BROADLY SAMPLED TREE OF EUKARYOTIC LIFE

Broadly Sampled Multigene Analyses Yield a Well-resolved Eukaryotic Tree of Life

Laura Wegener Parfrey[1†], Jessica Grant[2†], Yonas I. Tekle[2,6], Erica Lasek-Nesselquist[3,4], Hilary G. Morrison[3], Mitchell L. Sogin[3], David J. Patterson[5], Laura A. Katz[1,2,*]

[1]Program in Organismic and Evolutionary Biology, University of Massachusetts, 611 North Pleasant Street, Amherst, Massachusetts 01003, USA

[2]Department of Biological Sciences, Smith College, 44 College Lane, Northampton, Massachusetts 01063, USA

[3]Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, Massachusetts 02543, USA

[4]Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman Street, Providence, Rhode Island 02912, USA

[5]Biodiversity Informatics Group, Marine Biological Laboratory, 7 MBL Street, Woods Hole, Massachusetts 02543, USA

[6]Current address: Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut 06520, USA

[†]These authors contributed equally

*Corresponding author: L.A.K - Lkatz@smith.edu Phone: 413-585-3825, Fax: 413-585-3786

Keywords: Microbial eukaryotes, supergroups, taxon sampling, Rhizaria, systematic error, Excavata

An accurate reconstruction of the eukaryotic tree of life is essential to identify the innovations underlying the diversity of microbial and macroscopic (e.g. plants and animals) eukaryotes. Previous work has divided eukaryotic diversity into a small number of high-level 'supergroups', many of which receive strong support in phylogenomic analyses. However, the abundance of data in phylogenomic analyses can lead to highly supported but incorrect relationships due to systematic phylogenetic error. Further, the paucity of major eukaryotic lineages (19 or fewer) included in these genomic studies may exaggerate systematic error and reduces power to evaluate hypotheses. Here, we use a taxon-rich strategy to assess eukaryotic relationships. We show that analyses emphasizing broad taxonomic sampling (up to 451 taxa representing 72 major lineages) combined with a moderate number of genes yield a well-resolved eukaryotic tree of life. The consistency across analyses with varying numbers of taxa (88-451) and levels of missing data (17-69%) supports the accuracy of the resulting topologies. The resulting stable topology emerges without the removal of rapidly evolving genes or taxa, a practice common to phylogenomic analyses. Several major groups are stable and strongly supported in these analyses (e.g. SAR, Rhizaria, Excavata), while the proposed supergroup 'Chromalveolata' is rejected. Further, extensive instability among photosynthetic lineages suggests the presence of systematic biases including endosymbiotic gene transfer from symbiont (nucleus or plastid) to host. Our analyses demonstrate that stable topologies of ancient evolutionary relationships can be achieved with broad taxonomic sampling and a moderate number of genes. Finally, taxon-rich analyses such as presented here provide a method for testing the accuracy of relationships that receive high bootstrap support in phylogenomic analyses and enable placement of the multitude of lineages that lack genome scale data.

Perspectives on the structure of the eukaryotic tree of life have shifted in the past decade as molecular analyses provide hypotheses for relationships among the ~75 robust lineages of eukaryotes. These lineages are defined by ultrastructural identities (Patterson, 1999) – patterns of cellular and subcellular organization revealed by electron microscopy – and are strongly supported in molecular analyses (Parfrey et al., 2006; Yoon et al., 2008). Most of these lineages now fall within a small number of higher-level clades, the supergroups of eukaryotes (Simpson and Roger 2004; Adl et al., 2005; Keeling et al., 2005). Several of these clades—Opisthokonta, Rhizaria, and Amoebozoa—are increasingly well-supported by phylogenomic (Rodríguez-Ezpeleta et al., 2007a; Burki et al., 2008; Hampl et al., 2009) and phylogenetic analyses (Parfrey et al., 2006; Pawlowski and Burki, 2009), while support for 'Archaeplastida' predominantly comes from some phylogenomic analyses (Rodríguez-Ezpeleta et al., 2005; Burki et al., 2007) or analyses of plastid genes (Yoon et al., 2002; Parfrey et al., 2006). In contrast, support for 'Chromalveolata' and Excavata is mixed, often dependent on the selection of taxa included in analyses (Rodríguez-Ezpeleta et al., 2005; Parfrey et al., 2006; Rodríguez-Ezpeleta et al., 2007a; Burki et al., 2008; Hampl et al., 2009). Moreover, it is difficult to evaluate the overall stability of major clades of eukaryotes because phylogenomic analyses have 19 or fewer of the major lineages and hence do not sufficiently sample eukaryotic diversity (Rodríguez-Ezpeleta et al., 2007a; Burki et al., 2008; Hampl et al., 2009), while taxon-rich analyses with four or fewer genes yield topologies with poor support at deep nodes (Cavalier-Smith, 2004; Parfrey et al., 2006; Yoon et al., 2008).

