Cell

The Genome of *Naegleria gruberi* Illuminates Early Eukaryotic Versatility

Lillian K. Fritz-Laylin,[1,10] Simon E. Prochnik,[3,10] Michael L. Ginger,[4] Joel B. Dacks,[5,6] Meredith L. Carpenter,[1] Mark C. Field,[6] Alan Kuo,[3] Alex Paredez,[1] Jarrod Chapman,[3] Jonathan Pham,[7] Shengqiang Shu,[3] Rochak Neupane,[2] Michael Cipriano,[7] Joel Mancuso,[8] Hank Tu,[3,11] Asaf Salamov,[3] Erika Lindquist,[3] Harris Shapiro,[3] Susan Lucas,[3] Igor V. Grigoriev,[3] W. Zacheus Cande,[1] Chandler Fulton,[9] Daniel S. Rokhsar,[1,3,*] and Scott C. Dawson[7,*]

[1]Department of Molecular and Cell Biology
[2]Center for Integrative Genomics, 545 Life Sciences Addition
University of California, Berkeley, Berkeley, CA 94720, USA
[3]U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA
[4]School of Health and Medicine, Division of Biomedical and Life Sciences, Lancaster University, Lancaster LA1 4YQ, UK
[5]Department of Cell Biology, University of Alberta Edmonton, Alberta, Canada
[6]The Molteno Building, Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QT, UK
[7]Department of Microbiology, University of California, Davis, CA 95616, USA
[8]Gatan Inc., 5794 W. Las Positas Boulevard, Pleasanton, CA 94588, USA
[9]Department of Biology, Brandeis University, Waltham, MA 02454-9110, USA
[10]These authors contributed equally to this work
[11]Present address: Life Technologies, 850 Lincoln Center Drive, Foster City, CA 94404, USA
*Correspondence: dsrokhsar@gmail.com (D.S.R.), scdawson@ucdavis.edu (S.C.D.)
DOI 10.1016/j.cell.2010.01.032

SUMMARY

Genome sequences of diverse free-living protists are essential for understanding eukaryotic evolution and 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, analyzed in the context of other protists, reveals a remarkably complex ancestral eukaryote with a rich repertoire of cytoskeletal, sexual, signaling, 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 that 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 15 μm 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 (Figure 1) (Fulton, 1993). This change includes de novo assembly of an entire cytoplasmic microtubule cytoskeleton, including canonical basal bodies (Figure 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,

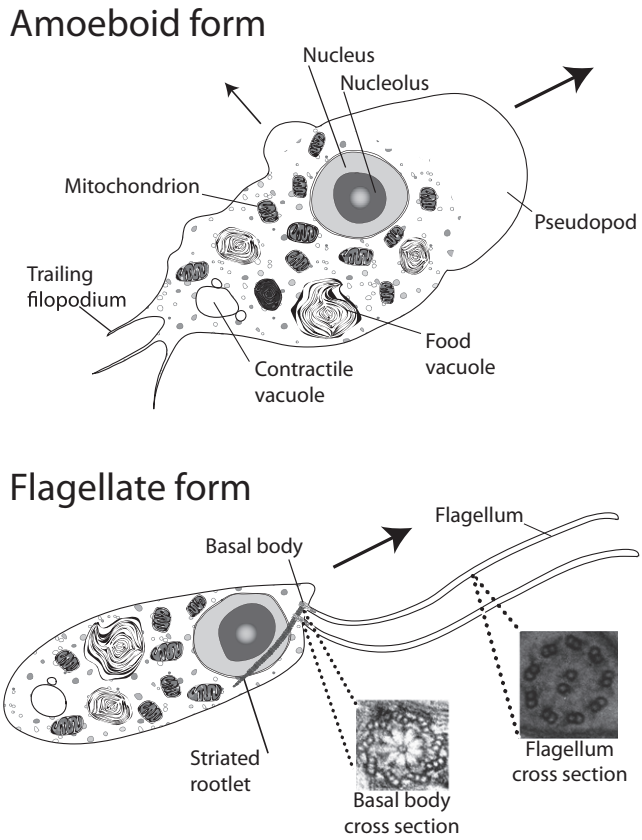


Figure 1. Schematic of *Naegleria* Amoeba and Flagellate Forms

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 that 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.

comprise the ancient and ecologically diverse clade termed "JEH" for Jakobids, Euglenozoa, Heterolobosea (Figure 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 (Figure 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 (Figure 2, Root A), by allowing analysis of free-living descendants of an early common ancestor (Figure 2, Root B), or by allowing analysis of free-living descendants of every major eukaryotic group via uniting JEH and POD (a monophyletic group encompassing Parabasalids, Oxymonads, and Diplomonads) into the Excavates (Figure 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; Figure 2) and perhaps in the ancestor of all eukaryotes. For example, *Naegleria* and humans (members of opisthokonts) diverged early (Figure 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 Figure 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 12 chromosomes (Figure S1A) and only 5.1% repetitive sequence (Extended Experimental Procedures and Table S1). The genome is a mosaic of heterozygous and homozygous regions (Figure S1A). Heterozygous

