fermentation pathways used by *N. gruberi* under hypoxic or anoxic conditions are shown. Solid arrows indicate individual enzyme-catalysed reactions, noting key nucleotide or co-enzyme interconversions. Predicted fermentation end-products are colored red. We cannot predict whether a NADH dehydrogenase transfers electrons directly from NADH for H_2 production (shown) or if electrons are transferred from NADH to 2Fe-2S ferredoxin first (Fig. S2). The HydE, HydF, and HydG Fe-hydrogenase maturation components (orange) are predicted to be mitochondrially-targetted. Question marks indicate uncertainty regarding whether (lower centre) an active mitochondrial complex I (mcI) pumps protons across the mitochondrial inner membrane, (lower right) a proton motive force is used for ATP generation, (upper right) ATP hydrolysis is used to generate mitochondrial membrane potential, and additionally (lower left), the co-substrate used by soluble fumarate reductase. See also Fig. S2 and S5, Tables S14-S16, and Text S3.

Figure 4. Phylogenetic distribution of selected genes associated with ameboid motility (AMs) and flagellar motility (FMs).

We show the presence (green) or absence (white) of genes listed at bottom in species indicated on the left (except for Amoebozoans because AM proteins must be present in at least one of *Dictyostelium* and *Entamoeba*). Glyphs at the side indicate species with flagellar and/or actin-based amoeboid locomotion. S.S.R., sphingomyelin-synthase-related protein. See also Fig. S4 and Tables S5, S6, S17, S18.

Figure 5. *Naegleria* signaling modules.

The *Naegleria* genome encodes GPCR and histidine kinase signaling; two modules missing in some parasites (dotted boxes). Predicted numbers of proteins are indicated. RGS, regulator of G-protein signaling; GEF, guanine nucleotide exchange factor; GAP, GTPase activating protein; PDE, phosphodiesterase; A/G cyclase, adenylate/guanylate cyclase; PLC-beta, phopholipase-C beta; IP3, inositol-1,4,5-triphosphate; PIP2, phosphatidylinositol-4,5-bisphosphate; PIP3, phosphatidylinositol-3,4,5-triphosphate; PTEN, phosphatase and tensin homologue; PI3K phosphatidylinositol-3-OH kinase. See also Fig. S3, Table S11 and Text S4.

Figure 6: Ancient origin and innovation in eukaryotic proteins.

Schematics of the four scenarios of protein origin we consider are along the bottom, and color-coded in the charts: ancient (blue), novel (green), addition of a eukaryote-specific protein domain (orange), and eukaryotic-specific fusion of two domains (red). The protein families that could be categorized are presented in (A) overview pie charts comparing the origins of protein families in ancient eukaryotes (top) and animals (bottom, from Putnam et al., 2007) and (B) stacked barcharts showing subsets of the ancient eukaryotic families divided by KOG function , omitting unknown and general KOG functions. prok, prokaryotic (i.e. archaea and/or bacteria); euk, eukaryotic; Trans, translational. See also Tables S4, S19-S23.

Tables

Table 1: Genome statistics from *Naegleria gruberi* and selected species. See also Fig. S1, and Tables S1-S3, S8-S10.

n.d. not determined.

Species	Genome Size (Mb.p.)	No. chromo-somes	%GC	Protein coding loci	% coding	% genes w/ introns	Introns per gene	Median intron length (b.p.)
Naegleria	41	>=12	33	15,727	57.8	36	0.7	60
Human	2851	23	41	23,328	1.2	83	7.8	20,383
Neurospora	40	7	54	10,107	36.4	80	1.7	72
Dictyostelium	34	6	22	13,574	62.2	68	1.3	236
Arabidopsis	140.1	5	36	26,541	23.7	80	4.4	55
Chlamydomonas	121	17	64	14,516	16.3	91	7.4	174
T. brucei	26.1	>100	46	9,152	52.6	~0 (1 total)	n.d.	n.d.
Giardia	11.7	5	49	6,480	71.4	~0 (4 total)	n.d.	n.d.

Figure 1

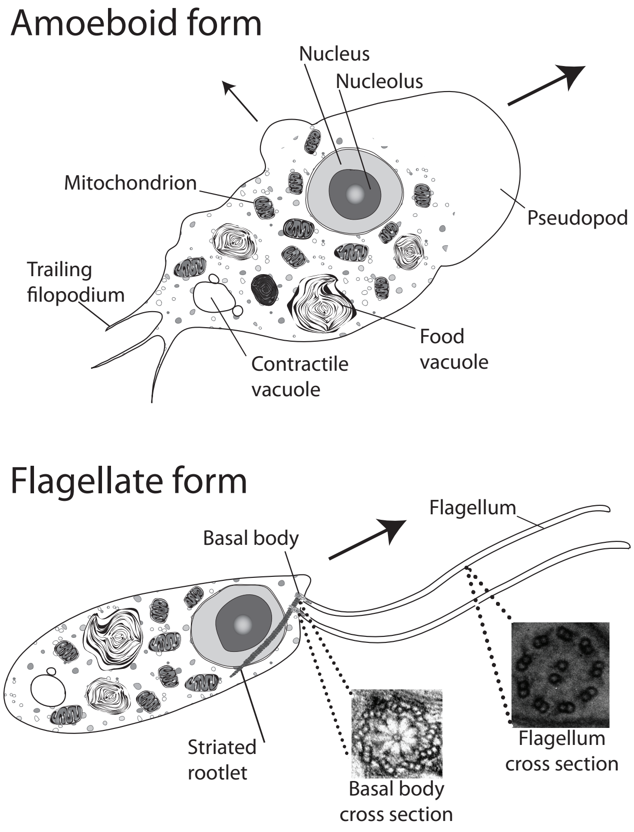

Amoeboid form
Nucleus
Nucleolus
Mitochondrion
Pseudopod
Trailing filopodium
Food vacuole
Contractile vacuole
Flagellate form
Flagellum
Basal body
Striated rootlet
Basal body cross section
Flagellum cross section

Figure 2

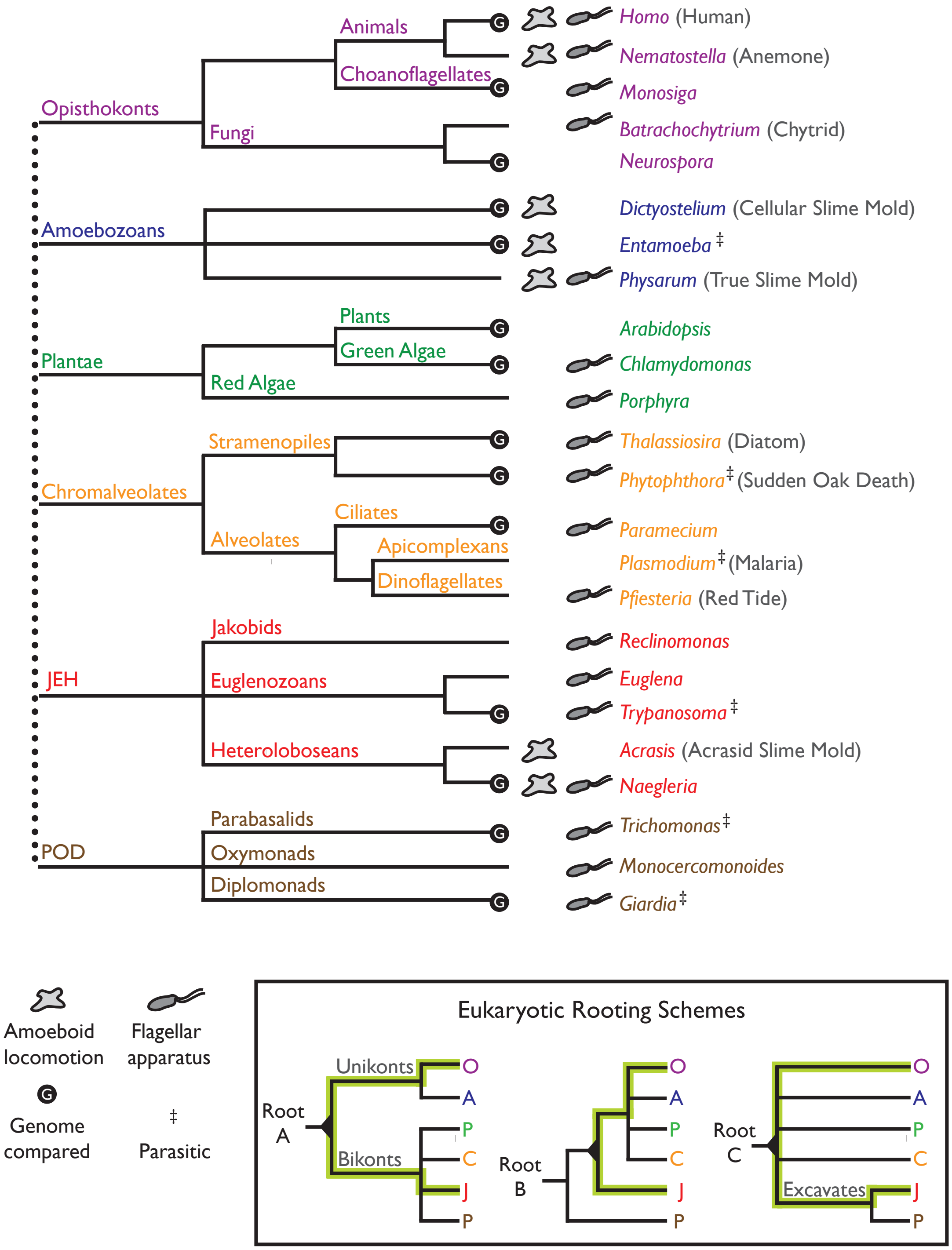
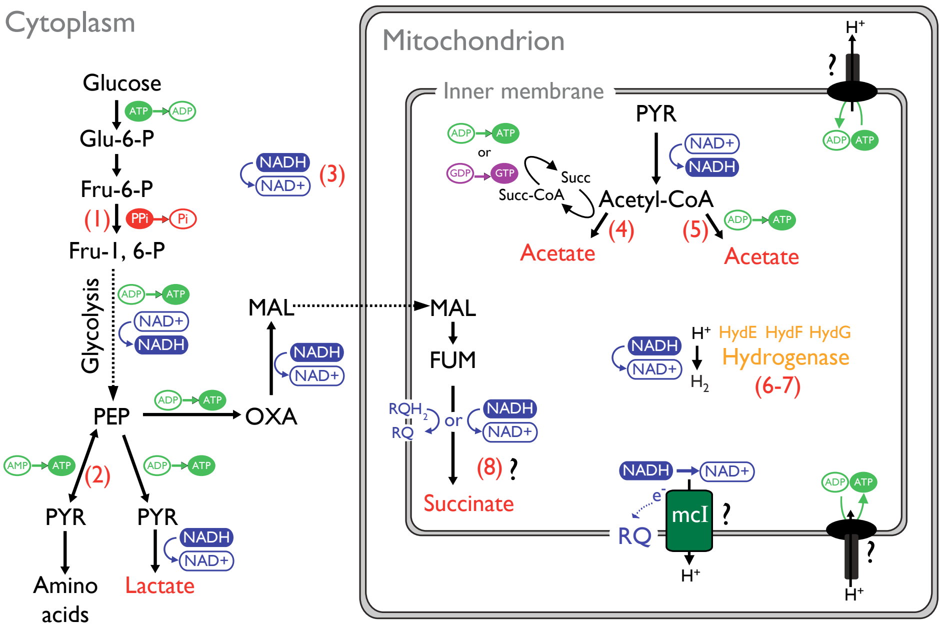

Animals
Choanoflagellates
Opisthokonts
Fungi
Homo (Human)
Nematostella (Anemone)
Monosiga
Batrachochytrium (Chytrid)
Neurospora
Amoebozoans
Dictyostelium (Cellular Slime Mold)
Entamoeba ‡
Physarum (True Slime Mold)
Plants
Green Algae
Plantae
Red Algae
Arabidopsis
Chlamydomonas
Porphyra
Stramenopiles
Chromalveolates
Ciliates
Alveolates
Apicomplexans
Dinoflagellates
Thalassiosira (Diatom)
Phytophthora ‡ (Sudden Oak Death)
Paramecium
Plasmodium ‡ (Malaria)
Pfiesteria (Red Tide)
Jakobids
JEH
Euglenozoans
Heteroloboseans
Reclinomonas
Euglena
Trypanosoma ‡
Acrasis (Acrasid Slime Mold)
Naegleria
Parabasalids
POD
Oxymonads
Diplomonads
Trichomonas ‡
Monocercomonoides
Giardia ‡
G
Amoeboid locomotion
Flagellar apparatus
G Genome compared
‡ Parasitic
Eukaryotic Rooting Schemes
Unikonts
Bikonts
Root A
Root B
Root C
Excavates
O
A
P
C
J
P

Figure 3

Click here to download Figure: FritzLaylin_Fig3.pdf

Figure 3



Naegleria enzymes pivotal for anaerobic fermentation in other protists:

(1) PPi-dependent phosphofructokinase

(2) Pyruvate phosphate dikinase

(3) NAD+-dependent oxidoreductases

(4) Acetate:succinate CoA transferase

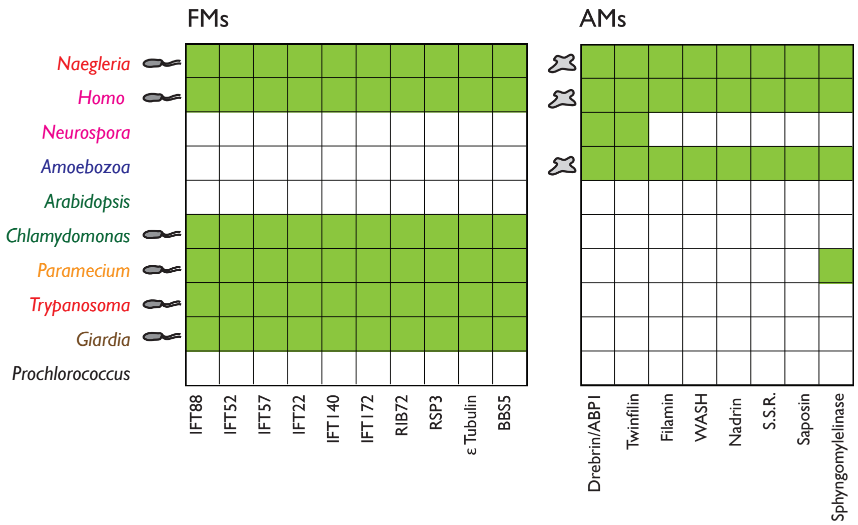
(5) Acetyl-CoA synthetase (ADP-forming)

(6) NADH dehydrogenase

(7) Fe-hydrogenase

(8) Soluble fumarate reductase

Figure 4



FMs

	IFT88	IFT52	IFT57	IFT22	IFT140	IFT172	RIB72	RSP3	ε Tubulin	BBS5
Naegleria	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Homo	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Neurospora										
Amoebozoa										
Arabidopsis										
Chlamydomonas	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Paramecium	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Trypanosoma	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Giardia	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Prochlorococcus										

AMs

	Drebrin/ABP1	Twinfilin	Filamin	WASH	Nadrin	S.S.R.	Saposin	Sphyngomyelinase
Naegleria	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Homo	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Neurospora	🟩	🟩						
Amoebozoa	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Arabidopsis								
Chlamydomonas								
Paramecium								🟩
Trypanosoma								
Giardia								
Prochlorococcus								

Figure 5

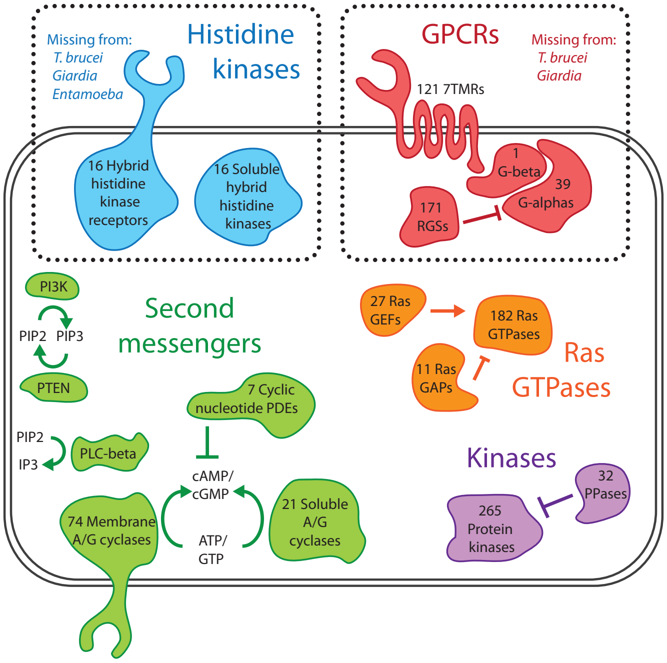
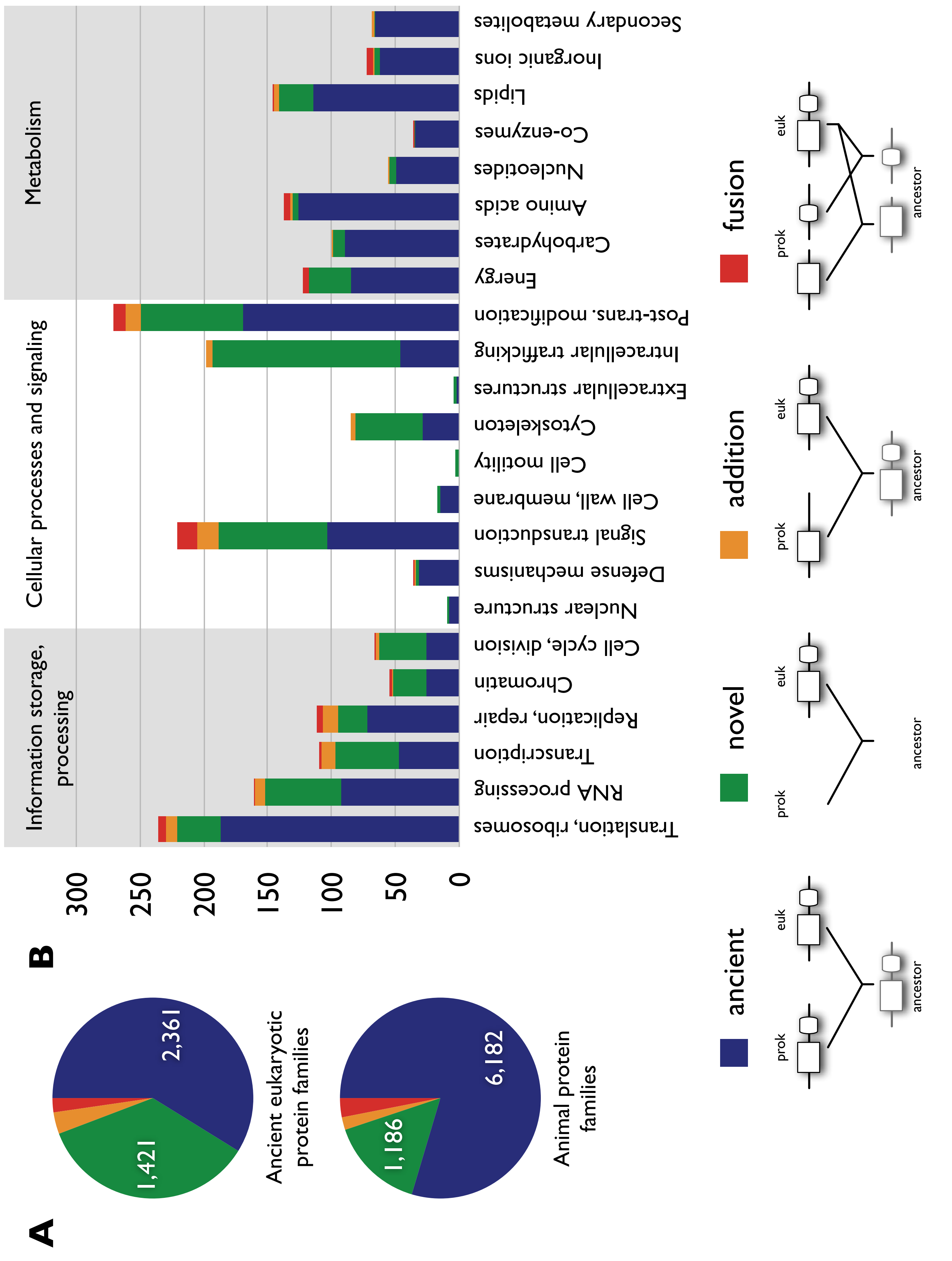

Missing from:
T. brucei
Giardia
Entamoeba
Histidine kinases
16 Hybrid histidine kinase receptors
16 Soluble hybrid histidine kinases
GPCRs
Missing from:
T. brucei
Giardia
121 7TMRs
1 G-beta
39 G-alphas
171 RGSs
PI3K
PIP2
PIP3
PTEN
Second messengers
PIP2
IP3
PLC-beta
7 Cyclic nucleotide PDEs
cAMP/ cGMP
ATP/ GTP
74 Membrane A/G cyclases
21 Soluble A/G cyclases
27 Ras GEFs
182 Ras GTPases
11 Ras GAPs
Ras GTPases
Kinases
32 PPases
265 Protein kinases

Figure 6

Click here to download Figure: FritzLaylin_Fig6.pdf



Inventory of Supplemental Information

Inventory of Supplemental Materials

The following Supplemental items are related to Table 1:
Fig. S1, and Tables S1-S3, S8-S10

The following Supplemental items are related to Figure 1:
Tables S7, S12, S13, and Text S2.

The following Supplemental items are related to Figure 2:
Text S1

The following Supplemental items are related to Figure 3:
Fig. S2 and S5, Tables S14-S16, and Text S3

The following Supplemental items are related to Figure 4:
Fig. S4 and Tables S5, S6, S17, S18.

The following Supplemental items are related to Figure 5:
Fig. S3, Table S11 and Text S4.

The following Supplemental items are related to Figure 6:
Tables S4, S19-S23.

SUPPLEMENTAL TEXT

Text S1 (related to Figure 2). Rooting the eukaryotic tree and major eukaryotic groups

The position of the eukaryotic root is a matter of controversy and great interest (Baldauf, 2003) with no clearly supported hypothesis at present. Two of the three main hypotheses (Fig. 2 insets) employ different strategies for determining the most basal branches in the eukaryotic tree: the first uses Archaeal sequences as an outgroup to define the deepest branches in the eukaryotic tree (Root B) (Yoon et al., 2008); in the second (Root A), the root has been inferred from a single character (Stechmann and Cavalier-Smith, 2002). The last hypothesis (Root C) relies on the deep "excavate" clade being monophyletic.

Determining the order of branches in a phylogenetic tree requires the use of an outgroup. By definition, the deepest branches correspond to the node in the tree closest to the outgroup.

We summarize the three principal hypotheses for the rooting of the eukaryotic tree and subsequent deepest branches as indicated on Fig. 2. Each would have important implications for the interpretation of the *Naegleria* genome data and is presented below.

Root A: "Unikont-bikont"

This hypothesis infers the root to lie between "unikonts" (animals, fungi, amoebozoa) and "bikonts" (plants, other protists) (Fig. 2) based on a single character rather than a molecular phylogeny, namely the fusion of the DHFR and TS genes (Stechmann and Cavalier-Smith, 2002). Under this rooting scheme, parsimonious arguments imply that features shared between *Naegleria* (a bikont) and any unikont were present in the last common eukaryotic ancestor. Features shared only with other bikonts either emerged subsequently, or were lost in the unikont lineage.

Root B: "POD first"

This hypothesis is based on phylogenetic trees made from concatenated protein alignments, rooted by using archaeal sequences as an outgroup. Support for this hypothesis derives from such rooted phylogenies and shows that microbial eukaryotes from the POD lineages (containing amitochondriate organisms) branched first, followed by those of JEH second (Fig. 2) (Arisue et al., 2005; Bapteste et al., 2002; Ciccarelli et al., 2006). Although early evolution and separation of the POD clade remains a possibility (e.g., see (Morrison et al., 2007)), all POD genome sequences are from parasitic microbal eukaryotes, and may therefore have undergone gene loss relative to their close free-living relatives. Thus, under the "POD first" rooting scenario, it is unclear whether features found in *Naegleria* and any other eukaryote were ancestral to all eukaryotes and lost in POD members, or emerged after divergence of the POD group from the eukaryotic stem and are thus ancestral to the other major eukaryotic groups.

Root C: "Excavate supergroup"

Morphological (Simpson, 2003) and molecular data (Burki et al., 2008; Rodriguez-Ezpeleta et al., 2007; Yoon et al., 2008) suggest the JEH clade is sister to the POD group (Parabasalids, Oxymonads and Diplomonads; also known as metamonads), together forming the currently controversial supergroup "excavates" (e.g. references (Burki et al., 2008; Yoon et al., 2008)). Since *Naegleria* is the first free-living excavate to be sequenced, it becomes uniquely informative about the last common eukaryotic ancestor under this hypothesis as its inclusion means that there is now a genome sequence from a free-living species from every major eukaryotic group. Current phylogenies indicating the monophyly of excavates are unrooted, leaving the possibility that the eukaryotic root could split this clade, with other eukaryotes emerging from within the excavates.

Text S2 (related to Figure 1): Hallmarks of eukaryotic cells

DNA replication and translation

Our understanding of eukaryotic DNA replication and translation has centered on components found in yeast and metazoan systems. However, *Giardia* and trypanosomes are missing many of these proteins (Berriman et al., 2005; El-Sayed et al., 2005a; Ivens et al., 2005; Morrison et al., 2007).This has raised speculations the earliest eukaryotes used simplified machinery, similar to that of Archaea (Best et al., 2004). However, Naegleria seems to contain DNA replication initiation proteins missing in trypanosomes (Table S8). Additionally, *Naegleria*'s transcriptional machinery is quite similar to that of yeast (containing 21 of 23 basal transcription factors) and not nearly as simplified as seen in parasitic protists such as *Giardia* (containing only two of the 23 basal transcription factors in yeast) and trypanosomes (Tables S9 and S10). Together, these data indicate that the last common ancestor to extant eukaryotes may have had a more complex DNA/RNA metabolism than suggested by looking only at the complement found in parasitic genomes.

Cytoskeleton

Naegleria contains two potentially autonomous microtubule cytoskeletons (mitotic and flagellar) (Fulton, 1970), as well as an extensive actin cytoskeleton. To determine if these structures are likely formed from canonical proteins, known microtubule and actin cytoskeleton genes were identified in the *Naegleria* genome by manual searches using Pfam domain annotations (Sonnhammer et al., 1998) and BLAST (Altschul et al., 1990) searches using homologs from a variety of genomes as queries. If no homolog was found, searches were repeated using other parameters and homologs. This analysis revealed that *Naegleria*'s genome contains almost all well-conserved actin and microtubule components (Tables S5 and S6, respectively).

To further define what kinds of actin related proteins and tubulins *Naegleria*'s genome encodes, a phylogenetic tree was constructed for each protein family (Fig. S4). Cytoskeletal motors (kinesins, myosins and dyneins) were also classified phylogenetically (Fig. S4). For details of phylogenetic tree construction, see below.

Intron presence and conserved intron position

Nearly 36% of *Naegleria* genes contain at least one intron (median size 61 b.p.); 17% contain multiple introns; 269 genes contain at least five. This is far more than typically found in parasitic protists (*e.g.*, only a single intron in the trypanosome *T. brucei* (Berriman et al., 2005)), yet less than other free-living protists, plants, and animals (Table 1). Analysis of conservation of intron position between *Naegleria*, a land plant (*Arabidopsis*), an animal (human) and a chlorophyte alga (*Chlamydomonas*) revealed 31 proteins for which a sequence could be aligned from each of the four species. 40 introns from *Naegleria* were contained in these alignments and 24 of these were not conserved in another species, suggesting they may be either *Naegleria* inventions or represent introns that have been lost in other lineags. The next highest category consists of 9 introns (22%) which are conserved in *Naegleria* and human only. 15 introns (37%) are conserved between at least *Naegleria* and human, and possibly other species, a very similar fraction when compared to Arabidopsis-human or *Chlamydomonas*-human. 16 (40%) of these introns were likely present in the last common eukaryotic ancestor as they are conserved in at least two species in this analysis of introns. These results suggest an extensive history of intron gain and loss in the JEH lineage. The presence of many introns in *Naegleria* is consistent with an intron-rich ancestor (Rogozin et al., 2005).

Membrane trafficking

Various comparative genomic and phylogenetic analyses have suggested that early eukaryotes possessed a complex membrane-trafficking system (Dacks and Doolittle, 2001; Dacks and Field, 2007)). Although lacking a visible Golgi apparatus, the presence of all major families involved in membrane-trafficking , including an extensive array of potential Golgi-associated factors (Table S12), strongly suggests that functional Golgi machinery exists in *Naegleria* (Dacks et al., 2003). Based on the number of Rab proteins, the complexity of *Naegleria*'s protein trafficking system may be more complex than that of the trypanosomatids.