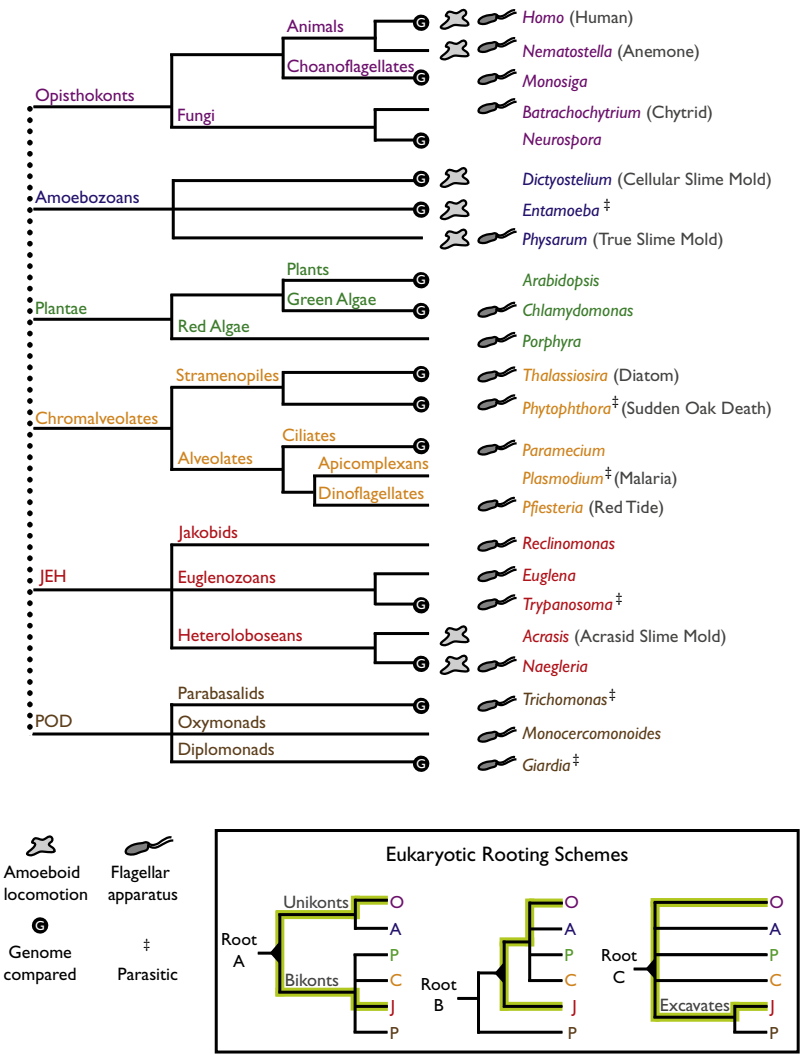


Figure 2. Consensus Cladogram of Selected Eukaryotes

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 Information). The branches connecting *Naegleria* to humans are highlighted in green, with a black triangle indicating their last common ancestor. See also Text S1.

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 (Figure 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 ~4000 copies of a sequenced extrachromosomal plasmid that encodes rDNA (Clark and Cross, 1987; Maruyama and Nozaki, 2007) and a 50 kb mitochondrial genome (GenBank accession number AF288092).

We predicted 15,727 protein-coding genes spanning 57.8% of the genome by combining ab initio and homology-based

Table 1. Genome Statistics from *Naegleria gruberi* and Selected Species

Species	Genome Size (Mbp)	No. Chromosomes	%GC	Protein-Coding Loci	% Coding	% Genes w/ Introns	Introns per Gene	Median Intron Length (bp)
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	9152	52.6	~0 (1 total)	ND	ND
Giardia	11.7	5	49	6480	71.4	~0 (4 total)	ND	ND

See also Figure S1 and Tables S1, S2, S3, S8, S9, and S10. ND, not determined.

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 5632 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 molybdopterin/thiamine biosynthesis family protein) (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 Figure S4), extensive meiotic, DNA replication, and transcriptional machinery (Tables S7, S8, S9, and 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. Although 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 that 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.

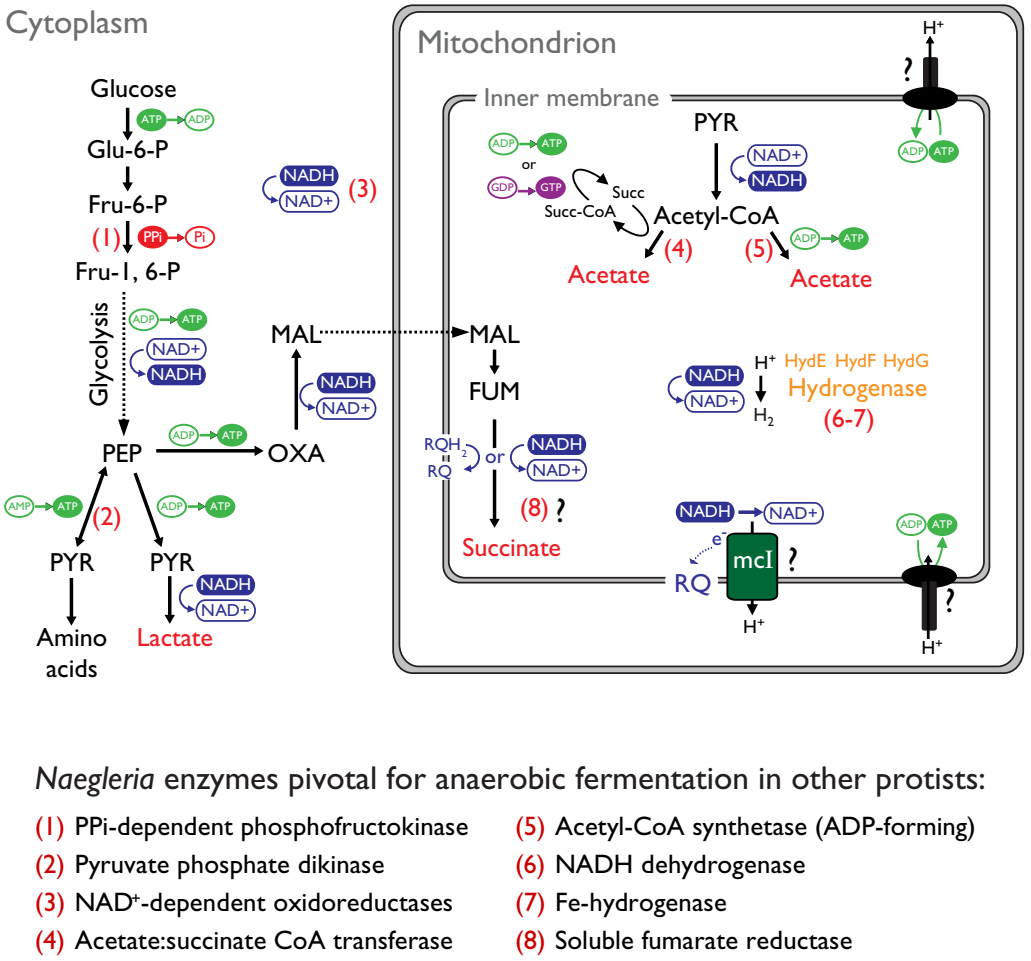
Metabolic Flexibility

Like many microbial eukaryotes, *Naegleria* oxidizes glucose, various amino acids, and fatty acids via the Krebs cycle and a branched mitochondrial respiratory chain using oxygen as a terminal electron acceptor (Figure S2; Table S14; Text S3). However, *Naegleria*'s genome also encodes features of an elaborate and sophisticated anaerobic metabolism (Figure 3; Figure S2; Text S3) including (1) 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); (2) an ability to use fumarate as an electron sink; and (3) genes encoding an Fe-hydrogenase and its associated maturation system. *Naegleria*'s anaerobic and aerobic metabolism parallels the recently discovered metabolic flexibility of another soil/pond dweller, the free-living alga *Chlamydomonas* (Figure 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 (Figure S2; Table S14) suggests that the organism is capable of oxidative phosphorylation. Many complex I subunits (NADH: ubiquinone oxidoreductase) are encoded by the mitochondrial genome (GenBank accession number NC_002573), but electrons can also be transferred to ubiquinone by two alternative NADH isoforms, succinate dehydrogenase (complex II) and electron transferring flavoprotein (Figure S2). Two terminal oxidases (cytochrome *c* oxidase and alternative oxidase) catalyze 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 that *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 repertoire of genes for both classic aerobic respiration and 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.,



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 3. A Model for Anaerobic Metabolism in *Naegleria*

Likely fermentation pathways used by *N. gruberi* under hypoxic or anoxic conditions are shown. Solid arrows indicate individual enzyme-catalyzed reactions, noting key nucleotide or coenzyme 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 (Figure S2). The HydE, HydF, and HydG Fe-hydrogenase maturation components (orange) are predicted to be mitochondrially targeted. Question marks indicate uncertainty regarding whether (lower center) 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 cosubstrate used by soluble fumarate reductase. See also Figure S2 and Figure S5, Tables S14, S15, and S16, and Text S3.

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 repertoire and an Fe-hydrogenase and maturation machinery) thus resides at the far end of this continuum of mitochondrial functions.