Text S3 (related to Figure 3). *N. gruberi* metabolism

(A) Classical aerobic metabolism

A complete TCA cycle is predicted. Thus, heterodimeric NAD$^+$-dependent isocitrate dehydrogenase is present. This is in contrast to the distantly related parasitic trypanosomatids where NAD$^+$-dependent isocitrate dehydrogenase is absent (van Weelden et al., 2005), and the TCA cycle is not considered to function, as a classical cycle PMID: 19542311.

For mitochondrial respiration four candidate NADH:ubiquinone oxidoreductases enzymes are present: a proton-pumping complex I (predominantly mitochondrially-encoded) and three alternative non-proton pumping enzymes. As in some fungi (Kerscher et al., 2001) and other organisms, it is likely that one or more of these enzymes functions in the mitochondrial matrix for electron transfer from NADH to ubiquinone. One or more of these NADH dehydrogenases is reasonably predicted to function at the outer face of the mitochondrial inner membrane, thus contributing to oxidation of cytosolic NADH.

The presence of peroxins and gene models encoding lipid-metabolising enzymes with either a canonical C-terminal PTS-1 or a N-terminal PTS-2 targeting signal indicate *N. gruberi* contains peroxisomes. In the distantly-related trypanosomatids, peroxisomes are involved in numerous pathways, most notably and uniquely carbohydrate metabolism (Michels et al., 2006). With the possible exception of a PTS-1 on one isoform of soluble fumarate reductase, no unexpected or novel peroxisomal enzymes were identified, suggesting peroxisomal metabolism in *N. gruberi* is more similar to that of animal, plant and yeast organelles (*i.e.* functioning primarily in lipid catabolism and anabolism), rather than the highly modified organelles seen in *Naegleria*'s distant "JEH" trypanosomatid relatives.

(B) Anaerobic metabolism

The presence of enzymes that are classically used in anaerobic metabolism (*e.g.* acetate:succinate CoA transferase activity, pyruvate phosphate dikinase, soluble (NADH-dependent) fumarate reductase) in trypanosomatids that are obligate aerobes (*e.g. Trypanosoma brucei* (van Weelden et al., 2003) and *Leishmania* (Van Hellemond et al., 1997)) indicates the presence of genes encoding enzymes suited for anaerobic fermentation is not necessarily a robust indicator of anaerobic or micraerophilic metabolism. However, in addition to these and other anaerobic traits, *N. gruberi* also contains Fe-hydrogenase, an oxygen-sensitive enzyme that is central to anaerobic fermentation in some organisms. Homologues of the three component Fe-hydrogenase maturation system that is present in the hydrogenosomes of *Trichomonas vaginalis* (Putz et al., 2006) and the chloroplast of *Chlamydomonas reinhardtii* (Posewitz et al., 2004) are also present in *N. gruberi*.

To investigate the possible location of the *Naegleria* Fe-hydrogenase and Fe-hydrogenase-associated maturases we used the sub-cellular localisation prediction tools Mitoprot (Claros and Vincens, 1996), Predotar (Small et al., 2004), PSORT II (Nakai and Horton, 1999), and TargetP 1.1 (Emanuelsson et al., 2000). For comparison, we also subjected the *bona fide* Fe-hydrogenase from *Blastocystis* (Stechmann et al., 2008) to the same analyses. Each sequence was analysed using parameters optimised for either yeast/animal or non-plant input queries (Table S15). Although the mitochondrial import sequences from the JEH member *Trypanosoma brucei* can be recognised by the mitochondrial matrix import apparatus of yeast, there are nonetheless differences in the efficiency with which yeast can recognise and process some trypanosome import signals (*e.g.* (Hausler et al., 1997)). Thus, the available prediction tools are not likely to be optimised for the recognition of mitochondrial import sequences in other JEH members, such as *N. gruberi*. Notwithstanding this caveat, the probability scores shown in Table S15 provide a good indication that *Naegleria*'s Fe-hydrogenase and associated maturases are likely to be mitochondrial proteins.

The prediction of two soluble fumarate reductases in *N. gruberi* suggests mitochondrial fumarate is used as an electron sink. It is likely that these soluble fumarate reductases use NADH as their electron donor, but we note that the soluble fumarate reductase from *Shewanella putrefaciens* uses a membrane-associated quinone as an electron source (Pealing et al., 1992)). Genes encoding enzymes required specifically for the synthesis of appropriate quinones (*i.e.* of lower redox potential than ubiquinone – *e.g.* rhodoquinone) are not known. Thus, confirmation of the electron donor(s) for fumarate reductases in *N. gruberi* will be dependent upon an analysis of quinones in the organism. If a reduced quinone provide electrons for fumarate reduactase activity in *N. gruberi*, then proton-pumping complex I activity is conceivably coupled to ATP production by ATP synthase under anaerobic conditions.

We found no evidence for the presence in *N. gruberi* of several other enzymes used in anaerobic metabolism by other eukaryotes. For example, we did not identify gene models encoding homologs of NADH-dependent trans-2-enoyl-CoA reductase (used for wax ester synthesis in anaerobic *Euglena gracilis*), alcohol dehydrogenase E, acetate kinase, pyruvate-formate lyase, or pyruvate-ferredoxin oxidoreductase (PFO). PFO is often found in anaerobic eukaryotes, and its activity is commonly correlated with anaerobic adaptation (replacement of the multi-subunit pyruvate dehydrogenase, which is present in *N. gruberi*) and Fe-hydrogenase function (through the delivery of electrons via reduced ferredoxin (Fig. S2)). However, the absence, thus far, of PFO from the anaerobic ciliate *Nyctotherus ovalis* suggests Fe-hydrogenase activity in the absence of PFO is unlikely to be without precedent (Boxma et al., 2005). We predict that Fe-hydrogenase activity in *N. gruberi* will either be coupled directly to NADH oxidation or that a naeglerial NADH dehydrogenase activity transfers electrons from NADH to 2Fe-2S ferredoxin, which then serves as the electron donor for Fe-hydrogenase.

Identification of the gene models encoding protein IDs 54727 and 47456 provided tentative evidence for the presence of an arginine dehydrolase pathway in *N. gruberi.* This pathway is present in the “POD” parasites *Giardia lamblia* and *T. vaginalis* and is thought to be significant for energy generation in these parasites (Brown et al., 1998; Yarlett et al., 1996). However, absence of a good gene model for ornithine transcarbamoylase, which catalyses the first step of the arginine dihydrolase pathway (protein id 74661 exhibits low homology to ornithine transcarbamoylase), means we are not confident about the presence of this typically prokaryotic and anaerobic route for ATP production in *N. gruberi.*

Finally, we considered the possibility of nitrate respiration as an alternative strategy for anaerobic energy production. Use of nitrate in anaerobic ATP production is commonplace in many bacteria and has been described in some eukaryotes, including a few fungi (Takasaki et al., 2004; Takaya et al., 1999; Zhou et al., 2002). Using query sequences that corresponded to either typically assimilatory nitrate reductases present in fungi, algae and (NAS class) prokaryotes (Campbell, 2001; Richardson et al., 2001) or the catalytic modules of respiratory bacterial nitrate reductases (NAR and NAP classes)

(Richardson et al., 2001) no evidence for nitrate-dependent respiration in *N. gruberi* was found. One gene model identified in these searches with E-value < 1E-20 is likely to correspond to a sulfite oxidase ortholog, and thus be involved in the catabolism of sulfur-containing amino acids or the detoxification of sulfite or SO_2 (Richardson et al., 2001).

Text S4 (related to Figure 5). Signaling proteins

Putative G-protein coupled receptors

Naegleria contains a 171 putative serpentine receptors, with 7-8 transmembrane domains (TMs) (as predicted by TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) and have no other predicted domain (via Pfam with E-value < 1E-3). Only one gene (JGI protein ID 72027) has homology to characterized G-protein coupled receptors, with a predicted CAR domain. *Naegleria* serpentine receptors likely signal through heterotrimeric G-proteins consisting of alpha, beta and gamma subunits. We predict 39 alpha subunits and one beta subunit (JGI protein ID 82063) in the *Naegleria* genome. We were not able to detect a gamma subunit, likely because these proteins have low complexity sequence, making it difficult to detect and assign orthology. *Naegleria* also contains 171 putative regulator of G-protein signaling proteins (Pfam domain PF00615), GTPase-accelerating proteins that can rapidly quench the G-protein coupled receptor signaling pathways (De Vries et al., 2000). We did not detect trimeric G-protein components in the predicted proteome of either *T. brucei* or *Giardia.*

Ras monomeric GTPases

Naegleria's genome encodes many proteins likely involved in cellular responses. In particular, *Naegleria* has more monomeric Ras-like GTPases than most sequenced microbial and multicellular eukaryotes (182 genes, 1.2% of the total). This includes many Rho and ras GTPases -- small GTPases canonically involved in cell motility, membrane trafficking, and differentiation, as well as the GTPase-activating proteins (GAPs) and GTP exchange factors (GEFs) that regulate them (Boureux et al., 2007) (Table S11).

Histidine Kinase Signaling

Naegleria contains 32 putative hybrid histidine kinases (contain both response regulator receiver domains, Pfam domain PF00072, and histidine kinase/gyrase/HSP90 domains using Pfam gathering thresholds, Table S11), 16 of which have a predicted TM helix http://www.cbs.dtu.dk/services/TMHMM/). Further, *Naegleria* has 27 protein sequences with a phospho-acceptor domain (PF00512). This domain is used for dimerization of Histidine kinases after activation (Hoch and Varughese, 2001). There is no evidence for histidine phosphotransferase domain containing proteins in *Naegleria*, leaving the next step of the pathway a mystery. We were not able to detect histiding kinase pathway proteins in *Entamoeba*, *T. brucei* or *Giardia*.

Guanylate and adenylate cylases

Naegleria contains 108 genes with an adenylate/guanylate cyclase domain (Pfam domain PF00211) (Fig. S3). This is the highest proportion of cyclase genes encoded in any of the 17 genomes we used to build protein families, except for *T. brucei*, where there is a large expansion of a single cyclase gene family within the variable coat protein regions (El-Sayed et al., 2005b) (Table S11). One might predict a correspondingly high number of cyclic phosphodiesterases (PDEs), however in the three genomes with the highest number of cyclases (*Naegleria*, *Trypanosome* and *Chlamydomonas*), the number of cyclic phosphodiesterases has not increased proportionally (Table S11).

Naegleria cyclases come in four types. The first, comprising 21 *Naegleria* sequences, is cytoplasmic with no predicted transmembrane sequence. Others, similar to trypanosomal cyclases, contain a single transmembrane helix. A third class contains two regions with multiple transmembrane helices. These are similar to human membrane bound cyclases, but while the human cyclases contain two cyclase domains that can dimerize, the *Naegleria* proteins only contain one cyclase domain. Finally, a fourth class of *Naegleria* cyclases contain a single multi-transmembrane region. Thus, *Naegleria* contains cyclases that are similar to those in trypanosomes (those with a single transmembrane pass), as well as some that are more similar to human sequences (with multiple regions containing multiple transmembrane helices). Furthermore, *Naegleria* cyclases often contain other domains (Fig. S3).

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Strains

High quality genomic DNA was prepared from an axenic culture of amoebae of *Naegleria gruberi* strain NEG-M (ATCC 30224) (Fulton, 1974), which was derived from clonal strain NEG (Fulton, 1970) as a clone able to grow in simplified axenic media. The amoebae were grown axenically in suspension in M7 medium (Fulton, 1974) from frozen stocks, and DNA was prepared from cells using Qiagen Genomic DNA Kit (Qiagen, USA).

Whole genome shotgun sequencing and sequence assembly

The initial sequence data set was generated from whole-genome shotgun sequencing (Weber and Myers, 1997) of four libraries. We used one library with an insert size of 2-3 kb (BCCH), one with an insert size of 6-8 kb (BCCI) and two fosmid libraries with insert sizes of 35-40 kb (BCCN, BGAG). We obtained reads as follows: 220,222 reads from the 2-3 kb insert libraries comprising 245 Mb of raw sequence, 261,984 reads from the 6-8 kb insert libraries comprising 263 Mb of raw sequence, and 52,608 reads from the 35-40 kb insert libraries comprising 54 Mb of raw sequence. The reads were screened for vector sequence using Cross_match (Ewing et al., 1998) and trimmed for vector and low quality sequences. Reads shorter than 100 bases after trimming were excluded from the assembly. This reduced the data set to 182,658 reads from the 2-3 kb insert libraries comprising 132 Mb of raw sequence, 245,457 reads from the 6-8 kb insert libraries comprising 193 Mb of raw sequence, and 43,514 reads from the 35-40 kb insert libraries comprising 26 Mb of raw sequence.

The trimmed read sequences were assembled using release 2.9 of JAZZ (Aparicio et al., 2002). A word size of 13 was used for seeding alignments between reads, with a minimum of 10 shared words required before an alignment between two reads would be attempted. The unhashability threshold was set to 50, preventing words present in the data set in more than 50 copies from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than about 97% identical. The genome size and sequence depth were initially estimated to be 35 Mb and 8.0X, respectively. The initial assembly contained 44.8 Mb of scaffold sequence, of which 5.9 Mb (13.1%) was gaps. There were 2,868 scaffolds, with a scaffold N/L50 of 38/384.3 kb, and a contig N/L50 of 77/148.6 kb. The assembly was then filtered to remove scaffolds < 1kb long as well as redundant scaffolds, where redundancy was defined as those scaffolds shorter than 5kb long with a greater than 80% identity to another scaffold greater than 5kb long.

After excluding redundant and short scaffolds, 41.1 Mb remained, of which 4.7 Mb (11.5%) was gaps. The filtered assembly contained 813 scaffolds, with a scaffold N/L50 of 33/401.6 kb, and a contig N/L50 of 69/157.7 kb. The sequence depth derived from the assembly was 8.6 ± 0.1.

To estimate the completeness of the assembly, the consensus sequences from clustering a set of 28,768 ESTs were BLAT-aligned (with default parameters) to the unassembled trimmed data set, as well as the assembly itself. 28,486 ESTs (99.0%) were more than 80% covered by the unassembled data and 28,502 ESTs (99.1%) had hits to the assembly.

Mitochondrial genome sequence (GenBank AF288092) was used to identify the 18 scaffolds belonging to the organelle genome; this sequence is available from the JGI Naegleria Genome Portal (http://www.jgi.doe.gov/naegleria/).

Heterozygosity

All *Naegleria* WGS reads from each of two libraries (BCCH, consisting of 182,658 reads with 3kb-insert and BCCI, consisting of 245,457 reads with 8kb-insert) were aligned to the genome with NCBI BLAST with parameters: -p blastn -e 1e-100 - F 'm D' -W 24. Only genomic positions where 6-8 WGS reads aligned were considered. The number of SNPs per 500 bp window was plotted and fitted to a geometric function [y(x) = A*p*(1-p)^x, with A = 0.708 +/- 0.003, p = 0.259 +/- 0.002] using gnuplot (Fig. S1D). The fit excluded the zero SNP bin which is an outlier and is consistent with regions of homozygosity on a heterozygous background. There were two classes of genomic region, those with 0.58% SNP rate (i.e. (1-p)/p = 2.87 SNPs per 500 bp) (70.8% of the genome) and those with ~0% (29.2% of the genome) (Fig. S1D).

cDNA library construction and EST sequencing

EST sequences were made from two samples: 1) asynchronous cells where some were differentiating into flagellates and others back into amoebae and 2) confluent amoeba grown in tissue culture flasks. Poly-A+ RNA was isolated from total RNA for each sample using the Absolutely mRNA Purification kit and manufacturer's instructions (Stratagene, La Jolla, CA). cDNA synthesis and cloning was a modified procedure based on the "SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning" (Invitrogen). 1-2 g of poly A+ RNA, reverse transcriptase SuperScript II (Invitrogen) and oligo dT-NotI primer:

5'- GACTAGTTCTAGATCGCGAGCGGCCGCCCTTTTTTTTTTTTTTT -3'

were used to synthesize first strand cDNA. Second strand synthesis was performed with E. coli DNA ligase, polymerase I, and RNaseH followed by end repair using T4 DNA polymerase. A SalI adaptor (5'- TCGACCCACGCGTCCG and 5'-CGGACGCGTGGG) was ligated to the cDNA, digested with NotI (NEB), and subsequently size selected by gel electrophoresis (using 1.1% agarose). Two size ranges of cDNA (0.6 - 2.0 kb.p. and > 2 kb.p.) were cut out of the gel for the amoeba sample and one size range (0.6 -2.0 kb.p.) for the flagellate sample. They were directionally ligated into the SalI and NotI digested vector pMCL200_cDNA. The ligation product was transformed into ElectroMAX T1 DH10B cells (Invitrogen).

Library quality was first assessed by randomly selecting 24 clones and PCR amplifying the cDNA inserts with the primers M13-F (GTAAAACGACGGCCAGT) and M13-R (AGGAAACAGCTATGACCAT). The number of clones without inserts was determined and 384 clones for each library were picked, inoculated into 384 well plates (Nunc) and grown for 18 hours at 37°C. Each clone was amplified using RCA then the 5' and 3' ends of each insert was sequenced using vector specific primers (forward (FW): 5'-ATTTAGGTGACACTATAGAA and reverse (RV) 5' – TAATACGACTCACTATAGGG) and Big Dye chemistry (Applied Biosystems). 44,544 EST reads were attempted from the 2 samples.

The JGI EST Pipeline begins with the cleanup of DNA sequences derived from the 5' and 3' end reads from a library of cDNA clones. The Phred software (Ewing and Green, 1998; Ewing et al., 1998) is used to call the bases and generate quality scores. Vector, linker, adapter, poly-A/T, and other artifact sequences are removed using Cross_match (Ewing and Green, 1998; Ewing et al., 1998), and an internally developed short pattern finder. Low quality regions of the read are identified using internally developed software, which masks regions with a combined quality score of less than 15. The longest high quality region of each read is used as the EST. ESTs shorter than 150 bp were removed from the data set. ESTs containing common contaminants such as *E. coli*, common vectors, and sequencing standards were also removed from the data set. There were 38,211 EST sequences left after filtering.

EST clustering was performed on 38,282 trimmed, high-quality ESTs (the 38,211 filtered and trimmed JGI EST sequences combined with the JGI ESTs combined with 71 EST sequences downloaded from GenBank (Benson et al., 2009) by making all-by-all pairwise alignments with MALIGN (Sobel and Martinez, 1986). ESTs sharing an alignment of at least 98% identity, and 150 bp overlap are assigned to the same cluster. These are relatively strict clustering cutoffs, and are intended to avoid placing divergent members of gene families in the same cluster. However, this could also have the effect of separating splice variants into different clusters. Optionally, ESTs that do not share alignments are assigned to the same cluster, if they are derived from the same cDNA clone. We made 4,873 EST clusters.

EST cluster consensus sequences were generated by running Phrap (Ewing and Green, 1998) on the ESTs comprising each cluster. All alignments generated by MALIGN {Sobel, 1986 #351 are restricted such that they will always extend to within a few bases of the ends of both ESTs. Therefore, each cluster looks more like a 'tiling path' across the gene, which matches well with the genome based assumptions underlying the Phrap algorithm. Additional improvements were made to the phrap assemblies by using the 'forcelevel 4' option, which decreases the chances of generating multiple consensi for a single cluster, where the consensi differ only by sequencing errors.

Generation of gene models and annotation

The genome assembly was annotated using the JGI Annotation Pipeline. First the 784 *N. gruberi* v.1 scaffolds were masked using RepeatMasker {Smit, 1996-2004 #289} and a custom repeat library of 123 putative transposable element-like sequences. Next, the EST and full-length cDNAs were clustered into 4,873 consensus sequences (see above) and aligned to the scaffolds with BLAT (Kent, 2002). Gene models were predicted using the following methods: i) *ab initio* (FGENESH (Salamov and Solovyev, 2000); ii) homology-based (FGENESH+ (Salamov and Solovyev, 2000) and Genewise (Birney et al., 2004), with both of these tools seeded by Blastx (Altschul et al., 1990) alignments of sequences from the 'nr' database from the National Center for Biotechnology Information (NCBI, Genbank) (Benson et al., 2009) to the *Naegleria* genome); and iii) mapping *N. gruberi* EST cluster consensus sequences to the genome (EST_map; http://www.softberry.com/) (Table S2).

Truncated Genewise models were extended where possible to start and stop codons in the surounding genome sequence. EST clusters, mapped to the genome with BLAT (Kent, 2002) were used to extend, verify, and complete the predicted gene models. The resulting set of models was then filtered, based on a scoring scheme which maximises completeness, length, EST support, and homology support, to produce a single gene model at each locus, and predicting a total of 15,753 models.

Only 13% of these gene models were seeded by sequence alignments with proteins in the nr database at NCBI (Benson et al., 2009) or *N. gruberi* EST cluster consensus sequences, while 86% were *ab initio* predictions (Table S2). Complete models with start and stop codons comprise 93% of the predicted genes. 30% are consistent with ESTs and 74% align with proteins in the nr database at GenBank (Benson et al., 2009) (Table S3).

Protein function predictions were made for all predicted gene models using the following collection of software tools: SignalP (http://www.cbs.dtu.dk/services/SignalP/), TMHMM (http://www.cbs.dtu.dk/services/TMHMM/), InterProScan (http://www.ebi.ac.uk/interpro/ (Quevillon et al., 2005)), and hardware-accelerated double-affine Smith-Waterman alignments (http://www.timelogic.com/decypher_sw.html) against SwissProt (http://www.expasy.org/sprot/), KEGG (http://www.genome.jp/kegg/), and KOG (http://www.ncbi.nlm.nih.gov/COG/). Finally, KEGG hits were used to map EC numbers (http://www.expasy.org/enzyme/), and Interpro and SwissProt hits were used to map GO terms (http://www.geneontology.org/).

Nearly half (45%) of the gene models have Pfam (Finn et al., 2008) domain annotations (Table S3). The average gene length is 1.65 kb.p. The average protein length is 492 aa. We predicted that 3,514 proteins (22%) possess a leader peptide, 3,439 proteins (22%) possess at least one transmembrane domain, and 2060 (13%) possess both.

Web-based interactive editing tools available through the JGI genome portal (http://www.jgi.doe.gov/naegleria/) were used to manually curate the automated annotations in three ways: i) to assess and if necessary correct, predicted gene structures. ii) to assign gene functions and report supporting evidence, and iii) to create, if necessary, new gene structures.

On 19 July 2007, the manually-annotated gene set was frozen to make a catalog. This set of 15,776 transcripts encoded by 15,727 genetic loci was used for all analyses in this paper. In a few cases, as noted in the main text, manual improvements to gene models were needed before detailed analysis was possible. As of May 15, 2008, 4,016 genes (25%) have been manually curated. All annotations, may be viewed at a JGI portal (http://www.jgi.doe.gov/naegleria/).

Simple and complex repeat analysis

Prior to our analysis, little was known about the repeat landscape in *Naegleria*. To investigate the repeats in the *Naegleria* genome, RepeatMasker (Smit et al., 1996-2004) was run on the genome with the options '-gccalc -species Eukaryota'. This masked 1.71% of the genome assembly, of which 1.32% are simple repeats or low-complexity. However, as *Naegleria* is not closely related to other organisms with sequenced genomes whose repeat sequences have used to build the RepeatMasker libraries, a de novo repeat finding program, RepeatScout (Price et al., 2005), was run on the assembly. This generated a library of 206 repeat sequences (Supplemental File 1). We classified these sequences into the following four categories where possible: i) those with homology to known TEs in the RepeatMasker library or rRNAs using RepeatMasker (Smit et al., 1996-2004), ii) those that overlap gene models, or ESTs or are annotated as tRNAs with tRNAscan-SE (Lowe and Eddy, 1997) or iii) are annotated with a Pfam domain from a manually curated list of Pfams that are associated with transposon proteins (TE-associated Pfam domain) with E-value < 1E-5 and iv) sequences annotated with any other Pfam domain (i.e. non TE-associated Pfam domains), which are likely repeats representing larger gene families. Sequences in category i) include both copia- and gypsy-like putative retrotransposons. Sequences that could not be classified include putative DNA transposons that are highly diverged from known transposable elements and have not been functionally characterized.This analysis increased detection of the non-genic repeat content of the genome to 2,068,185 (5.05%), after adding 548,091 nt covered by simple repeats predicted by RepeatMasker. In our RepeatScout repeat library, we include 151 potentially novel repeat sequences after filtering for overlap with known gene models and Pfam domains (Table S1). These sequences cover 1,380,214 nt (3.37%) of the genome.

Analysis of conserved intron position

To investigate the pattern of intron gain and loss, we looked for conservation of intron position in genes found in Naegleria and three other intron-rich species (averaging at least 5 per gene). We picked a land plant (*Arabidopsis*), an animal (human) and a chlorophyte alga (*Chlamydomonas*). We assembled sets of orthologous protein sequences in these four species by mutual best Smith-Waterman (Smith and Waterman, 1981) hits between *Naegleria* and each of the three species. Next we used CLUSTALW (Thompson et al., 2002) with default settings to make multiple sequence alignments of the protein sequences which were represented by an ortholog in all four species. We mapped the positions of introns from transcript sequence onto the protein sequence in each multiple sequence alignment and looked for introns in well-conserved regions of the alignment for which there was also EST support for all splice sites in *Naegleria*.

Determining lateral gene transfer

In order to identify potential lateral gene transfers from prokaryotes to the *Naegleria* genome, we used the following conservative protocol: we selected genes that have a blast hit to Bacteria or Archaea (E-value < 1E-10) and no hit to Eukarya (E-value < 1E-4) using NCBI blastp v.2.2.17(Altschul et al., 1990) against the nr database at GenBank (Posted date: Nov 9, 2009 5:57 PM) (Benson et al., 2009). This resulted in 191 candidate lateral transfer genes (CLTGs). We constructed a set of homologous sequences by collecting BLAST hits with E-value < 1E-4 to the *Naegleria* sequence in the nr database, as well as the *Naegleria* genome (July 7, 2007 frozen catalog, http://www.jgi.doe.gov/naegleria/). Seven CLTGs were discarded at this stage because they only had one or two bacterial homologs, leaving 184 CLGTs. We next built phylogenetic trees for each of these 184 genes to assess the likelihood of lateral gene transfer. Each set of homologs was aligned using MUSCLE (Edgar, 2004) with default settings, and the multiple sequence alignment was processed with GBLOCKS (Castresana, 2000) (using -b4=3 -k=y -p=s). Maximum likelihood phylogenetic trees were created using RAxML (raxmlHPC-PTHREADS-icc -f a -x 12345 -p 12345 -N 100 -m PROTGAMMAJTT) with 100 bootstrap runs. The bootstrap support values were added to the best scoring trees. In 45 of the 184 trees, the *Naegleria* CLTG lay within a known bacterial clade with strong (>75%) bootstrap support (Table S4). The remainder consisted of either i) a *Naegleria* CLTG grouping within a known bacterial clade with weak (50-75%) bootstrap support for the position of the *Naegleria* sequence or of the bacterial clades or ii) a *Naegleria* CLTG grouping with bacterial sequences, but forming a separate lineage outside known bacterial groups.

Protein trafficking proteins

Identification of proteins involved in protein trafficking

We performed searches against the filtered model set of *Naegleria* proteins at the JGI portal (using the BLOSUM45 matrix). Typically we searched with known trafficking protein sequences from *S. cerevisiae, H. sapiens* or *T. brucei*. *Naegleria* hits were blasted back against the genome of the query protein and against nr database at NCBI (Benson et al., 2009). Domains in *N. gruberi* predicted proteins were detected using CDDB at NCBI. In some instances, where the search strategy described above failed to identify a hit in *N. gruberi*, additional searches were performed using the Smith-Waterman algorithm (Smith and Waterman, 1981), implemented on the CLC Workbench V3.5.1 with CUBE hardware acceleration (CLC Bio, Denmark, www.clcbio.com) or were repeated at the JGI using the unfiltered gene model set. All *Naegleria* hits were subjected to reverse BLAST as before. Hits whose length was over 40% shorter or longer than the length of the query sequence were discarded in order to avoid misannotated gene models. For genes that constitute paralagous families (e.g. Rabs and SNAREs), all hits to the *N. gruberi* protein set were included and subjected to phylogenetic analysis.