Although it is clear that mitochondrion-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 (Figure S5 and Hug et al., 2009). This suggests that 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 proto-mitochondrion (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 that cytosolic Fe-hydrogenases existed early in eukaryotic biology. Although lateral gene transfer is a likely source of some organellar iron hydrogenases (e.g., ciliate Fe-hydrogenases; Boxma et al., 2007), other organellar Fe-hydrogenases could have arisen via retargeting 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

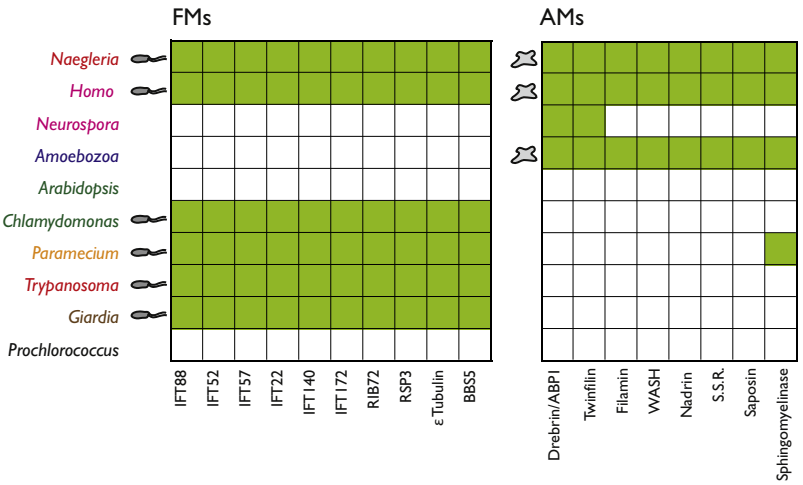


Figure 4. Phylogenetic Distribution of Selected Genes Associated with Ameboid Motility and Flagellar Motility

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 Figure S4 and Tables S5, S6, S17, S18.

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., 2006, 2007).

Conserved Amoeboid- and Flagellar-Motility Genes in the Eukaryotic Ancestor

Flagellar motility is found in every major eukaryotic group (Figure 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 eukaryotic root, Figure 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 (*f*lagellar-*m*otility-associated genes [FMs]) or amoeboid motility (*a*moeboid-*m*otility-associated genes [AMs]) (Figure 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) (Figure 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 that also play a role in nonmotile functions, such as actin, Arp2/3 (which nucleates actin filaments), or other general actin cytoskeletal components, as 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 protein (AM17), which may activate sphingomyelinase. (Sphingomyelinase is not in the AM set because it is found in the non-amoeboid *Paramecium*; Figure 4.) As sphingomyelin is enriched in the pseudopodia of human amoeboid cells (Jandak et al., 1990), we suggest that 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 (Figure 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 30 putative hybrid histidine kinases and 6 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*

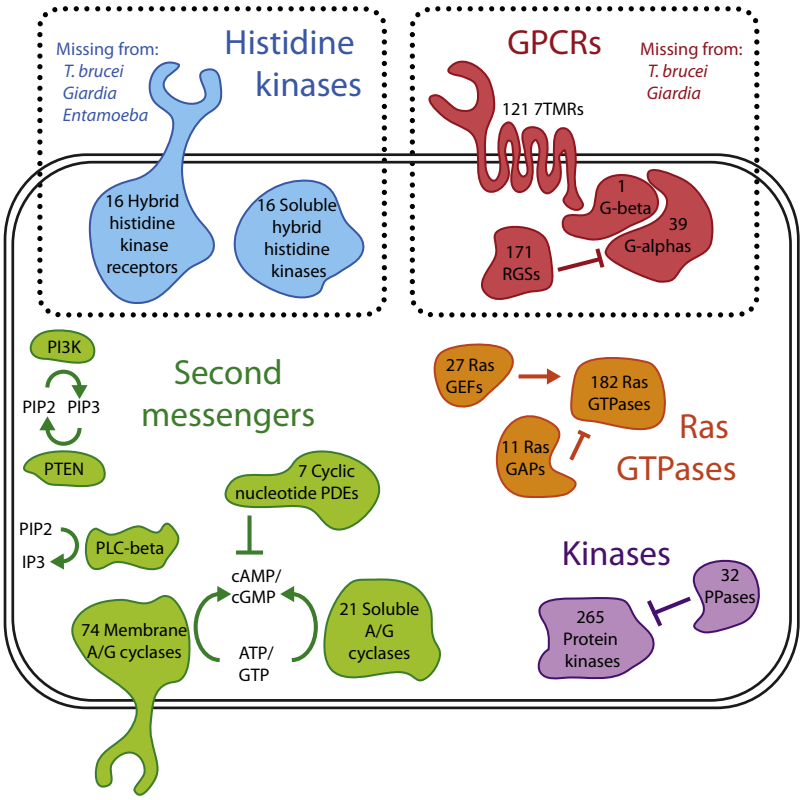


Figure 5. *Naegleria* Signaling Modules

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 homolog; PI3K phosphatidylinositol-3-OH kinase. See also Figure S3, Table S11, and Text S4.

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 (Figure 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 4133 ancient eukaryotic gene families, requiring (1) a minimum of one *Naegleria* protein and two orthologs and (2) 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 (Figure 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 1292 ancient eukaryotic gene families to the KOG analysis. Four hundred and eighty-one 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 4133 ancient gene families are present in at least three eukaryotic groups, and nearly half (1983) of the ancient gene families are present in all five major eukaryotic groups that include a genome sequence from a free-living species (Figure 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., 2005; 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 4133 ancient gene families, with substantially more losses observed in parasitic lineages (Table S20). Losses also likely occurred in the JEH lineage, as 1139 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 2424 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.

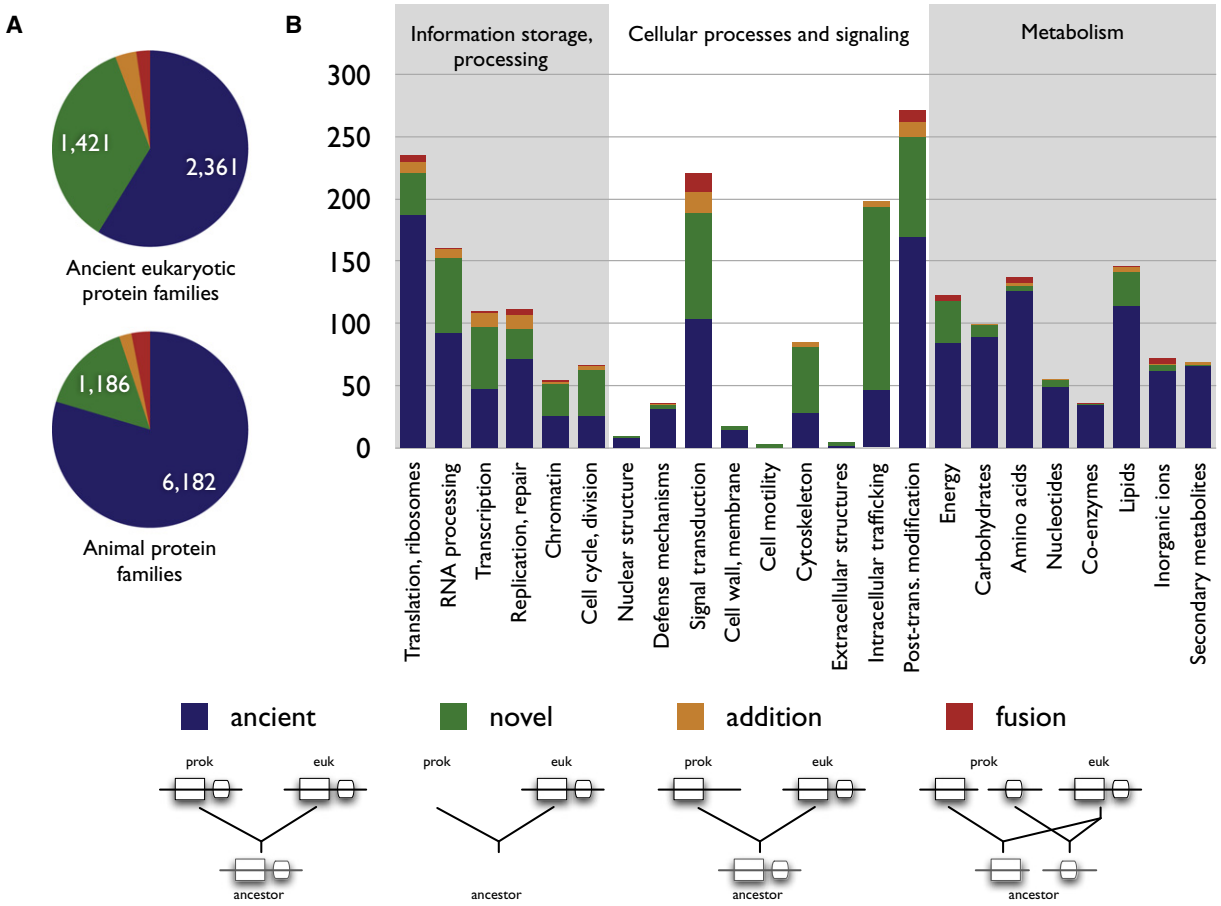


Figure 6. Ancient Origin and Innovation in Eukaryotic Proteins

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., archaeal and/or bacterial); euk, eukaryotic; Post-trans., post-translational. See also Tables S4, S19, S20, S21, S22, and S23.

Origin of Eukaryotic Genes

Which of these ancient gene families are shared with Archaea and/or Bacteria, and which are specific to eukaryotes? To investigate the origin of ancient eukaryotic gene families, we compared each of the 4133 families to prokaryotic (archaeal and bacterial) protein sequences. Approximately 57% (2361) 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"; Figure 6A). Conversely, 40% (1421) 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 1421 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, ubiquitin-based protein degradation, and, to a lesser extent, cytoskeletal and RNA-processing genes (Figure 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 (Figure 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 bacterial/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 an archaeal 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 (Figure 2), the genome of *Naegleria* reveals unexpectedly rich versatility in early eukaryotic ancestors, as 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 the Extended Experimental Procedures for further details of 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 comparisons 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 5107 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 4133 ancient eukaryotic gene families that contain a minimum of one *Naegleria* protein and two orthologs and also required at least one of the orthologs to 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.