Phylogenetic analysis

In order to classify putative membrane-trafficking factors into known types, the sequences were subjected to phylogenetic analysis. In the case of the Rabs, subgroup assignment was achieved by analysis using Neighbor-Joining trees constructed with the *N. gruberi* GTPase candidates and relevant sets of authenticated representative genes from selected taxa (Ackers et al., 2005; Pereira-Leal and Seabra, 2001). More precise analysis of subgroups was then performed using MrBayes(Ronquist and Huelsenbeck, 2003) or PhyML (Guindon and Gascuel, 2003) as appropriate. In all other cases, a combination of Bayesian analysis and maximum likelihood phylogeny was used. Alignments were built using CLUSTALW (Thompson et al., 1997), T-COFFEE (Notredame et al., 2000) or MUSCLE (Edgar, 2004) and improved manually. The model of protein sequence evolution was determined using PROTTEST (Abascal et al., 2005), incorporating corrections for rate variation among sites and invariable sites when relevant. Tree topologies and Bayesian posterior probability values were obtained using the program MrBayes (Ronquist and Huelsenbeck, 2003) with 1,000,000 generations and with the burn-in estimated graphically, excluding all trees prior to the plateau. Maximum likelihood bootstrap support values were determined from 100 pseudo-replicates using the programs RAxML (Stamatakis, 2006) and/or PhyML (Guindon and Gascuel, 2003).

Construction of protein families

As a pre-requisite to comparing the protein-coding potential of *Naegleria* to other organisms at the whole-genome scale, we constructed families of homologous proteins from all protein sequences that are found in both *Naegleria* and at least one other species from a wide a range of eukaryotes. Errors in gene prediction and large-scale species-specific gene losses can cause problems building protein families and drawing phylogenetic inferences from the families. To mitigate this, we chose a range of organisms to ensure that at least two species from every major eukaryotic group with genome sequence were included. Where several closely-related genome sequences were available, we chose manually- or well-annotated species to represent clades of interest. We also included a representative photosynthetic prokaryote, *Prochlorococcus marinus*.

Families of protein sequences were generated such that there is one family for each protein in the common ancestor of all the species which have proteins in the family, and that all the extant proteins descended from the ancestral protein are in the family. The predicted shared ancestry (homology) of family members should enable us to infer shared function, allowing functional annotations to be transferred among family members.

To create protein families, we first blasted [WU-BLASTP 2.0MP-WashU (Altschul et al., 1990)] each of the 15,727 protein sequences in *Naegleria* to all protein sequences in the animals human (Ensemble; Lander et al., 2001; Venter et al., 2001) and *Trichoplax adherens* (Srivastava et al., 2008); the choanoflagellate *Monosiga brevicollis* (King et al., 2008); the fungus *Neurospora crassa* (assembly v7.0; annotation v3.0, http://fungal.genome.duke.edu); the amoebae *Dictyostelium discoideum* (Eichinger et al., 2005) and *Entamoeba histolytica* (TIGR, http://www.tigr.org/tdb/e2k1/eha1/); the land plants *Arabidopsis thaliana* (Initiative, 2000) and *Physcomitrella patens* (assembly v.1

(Rensing et al., 2008); the green alga *Chlamydomonas reinhardtii* (Benson et al., 2009; Merchant et al., 2007); the oomycete *Phytophthora ramorum* (v1, (Joint Genome Institute); the diatoms *Thalassiosira pseudonana* (assembly v3.0 (Armbrust et al., 2004; Joint Genome Institute)) and *Phaeodactylum tricornutum* (assembly v2.0, Available at http://genome.jgi-psf.org/; the alveolate *Paramecium tetraurelia* (Paramecium DB release date 28-MCH-2007; http://paramecium.cgm.cnrs-gif.fr/); the euglenozoan *Trypanosoma brucei* (v4 genome; http://www.genedb.org/genedb/tryp/); the diplomonad *Giardia lamblia* (GMOD; http://www.giardiadb.org/giardiadb/); the parabasalid *Trichomonas vaginalis* (TIGR, http://www.tigr.org/tdb/e2k1/tvg/); and the cyanobacterium *Prochlorococcus marinus* strain MIT9313 (Joint Genome Institute).

Assignment of orthology was determined by the presence of a mutual best hit between two proteins, based on score with a cutoff of E-value < 1E-10. In creating individual protein families, we first generated all possible ortholog pairs consisting of one *Naegleria* protein and a protein from another organism. Next, paralogs that met certain criteria were added to each pair of proteins. A paralog from a given organism was added if its p-dist from the putative ortholog in the same organism (defined as 1 - the fraction of identical aligning amino acids in the proteins) was less than a certain fraction of the p-dist between the two orthologs in the pair. The fractions were chosen to be 0.5 for pairs of organisms involving two eukaryotes and 0.1 for *Naegleria* and the prokaryotic cyanobacterium. Two considerations led to the choice of these values. In order to assign function correctly, we wanted to include only ‘in-paralogs’ (i.e. paralogs that had duplicated after speciation) (Remm et al., 2001). Secondly, we previously determined that higher (less stringent) values led to the generation of protein families with >22,000 members that could not be analyzed further (Merchant et al., 2007). As a final step, all pair-wise families of two orthologs plus paralogs were merged if they contained the same *Naegleria* protein. This created 5,115 families of homologous proteins, with 5,107 families containing proteins from *Naegleria* and at least one other eukaryote and 8 families restricted to *Naegleria* and the cyanobacterium *Prochlorococcus*. Each individual family consists of one or more *Naegleria* paralog(s), mutual best hits to proteins of other species (orthologs) and any paralogs in each of those species. The set of protein families was used in subsequent phylogenetic profiling of proteins associated with amoeboid motility (AMs) or flagellar motility (FMs) (see below). To accomplish this, we built a software tool that allowed us to search for protein families containing any desired combination of species. The search results are called a ‘cut’ (see below) as it represents a phylogenetic slice through the collection of protein families.

The random gene duplication, subsequent divergence and loss that accompanies the evolution of gene families means that it is challenging and sometimes impossible to precisely assign orthology and paralogy between genes. The problem gets more difficult for larger families, which are statistically more likely to undergo mutations and old families that have had longer to diverge. As a result, mutual best hit relationships between sequences may not exist, preventing family construction, or may not be between correct proteins, leading to inclusion of non-homologous proteins in families.

Inferring the protein complement of the eukaryotic ancestor

We built 5,107 eukaryotic gene families (see above) that were founded on mutual best hits between *Naegleria* and other eukaryote(s). The subset of these families with deep phylogenetic distribution likely arose early in eukaryotic evolution, and perhaps were present in the eukaryotic ancestor, or earlier. We identified such a subset of 4,133 of the eukaryotic gene families by requiring that each family contain a minimum of one *Naegleria* protein and two orthologs, and that at least one of the orthologs be from another major eukaryotic group.

Our requirements for ancient gene families are conceptually similar to KOGs (clusters of orthologous groups), but with an additional requirement (see below). The KOGs are based on genes shared between several opisthokonts (represented in the KOG analysis by genomes from animals and fungi) (Fig. 2) and *Arabidopsis* (Tatusov et al., 2003). A subset of 3,285 KOGs are analogous to our ancient gene families as they are present in opisthokonts and a plant (crown KOGs) (i.e., those in *Arabidopsis*). These KOGs are presumably present in the ancestor of opisthokonts and plants (two major eukaryotic groups) and not just innovations in, for example, the animal lineage. However, by including proteins from species in more diverse groups (i.e., in addition to plants and opisthokonts) as well as *Naegleria*, we hoped to achieve a more robust analysis of ancient and/or ancestral eukaryotic proteins.

To predict protein function where possible, we assigned majority rule KOG annotations (Tatusov et al., 2003) to each family in two steps. First, each protein in the family was searched against the KOG sequence database (Tatusov et al., 2003) with RPS-BLAST (Altschul et al., 1990) and the best hit with E-value < 1E-5 was retained. Second, if the commonest KOG annotation in a protein family was in at least half the proteins in a family, that KOG was assigned to the family. Pfams were assigned using HMMer (Eddy, 1998) run on two TimeLogic DeCypher boards (http://www.timelogic.com) using E-value < 1E-5 and Pfam library v21 (Sonnhammer et al., 1998).

While it is possible that an ancestral eukaryotic protein could be present in more than one eukaryotic group due to inter-eukaryotic lateral gene transfer, this process is rare (Keeling and Palmer, 2008), and in addition 92% of the 4,133 ancient eukaryotic gene families are present in at least three major eukaryotic groups making lateral gene transfer an unparsimonious explanation for their presence.

Given the poorly resolved tree of eukaryotic groups, and consequent uncertainty about the position of the root (Ciccarelli et al., 2006; Rodriguez-Ezpeleta et al., 2007; Stechmann and Cavalier-Smith, 2002), means that some genes present in *Naegleria* and one other species from a sister group could have evolved after the ancestor of these two groups diverged from the rest of eukaryotes. For example, it is conceivable that JEH + POD shared an ancestor that diverged from the rest of eukaryotes (a prediction of the controversial excavate hypothesis (Burki et al., 2008; Hampl et al., 2009)), allowing evolution of lineage-specific gene families that are not present in other eukaryotic groups. Only nine families are found just in JEH and POD, suggesting negligible ancestry shared uniquely between these two groups.

The origin of eukaryotic genes

We asked whether each of the 4,133 ancient eukaryotic protein families we had constructed (see above) had been inherited from prokaryotes (i.e. from Archaea/Bacteria), or were eukaryotic inventions, or some combination of these two scenarios. To do this, we first constructed a "centroid" sequence for each of ancient protein family. We define the centroid of a protein family as the hypothetical protein sequence that maximizes the sum of BLAST alignment scores between the centroid and the protein sequences in the family. Thus, each centroid sequence act as a proxy for the ancestral protein sequence from which all extant sequences are descended. We next made a set of all prokaryotic proteins in the UniRef90 protein database at GenBank (Benson et al., 2009) with taxonomy ID = 2 (Bacteria) or 2157 (Archaea). Then we searched this set of prokaryotic proteins for homology to each centroid sequence. For the search, we used blastp [(NCBI version 2.2.15) with command line parameters -p blastp -m 9 -b 3 -v 3 and removed any hit with an E-value < 1E-6. If the centroid sequence had no hit to a prokaryotic protein it was classified as eukaryotic-specific (Fig 6, "novel"). We found 1,421 such "novel" protein families (Fig 6A).

In the following classification steps, we compared Pfam domain annotations in the eukaryotic centroid and prokaryotic sequences. For the classification of centroid sequences with a hit to a prokaryotic protein, we ran Interproscan (Quevillon et al., 2005) locally with the v23 library of Pfam HMMs (Finn et al., 2008) to assign Pfam domains to the centroid sequences and used the Pfam domain annotations from UniRef90 for the prokaryotic proteins.

We classified protein families as "ancient" if the centroid and the best hitting prokaryotic protein met any of the following criteria: i) neither sequence has a Pfam (Finn et al., 2008) domain; ii) the two sequences have the same combination of pairwise domains; iii) the two sequences have another simple pattern of domain gain/loss that does not imply novelty in the eukaryotic lineage. This class of ancient proteins has 2,361 protein families. The remaining protein families showed some degree of innovation in eukaryotes relative to their prokaryotic homologs. The first class had no homolog in prokaryotic genomes (1,421 "novel" families, Table S21). The second class had extra eukaryotic-specific domain(s) (140 "addition" families, Table S22). The third class had been formed by the fusion in eukaryotes of multiple ubiquitous domains into a single polypeptide (92 "fusion" families, Table S23). Some proteins showed domain innovations in both the second and third classes, in which case the commonest type of innovation was chosen.

Ties were left unclassified and joined the remaining 119 families with more complex evolutionary patterns. These proteins showed for example evidence of evolutionary splitting of multi-domain prokaryotic polypeptides into different proteins in eukaryotes, conceptually the opposite of the "fusion" category.

To investigate the putative functions encoded in the ancient, novel, addition and fusion classes of ancient eukaryotic proteins, majority-rule KOGs were assigned as described above (Fig. 6B).

Verification of Flagellar Motility-associated proteins (FMs)

We compared the proteins we had identified to a hand-curated list of 101 *Chlamydomonas* flagellar proteins that had been discovered by biochemical, genetic, and bioinformatic methods (Pazour et al., 2005). Of the 182 FM proteins, 34 are in families containing a characterized *Chlamydomonas* flagellar protein, and an additional 59 are in a family with a *Chlamydomonas* flagellar proteome protein (Pazour et al., 2005). Thus, at least 51% of the FlagellateCut genes are likely to encode proteins that localize to flagella.

Verification of Amoeboid Motility-associated proteins (AMs)

The search for proteins associated with amoeboid motility found 112 protein families containing 139 *Naegleria* proteins. 36 families contained proteins with homology (BLASTP E-value < 1E-10) to a protein in one or more non-amoeboid species from the list we had previously used to build the *Naegleria* protein families, and these 36 families were excluded from the AM gene set. In addition, 13 families were removed because their members belong to very large protein families (containing ≥ 245 members) and we reasoned that difficulties in assigning correct orthology in families this large (see above) made them unlikely to be true representatives of the AmoebaCut. This left 63 AM protein families containing 67 *Naegleria* proteins (Table S18). There is no way to estimate the false positive rate for this computational analysis as no experimental catalog of AMs is available for comparison.

Although the POD member *Trichomonas* has been described as "amoeboid", it does not undergo amoeboid locomotion, and was not used to define AM protein families. However, *Trichomonas* does possess seven of the AMs (Table S18), suggesting most AMs are involved in cell locomotion, and not simply amoeboid-like morphology.

Pfam domain assignment

For analysis of whole proteomes, Pfams were assigned using HMMer (Eddy, 1998) run on TimeLogic DeCypher boards (http://www.timelogic.com) E-value < 1E-5 and Pfam library v. 21 (Sonnhammer et al., 1998). However for manual examination of protein sequences, we used predictions from running Interproscan (Quevillon et al., 2005) with Pfam v. 23 as Interproscan implements the more accurate gathering threshold cutoffs for assigning domains.

Construction of large scale phylogenies

To classify the number and type of members of large paralogous gene families, we used maximum likelihood phylogenetic analyses (described below) to characterize *Naegleria* tubulins, actins/Arps, myosins, dyneins, kinesins and a singe Fe-Fe hydrogenase:

Tubulins

Homolog Gathering:

We searched for annotated tubulin superfamily sequences, primarily those utilized in previous studies (Dutcher, 2003; McKean et al., 2001)). For gamma, delta, epsilon, zeta, and eta tubulins, only one gene (if any) was present in a given genome. For alpha and beta tubulins, only one representative of each (based on annotated sequences) was selected from each non-*Naegleria* genome. The classification of tubulin family members is supported by bi-directional BLAST searches for *Naegleria* sequences.

Two potential *Naegleria* tubulin gene models (JGI protein IDs 88210 and 88211) were incomplete due to scaffold gaps and therefore not included in this analysis. In addition, two alpha tubulins (JGI protein IDs 39221 and 56065) and two beta tubulins (JGI protein IDs 56391 and 55423) were excluded from this analysis because their protein sequences were identical to JGI proteins 56236 and 83350, respectively.

Multiple sequence alignment:

Multiple sequence alignment was made with MUSCLE (Edgar, 2004) using default settings.

Phylogenetic tree construction:

The RtREV+F model was chosen by PROTTEST (Abascal et al., 2005) using the corrected Aikaike Information Criterion (AICc). A maximum likelihood tree was constructed using RAxML (7.0.2) (Stamatakis, 2006) with 100 bootstrap replicates at the CIPRES website (http://www.phylo.org).

Actins and Arps

Homolog Gathering:

The initial sequence set included those with actin-like domains (Pfam domain PF00022 with E-value < 1E-3) contained in human, *Naegleria gruberi, Monosiga brevicolis, Phytophthora ramorum, Physcomitrella patens, Trichoplax adherins, Tricomonas vaginalis, Trypanosoma brucei,* and *Thalassiosira pseudonana*. Additional *Naegleria* sequences were identified by performing BLAST searches against the genome proteome, and manually adding all sequences with E-value < 1E-3. To aid phylogenetic classification of subfamilies, we added sequences from existing multiple sequence alignments from Goodson *et. al.* (Goodson and Hawse, 2002).

Multiple sequence alignment:

Initial alignments were built using MAFFT (v. 6.611b) (Katoh et al., 2002) with the following parameters: BLOSUM45 substitution matrix, 4 retrees, 100 iterations. The resulting alignments were manually edited (including removal of poorly-aligning sequences, and repositioning of individual amino acids), and homologous positions were selected for use in phylogenetic analyses.

Phylogenetic tree construction:

Homologs were classified using bootstrapped maximum likelihood within CIPRES (www.phylo.org) with RAxML (7.0.4) using the following parameters: 100 bootstraps, JTT model of protein evolution, likelihood searches. Consensus phylogenetic trees are presented using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Myosin motor domain-containing proteins

Homolog gathering:

Multiple sequence alignment: Initial alignments were derived from a previously published phylogenetic analysis of myosin head domains (Foth et al., 2006) with refinements using MAFFT (v. 6.611b) (Katoh et al., 2002) with the following parameters: BLOSUM45 substitution matrix, 4 retrees, 100 iterations. The resulting alignments were manually edited (including removal of poorly-aligning sequences, and repositioning of individual amino acids), and homologous positions were selected for use in phylogenetic analyses.

Phylogenetic tree construction: Homologs were classified using bootstrapped maximum likelihood within CIPRES (www.phylo.org) with RAxML (7.0.4) using the following parameters:100 bootstraps, JTT model of protein evolution, likelihood searches. Consensus phylogenetic trees are presented using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Dynein heavy chain-containing proteins

Homolog gathering:

The initial sequence set included those with the dynein motor domain (Pfam domain PF03028 with E-value < 1E-3) contained in human, *Naegleria gruberi*, *Monosiga brevicolis*, *Phytophthora ramorum*, *Physcomitrella patens*, *Trichoplax adherens*, *Trichomonas vaginalis*, *Trypanosoma brucei*, and *Thalassiosira pseudonana*. Additional *Naegleria* sequences were identified by performing BLAST against the proteome, and manually adding all hits with E-value < 1E-3. To aid phylogenetic classification of subfamilies, we added sequences from existing multiple sequence alignments from Wickstead *et. al.* (Wickstead and Gull, 2007).

Multiple sequence alignment:

Initial alignments were built using MAFFT (v. 6.611b) (Katoh et al., 2002) with the following parameters: blosum 45 substitution matrix, 4 retrees, 100 iterations. The resulting alignment was manually edited (including removal of poorly-aligning sequences, and repositioning of individual amino acids), and homologous positions were selected for use in phylogenetic analysis.

Phylogenetic tree construction:

Homologs were classified using maximum likelihood within CIPRES (www.phylo.org) with RAxML (7.0.4) using the following parameters: JTT model of protein evolution, likelihood searches. Consensus phylogenetic trees are presented using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Kinesin head domain-containing proteins

Homolog gathering:

The initial sequence set included those with a kinesin motor domain (domain PF00225 with E-value < 1E-3) contained in human, *Naegleria gruberi*, *Monosiga brevicolis*, *Phytophthora ramorum*, *Physcomitrella patens*, *Trichoplax adherens*, *Trichomonas vaginalis*, *Trypanosoma brucei*, and *Thalassiosira pseudonana*. Additional *Naegleria* sequences were identified by BLAST searches against the genome, and manually curating all sequences with an E-value < 1E-3. To aid phylogenetic classification of subfamilies, we added sequences from existing multiple sequence alignments from Wickstead *et. al.* (Wickstead and Gull, 2006)

Multiple sequence alignment:

Initial alignments were built using MAFFT (v. 6.611b) (Katoh et al., 2002) with the following parameters: BLOSUM45 substitution matrix, 4 retrees, 100 iterations. The resulting alignment was manually edited (including removal of poorly-aligning sequences, and repositioning of individual amino acids), and homologous positions were selected for use in phylogenetic analysis.

Phylogenetic tree construction:

Homologs were classified using bootstrapped maximum likelihood within CIPRES (http://www.phylo.org) with RAxML (7.0.4) using the following parameters:100 bootstraps, JTT model of protein evolution, likelihood searches. Consensus phylogenetic trees are presented using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Fe-Hydrogenases

Homolog gathering:

Hydrogenase homologs were collected by searching the nr database at NCBI (Benson et al., 2009) with BLAST. After manual curation, the top 247 hits were selected for analysis.

Multiple sequence alignment:

Initial alignments were built using MAFFT (v. 6.611b) (Katoh et al., 2002) with the following parameters: BLOSUM45 substitution matrix, 4 retrees, 100 iterations. The resulting alignment was manually edited (including removal of poorly-aligning sequences, and repositioning of individual amino acids), and homologous positions were selected for use in phylogenetic analysis.

Phylogenetic tree construction:

Homologs were classified using bootstrapped maximum likelihood within CIPRES (www.phylo.org) with RAxML (7.0.4) using the following parameters:100 bootstraps, JTT model of protein evolution, likelihood searches. Consensus phylogenetic trees are presented using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

SUPPLEMENTAL REFERENCES

Abascal, F., Zardoya, R., and Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. Bioinformatics *21*, 2104-2105.

Ackers, J.P., Dhir, V., and Field, M.C. (2005). A bioinformatic analysis of the RAB genes of *Trypanosoma brucei*. Mol Biochem Parasitol *141*, 89-97.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.* (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science *297*, 1301-1310.

Arisue, N., Hasegawa, M., and Hashimoto, T. (2005). Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. Mol Biol Evol *22*, 409-420.

Baldauf, S.L. (2003). The deep roots of eukaryotes. Science *300*, 1703-1706.

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, 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*, 1414-1419.

Best, A.A., Morrison, H.G., McArthur, A.G., Sogin, M.L., and Olsen, G.J. (2004). Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. Genome Res *14*, 1537-1547.

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res *14*, 988-995.

Boureux, A., Vignal, E., Faure, S., and Fort, P. (2007). Evolution of the Rho family of ras-like GTPases in eukaryotes. Mol Biol Evol *24*, 203-216.

Brown, D.M., Upcroft, J.A., Edwards, M.R., and Upcroft, P. (1998). Anaerobic bacterial metabolism in the ancient eukaryote *Giardia duodenalis*. Int J Parasitol *28*, 149-164.

Campbell, W.H. (2001). Structure and function of eukaryotic NAD(P)H:nitrate reductase. Cell Mol Life Sci *58*, 194-204.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol *17*, 540-552.

Claros, M.G., and Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem *241*, 779-786.

Dacks, J.B., and Doolittle, W.F. (2001). Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help. Cell *107*, 419-425.

De Vries, L., Zheng, B., Fischer, T., Elenko, E., and Farquhar, M.G. (2000). The regulator of G protein signaling family. Annu Rev Pharmacol Toxicol *40*, 235-271.

Dutcher, S.K. (2001). The tubulin fraternity: alpha to eta. Curr Opin Cell Biol *13*, 49-54.

Dutcher, S.K. (2003). Long-lost relatives reappear: identification of new members of the tubulin superfamily. Curr Opin Microbiol *6*, 634-640.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

El-Sayed, N.M., Myler, P.J., Bartholomeu, D.C., Nilsson, D., Aggarwal, G., Tran, A.N., Ghedin, E., Worthey, E.A., Delcher, A.L., Blandin, G.*, et al.* (2005a). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. Science *309*, 409-415.

El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E.A., Hertz-Fowler, C.*, et al.* (2005b). Comparative genomics of trypanosomatid parasitic protozoa. Science *309*, 404-409.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol *300*, 1005-1016.

Ensembl. http://www.ensembl.org/

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res *8*, 186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res *8*, 175-185.

Foth, B.J., Goedecke, M.C., and Soldati, D. (2006). New insights into myosin evolution and classification. Proc Natl Acad Sci U S A *103*, 3681-3686.

Goodson, H.V., and Hawse, W.F. (2002). Molecular evolution of the actin family. J Cell Sci *115*, 2619-2622.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol *52*, 696-704.

Hausler, T., Stierhof, Y.D., Blattner, J., and Clayton, C. (1997). Conservation of mitochondrial targeting sequence function in mitochondrial and hydrogenosomal proteins from the early-branching eukaryotes *Crithidia*, *Trypanosoma* and *Trichomonas*. Eur J Cell Biol *73*, 240-251.

Hemschemeier, A., Fouchard, S., Cournac, L., Peltier, G., and Happe, T. (2008). Hydrogen production by *Chlamydomonas reinhardtii*: an elaborate interplay of electron sources and sinks. Planta *227*, 397-407.

Hemschemeier, A., and Happe, T. (2005). The exceptional photofermentative hydrogen metabolism of the green alga *Chlamydomonas reinhardtii*. Biochem Soc Trans *33*, 39-41.

Hoch, J.A., and Varughese, K.I. (2001). Keeping signals straight in phosphorelay signal transduction. J Bacteriol *183*, 4941-4949.

Horvath, A., Kingan, T.G., and Maslov, D.A. (2000). Detection of the mitochondrially encoded cytochrome c oxidase subunit I in the trypanosomatid protozoan *Leishmania tarentolae*. Evidence for translation of unedited mRNA in the kinetoplast. J Biol Chem *275*, 17160-17165.

Initiative, T.A.G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796-815.

Ivens, A.C., Peacock, C.S., Worthey, E.A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M.A., Adlem, E., Aert, R., *et al.* (2005). The genome of the kinetoplastid parasite, *Leishmania major*. Science *309*, 436-442.

Joint Genome Institute. http://www.jgi.doe.gov/

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res *30*, 3059-3066.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kerscher, S.J., Eschemann, A., Okun, P.M., and Brandt, U. (2001). External alternative NADH:ubiquinone oxidoreductase redirected to the internal face of the mitochondrial inner membrane rescues complex I deficiency in *Yarrowia lipolytica*. J Cell Sci *114*, 3915-3921.

King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., *et al.* (2008). The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature *451*, 783-788.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res *25*, 955-964.

Malik, S.B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M., and Logsdon, J.M., Jr. (2008). An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. PLoS ONE *3*, e2879.

McKean, P.G., Vaughan, S., and Gull, K. (2001). The extended tubulin superfamily. J Cell Sci *114*, 2723-2733.

Michels, P.A., Bringaud, F., Herman, M., and Hannaert, V. (2006). Metabolic functions of glycosomes in trypanosomatids. Biochim Biophys Acta *1763*, 1463-1477.

Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci *24*, 34-36.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol *302*, 205-217.

Pazour, G.J., Agrin, N., Leszyk, J., and Witman, G.B. (2005). Proteomic analysis of a eukaryotic cilium. J Cell Biol *170*, 103-113.

Pealing, S.L., Black, A.C., Manson, F.D., Ward, F.B., Chapman, S.K., and Reid, G.A. (1992). Sequence of the gene encoding flavocytochrome c from *Shewanella putrefaciens*: a tetraheme flavoenzyme that is a soluble fumarate reductase related to the membrane-bound enzymes from other bacteria. Biochemistry *31*, 12132-12140.

Pereira-Leal, J.B., and Seabra, M.C. (2001). Evolution of the Rab family of small GTP-binding proteins. J Mol Biol *313*, 889-901.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21 Suppl 1*, i351-358.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res *33*, W116-120.

Remacle, C., Barbieri, M.R., Cardol, P., and Hamel, P.P. (2008). Eukaryotic complex I: functional diversity and experimental systems to unravel the assembly process. Mol Genet Genomics *280*, 93-110.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol *314*, 1041-1052.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., *et al.* (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science *319*, 64-69.

Richardson, D.J., Berks, B.C., Russell, D.A., Spiro, S., and Taylor, C.J. (2001). Functional, biochemical and genetic diversity of prokaryotic nitrate reductases. Cell Mol Life Sci *58*, 165-178.

Riviere, L., van Weelden, S.W., Glass, P., Vegh, P., Coustou, V., Biran, M., van Hellemond, J.J., Bringaud, F., Tielens, A.G., and Boshart, M. (2004). Acetyl:succinate CoA-transferase in procyclic *Trypanosoma brucei*. Gene identification and role in carbohydrate metabolism. J Biol Chem *279*, 45337-45346.

Rogozin, I.B., Sverdlov, A.V., Babenko, V.N., and Koonin, E.V. (2005). Analysis of evolution of exon-intron structure of eukaryotic genes. Brief Bioinform *6*, 118-134.

Ronquist, F., and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics *19*, 1572-1574.

Ruiz, F., Krzywicka, A., Klotz, C., Keller, A., Cohen, J., Koll, F., Balavoine, G., and Beisson, J. (2000). The SM19 gene, required for duplication of basal bodies in *Paramecium*, encodes a novel tubulin, eta-tubulin. Curr Biol *10*, 1451-1454.

Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in *Drosophila* genomic DNA. Genome Res *10*, 516-522.

Schimanski, B., Nguyen, T.N., and Gunzl, A. (2005). Characterization of a multisubunit transcription factor complex essential for spliced-leader RNA gene transcription in *Trypanosoma brucei*. Mol Cell Biol *25*, 7303-7313.

Simpson, A.G. (2003). Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). Int J Syst Evol Microbiol *53*, 1759-1777.

Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004). Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics *4*, 1581-1590.

Smit, A.F.A., Hubley, R., and Green, P. (1996-2004). RepeatMasker Open-3.0. . http://www.repeatmasker.org.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Sobel, E., and Martinez, H.M. (1986). A multiple sequence alignment program. Nucleic Acids Res *14*, 363-374.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucl Acids Res *26*, 320-322.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L., *et al.* (2008). The *Trichoplax* genome and the nature of placozoans. Nature *454*, 955-960.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics *22*, 2688-2690.

Takasaki, K., Shoun, H., Yamaguchi, M., Takeo, K., Nakamura, A., Hoshino, T., and Takaya, N. (2004). Fungal ammonia fermentation, a novel metabolic mechanism that couples the dissimilatory and assimilatory pathways of both nitrate and ethanol. Role of acetyl CoA synthetase in anaerobic ATP synthesis. J Biol Chem *279*, 12414-12420.

Takaya, N., Suzuki, S., Kuwazaki, S., Shoun, H., Maruo, F., Yamaguchi, M., and Takeo, K. (1999). Cytochrome p450nor, a novel class of mitochondrial cytochrome P450 involved in nitrate respiration in the fungus Fusarium oxysporum. Arch Biochem Biophys *372*, 340-346.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics *Chapter 2*, Unit 2 3.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res *25*, 4876-4882.

van Hellemond, J.J., van der Meer, P., and Tielens, A.G.M. (1997). *Leishmania infantum* promastigotes have a poor capacity for anaerobic functioning and depend mainly on respiration for their energy generation. Parasitology *114*, 351-360.

van Weelden, S.W., Fast, B., Vogt, A., van der Meer, P., Saas, J., van Hellemond, J.J., Tielens, A.G., and Boshart, M. (2003). Procyclic Trypanosoma brucei do not use Krebs cycle activity for energy generation. J Biol Chem *278*, 12854-12863.

van Weelden, S.W., van Hellemond, J.J., Opperdoes, F.R., and Tielens, A.G. (2005). New functions for parts of the Krebs cycle in procyclic *Trypanosoma brucei*, a cycle not operating as a cycle. J Biol Chem *280*, 12451-12460.

Vaughan, S., Attwood, T., Navarro, M., Scott, V., McKean, P., and Gull, K. (2000). New tubulins in protozoal parasites. Curr Biol *10*, R258-259.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., *et al.* (2001). The Sequence of the Human Genome. Science *291*, 1304-1351.

Weber, J.L., and Myers, E.W. (1997). Human whole-genome shotgun sequencing. Genome Res *7*, 401-409.

Wickstead, B., and Gull, K. (2006). A "holistic" kinesin phylogeny reveals new kinesin families and predicts protein functions. Mol Biol Cell *17*, 1734-1743.

Wickstead, B., and Gull, K. (2007). Dyneins across eukaryotes: a comparative genomic analysis. Traffic *8*, 1708-1721.

Yarlett, N., Martinez, M.P., Moharrami, M.A., and Tachezy, J. (1996). The contribution of the arginine dihydrolase pathway to energy metabolism by *Trichomonas vaginalis*. Mol Biochem Parasitol *78*, 117-125.

Zhou, Z., Takaya, N., Nakamura, A., Yamaguchi, M., Takeo, K., and Shoun, H. (2002). Ammonia fermentation, a novel anoxic metabolism of nitrate by fungi. J Biol Chem *277*, 1892-1896.

FigS1, Legend and low resolution image

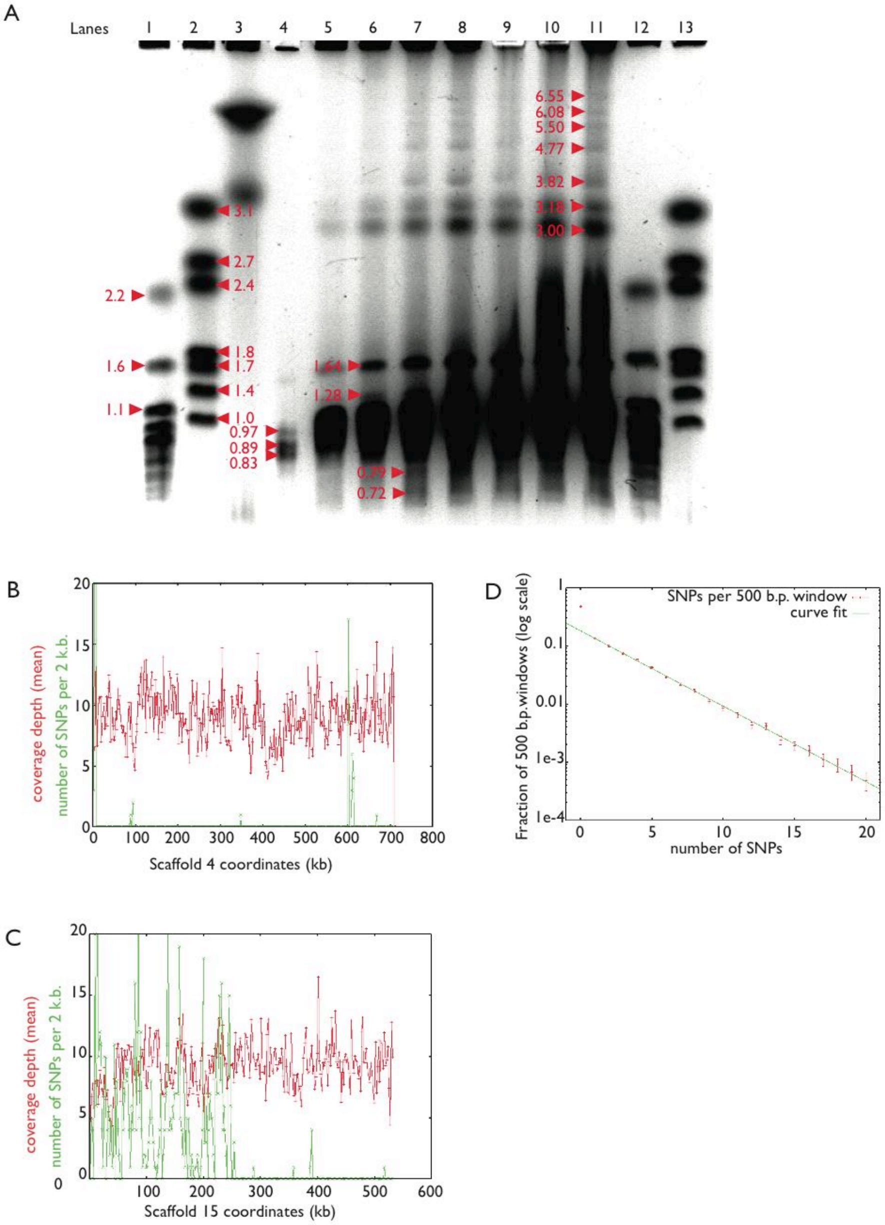


Figure S1 (related to Table 1). Electrophoretic karyotype, heterozygosity of *Naegleria gruberi*

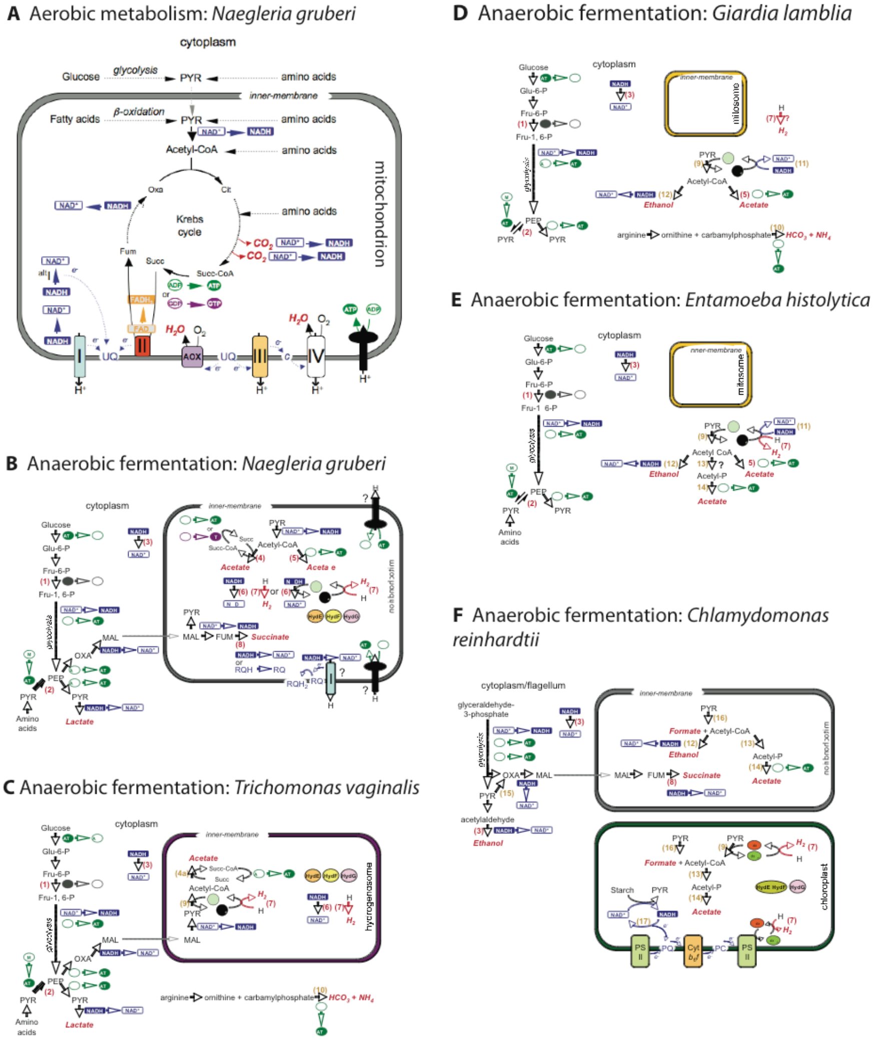
(A) Pulsed field electrophoresis gel of *Naegleria gruberi*, strain NEG-M (lanes 4-11), with the amount of DNA loaded increasing left to right. Lanes 1-3 contain markers with chromosome sizes indicated (*Saccharomyces cerevisiae* in lane one, and *Hansenula wingei* in lane two, and *Schizosaccharomyces pombe* in the third lane). *Naegleria* chromosome sizes are indicated, and range from ~0.7 to ~6.6 Mb. We estimate the total genome size to be 42 Mb.

(B,C) Variations in heterozygosity and sequence depth in the *Naegleria* assembly. Depth of sequence coverage is shown (red) with number of SNPs per 2 kb window (green) along scaffold 4 (B) and scaffold 15 (C). Blocks of homozygous sequence in the genome include very long regions (hundreds of kilobases up to megabases) and have very uniform levels of homozygosity, with zero or near zero counts of SNPs in two kb windows (B). This is in stark contrast to the background level seen over the rest of the genome, seen for example at the 5’ end of scaffold 15 at coordinates 0 to approximately 250 kb (C). The uniformity of sequence read depth rules out the explanation that random statistical noise is responsible for the homozygosity seen in these blocks (B,C).

(D) Geometric distribution of the number of single nucleotide polymorphisms in the *Naegleria* genome

We show the distribution of the number of single nucleotide polymorphisms per 500 base pair window at bases sampled between 6 and 8 times in the shotgun data in red. A curve fit to the data using y(x) = A*p*(1-p)^x with A = 0.708 +/- 0.003, p = 0.259 +/- 0.002 is shown in green.

FigS2, Legend and low resolution image



A Aerobic metabolism: *Naegleria gruberi*

B Anaerobic fermentation: *Naegleria gruberi*

C Anaerobic fermentation: *Trichomonas vaginalis*

D Anaerobic fermentation: *Giardia lamblia*

E Anaerobic fermentation: *Entamoeba histolytica*

F Anaerobic fermentation: *Chlamydomonas reinhardtii*

Figure S2 (related to Figure 3). Predicted canonical aerobic metabolism for *Naegleria gruberi,* and anaerobic fermentation in *Naegleria gruberi* and other protists

(A) *Naegleria gruberi* has canonical aerobic metabolism

Glucose, amino acids and fatty acids can be all be used as carbon sources for energy metabolism.

Metabolite abbreviations:

Cit, citrate; Fum, fumarate; Oxa, oxaloacetate; PYR, pyruvate; Succ, succinate; Succ-CoA, succinyl-CoA.

Abbreviations for mitochondrial respiratory enzymes:

I, NADH:ubiquinone oxidoreductase; II, succinate dehydrogenase; III ubiquinol:cytochrome c oxidoreductase; IV, cytochrome *c* oxidase. AOX, alternative oxidase; ^{alt}I, , alternative NADH dehydrogenase.

(B) Predicted pathways for anaerobic fermentation in *N. gruberi.* Reactions involved in the hydrolysis or production of nucleotide tri-phosphates, and the oxidation or reduction of NAD$^+$ or NADH are highlighted. Enzymes distributed widely in anaerobes/microaerophiles, but more generally not found in aerobic eukaryotes are numbered in red.: The predicted presence in mitochondria of three proteins (HydE, HydF, HydG) required for Fe-hydrogenase maturation is shown. Uncertainties regarding the possible functions of complex I and ATP synthase (denoted by question marks) in the putative anaerobic/microaerophillic metabolism of *N. gruberi* are summarized in Text S3.

(C-F) Anaerobic fermentation in other protists. Comparisons are made with those protists where biochemical evidence of anaerobic metabolism is augmented by the availability of a sequenced nuclear genome. In the microaerophilic parasites *T. vaginalis*, *G. lamblia*, and *E. histolytica* mitochondrial degeneracy is observed. The recently characterised anaerobic metabolism of *C. reinhardtii* (E) is used as a response to either dark anaerobic conditions or nutrient (sulphur) deprivation, and is distributed across three sub-cellular compartments: cytosol, mitochondrion, and chloroplast (Atteia et al., 2006; Hemschemeier et al., 2008; Hemschemeier and Happe, 2005; Mus et al., 2007). Enzymes characteristic of anaerobic metabolism, but not found in *N. gruberi* are numbered in yellow.

Red, italics: predicted (A) or known (B-E) end-products of anaerobic fermentation.

Enzymes highlighted:

(1) PP_i-dependent phosphofructokinase

(2) pyruvate phosphate dikinase

(3) NADH-dependent dehydrogenases (of unknown substrate specifities)

(4) Acetate:succinate CoA transferase (type I and type II families in *N. gruberi* (Riviere et al., 2004; van Grinsven et al., 2008); type II family only in *T. vaginalis* (van Grinsven et al., 2008))

(5) putative acetyl-CoA synthetase (ADP-forming family (Sanchez et al., 2000))

(6) soluble NADH dehydrogenase

(7) Fe-hydrogenase

(8) soluble fumarate reductase

(9) pyruvate:ferredoxin oxidoreductase

(10) carbamate kinase (from the arginine dihydrolase pathway)

(11) NADH oxidase

(12) alcohol dehydrogenase E

(13) phosphotransacetylase

(14) acetate kinase

(15) pyruvate carboxylase

(16) pyruvate:formate lyase

(17) predicted, but as yet unidentified oxidoreductase (Hemschemeier and Happe, 2005).

Additional abbreviations to those defined in (A):

Glu-6-P, glucose-6-phosphate; Fru-6-P, fructose-6-phosphate; Fru-1,6-P, fructose-1,6-bisphosphate; PEP, phosphoenolpyruvate; MAL, malate; fdx/fdx$_{red}$, oxidised ferredoxin/reduced ferredoxin.

FigS3, Legend and low resolution image

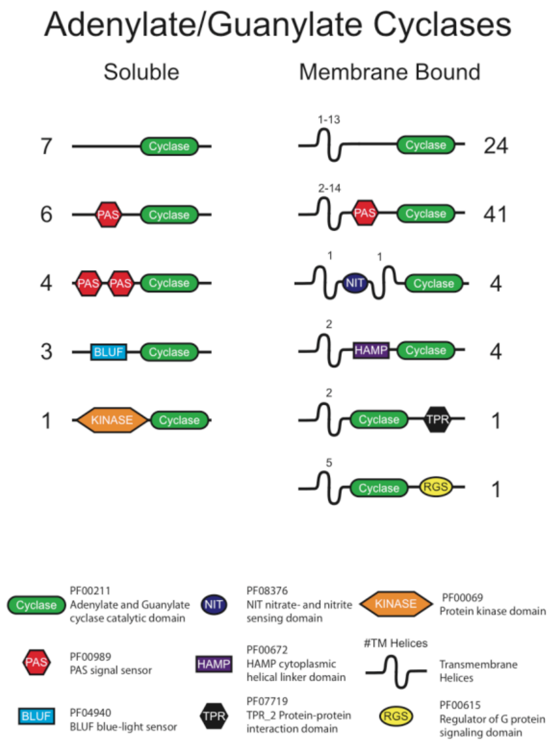


Figure S3 (related to Figure 5). Cyclases in *Naegleria*

Diagram of the 96 sequences in *Naegleria* with Pfam domain PF00211 (adenylate and guanylate cyclase catalytic domain) predicted with E-value < 1E-3, and confirmed using gathering thresholds. Note that using a gathering threshold alone predicts 108 *Naegleria* cyclases. Presence and number of transmembrane helices and other predicted (E-value < 1E-10) Pfam domains are also indicated.

Table of Contents

A Actin/arp phylogeny

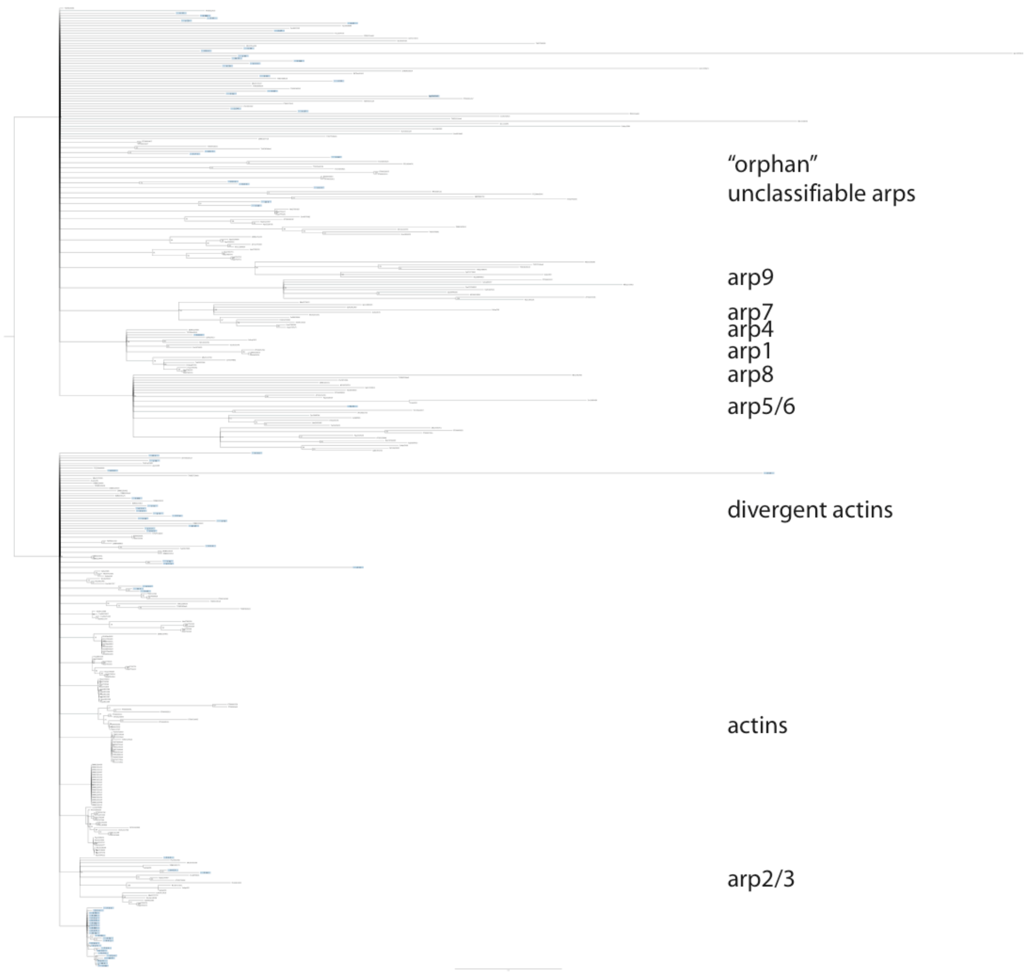


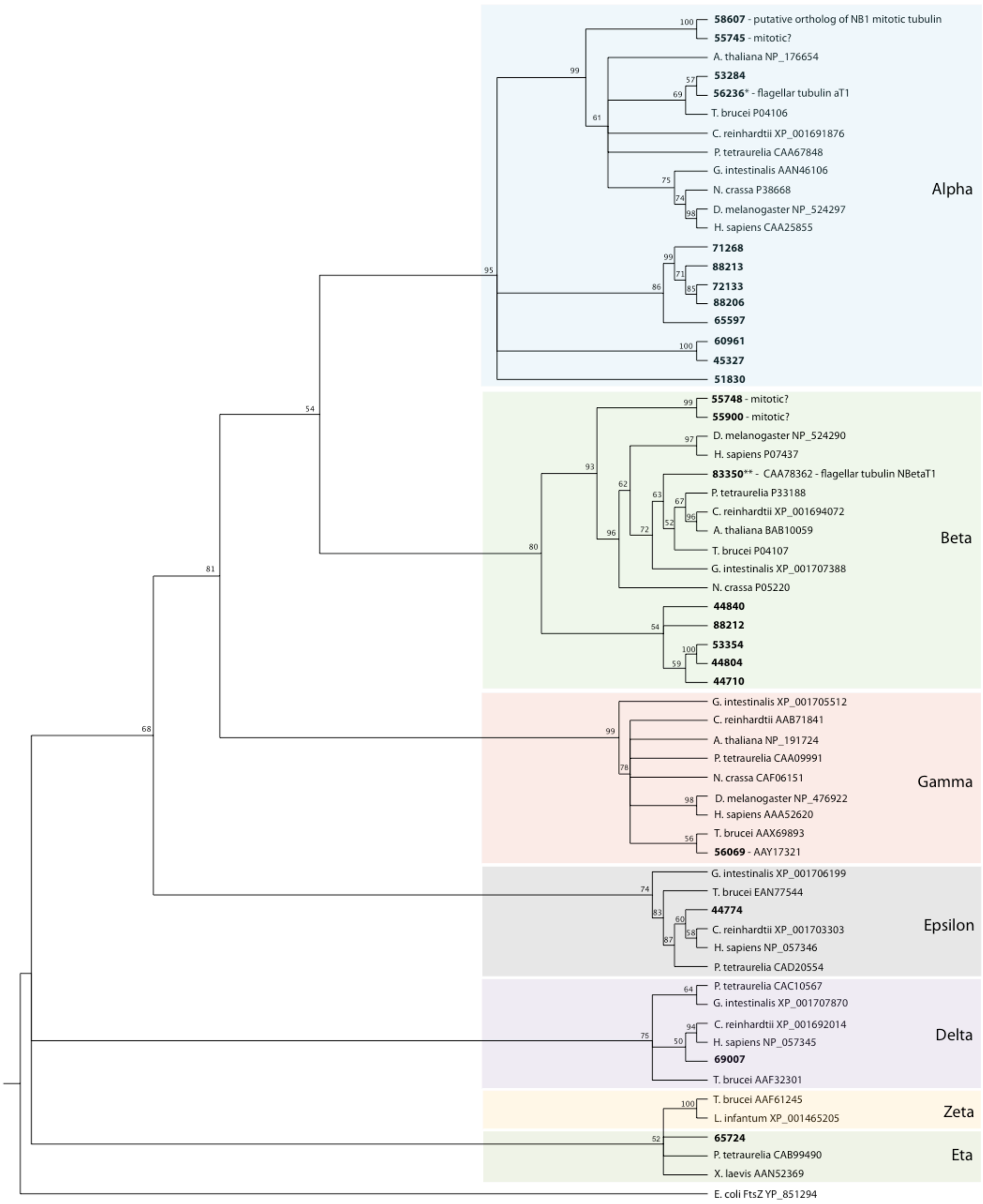
Figure S4 (Related to Figure 4).

(A) Actin/Arp phylogeny

Phylogenetic analysis of the 78 *N. gruberi* actins and actin-related proteins (Arps) was performed (with 340 homologous positions of 422 diverse taxa) using the JTT amino acid substitution model within RAxML (see Supplemental Experimental Procedures). 100 bootstrap replicates were performed and nodes with >50 bootstrap support are indicated. The *Naegleria* actin and Arp homologs are shown in red, and clades of actins/Arps are shown in the tree and below:

Actin/Arp subfamily	*Naegleria* JGI protein IDs used in phylogenetic analysis
Canonical Actin:	74513, 56150, 82392, 88138, 55502, 56113, 55094, 56107, 82840, 56335, 55154, 55489, 44432, 60612, 54819, 55286, 77652, 88136, 49788, 59270, 33387, 48298, 54894, 83258
Additional Actins	44350, 29917, 65595, 30052, 67159
Actin-like	35386, 60433, 29087, 60797, 32902, 33902, 72728, 44817, 30071, 80160, 69091, 60876, 47526, 72694, 60310, 60993
Arp2/3	50292, 82653,65498
Arp5/6	72200
Arp1	60816
Orphan Arps	60995, 54418, 31967, 60869, 53573, 30796, 74761, 44581, 50297, 32634, 70952, 29502, 33847, 70153, 44602, 74378, 88141, 30098, 29311, 32689, 48860, 46504, 32125, 74221, 49873, 73491, 44886, 33917

B Tubulin phylogeny



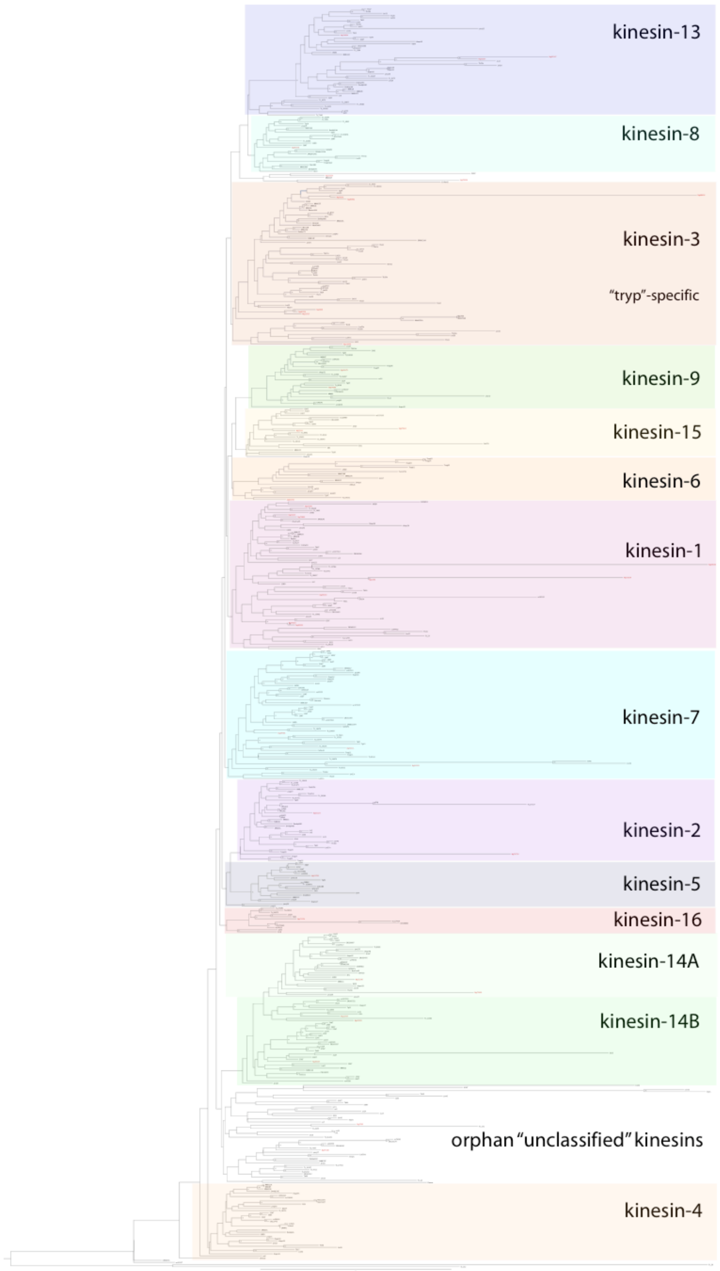
*Models with protein IDs 56065 and 39221 share identical protein sequence
**Models with protein IDs 56391 and 55423 share identical protein sequence

(B) Tubulin phylogeny

A phylogenetic tree of the tubulin superfamily, including all 24 non-redundant *Naegleria* tubulin sequences with complete gene models (models with protein IDs 88210 and 88211, which are incomplete due to scaffold gaps, were not included; in addition, 4 tubulins with redundant protein sequences were not included, as described in Supplemental Experimental Procedures). This maximum likelihood tree was created with RAxML using the JTT amino acid model, 1000 rapid bootstrap replicates, and *E. coli* FtsZ as the outgroup (see Supplemental Experimental Procedures). *Naegleria* sequences are identified by their protein ID (bold), and all other sequences by the species and GenBank accession number. Bootstrap values above 50% are shown; nodes with bootstrap values below 50% were collapsed into polytomies.