Although 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 4133 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 4133 ancient eukaryotic protein families (see above) had been inherited from prokaryotes (i.e., from Archaea/Bacteria), were eukaryotic inventions, or were some combination of these two scenarios, we first constructed a "centroid" sequence for each 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 (Figure 6A, "novel"). We found 1421 such "novel" protein families.

In the following classification steps, we compared Pfam domain annotations in the eukaryotic centroid and prokaryotic sequences. We classify protein families as "ancient" if the centroid and the best hitting prokaryotic protein meet any of the following criteria: (1) neither sequence has a Pfam (Finn et al., 2008) domain; (2) the two sequences have the same combination of pairwise domains; (3) 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 2361 protein families. The remaining protein families showed some degree of innovation in eukaryotes relative to their prokaryotic homologs. The first class had no homologs in prokaryotic genomes (1421 "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 multidomain prokaryotic polypeptides into different proteins in eukaryotes, conceptually the opposite of the "fusion" category.

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

Generation of Flagellar-Motility-Associated Proteins

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 (*f*lagellar-*m*otility-associated proteins).

Generation of Amoeboid-Motility-Associated Proteins

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 organisms that have no amoeboid movement (*Prochlorococcus*, *Arabidopsis*, *Physcomitrella*, Diatom, *Paramecium*, Trypanosome, *Giardia*, *Chlamydomonas*) (Table S18).

ACCESSION NUMBERS

The genome assembly, predicted gene models, and annotations have been deposited at DDBJ/EMBL/GenBank under accession number ACER00000000.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, four Supplemental Texts, and 23 tables and can be found with this article online at doi:10.1016/j.cell.2010.01.032.

ACKNOWLEDGMENTS

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 M.L.G., a Tien Scholars Fellowship in Environmental Sciences and Biodiversity to L.K.F.-L., a Wellcome Trust and Parke-Davis Fellowship, as well as start-up funding from the U. of Alberta, to J.B.D., and funding by Wellcome Trust for M.C.F.

Received: August 28, 2009
Revised: November 17, 2009
Accepted: January 15, 2010
Published: March 4, 2010

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–D31.

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–S171.

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 Protozool. *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–D288.

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. USA *74*, 2157–2161.

Fulton, C. (1970). Amebo-flagellates as research partners: The laboratory biology of *Naegleria* and *Tetramitus*. Methods Cell Physiol. *4*, 341–346.

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. J. Eukaryot. Microbiol. *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. USA *106*, 3859–3864.

Hartman, H., and Fedorov, A. (2002). The origin of the eukaryotic cell: a genomic investigation. Proc. Natl. Acad. Sci. USA *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. *27*, 311–324.

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. J. Protozool. *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, CT: 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. USA *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.

Extended 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 a simplified axenic medium. 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 tends 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 2868 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 <1 kb long as well as redundant scaffolds, where redundancy was defined as those scaffolds shorter than 5 kb long with a greater than 80% identity to another scaffold greater than 5 kb 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 accession number 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 3 kb insert and BCCI, consisting of 245,457 reads with 8 kb 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 (Figure 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) (Figure 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, USA). 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′-GACTAGTTCTAGATCGCGAGCGGCCGCCCTTTTTTTTTTTTTT-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 kbp and >2 kbp) were cut out of the gel for the amoeba sample and one size range (0.6–2.0 kbp) 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 hr 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′- ATTTAGGTGACAC-TATAGAA and reverse (RV) 5′-TAATACGACTCACTATAGGG) and Big Dye chemistry (Applied Biosystems). 44,544 EST reads were attempted from the two 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, adaptor, 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 et al., 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 4873 consensus sequences (see above) and aligned to the scaffolds with BLAT (Kent, 2002). Gene models were predicted using the following methods: (1) ab initio (FGENESH [Salamov and Solovyev, 2000]); (2) 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 (3) 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 maximizes 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 kbp. The average protein length is 492 aa. We predicted that 3514 proteins (22%) possess a leader peptide, 3439 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: (1) to assess and if necessary correct, predicted gene structures; (2) to assign gene functions and report supporting evidence; and (3) 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, 4016 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. We classified these sequences into the following four categories where possible: (1) those with homology to known TEs in the RepeatMasker library or rRNAs using RepeatMasker (Smit et al., 1996-2004), (2) those that overlap gene models, or ESTs or are annotated as tRNAs with tRNAscan-SE (Lowe and Eddy, 1997) or (3) 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 (1) 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 (2) a *Naegleria* CLTG grouping with bacterial sequences, but forming a separate lineage outside known bacterial groups.

Protein Trafficking Proteins

We performed searches against the filtered model set of *Naegleria* proteins at the JGI portal (using the BLOSUM45 matrix). Typically we searched with known protein 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, http://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 "n-paralogs" (i.e., paralogs that had duplicated after speciation) (Remm et al., 2001). Second, 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 5115 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 5107 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 4133 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) (Figure 2) and *Arabidopsis* (Tatusov et al., 2003). A subset of 3285 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 4133 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), 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 4133 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 (Figure 6, "novel"). We found 1421 such "novel" protein families (Figure 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 classify protein families as "ancient" if the centroid and the best hitting prokaryotic protein meet any of the following criteria: (1) neither sequence has a Pfam (Finn et al., 2008) domain; (2) the two sequences have the same combination of pairwise domains; (3) 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 2361 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 (1421 "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 (Figure 6B).