The classification of subfamilies (alpha through eta) is based on previously published annotations for non-*Naegleria* sequences, and supported by bi-directional BLAST searches for *Naegleria* sequences. As expected and based on the wide phylogenetic distribution of these proteins in flagellate organisms, the *Naegleria* genome contains homologs of alpha, beta, gamma, delta, and epsilon tubulin. *Naegleria* does not appear to have a homolog of zeta tubulin (Vaughan et al., 2000), suggesting that this tubulin family member is unique to the Trypanosomatids. However, based on bi-directional BLAST searches—though not well-resolved on this tree—*Naegleria* has a homolog of eta tubulin, which has been shown to be involved in basal body assembly (Ruiz et al., 2000) and is also found in *Chlamydomonas reinhardtii*, *Paramecium tetraurelia,* and possibly *Xenopus laevis* (its “cryptic tubulin” clusters with this group) (Dutcher, 2001; McKean et al., 2001).

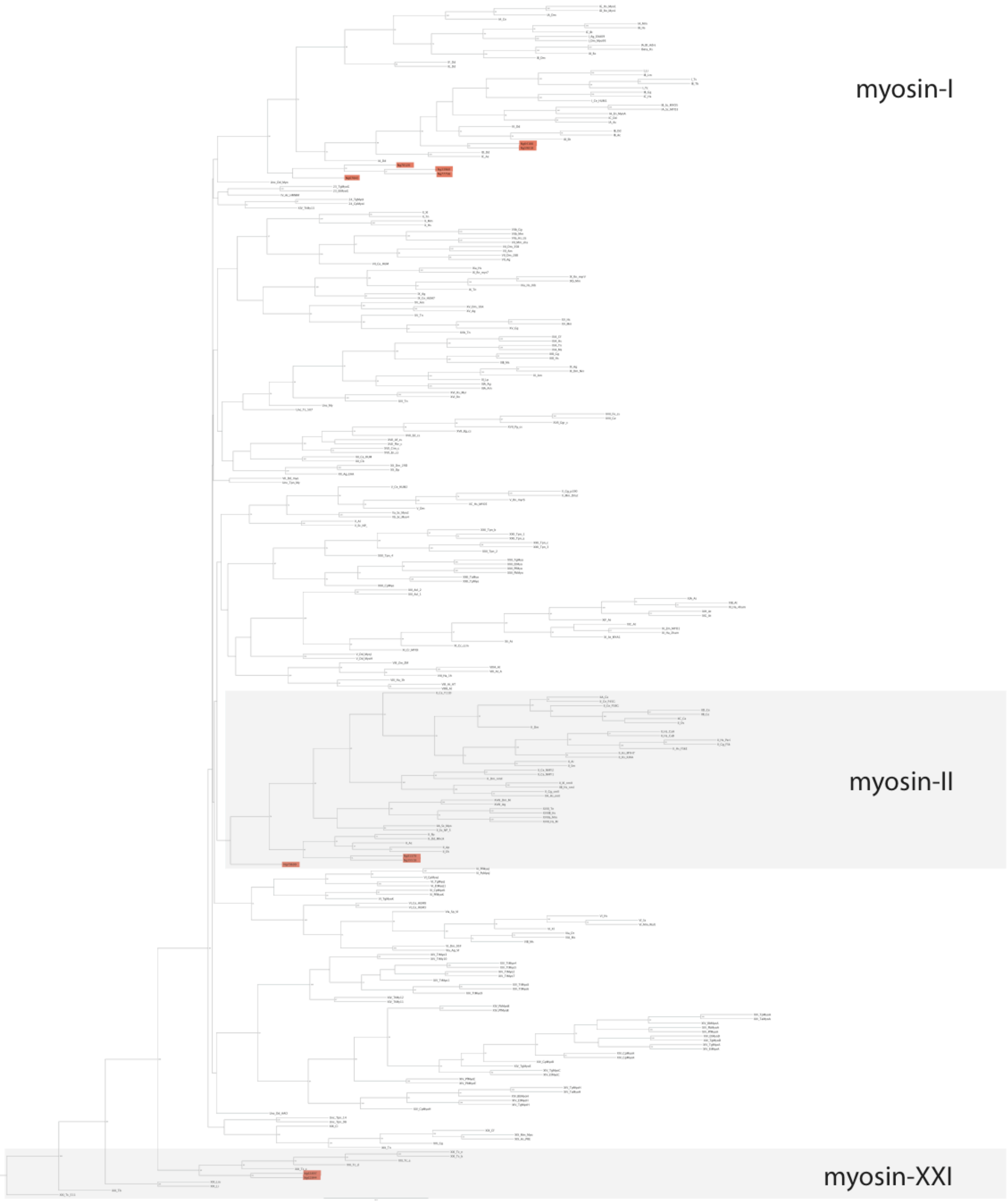
C Kinesin Phylogeny



(C) Kinesin motor domain phylogeny

Phylogenetic analysis of the 41 *N. gruberi* kinesin motor domains was performed (with 267 homologous positions of 583 diverse taxa) using the JTT amino acid substitution model within RAxML (see Supplemental Experimental Procedures). 100 bootstrap replicates were performed and nodes with >50% bootstrap support are indicated. *Naegleria* homologs group within the majority of canonical kinesin families, and these kinesin homologs are shown in red (and in Table S6). Several *Naegleria* kinesin-3 homologs group with strong support in the previously trypanosome-specific kinesin-3 subfamily in support of the JEH grouping.

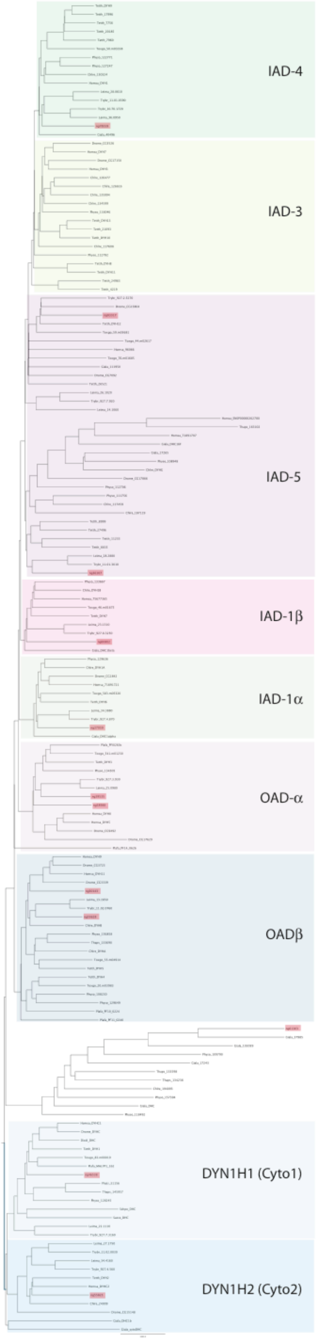
D Myosin phylogeny



(D) Myosin motor domain phylogeny

Phylogenetic analysis of the 11 *N. gruberi* myosin heavy chain homologs was performed (with 646 homologous positions of 278 diverse taxa) using the JTT amino acid substitution model within RAxML (see Supplemental Experimental Procedures). 100 bootstrap replicates were performed and nodes with >50% bootstrap support are indicated. *Naegleria* homologs within canonical myosin heavy chain families are shown in red (and in Table S5). Several *Naegleria* proteins group in the previously trypanosome-specific XXI myosin family with strong bootstrap support.

E Dynein phylogeny



(E) Dynein motor domain phylogeny

Phylogenetic analysis of the 12 *N. gruberi* dynein heavy chain homologs was performed (with 2596 homologous positions of 158 diverse taxa) using the JTT amino acid substitution model in RAxML (see Supplemental Experimental Procedures). The dynein heavy chain homologs present in inner arm dyneins, outer arm dyneins, and cytoplasmic/inner flagellar transport dyneins are shown in red (and in Table S6).

Table S17 (related to Figure 4). Flagellar motility associated proteins (FMs)

Flagellar-motility associated proteins (FMs) were identified as described in Supplemental Experimental Protocols. Those families with characterized *Chlamydomonas* homologs include the gene name from Version 3.0 of the *Chlamydomonas* genome (http://www.jgi.doe.gov/chlamy). *ath Arabidopsis thaliana,* ppa *Physcomitrella patens*, pra *Phytophthora ramorum*, tps *Thalassiosira pseudonana*, ptr *Phaeodactylum tricornutum*, ddi *Dictyostelium discoideum,* ncr *Neurospora crassa,* hsa human, tad *Trichoplax adherens*, mbr *Monosiga brevicollis*, pte *Paramecium tetraurelia*, tbr *Trypanosoma brucei*, gla *Giardia lamblia*, ehi *Entamoeba histolytica*, tva *Trichomonas vaginalis*, cre *Chlamydomonas reinhardtii.*

Table S17 (related to Figure 4). Flagellar motility associated proteins

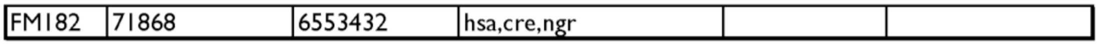
Name	*Naegleria* JGI protein ID	Gene family (cluster ID)	Species with genes in family	*Chlamy-domonas* homolog	Other homologs
FM1	63280	6550330	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	IFT88	IFT88
FM2	65383	6550366	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	LF4	
FM3	81047	6550418	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	RIB72	
FM4	81229	6550938	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	FAP32	
FM5	59637	6551401	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	FAP52	
FM6	77715	6551416	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	BLD1	IFT52
FM7	61993	6552659	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	FAP259	
FM8	31069	6552726	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	SEH1, MOT47	
FM9	1424	6552828	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	FAP250	
FM10	79456	6553116	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	ARL3	
FM11	82851	6553427	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	BUG21	PACRG
FM12	71898	6552987	pra,hsa,ppa,mbr,tad,tps,pte,tbr,gla,cre,ngr	DIP13	
FM13	68117	6550932	pra,hsa,ppa,mbr,tad,tps,pte,tbr,tva,cre,ngr	FAP50	
FM14	49668	6552299	pra,hsa,ppa,mbr,tad,tps,pte,tbr,cre,ngr		
FM15	63939	6550894	pra,hsa,ppa,mbr,tad,tps,gla,tva,cre,ngr	FLA2/FLA8	
FM16	66643	6550571	pra,hsa,ppa,mbr,tad,ptr,pte,cre,ngr	FAP215	
FM17	80690	6549767	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	DYF13	
FM18	78704	6549988	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	HY3	Hydin
FM19	45002	6550401	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	IFT57	
FM20	71180	6551150	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr		
FM21	30192	6551402	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP198	
FM22	64930	6551455	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	RSP3	
FM23	82719	6551498	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	IDA4	
FM24	3580	6551596	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	MOT15	

FM25	29177	6551944	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	TCTEX1	
FM26	50399	6551960	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP60	
FM27	44774	6552071	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	BLD2	Epsilon tubulin
FM28	48798	6552126	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	IFT140	
FM29	78559	6552188	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	DHC2	
FM30	2066	6552209	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP184	
FM31	54982	6552426	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP253	
FM32	32701	6552870	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP118	
FM33	30562	6552881	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	PF16	
FM34	29690	6552903	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP66	
FM35	63764	6553257	pra,hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	IFT172	
FM36	79290	6550473	pra,hsa,ppa,mbr,tad,pte,tbr,gla,cre,ngr	FAP82	
FM37	68996	6550170	pra,hsa,ppa,mbr,tad,pte,tbr,tva,cre,ngr		Sas-6
FM38	70274	6550190	pra,hsa,ppa,mbr,tad,pte,tbr,tva,cre,ngr	FAP70	
FM39	77945	6550628	pra,hsa,ppa,mbr,tad,pte,tbr,tva,cre,ngr	IFT80	
FM40	61313	6552455	pra,hsa,ppa,mbr,tad,pte,tbr,tva,cre,ngr	FAP57	
FM41	79626	6551170	pra,hsa,ppa,mbr,tad,pte,tbr,cre,ngr	FAP116	
FM42	69007	6552725	pra,hsa,ppa,mbr,tad,pte,gla,tva,cre,ngr	UNI3	Delta tubulin
FM43	29002	6552496	pra,hsa,ppa,mbr,tad,tbr,tva,cre,ngr	FAP146	
FM44	33676	6551289	pra,hsa,ppa,mbr,tad,gla,cre,ngr	POC1	
FM45	31544	6552579	pra,hsa,ppa,mbr,tad,cre,ngr	RAB23	
FM46	68950	6550567	pra,hsa,ppa,mbr,ptr,tbr,ehi,tva,cre,ngr		LAG1
FM47	74561	6551926	pra,hsa,ppa,mbr,pte,gla,tva,cre,ngr	FAP134	
FM48	33146	6553920	pra,hsa,ppa,tad,tps,pte,tbr,gla,tva,cre,ngr	FAP67	
FM49	80717	6551160	pra,hsa,ppa,tad,tps,pte,tbr,cre,ngr	MOT45	
FM50	62959	6550378	pra,hsa,ppa,tad,tps,tbr,cre,ngr	MBO2	

FM51	77902	6550398	pra,hsa,ppa,tad,ptr,cre,ngr	DAT1	
FM52	63921	6552226	pra,hsa,ppa,tad,pte,tbr,gla,tva,cre,ngr	MOT17	
FM53	62977	6551500	pra,hsa,ppa,tad,pte,tbr,tva,cre,ngr	IFT20	
FM54	380	6552004	pra,hsa,ppa,tad,pte,tbr,tva,cre,ngr	FAP59	
FM55	57343	6552331	pra,hsa,ppa,tad,pte,tbr,tva,cre,ngr	IDA7	
FM56	65518	6553128	pra,hsa,ppa,tad,pte,tbr,tva,cre,ngr	MOT16	SPATA4
FM57	33361	6552781	pra,hsa,mbr,tad,tps,ptr,tbr,cre,ngr	MOT (ECH1)	
FM58	78637	6550142	pra,hsa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	ODA9	
FM59	60431	6550351	pra,hsa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	ODA6	
FM60	79232	6550351	pra,hsa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	ODA6	
FM61	81548	6551027	pra,hsa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	ODA1	
FM62	74922	6553051	pra,hsa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	DLC1	
FM63	54720	6553051	pra,hsa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	DLC1	
FM64	44967	6549754	pra,hsa,mbr,tad,tps,pte,tbr,gla,cre,ngr	FAP127	
FM65	64648	6549959	pra,hsa,mbr,tad,tps,pte,tbr,gla,cre,ngr	KLP1	
FM66	60926	6550727	pra,hsa,mbr,tad,tps,pte,tbr,tva,cre,ngr	RABL2A	
FM67	64053	6551279	pra,hsa,mbr,tad,tps,pte,tbr,tva,cre,ngr	IFT81	
FM68	52666	6552934	pra,hsa,mbr,tad,tps,pte,tbr,tva,cre,ngr	MKS1	
FM69	78645	6553047	pra,hsa,mbr,tad,tps,pte,tbr,cre,ngr	PDE14	
FM70	79669	6553456	pra,hsa,mbr,tad,tps,pte,tva,cre,ngr	FLA3	Kinesin-associated protein 3
FM71	72811	6551092	pra,hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FBB17	
FM72	64818	6551191	pra,hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr	XRP2	
FM73	34252	6551275	pra,hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr	BBS5	
FM74	80979	6551366	pra,hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr	BBS8	
FM75	29188	6551631	pra,hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP251	
FM76	46605	6551986	pra,hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP91	

FM104	68814	6552786	pra,hsa,tad,pte,cre,ngr	SSA3	
FM105	80259	6551480	pra,hsa,tad,tbr,cre,ngr	D1bLIC	
FM106	49289	6552258	pra,hsa,tad,cre,ngr		
FM107	71505	6552992	pra,hsa,tad,cre,ngr	GSTS1	
FM108	70195	6552992	pra,hsa,tad,cre,ngr	GSTS1	
FM109	70247	6552992	pra,hsa,tad,cre,ngr	GSTS1	
FM110	75317	6552992	pra,hsa,tad,cre,ngr	GSTS1	
FM111	56805	6553062	pra,hsa,tad,cre,ngr		
FM112	78620	6550198	pra,hsa,pte,tbr,tva,cre,ngr	FAP36	
FM113	49798	6551425	pra,hsa,pte,gla,cre,ngr	RSP4	
FM114	73137	6550379	pra,hsa,pte,cre,ngr	CAH1	
FM115	67854	6551362	hsa,ppa,mbr,tad,tps,pte,tbr,cre,ngr	FAP45	
FM116	68477	6551157	hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP65	
FM117	70051	6551331	hsa,ppa,mbr,tad,pte,tbr,gla,tva,cre,ngr		
FM118	81845	6552075	hsa,ppa,mbr,tad,tbr,cre,ngr	DHC6	
FM119	63304	6549916	hsa,ppa,mbr,tad,gla,tva,cre,ngr	BOP5	
FM120	4843	6551486	hsa,ppa,mbr,tps,ptr,cre,ngr	CYN40	
FM121	70995	6553235	hsa,ppa,mbr,tps,ptr,cre,ngr		
FM122	83269	6551165	hsa,ppa,mbr,pte,cre,ngr	AAH1	
FM123	32341	6552151	hsa,ppa,tad,tps,pte,tbr,gla,tva,cre,ngr		
FM124	29888	6554041	hsa,ppa,tad,pte,tbr,gla,tva,cre,ngr	FAP44	
FM125	50227	6549899	hsa,ppa,tad,pte,tbr,cre,ngr	FAP14	
FM126	80274	6550509	hsa,ppa,tad,ehi,cre,ngr		Sirtuin
FM127	66079	6552202	hsa,ppa,tad,cre,ngr	TRXm	
FM128	4868	6551151	hsa,ppa,ptr,cre,ngr	DNJ29	
FM129	72718	6552619	hsa,ppa,pte,tbr,tva,cre,ngr	MOT39	
FM130	4931	6553861	hsa,ppa,tbr,cre,ngr		
FM131	64631	6552548	hsa,ppa,tva,cre,ngr	FAP269	
FM132	81521	6551060	hsa,mbr,tad,tps,ptr,tva,cre,ngr		
FM133	80835	6552017	hsa,mbr,tad,tps,pte,tbr,gla,tva,cre,ngr	SSA11	
FM134	77673	6551871	hsa,mbr,tad,tps,pte,tbr,tva,cre,ngr		
FM135	62841	6552058	hsa,mbr,tad,tps,pte,tbr,tva,cre,ngr		MKS3
FM136	30379	6553447	hsa,mbr,tad,tps,pte,tbr,tva,cre,ngr		
FM137	4601	6551499	hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr	FAP9	

FM138	74042	6551732	hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr		
FM139	50561	6552523	hsa,mbr,tad,pte,tbr,gla,tva,cre,ngr		
FM140	61232	6552767	hsa,mbr,tad,pte,tbr,tva,cre,ngr	MOT37	
FM141	65873	6553468	hsa,mbr,tad,pte,tbr,tva,cre,ngr	FAP161	
FM142	73596	6554034	hsa,mbr,tad,pte,tbr,tva,cre,ngr	FAP61	
FM143	80404	6552775	hsa,mbr,tad,pte,ehi,cre,ngr		
FM144	62107	6551340	hsa,mbr,tad,pte,cre,ngr	POC16	
FM145	57344	6553502	hsa,mbr,tad,tbr,tva,cre,ngr		
FM146	66608	6553089	hsa,mbr,ptr,pte,cre,ngr		
FM147	68057	6552972	hsa,mbr,tbr,cre,ngr		
FM148	79419	6550542	hsa,mbr,cre,ngr	FOX1	
FM149	73977	6550542	hsa,mbr,cre,ngr	FOX1	
FM150	80346	6552423	hsa,mbr,cre,ngr		
FM151	70654	6553513	hsa,mbr,cre,ngr		
FM152	83064	6551733	hsa,tad,tps,pte,tbr,gla,tva,cre,ngr	RIB43a	
FM153	65759	6553379	hsa,tad,tps,pte,tbr,cre,ngr		TECT3
FM154	73664	6550305	hsa,tad,tps,cre,ngr		
FM155	62591	6552981	hsa,tad,ptr,tva,cre,ngr		
FM156	71996	6552728	hsa,tad,ptr,cre,ngr	MOT50	
FM157	54684	6552728	hsa,tad,ptr,cre,ngr	MOT50	
FM158	67231	6550823	hsa,tad,pte,tbr,cre,ngr	PTP1	
FM159	71676	6553723	hsa,tad,pte,tbr,cre,ngr	FAP119	
FM160	4690	6550250	hsa,tad,pte,gla,cre,ngr	FAP111	
FM161	29577	6553164	hsa,tad,pte,cre,ngr	POC12	MKS1
FM162	59473	6553478	hsa,tad,pte,cre,ngr	PSO2	
FM163	48518	6551660	hsa,tad,tbr,gla,tva,cre,		
FM164	82958	6553096	hsa,tad,tbr,tva,cre,ngr		
FM165	29126	6550596	hsa,tad,tbr,cre,ngr		
FM166	70275	6553729	hsa,tad,gla,tva,cre,ngr		
FM167	58252	6552224	hsa,tad,cre,ngr		
FM168	71452	6553949	hsa,tad,cre,ngr		
FM169	73917	6552135	hsa,tps,cre,ngr	MOT51	
FM170	67664	6552862	hsa,tps,cre,ngr		
FM171	73885	6554672	hsa,ptr,pte,cre,ngr	PKHD1-2	
FM172	80536	6549888	hsa,ptr,tva,cre,ngr		
FM173	82475	6554227	hsa,ptr,cre,ngr	GSTS3	
FM174	31511	6553580	hsa,pte,cre,ngr		
FM175	78247	6554247	hsa,pte,cre,ngr	PSK1	
FM176	78184	6553815	hsa,tbr,cre,ngr	FKB12	
FM177	59563	6553039	hsa,gla,cre,ngr		
FM178	73058	6552889	hsa,ehi,cre,ngr		
FM179	78958	6554233	hsa,tva,cre,ngr	CYG11	
FM180	68774	6554233	hsa,tva,cre,ngr	CYG11	
FM181	66783	6554233	hsa,tva,cre,ngr	CYG11	



FM182	71868	6553432	hsa,cre,ngr		

Table S18 (related to Figure 4). Amoeboid motility associated proteins (AMs)

Amoeboid-motility associated proteins (AMs) were identified as described in Supplemental Experimental Procedures. Proteins encoded by multiple *Naegleria* paralogs are noted with multiple JGI ids in the second column. Red text is used to indicate AM gene families with homologs in *Trichomonas vaginalis*. Species abbreviations as in Table S17.

Name	Naegleria JGI protein ID(s)	Protein family (cluster ID)	species in cluster	Manual annotation of molecular function	PFAM domains (1e-10)
Actin Binding					
AM1	76225; 81173	6552646	ddi,ncr,hsa,mbr,tad,tva,ngr	Actin Binding	PF00307: Calponin homology (CH) domain|
AM2	82236	6550672	ddi,hsa,mbr,tad,ehi,ngr	Actin Binding (Drebrin/ABP-1)+1:65536	PF00018: SH3 domain|PF07653: Variant SH3 domain|
AM3	80016	6553037	hsa,tad,ehi,ngr	Actin Binding (Filamin)	PF00307: Calponin homology (CH) domain(2) | PF00630: Filamin/ABP280 repeat (4)
AM4	58328	6553194	ddi,ncr,hsa,mbr,tad,ehi,ngr	Actin Binding (twinfilin)	PF00241: Cofilin/tropomyosin-type actin-binding protein|
AM5	77687	6554206	ddi,hsa,tad,ngr	Actin Binding (Wash)	no PFAM
Signaling					
AM6	47789	6553607	hsa,ehi,ngr	Signalling	no PFAM | 7TMs predicted (TMHMM)
AM7	80282	6551531	ddi,ncr,hsa,mbr,ngr	Signalling	PF00018: SH3 domain (2) | PF07653: Variant SH3 domain (2)
AM8	67958	6554192	ddi,hsa,ehi,ngr	Signalling	PF04664: Opioid growth factor receptor (OGFr) conserved region|
AM9	71270	6554074	ddi,hsa,ngr	Signalling	PF07690: Major Facilitator Superfamily|
GAP					
AM10	80615	6553827	pra,hsa,ehi,ngr	GAP	PF00616: GTPase-activator protein for Ras-like GTPase|
AM11	81714	6549864	ddi,pra,ncr,hsa,mbr,tad,ehi,tva,ngr	GAP	PF02145: Rap/ran-GAP|
AM12	78320	6552715	ddi,ncr,hsa,mbr,tad,ehi,ngr	GAP (Nadrin)	PF00620: RhoGAP domain|
GEF					
AM13	50007; 68966	6550519	hsa,mbr,ehi,ngr	GEF	PF00618: Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif|PF00617: RasGEF domain|
Membrane					
AM14	57266	6554111	hsa,ehi,ngr	Membrane	PF00169: PH domain (3)
AM15	4009	6551952	ddi,ncr,hsa,mbr,tad,ngr	Membrane	PF04191: Phospholipid methyltransferase|
AM16	48624	6550972	hsa,tad,ehi,ngr	Membrane (Sphyngomyelin synthase-related)	no PFAM | 6 transmembrane domains predicted by TMHMM
AM17	82816	6552197	ddi,hsa,tad,ngr	Membrane (Saposin-B)	PF05184: Saposin-like type B, region 1 (3) | PF03489: Saposin-like type B, region 2 (3)
AM18	74044	6554134	ddi,hsa,ngr	Membrane	PF00754: F5/8 type C domain|
Cytoskeletal					
AM19	68732	6551755	ddi,pra,hsa,mbr,tad,ngr	Cytoskeletal	PF04912: Dynamitin|
Vesicle					
AM20	66720	6550251	hsa,ehi,ngr	Vesicle	PF02750: Synapsin, ATP binding domain|
AM21	62049	6552718	ddi,hsa,tad,ngr	Vesicle	no PFAM
AM22	78255	6554714	ddi,hsa,ngr	Vesicle	no PFAM
Protein Trafficking					
AM23	80788	6552115	ddi,hsa,tad,ngr	Protein Trafficking	no PFAM
Protein Turnover					
AM24	58872	6553361	ddi,ncr,hsa,mbr,ngr	Protein Turnover	no PFAM
AM25	65046	6553500	ddi,pra,hsa,ngr	Protein Turnover	no PFAM
Protein Interaction					
AM26	81452	6553306	hsa,tad,ehi,ngr	Protein Interaction	PF01436: NHL repeat (5)
Cell Cycle					
AM27	29264	6553985	ddi,hsa,tad,ngr	Cell Cycle	PF04005: Hus1-like protein|
AM28	58254	6553143	ddi,pra,hsa,tad,ngr	Cell Cycle	no PFAM
Metabolism					
AM29	65213	6550836	ddi,ncr,hsa,tad,ngr	Metabolism	PF06052: 3-hydroxyanthranilic acid dioxygenase|
AM30	81411	6549768	ddi,hsa,mbr,tad,ngr	Metabolism	PF03301: Tryptophan 2,3-dioxygenase|
AM31	78567	6553649	ddi,hsa,mbr,tad,ngr	Metabolism	no PFAM
AM32	69774	6552632	ddi,hsa,mbr,ngr	Metabolism	PF03632: Glycosyl hydrolase family 65 central catalytic domain|
AM33	78233	6553977	ddi,hsa,mbr,ngr	Metabolism	PF01229: Glycosyl hydrolases family 39|
AM34	78308	6554099	ddi,hsa,mbr,ngr	Metabolism	no PFAM
AM35	54990; 33467	6554340	ddi,hsa,ngr	Metabolism	PF03747: ADP-ribosylglycohydrolase| (not found in 54990)
Nucleic Acid Metabolism					
AM36	61798	6554539	ddi,pra,hsa,mbr,ngr	Nucleic Acid Metabolism	no PFAM
AM37	71340	6553262	ddi,pra,hsa,tad,tva,ngr	Nucleic Acid Metabolism	PF04858: TH1 protein|
AM38	53469	6549994	ddi,pra,hsa,tad,ngr	Nucleic Acid Metabolism	PF02144: Repair protein Rad1/Rec1/Rad17|
AM39	61854	6551921	ddi,pra,hsa,tad,ngr	Nucleic Acid Metabolism	PF00533: BRCA1 C Terminus (BRCT) domain (6)
AM40	56696	6552937	ddi,pra,hsa,ngr	Nucleic Acid Metabolism	PF05625: PAXNEB protein|
AM41	79767	6553442	ddi,ncr,hsa,ehi,ngr	Nucleic Acid Metabolism	PF02891: MIZ zinc finger|