Verification of Flagellar-Motility-Associated Proteins

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

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 $\geq$ 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 ligning 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: 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 maximum likelihood within CIPRES (http://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 (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/).

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

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.

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–i358.

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–W120.

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. Nucleic 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., 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.

Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinform. *23*, 2.3.1–2.3.22.

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–R259.

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.

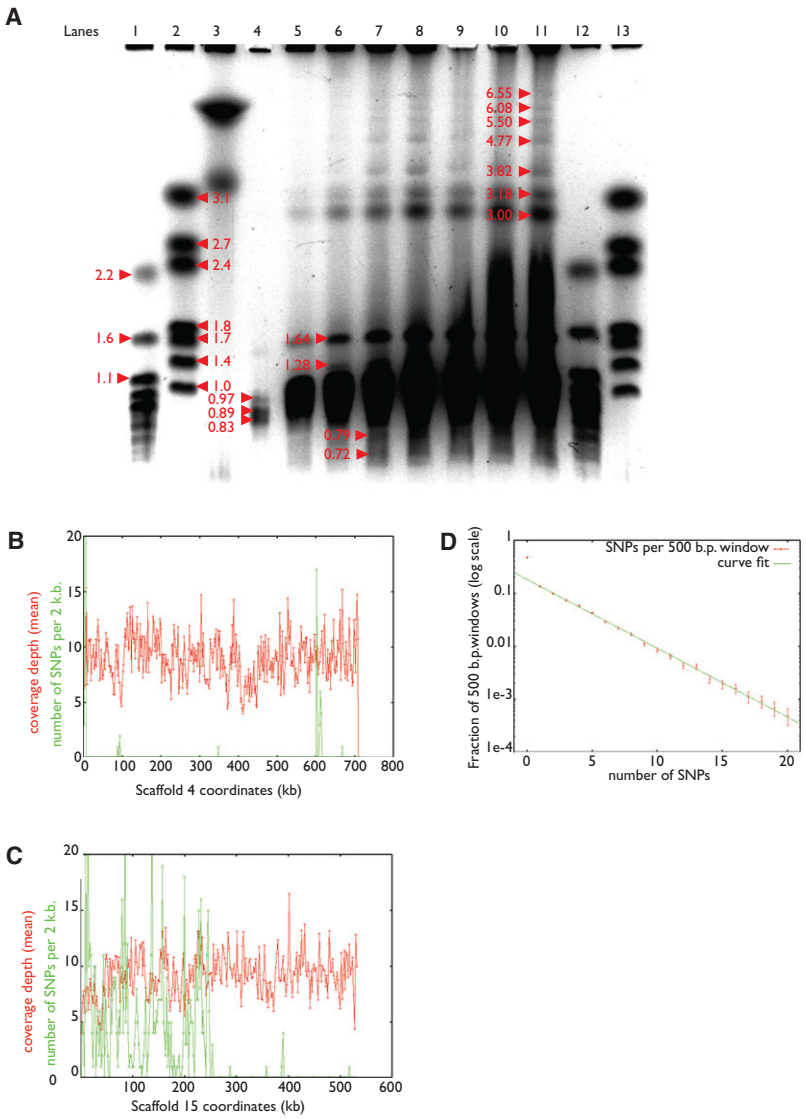


Figure S1. Electrophoretic Karyotype, Heterozygosity of *Naegleria gruberi*, Related to Table 1

(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 1, and *Hansenula wingei* in lane 2, 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 and 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 and 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.