AM42	77967	6554210	ddi,hsa,tad,ehi,ngr	Nucleic Acid Metabolism	PF06978: Ribonucleases P/MRP protein subunit POP1|
AM43	61462	6553485	ddi,hsa,tad,ngr	Nucleic Acid Metabolism	no PFAM
AM44	67690	6554441	ddi,hsa,ngr	Nucleic Acid Metabolism	no PFAM
Unknown					
AM45	74247	6553826	ddi,pra,hsa,tad,ehi,tva,ngr	Unknown	PF07258: HCaRG protein|
AM46	79980	6551166	ddi,pra,hsa,tad,tva,ngr	Unknown	no PFAM
AM47	80574	6550569	ddi,pra,hsa,tad,ngr	Unknown	no PFAM
AM48	5651	6549995	ddi,hsa,mbr,tad,ngr	Unknown	no PFAM
AM49	69245	6553024	ddi,hsa,mbr,tad,ngr	Unknown	PF07258: HCaRG protein|
AM50	67354	6553307	ddi,hsa,mbr,tad,ngr	Unknown	no PFAM
AM51	81535	6553370	ddi,hsa,mbr,tad,ngr	Unknown	no PFAM
AM52	65831	6553163	ddi,hsa,mbr,ngr	Unknown	no PFAM
AM53	81892	6551868	ddi,hsa,tad,tva,ngr	Unknown	no PFAM
AM54	45670	6553077	ddi,hsa,tad,tva,ngr	Unknown	no PFAM
AM55	80291	6552024	ddi,hsa,tad,ngr	Unknown	no PFAM
AM56	68270	6552063	ddi,hsa,tad,ngr	Unknown	PF07258: HCaRG protein|
AM57	75398	6552274	ddi,hsa,tad,ngr	Unknown	no PFAM
AM58	63144	6552665	ddi,hsa,tad,ngr	Unknown	no PFAM
AM59	81040	6552879	ddi,hsa,tad,ngr	Unknown	PF07742: BTG family|
AM60	79421; 4350	6554251	ddi,hsa,tad,ngr	Unknown	no PFAM
AM61	5358	6551639	ddi,hsa,ngr	Unknown	no PFAM
AM62	62278	6553655	ddi,hsa,ngr	Unknown	no PFAM
AM63	69376	6554699	ddi,hsa,ngr	Unknown	no PFAM

Tables S14-S16, FigS5 legend and low resolution image

Table of Contents

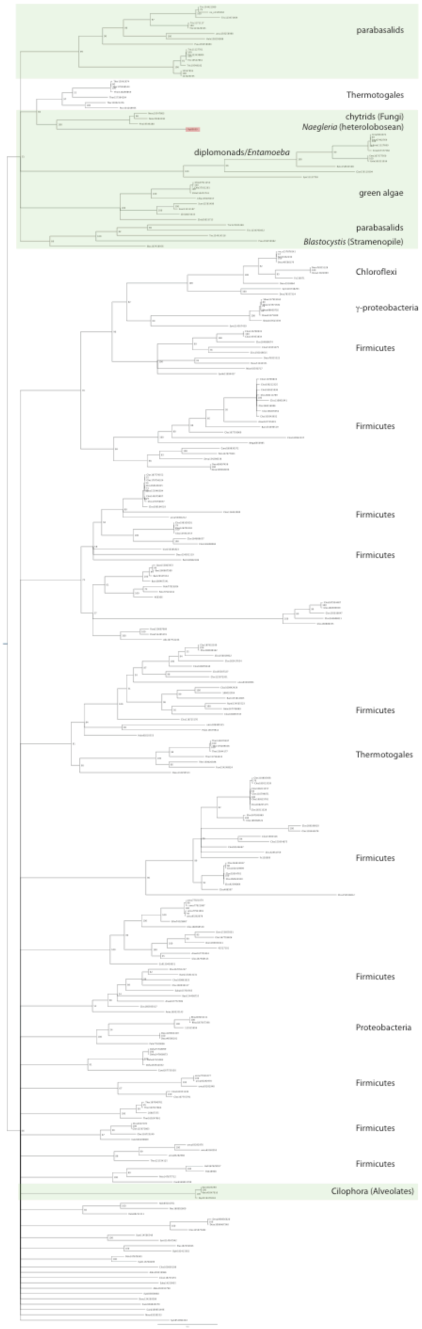

parabasalids
Thermotogales
chytrids (Fungi)
Naegleria (heterolobosean)
diplomonads/*Entamoeba*
green algae
parabasalids
Blastocystis (Stramenopile)
Chloroflexi
γ-proteobacteria
Firmicutes
Firmicutes
Firmicutes
Firmicutes
Firmicutes
Thermotogales
Firmicutes
Firmicutes
Proteobacteria
Firmicutes
Firmicutes
Firmicutes
Cilophora (Alveolates)

Figure S5 (related to Figure 3). Fe-Fe hydrogenase phylogeny

Phylogenetic analysis of the *N. gruberi* Fe-Fe hydrogenase was performed (with 577 homologous positions of 248 diverse eukaryotic and bacterial hydrogenases) using the JTT amino acid substitution model within RAxML (see Supplemental Experimental Procedures) with 100 bootstrap replicates. Values are shown for nodes with >50% bootstrap support. The *Naegleria* sequence is in red, and phylogenetic clades of hydrogenases are indicated in blue (eukaryotic) and gray (bacterial). The partial sequence (GenBank Accession Number CAD12183.1) of another putative Heterolobosean hydrogenase from the hydrogenosome-containing *Psalteriomonas lantena* was not included in the analysis. All 114 amino acids of the *P. lanterna* sequence show 60% identity with the *N. gruberi* hydrogenase. Hydrogenosomal (H), mitochondrial (M) and cytoplasmic (C) localization of hydrogenases are shown.

Table S14 (related to Figure 3). Core energy metabolism proteins

The *Naegleria* genome was searched manually for proteins that make up core metabolic pathways that have been biochemically characterised in one or more protists and fungi. Many of the proteins with a typical mitochondrial respiratory chain [*i.e.* complexes I-IV and ATP synthase (complexV)] are encoded in mitochondrial genome, rather than in the nucleus (NC_002573). This number of mitochondrially-encoded respiratory chain components is higher than that observed in many eukaryotes, but is lower than that observed in the mitochondrial genome of *Reclinomonas americana* (another JEH protist). Putative trypanosomatid-specific accessory components for complex IV (cytochrome *c* oxidase) (Horvath et al., 2000) were not evident in the *Naegleria* genome.

[a] The fourteen core subunits common to both prokaryotic and eukaryotic complex I enzymes were used for the analysis discussed here (Remacle et al., 2008).

Core pathways	**Naegleria protein IDs**
Glycolysis	
Glucokinase	81163
Other sugar kinases	69011; 2897; 34493; 68410
Glucose-6-phosphate isomerase	30686
PP_i-dependent phosphofructokinase	35679
Aldolase (class I)	56383
Glyceraldehyde-3-phosphate dehydrogenase	53883
Triosephosphate isomerase	29287
Phosphoglycerate kinase	81218
Phosphoglycerate mutase(PGAM)	72581; 52804 (PGAM-like)
enolase	60351
Pyruvate phosphate dikinase	36352; 59363
Pyruvate kinase	35453; 76757; 36690
Lactate dehydrogenase	3825; 48420; 51010; 75708
Glycerol kinase	38161
Glycerol-3-phosphate dehydrogenase	34539; 29597; 80825
Pentose phosphate pathway	
Glucose-6-phosphate dehydrogenase	30686
Phosphogluconate dehydrogenase	30694
Transaldolase	73024
Transketolase	44342; 6095
Ribose-5-phosphate isomerase	38157
Phosphoribosylpyrophosphate synthetase	60335; 34278
Regulatory enzyme	
Phosphofructokinase-2/fructose-2, 6-bisphosphatase	38553
Adenylate kinase	81301; 59535; 58410; 31874; 59363; 68635; 72729
Pyruvate-Acetate metabolism	
Pyruvate dehydrogenase (and related complexes *e.g.* α-Ketoglutarate dehydrogenase complex)	39315; 56281; 73427; 1128; 38032; 60828; 38237; 59476
Phosphoenolpyruvate carboxykinase	38463
Malic enzyme	59395; 76270; 81494
Malate dehydrogenase	31160; 60960; 83065
Acetyl-CoA synthetase (ADP-forming)	82174
Acetate:succinate CoA transferase (putative)	78694; 38428
Mitochondrial fatty acid β-oxidation	
Trifunctional enzyme	29546
Krebs cycle	
Citrate synthase	38914; 54230; 82269
Aconitase	38693; 30116; 59586
Isocitrate dehydrogenase (NAD^+-dependent)	80807; 70009
Isocitrate dehydrogenase (NADP-dependent)	82731
α-Ketoglutarate dehydrogenase complex	see pyruvate dehydrogenase
Succinyl-CoA synthetase	29455; 79109; 83245
Succinate dehydrogenase (SDH1)	44665
Succinate dehydrogenase (SDH2)	mitochondrial
Fumarase (class I and Class II)	34693; 83307
Malate dehydrogenase	31160; 60960; 83065
Mitochondrial respiratory chain	
Complex I (core sub-units only)[a]	

NuoA (bacterial nomenclature is used here)	Mitochondrial genome
NuoB	33757; 36743; 30514
NuoC	Mitochondrial genome
NuoD	Mitochondrial genome
NuoE	69707
NuoF	58165
NuoG	Mitochondrial genome
NuoH	Mitochondrial genome
NuoI	Mitochondrial genome
NuoJ	Mitochondrial genome
NuoK	Mitochondrial genome
NuoL	Mitochondrial genome
NuoM	Mitochondrial genome
NuoN	Mitochondrial genome
Complex II	
SDH1	44665
SDH2	Mitochondrial genome
Complex III	
Processing peptidases	82210; 58349
Rieske Fe-S protein	31585
Cytochrome *c*$_1$	53169
Another core sub-unit	81117
Cytochrome *c*	77897
Complex IV (COX1-3)	Mitochondrial genome
Other key proteins	
Alternative NADH dehydrogenase	72836; 81197; 51352
Electron transferring flavoprotein	56308; 75058
ETF:Q oxidoreductase	38537
Alternative oxidase	81108; 30919; 76066
Superoxide dismutase	35997; 81995; 75082; 69926; 4996
Soluble fumarate reductase	82312; 79044
Hydrogenase	
Fe-hydrogenase	80699
Maturation Factor HydE	81640
Maturation Factor HydF	65416
Maturation Factor HydG	81639
Fe-S cluster assembly	
cysteine desulphurase	44858
ISU1/ISU2/NifU	32298
NFU	3509
ISA1/ISA2	possible homologs only
Ferredoxin	31742; 81802
Ferredoxin reductase	1460
Frataxin	63017
Erv1	5453
NAR	47235
Possible arginine dehydrolase pathway	
Ornithine transcarbamoylase	No clear homolog
Arginine deiminase	47456
Carbamate kinase	54727

Alcohol dehydrogenase/oxidoreductase family proteins: 56035; 75143; 71114; 55836; 67275; 80400; 60616; 59126; 51101; 59049; 51515; 75508; 51218; 69127; 72777; 55836; 75143; 56126; 73866; 74818

Table S15 (related to Figure 3). Mitochondrial transit peptide predictions for hydrogenase module components

To investigate the possible location of the *Naegleria* Fe-hydrogenase and Fe-hydrogenase-associated maturases we used the sub-cellular localisation prediction tools Mitoprot (Claros and Vincens, 1996), Predotar (Small et al., 2004), PSORT II (Nakai and Horton, 1999), and TargetP 1.1 (Emanuelsson et al., 2000) (see below) For comparison, we also subjected the *bona fide* Fe-hydrogenase from *Blastocystis* (Stechmann et al., 2008). Extremely high confidence predictions for mitochondrial targeting are in bold italics.

	Likelihood of mitochondrial targeting			
Protein	*Mitoprot*	*Predotar*$^{\text{animal/fungal seq}}$	*PSORT II*	*TargetP 1.1*
Fe-hydrogenase (JGI peptide ID, 80699)	***0.92***	***0.51***	0.22	0.71
HydE (JGI peptide ID 81640)	***0.78***	***0.78***	0.48	0.57
HydF (JGI peptide ID 65416)	***0.97***	0.02	0.52	***0.78***
HydG (JGI peptide ID 81639)	***0.98***	***0.68***	0.22	***0.96***
Blastocystis sp. Fe-Hydrogenase (ACD10930)	***0.93***	***0.91***	0.61	***0.83***

Table S16 (related to Figure 3). Phylogenetic distribution of core biosynthetic pathways

	free -living soil dwellers			excavate parasites		
	Ng	*Dd*	*Cr*	*Tb*	*Gl*	*Tv*
Purine biosynthesis	-	+	+	-	-	-
Pyrimidine biosynthesis	+	+	+	+	-	-
Gluconeogenesis	-[a]	+	+	+	-	-
Glycogen metabolism	-	+	-	-	?	?
Glyoxylate cycle	-	+	+	-	-	-
Fatty acid biosynthesis	-[b]	-[b]	+	-[c]	-	-
Mitochondrial type II fatty acid biosynthesis	+	+	+	+	-	-
Sterol biosynthesis	+	+	+	+	-	-
Polyketide biosynthesis	+	+++[d]	+	-	-	-
Heme biosynthesis	-[e]	+	+	-	-	-
Shikimate pathway[f]	-	-	+	-	-	-

Ng, Naegleria gruberi
Dd, Dictyostelium discoideum
Cr, Chlamydomonas reinhardtii
Tb, Trypanosoma brucei
Gl, Giardia lamblia
Tv, Trichomonas vaginalis

[a] Absence of a homolog from any of the four known classes of fructose-1,6-bisphosphatase.
[b] Yet requires no lipid in axenic culture medium.
[c] **Not** a fatty acid auxotroph – uses type III pathway for bulk fatty acid biosynthesis.
[d] Denotes large expansion of polyketide synthase family in *Dd.*
[e] Although *Ng* does contain ferrochelatase (indicating an ability to insert Fe into a pre-formed [scavenged] porphyrin ring) and an O_2-independent coproporhyrinogen oxidase homolog. A function for the latter enzyme is not immediately apparent.
[f] Pathway is found in the majority of fungi (an exception is the microsporidian *Encephlitozoon cuniculi*)

Table of Contents:

Table S1 (related to Table 1). Sumary of *de novo* repeats generated by RepeatScout

Annotation	Number of sequences in RepeatScout library	Total coverage in genome (bp)
Contains TE-associated Pfam domain	2	36,078 (0.09%)
Homology to known TEs	6	98,498 (0.24%)
Satellite	1	5,304 (0.01%)
Contains non TE-associated Pfam domain	20	534,350 (1.30%)
Unknown complex repeats	151	1,380,214 (3.37%)
rRNA	4	56,685 (0.14%)
tRNA	22	90,892 (0.22%)
Total	206	2,202,021 (5.38%)

Table S2 (related to Table 1). Genes predicted by automated annotation, classified by method.

Method	*N. gruberi* v.1
Total models	15,753 (100%)
Homology with proteins in GenBank nr database	2,031 (13%)
ab initio prediction	13,553 (86%)
EST cluster consensus sequence	169 (1%)

Table S3 (related to Table 1). Supporting evidence for gene models.

Evidence	*N. gruberi* v.1
Complete models	1,4615 (93%)
Models with EST alignment	4,669 (30%)
Models with homology in GenBank nr database	11,587 (74%)
Models with Swissprot alignment	8,452 (54%)
Models with Pfam domain	7,074 (45%)

Table of Contents

Table S7 (related to Figure 1). Putative meiosis genes

We searched for *Naegleria* homologs of known meiosis genes (Malik et al., 2008) using bidirectional BLAST searches. We indicate presence (+) of one or more homologs and absence (-) of a homolog from a genome with good sequence coverage. Blank cells indicate cases in which a homolog was not identified and this could have been a result of incomplete genome sequence.

Organism	Spo11*	Hop1*	Hop2*	Mnd1*	Dmc1*	Rad51	Msh4,5*	Msh2,6	Mre11	Rad50	Rad52	Mlh1	Mlh2	Mlh3	Pms1/2
Bacteria	-	-	-	-	-	RecA	MutS		-	-	-	MutL			
Archaea	TopoVI	-	-	-	-	RadA			SbcD	SbcC	-				
Giardia	+	+	+	+	++	-	-	++	+	+	+	+	+	-	+
Trichomonas	+	+	++	+	+	++	++	++	+	++	-	+++	++	+	+
Trypanosoma§	+	+	+	+	+	+	++	++	+	+	-	+	-	+	+
Naegleria	**50278 81222**	**69651**	**75325 3930**	**65192**	**nf**	**45247 88254**	**442 75725**	**29947 45760**	**51432**	**70670**	**51673**	**73051**	**nf**	**70637**	**75182 72513**
Entamoeba	++	-	+	+	+	+	++	++	+	+	+	+	-	+	++
Tetrahymena	+	+	+	+	+	+	+	++				+			+
Phytopthera	+			+	+	+		++	+	+	+	+			+
Chlamydomonas	++		+	+	+	+	+	+	+	+		+			
Arabadopsis	++++	+	+	+	+	+	++	++	+	+	-	+	-	+	+
Dictyostelium			+	+		+	++	++	+	+	++	+		+	+
Neurospora	+	+	-	-	-	+	++	++	+	+	+	+	+	+	+
Homo	+	+	+	+	+	+	++	++	+	+	+	+	+	+	+

Legend:
*Meiosis-specific genes (Ramesh et al. 2005)
§Genetic evidence for meiosis
Organism known to undergo meiosis
nf = not found

Table S8 (related to Figure 1). DNA replication components

Naegleria's DNA replication machinery was identified via manual genome searches. Swissprot sequences from *Saccharomyces* and human were used as queries. *Naegleria* JGI protein IDs are listed in the third column. Yellow highlight indicates presence in *Naegleria* but absence in trypanosomes. Text in red indicates cases of unclear homology. Presence (+) or absence (-) in the *Giardia* (Morrison et al., 2007) and *Trypanosome* genome (El-Sayed et al., 2005a) is indicated.

Fuctional Description		**Subunit**	***Naegleria***	***Giardia***	***Saccharomyces***	***Trypanosomes***
		Replication Initiation				
Origin Recognition Complex (ORC) binds chromatin at replication origins and serves as the foundation for assembly of the DNA replication pre-replicative complex.		ORC1/Cdc6	88228	**+**	**+**	**+**
		ORC4	45184	**+**	**+**	**-**
		ORC2	88227	**-**	**+**	**-**
		ORC3	-	-	**+**	-
		ORC5	-	-	**+**	-
		ORC6	88231	**-**	**+**	**-**
DNA replication licensing factor, required for pre-replication complex assembly		cdt1	-	-	**+**	-
ORC activation of the origin DNA leads to the binding of the MCM 1-7 proteins to the unwound origin as a ring-shaped heterohexamer . The MCM complex translocates along the DNA with the replication fork during S phase		mcm2	73646	**+**	**+**	**+**
		mcm3	36839	**+**	**+**	**+**
		mcm4	415	**+**	**+**	**+**
		mcm5	29938	**+**	**+**	**+**
		mcm6	76348	**+**	**+**	**+**
		mcm7	75829	**+**	**+**	**+**
Recruited to MCM pre-RC complexes; promotes release of MCM, recruits elongation machinery.		cdc45	47967	**-**	**+**	**+**
Replication protein A (RPA) is a heterotrimeric single-stranded DNA-binding protein. Plays essential roles in DNA replication, nucleotide excision repair, and homologous recombination. Sequences of RPA1-3 are related.		RPA-like	79436 58854 45814	**-**	**+**	**+ 28 kDa** **+ 51 kDa** **- 14kDa**
MCM1 belongs to the MADS box transcription factor family, and binds and stimulates ARSs.		mcm1	81157?	**-**	**+**	**n/d**
Required for efficient initiation of DNA replication. May act as docking platform for DNA polymerase.		mcm10	-	-	**+**	-
DNA replication licensing factor, required for pre-replication complex assembly		sid2	-	-	**+**	**n/d**
GINS complex (Sld5p, Psf1p, Psf2p, Psf3p), which is localized to DNA replication origins and implicated in assembly of the DNA replication machinery. GINS associates with the MCM2-7 complex and Cdc45 to activate the eukaryotic minichromosome maintenance helicase.		psf2	88232	**-**	**+**	**n/d**
		sld5	-	-	**+**	**n/d**
		psf1	-	-	**+**	**n/d**
		psf3	62119	**-**	**+**	**n/d**
		Replication				
Polymerase alpha: 4 components make up primase and polymerase activity (lagging strand synthesis).	p180 polymerase	DPOLA_HUMAN	61128	n/d	n/d	**+**
	p70 promasome assembly	DPOA2_HUMAN	80762	n/d	n/d	**-**
	p58 (primase large subunit)	PRI2_HUMAN	70379	n/d	n/d	**+**
	p48 (primase small subunit)	PRI1_HUMAN	73667	n/d	n/d	**+**
Replicative DNA polymerase	delta polymerase 125 kd subunit	DPOD1_HUMAN	44642	n/d	n/d	**+**
	delta polymerase 66 kd subunit	DPOD2_HUMAN	71662	n/d	n/d	**+**
	delta polymerase 48 kd subunit	DPOD3_HUMAN	-	n/d	n/d	**-**
	delta polymerase 12 kd subunit	DPOD4_HUMAN	-	n/d	n/d	**-**
	epsilon polymerase subunit 1	DPOE1_HUMAN	59151	n/d	n/d	**+**
	epsilon polymerase subunit 2	DPOE2_HUMAN	63811	n/d	n/d	**+**
	epsilon polymerase subunit 3	DPOE3_HUMAN	53490	n/d	n/d	**-**
	epsilon polymerase subunit 4	DPOE4_HUMAN	**54797**	n/d	n/d	**-**
PCNA (processivity factor)		PCNA_HUMAN	35238	n/d	n/d	**+**
RFC (PCNA loader)	145	RFC1_HUMAN	66154	n/d	n/d	**+**
	40	RFC2_HUMAN	68675	n/d	n/d	**+**
	38	RFC3_HUMAN	29162	n/d	n/d	**+**
	37	RFC4_HUMAN	44707	n/d	n/d	**+**
	36.5	RFC5_HUMAN	29498	n/d	n/d	**+**

Table S9 (related to Figure 1). RNA polymerase II subunits

Naegleria's RNA polymerase II subunits were identified via manual genome searches. SWISSPROT (http://ca.expasy.org/sprot/) sequences from *Saccharomyces* and human were used as queries as indicated in the third and fifth columns, respectively. The *Naegleria* homologs identified in each search are indicated in the fourth and sixth columns. *Naegleria* JGI protein IDs are listed in the third column. Yellow highlight indicates presence in *Naegleria* but absence in trypanosomes. Red indicates cases of unclear homology. Presence (+) or absence (-) in *Giardia* (Morrison et al., 2007), *Trichomonas* (Carlton et al., 2007), *Entamoeba* (Loftus et al., 2005) and *Trypanosome* genomes (Berriman et al., 2005; El-Sayed et al., 2005a; Ivens et al., 2005) is indicated.

Subunit	Fuctional Description of Yeast Homolog (From yeastgenome.org)	Yeast sequence	*Naegleria* (with yeast sequence)	Human Sequence	*Naegleria* (with human sequence)	*Giardia*	*Trichomonas*	*Entamoeba*	*Saccharomyces*	*Trypanosomes*
Transcription: RNA Polymerase 2 Subunits										
RNAPII B3	RNA polymerase II third largest subunit B44, part of central core; similar to prokaryotic alpha subunit	RPB3_YEAST	61034 29098	RPB3_HUMAN	61034 29098	+	+	+	+	+
RNAPII B5	RNA polymerase subunit ABC27, common to RNA polymerases I, II, and III; contacts DNA and affects transactivation	RPAB1_YEAST	29724	RPAB1_HUMAN	29724	+	+	+	+	+
RNAPII B6	RNA polymerase subunit ABC23, common to RNA polymerases I, II, and III; part of central core; similar to bacterial omega subunit	RPAB2_YEAST	33993	RPAB2_HUMAN	33993	+	+	+	+	+
RNAPII B10 (beta)	RNA polymerase subunit ABC10-beta, common to RNA polymerases I, II, and III	RPAB5_YEAST	88233	RPAB5_HUMAN	88233	+	+	-	+	+
RNAPII B11	RNA polymerase II subunit B12.5; part of central core; similar to Rpc19p and bacterial alpha subunit	RPB11_YEAST	29567 80573	RPB11_HUMAN	29567 80573	+	+	+	+	+
RNAPII B1	Largest subunit	RPB1_YEAST	51024 58671 35014 (III?) 79527(I?)	RPB1_HUMAN	51024 58671 35014 (III?) 79527(I?)	-	+	+	+	+
RNAPII B2	RNA polymerase II second largest subunit B150, part of central core; similar to bacterial beta subunit	RPB2_YEAST	59892 55898 (I?) 60598 (III?)	RPB2_HUMAN	59892 55898 (I?) 60598 (III?)	-	+	-	+	+
RNAPII B7	RNA polymerase II subunit B16; forms two subunit dissociable complex with Rpb4p	RPB7_YEAST	74650 29266(III?)	RPB7_HUMAN	74650 29266(III?)	-	+	+	+	-
RNAPII B8	RNA polymerase subunit ABC14.5, common to RNA polymerases I, II, and III	RPAB3_YEAST	71611	RPAB3_HUMAN	71611	-	+	+	+	+
RNAPII B4	RNA polymerase II subunit B32; forms two subunit dissociable complex with Rpb7p; involved recruitment of 3'-end processing factors to transcribing RNA polymerase II complex and in export of mRNA to cytoplasm under stress conditions	RPB4_YEAST	-	RPB4_HUMAN	64024	-	-	-	+	-
RNAPII B9	RNA polymerase II subunit B12.6; contacts DNA; mutations affect transcription start site; involved in telomere maintenance	RPB9_YEAST	72143	RPB9_HUMAN	72143	-	-	+	+	+
RNAPII B12	RNA polymerase subunit, found in RNA polymerase complexes I, II, and III (ABC10-alpha)	RPAB4_YEAST	61926	RPAB4_HUMAN	61926	-	-	-	+	-

Table S10 (related to Figure 1). Basal transcription factors

Naegleria’s basal transcription factors were identified via manual genome searches. SWISSPROT (http://ca.expasy.org/sprot/) sequences from *Saccharomyces* and/or human were used as queries as indicated in the second column. *Naegleria* JGI protein IDs are listed in the third column. Yellow highlight indicates presence in *Naegleria* but absence in trypanosomes. Text in red indicates cases of unclear homology. Presence (+) or absence (-) in *Giardia, Trichomonas, Entamoeba* and *Trypansome* genomes (as published in their genome papers (Berriman et al., 2005; Carlton et al., 2007; El-Sayed et al., 2005a; Ivens et al., 2005; Loftus et al., 2005; Morrison et al., 2007) is indicated. * Divergent TFIIA1 and 2 in *Trypansoma were* identified by biochemical methods (Schimanski et al., 2005).

			Giardia	***Trichomonas***	***Entamoeba***	***Saccharomyces***	***Trypanosomes***
Basal Transcription Factors							
TBP	TBP_YEAST	78851	+	+	+	+	+
TFIIH2	SSL1_YEAST	34964	+	+	+	+	N/F
TFIID1	TAF1_YEAST	81689 48788	-	+	+	+	N/F
TFIID2	TAF2_YEAST	65369	-	+	+	+	N/F
TFIID4	TAF5_YEAST	81220	-	+	+	+	N/F
TFIID5	TAF6_YEAST	1097	-	+	-	+	N/F
TFIID6	TAF7_YEAST TAF7_HUMAN TVAG_040830	-	-	+	-	+	N/F
TFIID7	TAF9_YEAST	4666	-	+	+	+	N/F
TFIID8	TAF10_YEAST	71677	-	+	-	+	N/F
TFIIE1	T2EA_YEAST T2EA_HUMAN	61798	-	+	-	+	-
TFIIH3	TF2H3_HUMAN	68603	-	+	+	+	-
TFIIH4	TFB2_YEAST	68362	-	+	+	+	-
TFIID9	TAF11_YEAST	45427	-	-	-	+	N/F
TFIID10	TAF12_YEAST TAF12_HUMAN	88234	-	-	-	+	N/F
TFIID11	TAF13_YEAST	5555	-	-	-	+	N/F
TFIIB	TF2B_YEAST	58428	-	-	-	+	-
TFIIA1	TOA1_YEAST TF2AY_HUMAN	88235	-	-	-	+	+*
TFIIA2	TOA2_YEAST T2AG_HUMAN	63919	-	-	-	+	+*
TFIIF1	T2FA_YEAST T2FA_HUMAN DDBDRAFT_0205751	-	-	-	-	+	-
TFIIF2	T2FB_YEAST	77700	-	-	-	+	-
TFIIF3	TAF14_YEAST (not in human)	77965	-	-	-	+	N/F
TFIIE2	T2EB_YEAST T2EB_HUMAN DDBDRAFT_0187408	77884 76715	-	-	-	+	-
TFIIH1	TFB1_YEAST	61113	-	-	-	+	N/F

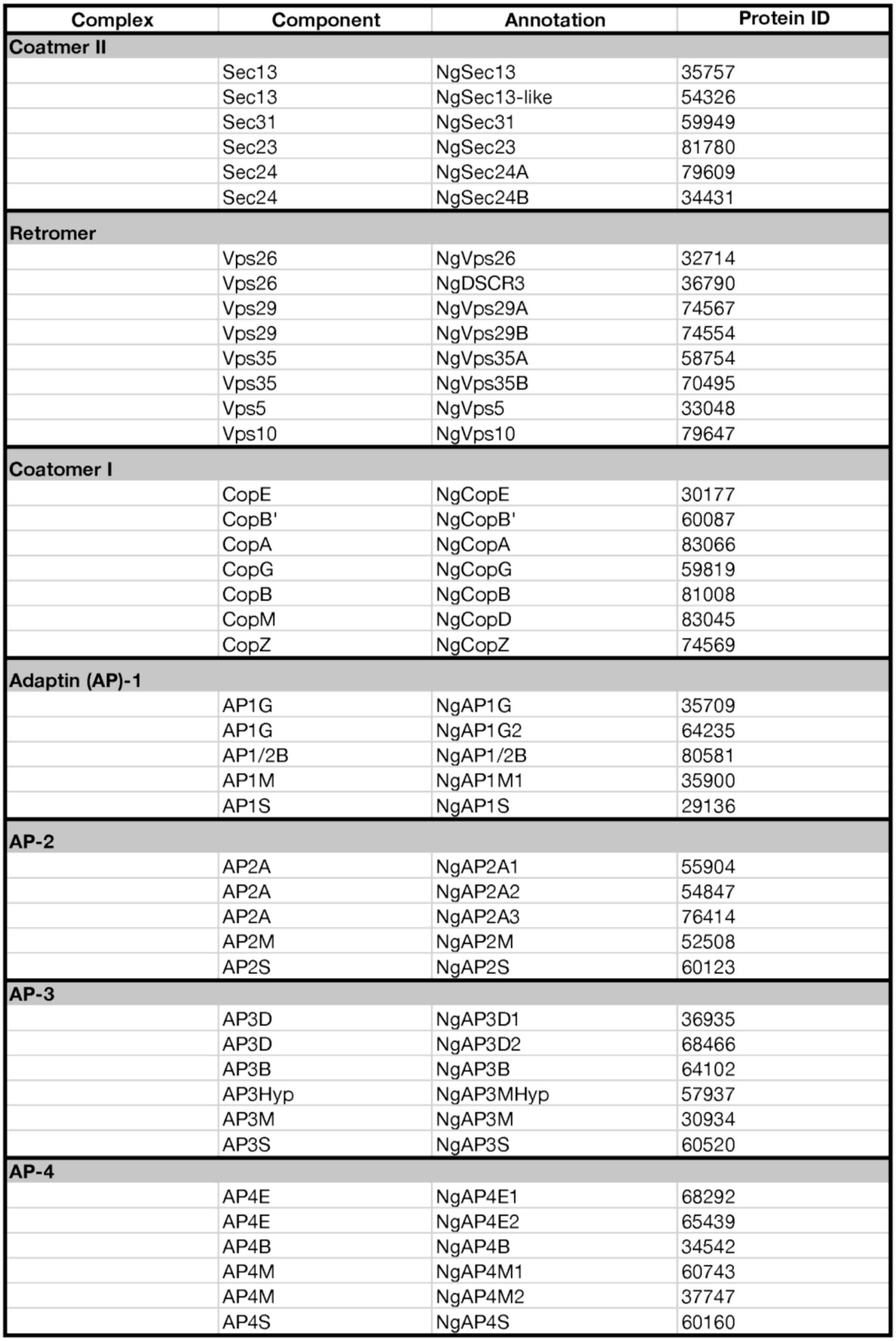
Table S11 (related to Figure 1). Signaling components across eukaryotes

The left-hand column contains the number of each of 56 Pfam domains with signaling functions in *Naegleria*, as determined using gathering thresholds. The central columns contain an estimate of the numbers of the indicated domain in the species, normalized to the number of predicted loci in that organism (E-value < 1E-10, and normalized by dividing the Pfam counts by the number of loci in the genome, then multiplied by 10,000 for readability). The highest frequency is shaded pink, the second highest yellow. The right hand column contains brief descriptions of the Pfam domain.

		Normalized counts in genomes (using evalue cutoffs of 1e-10)																	
Naegleria (Total Counts)		Human	Monosiga	Neurospora	Dictyostelium	Entamoeba	Arabidopsis	Physcomitrella	Chlamydomonas	Phytophtora	Thalassiosira	Phaeodactylum	Paramecium	Naegleria	Trypanosome	Trichomonas	Giardia	Prochlorococcus	
	Cyclic Nucleotide Signalling																		
108	Cyclase	7.7	17.4	1	2.9	0	0	0	40	1.3	10.5	13	9.1	51.5	65.6	15.4	0	0.9	PF00211: Adenylate and Guanylate cyclase catalytic domain
7	cAMP-ase	9	6.5	1	2.2	0	0	0	15.8	7	6.1	0	10.8	4.5	5.5	5.9	1.5	0	PF00233: 3'5'-cyclic nucleotide phosphodiesterase
7	cAMP Bind	0	0	0	0	0	0	0	6.9	0	0	0	90.6	4.5	5.5	1	1.5	0	PF00027: Cyclic nucleotide-binding domain
	PIP Signalling																		
14	P3/P4 Kinase	7.7	15.2	6.9	11.8	16.4	6.8	3.8	4.8	13.3	7.9	6	8.8	8.9	9.8	8.2	6.2	0	PF00454: Phosphatidylinositol 3- and 4-kinase
7	P4/P5kinase	3.9	5.4	2	4.4	2	5.7	1.5	2.1	10.2	2.6	2	7.3	4.5	3.3	0.3	1.5	0	PF01504: Phosphatidylinositol-4-phosphate 5-Kinase
11	P3K access.	3.9	4.3	2	5.9	8.2	0.8	0.3	0.7	1.9	0.9	0	1.5	4.5	1.1	1.3	0	0	PF00613: Phosphoinositide 3-kinase family, accessory domain (PIK domain)
3	PLPC, gamma	6.4	4.3	3	0.7	0	2.6	1.8	0.7	0	1.8	1	1.5	1.9	1.1	0	0	0	PF00387: Phosphatidylinositol-specific phospholipase C, Y domain
3	PLPC, x	6.4	6.5	4	0.7	0	3.4	1.8	0.7	0	1.8	2	1.5	1.9	2.2	0	0	0	PF00388: Phosphatidylinositol-specific phospholipase C, X domain
	Calcium Signalling																		
76	C2	43.3	18.5	7.9	19.9	18.4	31.3	14.3	9.6	15.9	6.1	8	7.6	34.3	8.7	6.9	0	0	PF00168: C2 domain (ca+ dependent membrane targetting)
70	EF	36.4	38.1	6.9	25	29.7	40.7	21.1	34.4	36.2	25.5	15	81.5	34.3	18.6	10.6	9.3	0	PF00036: EF hand
33	Ion Flux	7.7	9.8	10.9	8.1	8.2	12.8	7.6	12.4	10.8	10.5	12	10.6	7.6	9.8	1.7	3.1	0.4	PF00122: E1-E2 ATPase superfamily of cation transport enzymes mediate membrane flux of all common biologically relevant cations.
21	Calmod Bind	7.3	7.6	2	2.9	0	8.7	2.5	7.6	12.7	5.3	4	4.8	5.1	9.8	0.3	1.5	0	PF00612: IQ calmodulin-binding motif
6	Ca/Na pump	3.9	3.3	6.9	2.2	1	4.9	3.3	2.8	3.2	5.3	4	1.5	3.8	0	0.3	0	0.4	PF01699: Sodium/calcium exchanger protein
	Hetrotrimeric G-Proteins																		
39	Galpha	6.4	3.3	3	8.1	1	1.1	0	0	0.6	1.8	1	0	7.6	0	1	0	0	PF00503: G-protein alpha subunit
171	G regulator	11.1	1.1	1	2.9	0	0	0	0	1.9	1.8	0	0	1.9	0	0.2	0	0	PF00615: Regulator of G protein signaling domain
	Small G-Proteins																		
182	Ras	53.6	45.7	20.8	78.1	150	27.1	15.9	11	22.2	21.1	15	58	96	26.2	55.5	18.5	0	PF00071: Ras family
11	RasGap	5.6	7.6	4	7.4	5.1	0	0	0	0.6	0	0	0	6.4	0	0	0	0	PF00616: GTPase-activator protein for Ras-like GTPase
27	RasGEF	10.7	8.7	4	21.4	22.5	0	0	0	1.3	0	0	0	14.6	0	2	0	0	PF00617: RasGEF domain
19	RasGEF Nterm	2.6	0	2	5.2	6.1	0	0	0	0	0	0	0	9.5	0	0.2	0	0	PF00618: Guanine nucleotide exchange factor for Ras-like GTPases; N-terminal motif
33	ARF	13.7	15.2	6.9	15.5	15.3	9	8.3	10.3	8.3	7	6	8.6	20.3	16.4	4.9	7.7	0	PF00025: ADP-ribosylation factor family
8	ArfGAP	15	10.9	4	8.8	10.2	6.4	4	3.4	4.4	5.3	5	2.8	4.5	5.5	3.4	4.6	0	PF01412: Putative GTPase activating protein for Arf
4	Sec7/ARF GEF	6.4	6.5	3	5.2	5.1	3	2	2.1	2.5	4.4	4	3	1.9	2.2	1.8	3.1	0	PF01369: Sec7 domain (The Sec7 domain is a guanine-nucleotide-exchange-factor (GEF) for the PF00025 family)
21	RhoGEF	24.4	13	3	30.9	47.1	0	0	0	0.6	0	0	4	12.7	0	1.7	0	0	PF00621: RhoGEF domain
25	RhoGap	26.6	28.3	5.9	27.3	55.3	3.8	2	0	0.6	0	0	2.8	15.3	0	3.9	0	0	PF00620: RhoGAP domain
4	G-Binding	3.9	2.2	0	0.7	2	0.4	1.3	0.7	2.5	0.9	1	4.3	1.9	0	0	0	0	PF02263: Guanylate-binding protein, N-terminal domain
25	GTPase	5.6	17.4	11.9	13.3	9.2	9.8	8.8	14.5	10.8	21.9	20.9	7.6	14	16.4	1.2	12.3	4	PF01926: GTPase of unknown function
	Phosphate Signalling																		
265	Kinase	190	305	82.1	169	305	341	147	203	213	105	95.8	541	187	173	133	239	0	PF00069: Protein kinase domain
89	Tyr. Kinase	142	219	40.6	121	211	291	118	137	156	29.9	39.9	315	96.6	86.3	77.9	97.2	0	PF07714: Protein tyrosine kinase
47	Calcineurin	10.3	29.4	16.8	16.2	50.1	22.6	13.6	9.6	19.7	12.3	16	21.2	15.9	29.5	25.5	24.7	1.8	PF00149: Calcineurin-like phosphoesterase
32	phosphatase	12	14.1	7.9	9.6	18.4	10.9	6	9	12.7	8.8	8	5	14	9.8	2.3	3.1	0.4	PF03372: Endonuclease/Exonuclease/phosphatase family
32	Phophastase	11.6	8.7	4	10.3	16.4	1.5	2	5.5	5.7	4.4	1	26.5	13.4	9.8	0.8	3.1	0	PF00782: Dual specificity phosphatase, catalytic domain
7	Histidine PPAse	1.3	1.1	2	0	3.1	0	0	0.7	0.6	0	0	2.5	3.8	2.2	3.4	4.6	0	PF00328: Histidine acid phosphatase
5	14-3-3	4.3	1.1	2	1.5	3.1	5.3	2.8	1.4	0.6	0.9	1	3	3.2	2.2	3.7	1.5	0	PF00244: 14-3-3 protein (phosphoserine/threonine binding modules)
10	FHA	7.7	1.1	4.9	8.1	1	3.8	2.8	3.4	4.4	2.6	3	5.8	3.2	1.1	0.3	3.1	0	PF00498: FHA domain: phosphopeptide recognition domain found in many regulatory proteins with specificity for phosphothreonine-containing epitopes but will also recognise phosphotyrosine with relatively high affinity.
	Histidine Kinase																		
32	RRR	0	2.2	10.9	12.5	0	12.4	26.4	4.1	3.2	4.4	11	45.7	17.2	0	1.5	0	4	PF00072: Response regulator receiver domain
27	HisKinase	0	1.1	7.9	9.6	0	3.8	19.4	4.1	2.5	0	3	4.8	12.1	0	1	0	0.9	PF00512: His Kinase A (phosphoacceptor) domain (This entry represents the dimerisation and phosphoacceptor domain found in histidine kinases.)
	Sensors																		
50	PAS	3	0	0	0.7	0	1.9	2.8	0	3.2	2.6	1	0	3.8	0	0	0	0	PF00989: PAS fold (signal sensor domain, often w/ PAC domain)
4	Nitrase sense	0	0	0	0	0	0	0	0	0	0	0	0	2.5	0	0	0	0	PF08376: Nitrate and nitrite sensing
4	BLUF	0	0	0	0	0	0	0	0	0	0	0	0	1.9	0	0	0	0	PF04940: Sensors of blue-light using FAD

Table S12 (related to Figure 1). Protein trafficking genes in *Naegleria*

Putative *Naegleria* orthologs and paralogs were classified by reciprocal BLAST searches, followed by phylogenetic analysis to define subfamily or identity where appropriate (see below). *Naegleria* homologs are indicated by their JGI protein ID, and are grouped by predicted membership in complexes or functional systems. BLAST searches were used to predict function. Abbreviated gene names from this are shown under “Annotation”.



Complex	Component	Annotation	Protein ID
Coatmer II			
	Sec13	NgSec13	35757
	Sec13	NgSec13-like	54326
	Sec31	NgSec31	59949
	Sec23	NgSec23	81780
	Sec24	NgSec24A	79609
	Sec24	NgSec24B	34431
Retromer			
	Vps26	NgVps26	32714
	Vps26	NgDSCR3	36790
	Vps29	NgVps29A	74567
	Vps29	NgVps29B	74554
	Vps35	NgVps35A	58754
	Vps35	NgVps35B	70495
	Vps5	NgVps5	33048
	Vps10	NgVps10	79647
Coatomer I			
	CopE	NgCopE	30177
	CopB'	NgCopB'	60087
	CopA	NgCopA	83066
	CopG	NgCopG	59819
	CopB	NgCopB	81008
	CopM	NgCopD	83045
	CopZ	NgCopZ	74569
Adaptin (AP)-1			
	AP1G	NgAP1G	35709
	AP1G	NgAP1G2	64235
	AP1/2B	NgAP1/2B	80581
	AP1M	NgAP1M1	35900
	AP1S	NgAP1S	29136
AP-2			
	AP2A	NgAP2A1	55904
	AP2A	NgAP2A2	54847
	AP2A	NgAP2A3	76414
	AP2M	NgAP2M	52508
	AP2S	NgAP2S	60123
AP-3			
	AP3D	NgAP3D1	36935
	AP3D	NgAP3D2	68466
	AP3B	NgAP3B	64102
	AP3Hyp	NgAP3MHyp	57937
	AP3M	NgAP3M	30934
	AP3S	NgAP3S	60520
AP-4			
	AP4E	NgAP4E1	68292
	AP4E	NgAP4E2	65439
	AP4B	NgAP4B	34542
	AP4M	NgAP4M1	60743
	AP4M	NgAP4M2	37747
	AP4S	NgAP4S	60160

SNARE			
	Qa	Not resolved-NgQa1	34956
	Qa	Not resolved-NgQa2	68648
	Qa	Not resolved-NgQa3	61224
	Qa	Not resolved-NgQa4	58865
	Qa	Not resolved-NgQa5	80924
	Qa	Not resolved-NgQa6	61311
	Qa	Not resolved-NgQa7	77849
	Qa	NgSyn5	72316
	Qa	Not resolved-NgQa8	59068
	Qa	Not resolved-NgQa9	78202
	Qa	Not resolved-NgQa10	57153
	Qa	Not resolved-NgQa11	76330
	Qa	Not resolved-NgQa12	68854
	Qb	Not resolved-NgQb1	81391
	Qb	Not resolved-NgQb2	75545
	Qb	Not resolved-NgQb3	67259
	Qb	GOS2B	70268
	Qb	GOSR1	73430
	Qb	Not resolved-NgQb4	53114
	Qb	Bet1-like	61410
	Qb	NPSN	74626
	Qb	Not resolved-NgQb5	73677
	Qc	Not resolved-NgQc1	62459
	Qc	Not resolved-NgQc2	61952
	Qc	Not resolved-NgQc3	68071
	Qc	Not resolved-NgQc4	64514
	Qc	Not resolved-NgQc5	64911
	R	NgVAMP7A	29713
	R	NgVAMP7B	4072
	R	NgSYB-like	32174
	R	NgVAMP7C	82192
	R	NgVAMP7D	68250
	R	NgVAMP7E	71254
	R	NgVAMP7F	72497
	R	NgVAMP7G	69151
	R	NgVAMP7H	78471
	R	NgVAMP7I	69037
	R	NgVAMP7J	71222
	R	NgSec22	44614
	R	NgYkt6A	69478
	R	NgYkt6B	36593
SM proteins			
	Sec1	NgSec1A	80728
	Vps45	NgVps45A	56416
	Vps45	NgVps45B	34061
	Vps33	NgVps33A	79862
	Vps33	NgVps33B	82244
	Vps33	NgVps33C	29012
	Sly1	NgSly1	692
Golgi protein			
	GRASP	NgGRASP	62049
	p115	Ngp115	49429

Clathrin			
	AP180	NgAP180	56097
	Clathrin light chain	NgCLC	57212
	Clathrin heavy chain	NgCHC	31358
Conserved oligomeric Golgi complex (COG)			
	COG1	NgCOG1	66558
	COG3	NgCOG3	61868
	COG6	NgCOG6	73153
Dsl1			
	no subunits found		
Dynamin			
	Dynamin like-A	NgDnmA	82955
	Dynamin like-B	NgDnmB	29431
Other adaptors			
	epsinR	NgEpsR	69468
	eps15	NgEps15	64962
ESCRT0			
	no subunits found		
ESCRTI			
	Vps23	NgVps23	73037
	Vps28	NgVps28	72174
	Vps37	NgVps37	75751
ESCRTII			
	Vps22	NgVps22	74336
	Vps25	NgVps25	69867
	Vps36	NgVps36	70476
ESCRTIII			
	Vps2	NgVps2	39042
	Vps20	NgVps20	33227
	Vps24	NgVps24	75958
	Vps32	NgVps32	81519
ESCRTIII-associated			
	Rim20	NgRim20	79832
	Vps4	NgVps4	75791
	Vps31	NgVps31	79832
	Vps46	NgVps46	33610
	Vps60	NgVps60	33349
Exocyst			
	Sec3	NgSec3	63977
	Sec5	NgSec5	69336
	Sec6	NgSec6	63187
	Sec8	NgSec8	69336
	Sec10	NgSec10	61580
	Sec15	NgSec15	75634
	Exo70	NgExo70	56566
Golgi-associated retrograde protein complex (GARP)			
	Vps52	NgVps52	73029
	Vps54	NgVps54	62467
p67 (LAMP analogue)			
	p67 (lysosomal protein)	Ngp67	64562

Homotypic fusion and vacuole protein sorting (HOPS) complex			
	Vps11	NgVps11	81916
	Vps16	NgVps16	65718
	Vps18	NgVps18	61970/61873
	Vps33	NgVps33D	67882
	Vps39	NgVps39	38867
	Vps41	NgVps41	71662
Transport protein particle (TRAPP)-I			
	Bet3	NgBet3	75444
	Bet5	NgBet5	31091
	Trs20	NgTrs20	30765
	Trs23	NgTrs23	31037
	Trs31	NgTrs31	4684
	Trs33	NgTrs33	32491
TRAPP-II			
	no subunit recovered		
Endosomal PI 3,5-kinase			
	Fab1	NgFab1	78054
Endosomal PI 3-kinase			
	Vps34	NgVps34	67703
Rabs			
	Rab1	NgRab1	55383
	Rab2	NgRab2	44714
	Rab4	NgRab4	71359
	Rab11	NgRab11A	60727
	Rab11	NgRab11B	35122
	Rab11	NgRab11C	56963
	Rab14	NgRab14	59420
	Rab5	NgRab5	55970
	Rab21	NgRab21	82940
	Rab6	NgRab6	35987
	Rab28	NgRab28	33099
	Rab34/36	NgRab34/36	75713
	Rab7	NgRab7A	82544
	Rab7	NgRab7B	71436
	Rab8	NgRab8	30231
	Rab18	NgRab18A	30714
	Rab1B	NgRab18B	76807
	Rab32	NgRab32A	60792
	Rab32	NgRab32B	56124
	Rab29	NgRab29	4014
	Rab23	NgRab23	315441
	RabTbX3	NgRabTbX3	609261
	Ran	NgRabRanA	32121
	Ran	NgRabRanB	37563
	Rab (unclassified)	NgRabX1	62685
	Rab (unclassified)	NgRabX2	711913
	Rab (unclassified)	NgRabX3A	70677
	Rab (unclassified)	NgRabX3B	83236
	Rab (unclassified)	NgRabX3C	34455
	Rab (unclassified)	NgRabX4A	66688
	Rab (unclassified)	NgRabX4B	76956
	Rab (unclassified)	NgRabX4A	4275
	Rab (unclassified)	NgRabX5	32393

Table S13 (related to Figure 1). RNAi machinery of *Naegleria*

To identify potential *Naegleria* RNAi genes, the genome was searched (using BLASTP at the JGI genome portal, http://www.jgi.doe.gov/naegleria/) with genes from various eukaryotes (including human and Arabidopsis).

Gene	JGI Protein ID of *Naegleria* homolog
Dicer	62031
Argonaute	70125
RNA-dependent RNA polymerase	67488

Table of Contents

Table S19 (related to Figure 6). Phylogenetic distribution of core eukaryotic proteins without Pfam or KOG annotations

We made 4,133 ancient eukaryotic protein familes. Of these, 481 have no Pfam or KOG annotations. The phylogenetic distribution of these protein families among major eukaryotic groups is shown with a letter showing presence and (-) showing absence. J JEH, C chromalveolates, P plants, A amoebozoa, O opisthokonts.

distribution in major eukaryotic groups	number of families
`J----`	3
`J---O`	20
`J--A-`	16
`J--AO`	38
`J-P--`	18
`J-P-O`	18
`J-PA-`	14
`J-PAO`	20
`JC---`	27
`JC--O`	51
`JC-A-`	31
`JC-AO`	34
`JCP--`	28
`JCP-O`	82
`JCPA-`	16
`JCPAO`	65

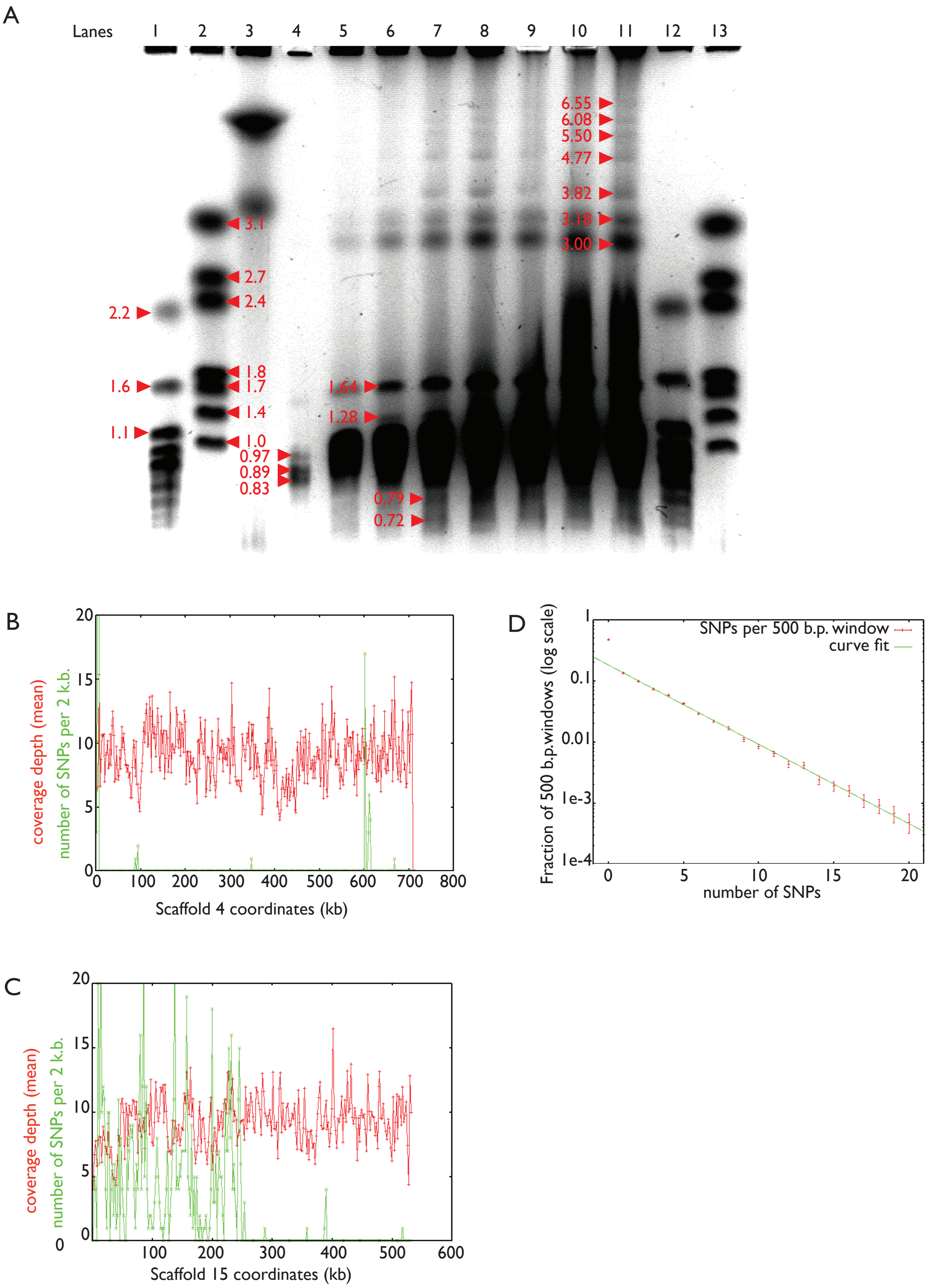
Table S20 (related to Figure 6). Losses of core eukaryotic genes in all major clades

Numbers of gene families shared between JEH and other eukaryotic groups are shown. We also show % loss relative to JEH. 3,784 familes are found in *Naegleria* and at least two other eukaryotic groups excluding POD (Ngr +2). 1,983 families are found in *Naegleria* and at least four other eukaryotic groups excluding POD (Ngr +4). In both cases, we consider the following major eukaryotic groups: Chromalveolates, Opisthokonts, Plants, Amoebozoa (Fig. 2). These are in addition to JEH (containing *Naegleria*). Membership in POD was not a search criterion, but numbers of families with POD members are shown. Ngr *Naegleria gruberi*

Eukaryotic group	Number of clusters containing proteins from Naegleria and one other eukaryotic group and at least three mutal best BLAST hits	% loss relative to JEH	Ngr + 2	Ngr + 4
JEH	4,133	0	3,784	1,983
Trypanosomes	1,709	59	1,631	1,179
POD	1,713	59	1,572	1,112
Amoebozoa	2,842	31	2,799	1,983
Opisthokonts	3,489	16	3,371	1,983
Plants	3,204	22	3,116	1,983
Chromalveolates	3,284	21	3,195	1,983