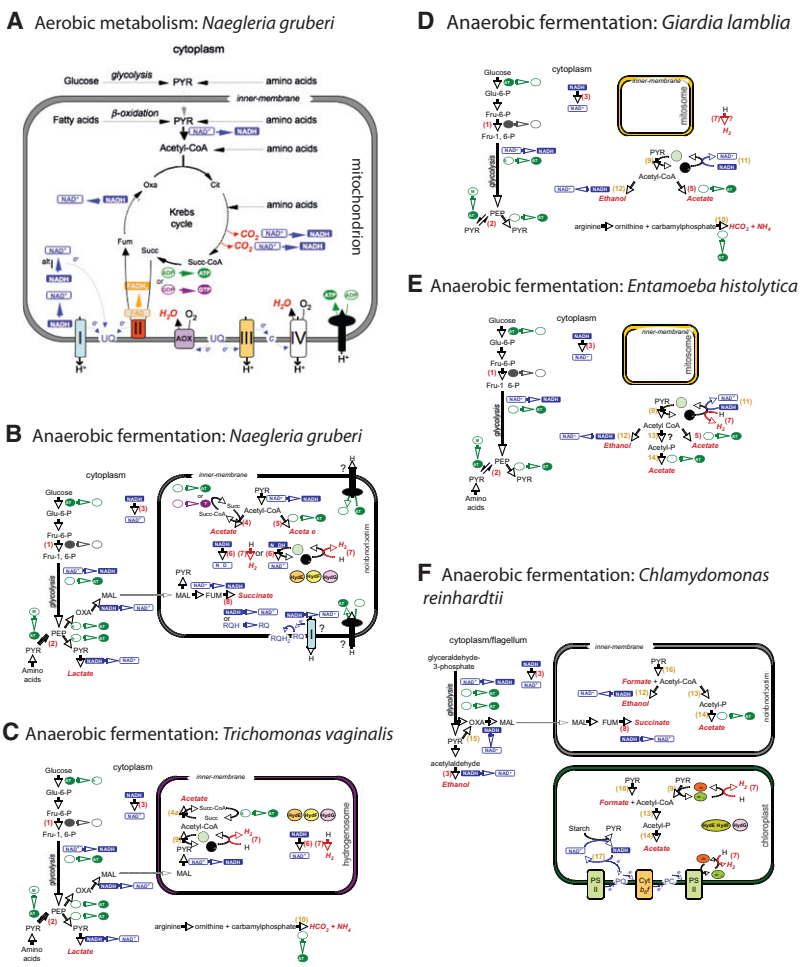


Figure S2. Predicted Canonical Aerobic Metabolism for *Naegleria gruberi* and Anaerobic Fermentation in *Naegleria gruberi* and Other Protists, Related to Figure 3

(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 triphosphates, 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 subcellular 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.

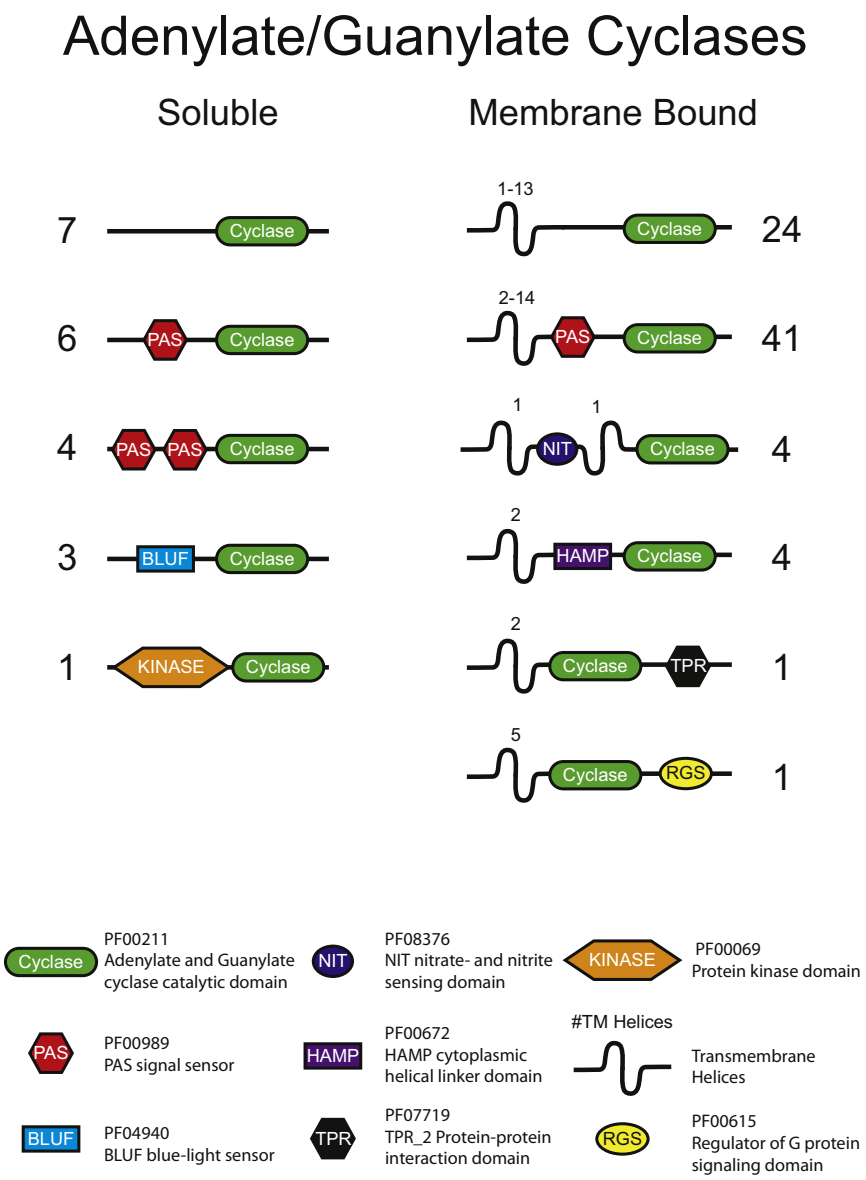


Figure S3. Cyclases in *Naegleria*, Related to Figure 5

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.