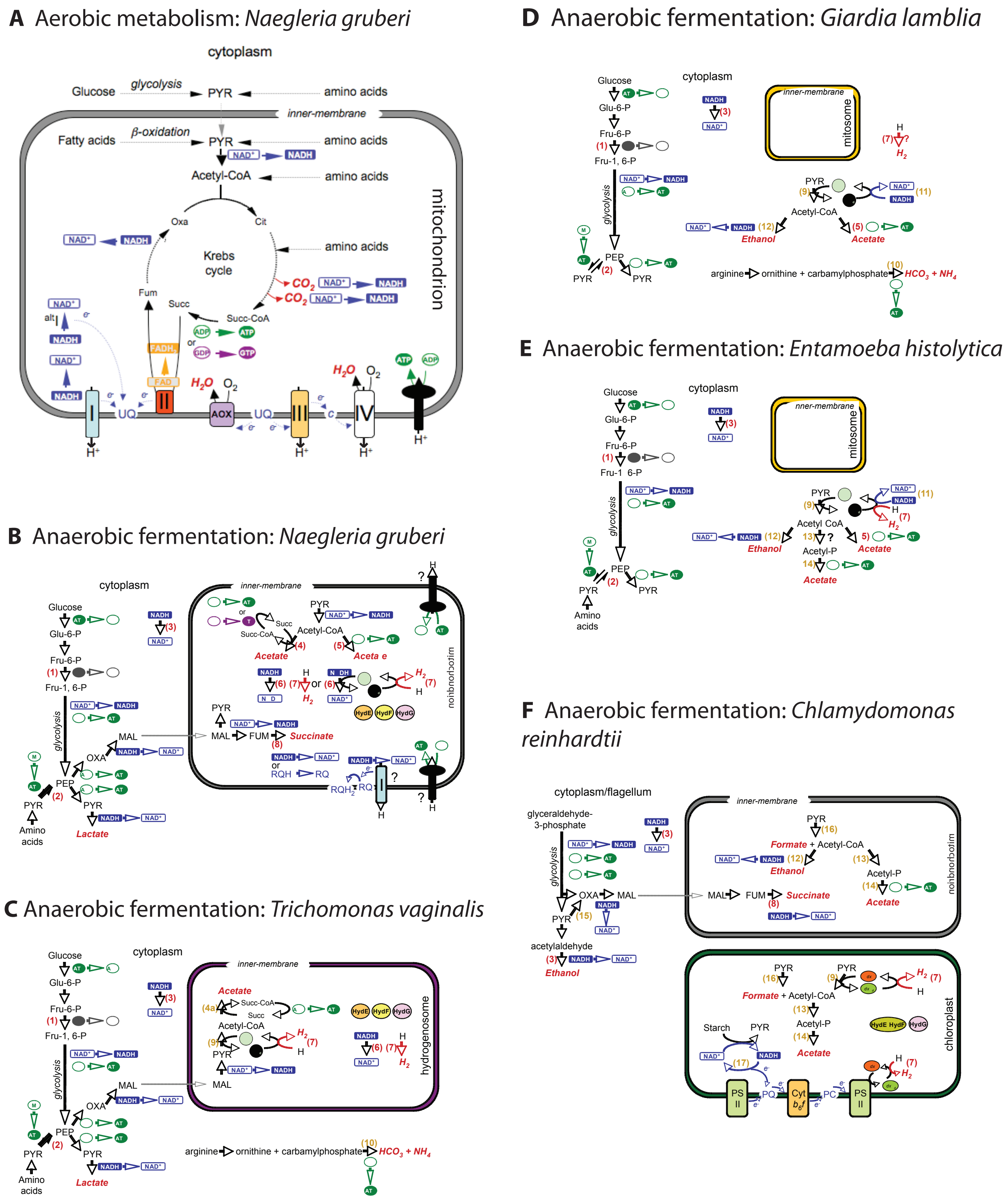
FigS1, high resolution

Click here to download Supplemental Figure: FritzLaylin_FigS1.pdf



FigS2, high resolution

Click here to download Supplemental Figure: FritzLaylin_FigS2.pdf



A Aerobic metabolism: *Naegleria gruberi*

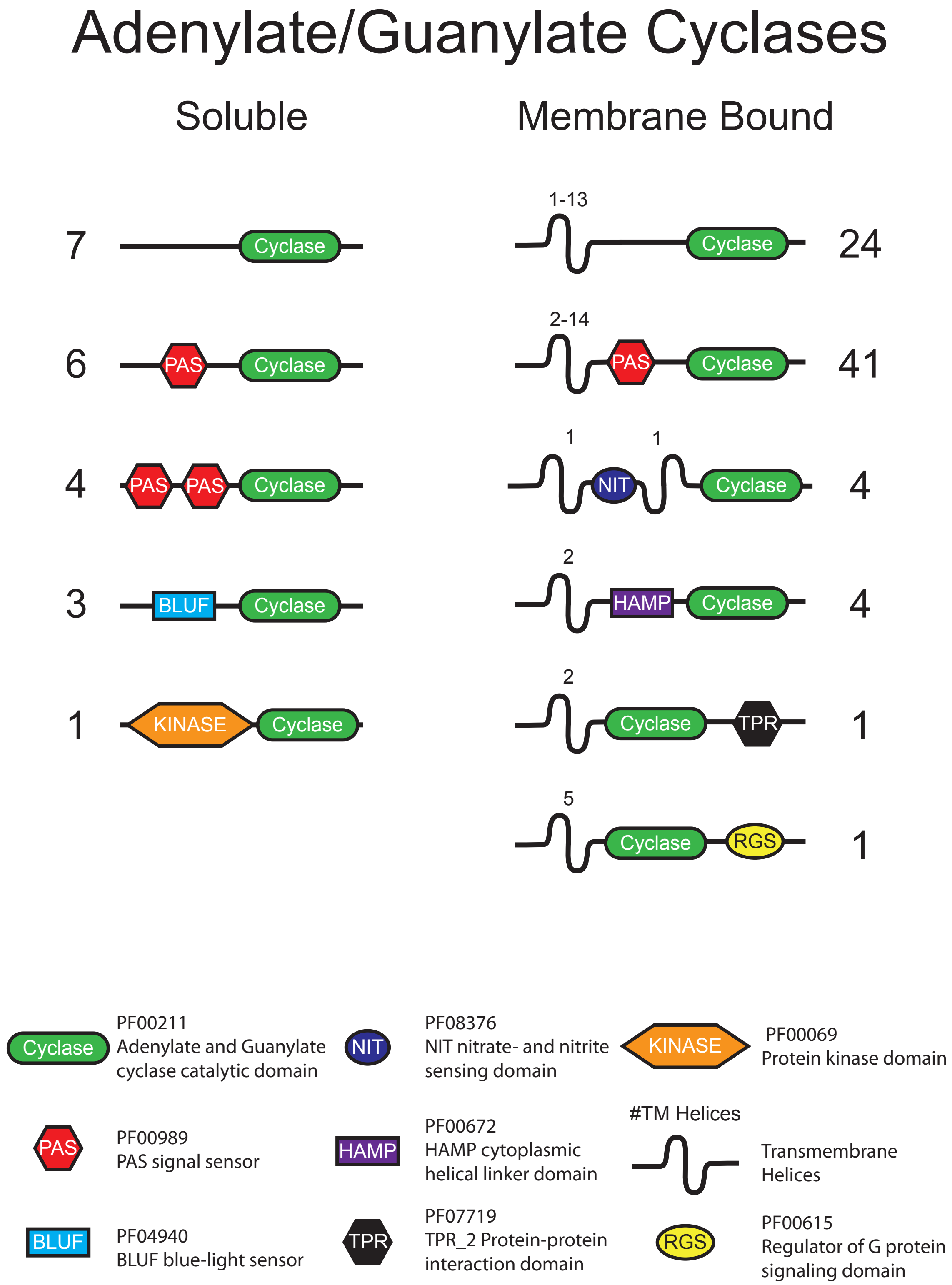
B Anaerobic fermentation: *Naegleria gruberi*

C Anaerobic fermentation: *Trichomonas vaginalis*

D Anaerobic fermentation: *Giardia lamblia*

E Anaerobic fermentation: *Entamoeba histolytica*

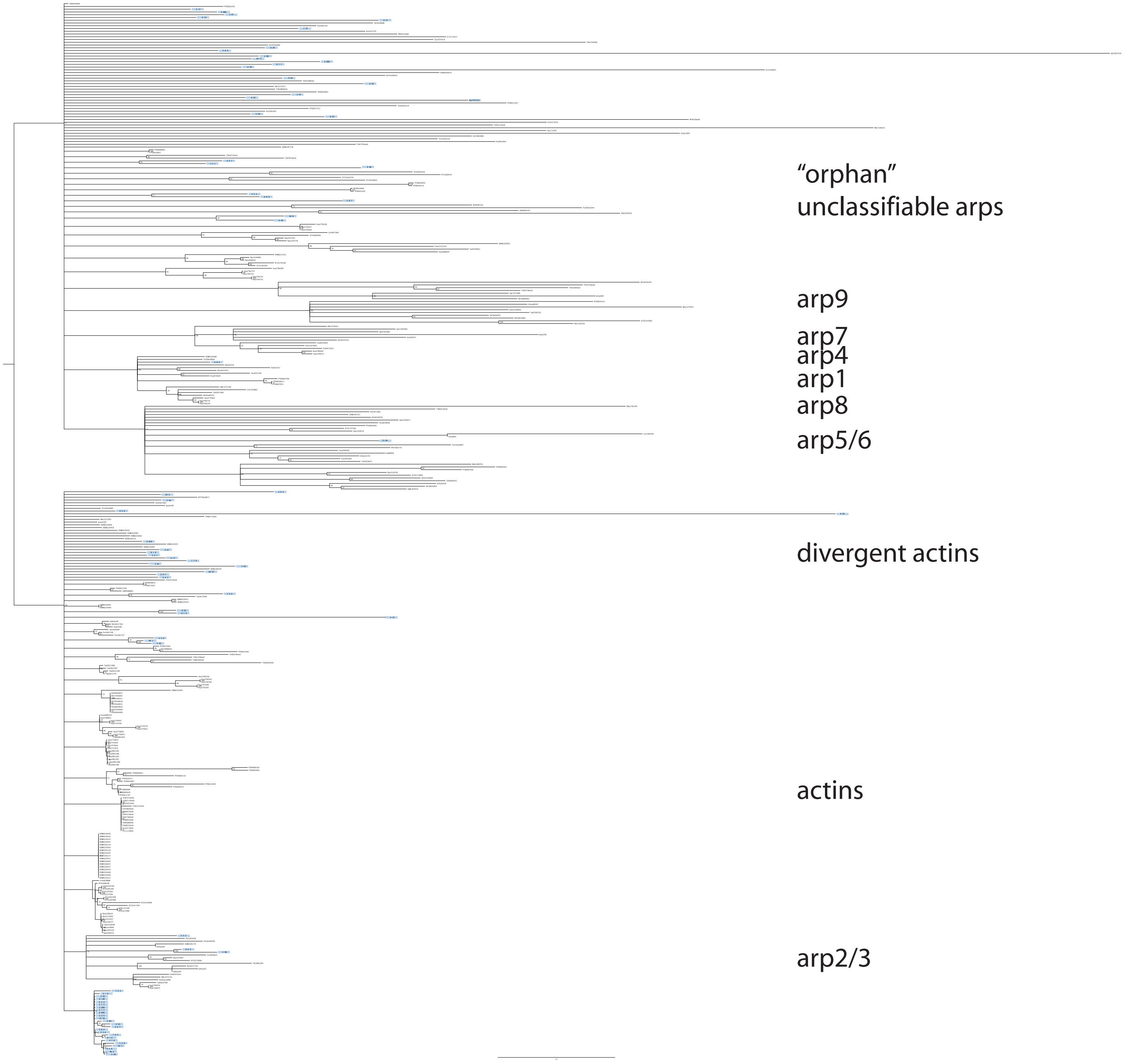
F Anaerobic fermentation: *Chlamydomonas reinhardtii*


Adenylate/Guanylate Cyclases
Soluble
Membrane Bound
7
Cyclase
6
PAS
Cyclase
4
PAS
PAS
Cyclase
3
BLUF
Cyclase
1
KINASE
Cyclase
1-13
Cyclase
24
2-14
PAS
Cyclase
41
1
1
NIT
Cyclase
4
2
HAMP
Cyclase
4
2
Cyclase
TPR
1
5
Cyclase
RGS
1
Cyclase
PF00211
Adenylate and Guanylate cyclase catalytic domain
NIT
PF08376
NIT nitrate- and nitrite sensing domain
KINASE
PF00069
Protein kinase domain
PAS
PF00989
PAS signal sensor
HAMP
PF00672
HAMP cytoplasmic helical linker domain
#TM Helices
Transmembrane Helices
BLUF
PF04940
BLUF blue-light sensor
TPR
PF07719
TPR_2 Protein-protein interaction domain
RGS
PF00615
Regulator of G protein signaling domain

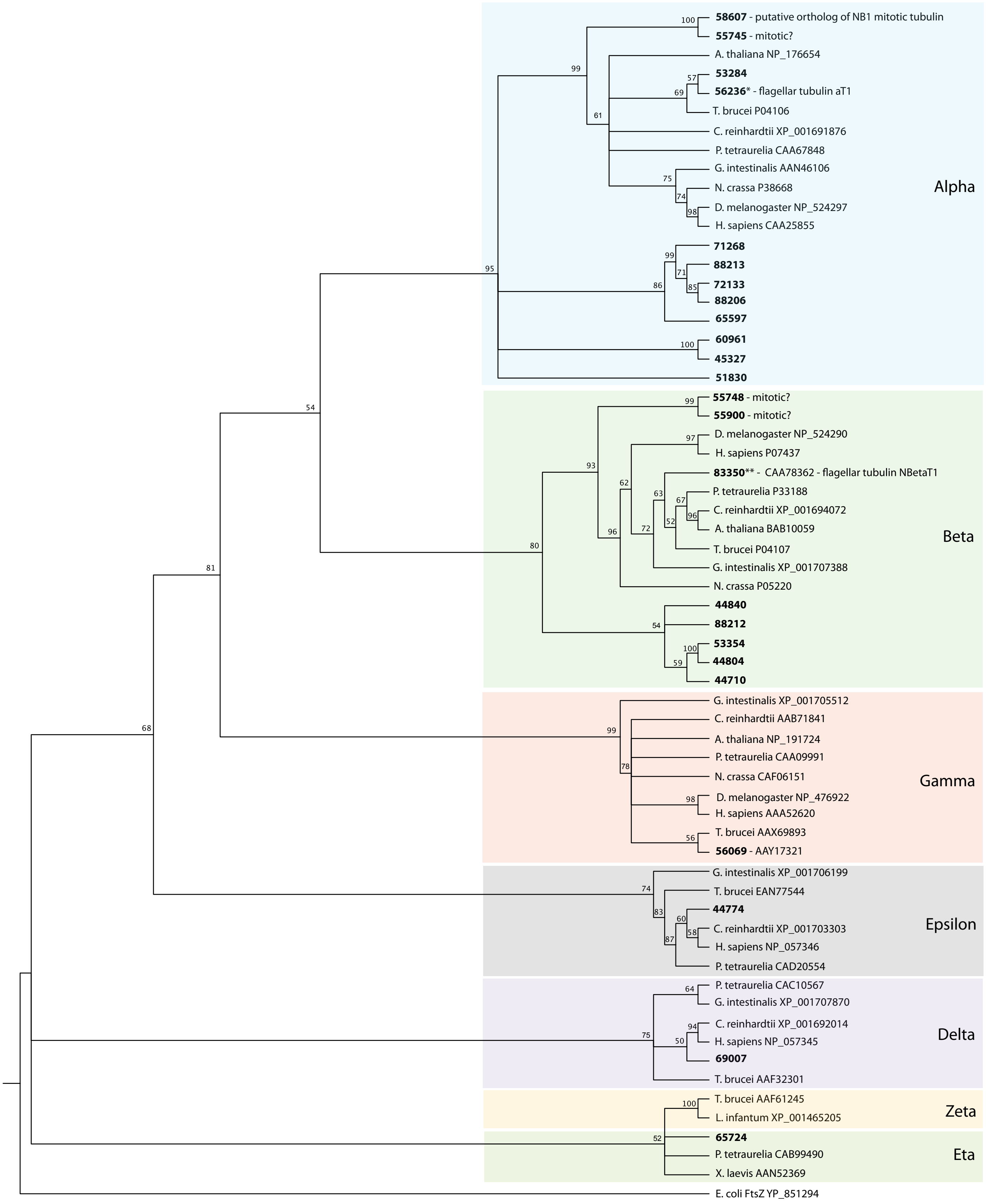
FigS4a, high resolution

Click here to download Supplemental Figure: FritzLaylin_FigS4a.pdf

A Actin/arp phylogeny

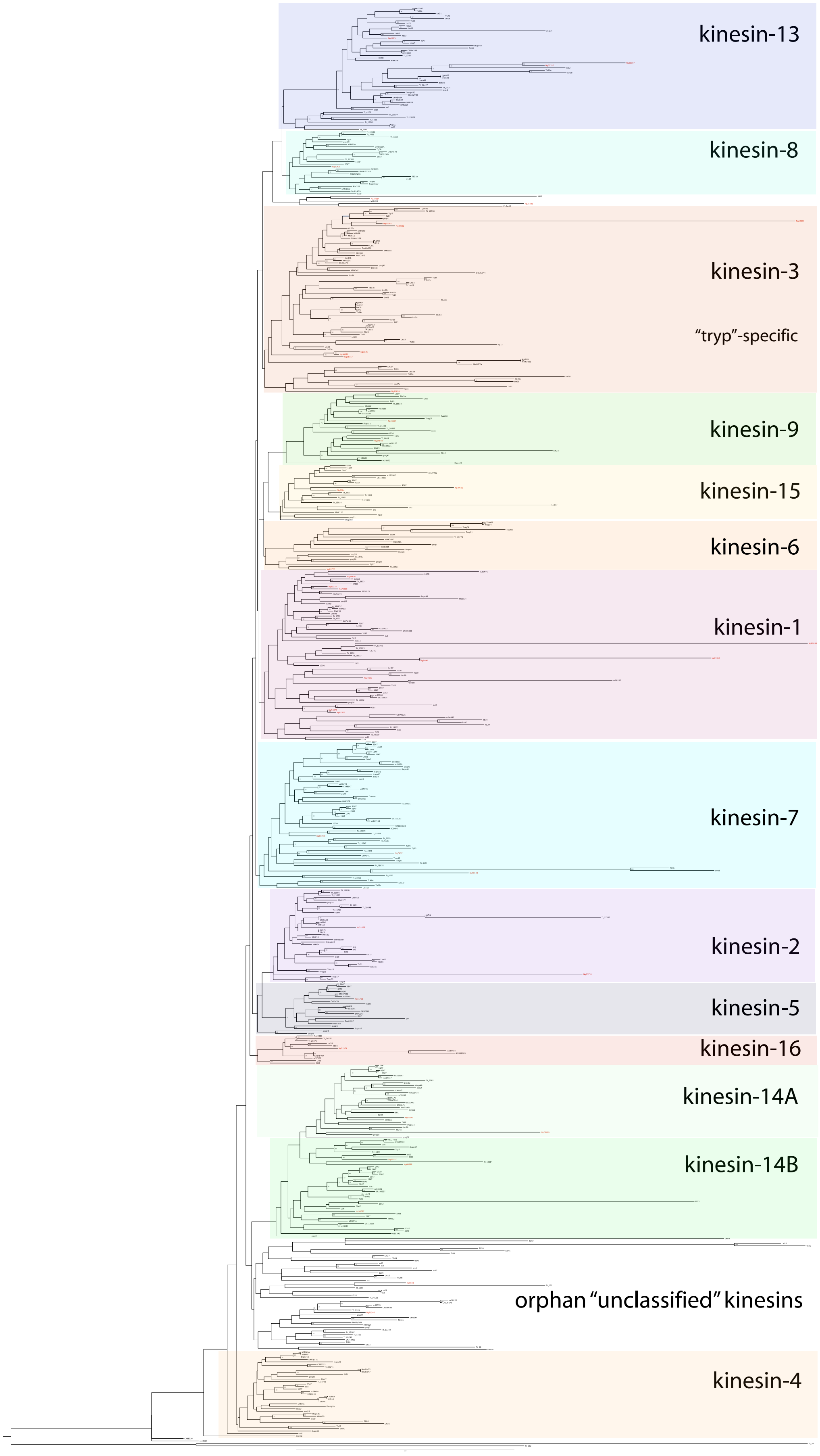


B Tubulin phylogeny

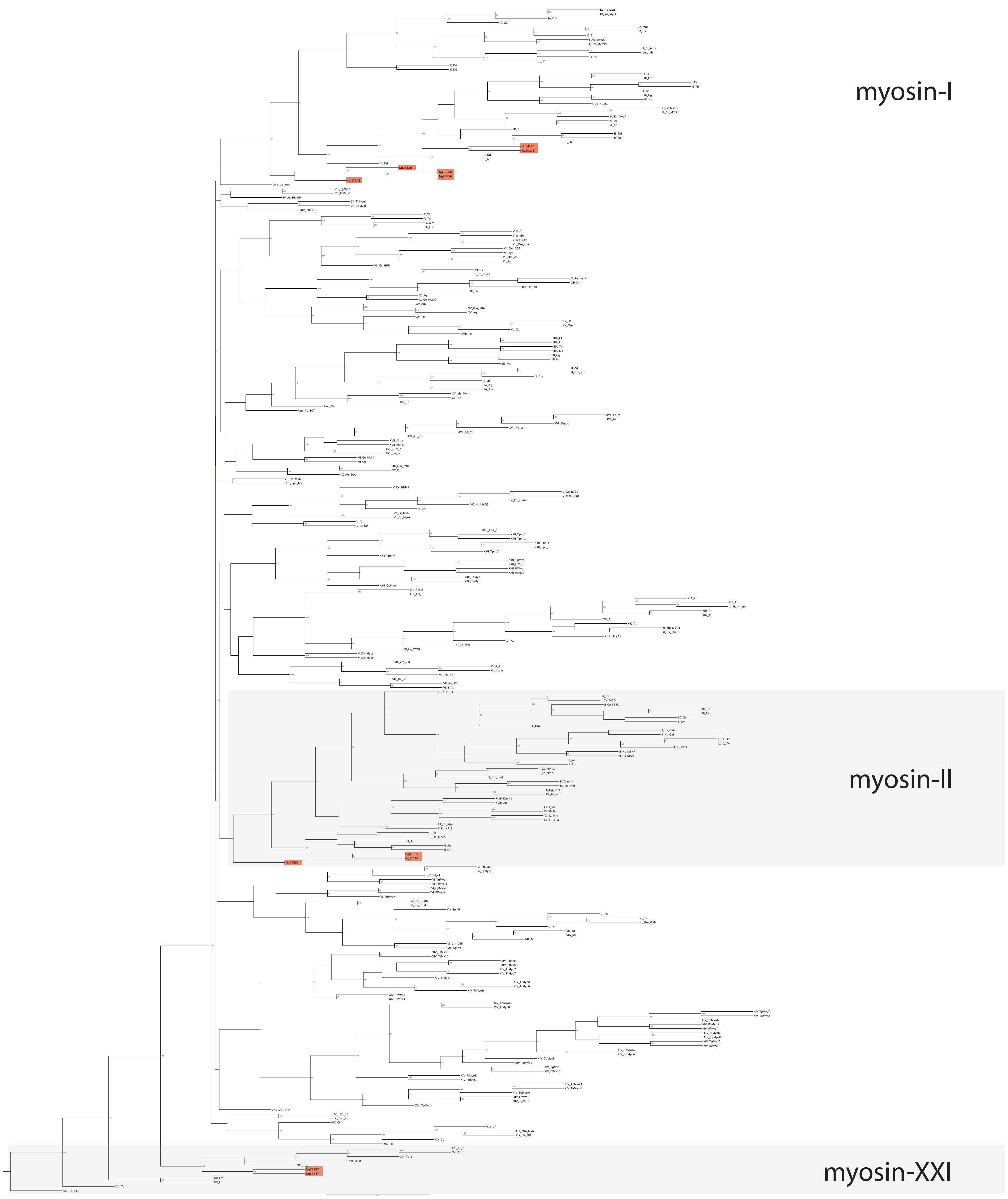


*Models with protein IDs 56065 and 39221 share identical protein sequence
**Models with protein IDs 56391 and 55423 share identical protein sequence

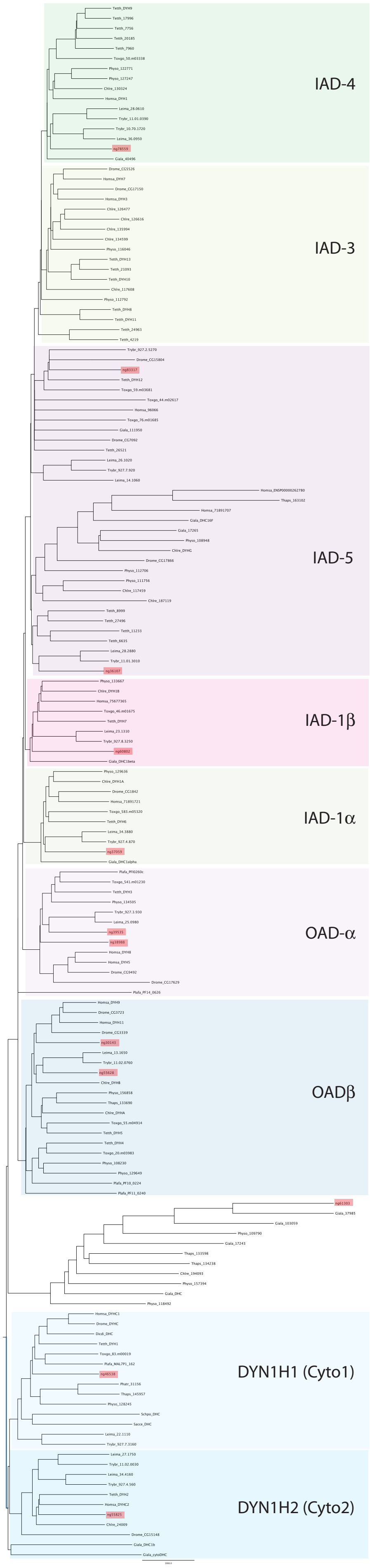
C Kinesin Phylogeny



D Myosin phylogeny

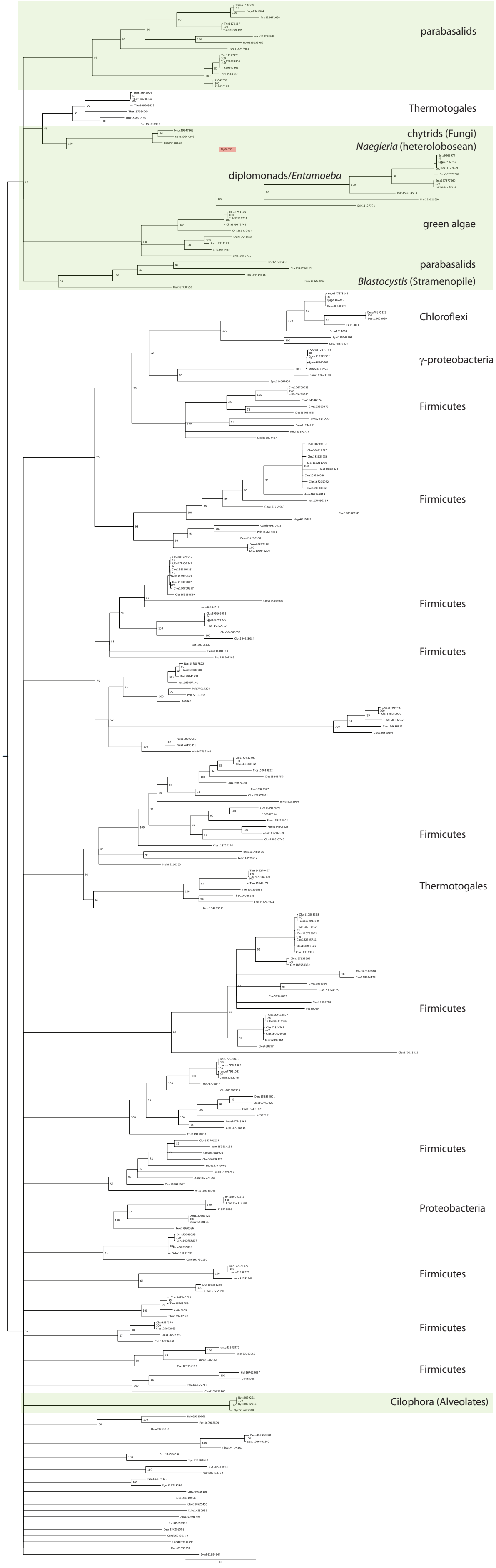


E Dynein phylogeny



FigS5, high resolution

Click here to download Supplemental Figure: FritzLaylin_FigS5.pdf


parabasalids
Thermotogales
chytrids (Fungi)
Naegleria (heterolobosean)
diplomonads/*Entamoeba*
green algae
parabasalids
Blastocystis (Stramenopile)
Chloroflexi
γ-proteobacteria
Firmicutes
Firmicutes
Firmicutes
Firmicutes
Firmicutes
Thermotogales
Firmicutes
Firmicutes
Proteobacteria
Firmicutes
Firmicutes
Firmicutes
Cilophora (Alveolates)
0.1