Estimating the relationships of the major lineages of eukaryotes is difficult because of both the ancient age of eukaryotes (1.2 to 1.8 billion years; Knoll et al., 2006) and complex gene histories that include heterogeneous rates of molecular evolution and complex patterns of paralogy (Maddison 1997; Gribaldo and Philippe 2002; Tekle et al., 2009). A further issue obscuring eukaryotic relationships is the chimeric nature of the eukaryotic genome—not all genes are vertically inherited due to lateral gene transfer (LGT) and endosymbiotic gene transfer (EGT) —that can also mislead efforts to reconstruct phylogenetic relationships (Andersson 2005; Rannala and Yang 2008; Tekle et al., 2009). This is especially true among photosynthetic lineages that comprise 'Chromalveolata' and 'Archaeplastida' where a large portion of the host genome (~8-18%) is derived from the plastid through EGT (Martin and Schnarrenberger 1997; Martin et al., 2002; Lane and Archibald 2008; Moustafa et al., 2009; Tekle et al., 2009).

There is a long-standing debate among systematists as to the relative benefits of increasing gene or taxon sampling (Hillis et al., 2003; Cummings and Meyer 2005; Rokas and Carroll 2005). Both approaches improve phylogenetic reconstruction by alleviating either stochastic or systematic phylogenetic error (e.g. Rokas and Carroll, 2005; Hedtke et al., 2006). Stochastic error results from too little signal in the data (e.g. single to few gene trees) to estimate relationships and results in poorly resolved trees with low bootstrap support, especially at deep levels (Swofford et al., 1996; Rokas and Carroll, 2005). The problems of stochastic error are amplified for deep relationships, such as relationships among major clades of eukaryotes (Roger and Hug, 2006). Many researchers opt to increase the number of genes, exemplified by phylogenomic studies, which alleviates stochastic error and yields well-resolved trees that are highly supported (Rokas and Carroll 2005; Burki et al., 2007; Hampl et al., 2009). However, analyses of many genes are still vulnerable to systematic error and often include very few lineages.

Systematic error results from biases in the data that mislead phylogenetic reconstruction, yielding incorrect sister group relationships that do not reflect historical relationships; the most well known of these is long-branch attraction (Felsenstein, 1978). Incongruence can also arise from conflicts between gene trees and species trees resulting from population genetic processes or the chimeric nature of eukaryotic genomes (Maddison, 1997; Rannala and Yang, 2008). Systematic errors can be detected and eliminated by several methods that are often combined, including using more realistic models of sequence evolution (e.g. Rodríguez-Ezpeleta et al., 2007), removing rapidly evolving genes and/or taxa that cause errors (Brinkmann et al., 2005), and by increasing taxonomic sampling (Zwickl and Hillis, 2002; Hedtke et al., 2006). Increased taxon sampling has been shown to improve phylogenetic accuracy even when the additional taxa contain large amounts of missing data (Philippe et al., 2004; Wiens 2005; Wiens and Moen 2008). In contrast, the abundance of data in phylogenomic studies can yield highly supported, but incorrect relationships caused by these systematic biases (Phillips et al., 2004; Hedtke et al., 2006; Jeffroy et al., 2006; Rokas and Chatzimanolis 2008). Taxon-rich analyses provide a method for testing the accuracy of relationships that receive high bootstrap support in phylogenomic analyses (Zwickl and Hillis, 2002; Heath et al., 2008).

Here we assess the eukaryotic tree of life by analyzing 16 genes from a broadly sampled dataset that includes 451 diverse taxa from 72 lineages. We aim to overcome both stochastic and systematic phylogenetic error by assessing two measures of clade robustness: 1) statistical support (bootstrap), and 2) the stability of clades across analyses with varying numbers of taxa and levels of missing data. We demonstrate that extensive taxon sampling coupled with selection of a modest number of well sampled genes counteracts systematic error and correctly places many rapidly evolving lineages without the removal of genes or taxa. Further, this approach enables us to place the numerous lineages that have only a few genes sequenced, and to assess support for the hypothesized clades of eukaryotes with a more inclusive sampling of diverse eukaryotes.

Methods

Gene Sequencing

Ovammina opaca and *Ammonia* sp. T7 were collected from a salt marsh on Cabretta Island, GA with assistance from Susan T. Goldstein (University of Georgia). DNA was isolated from 60 cells each that were individually picked, washed, and purged of food items overnight using a plant DNeasy kit (Qiagen). *Gromia* sp. Antarctica DNA was isolated from one cell undergoing gametogenesis and generously provided by Sam Bowser and Andrea Habura (Wadsworth Center). DNA for all other taxa was obtained from ATCC (Table S1; supplementary material is available online at http://www.sysbio.oxfordjournals.org) and accessions have been photodocumented (eutree.lifedesks.org). Small subunit ribosomal DNA (ssu-rDNA) was amplified with previously described primers (Medlin et al., 1988) and three additional primers were used to generate overlapping sequences from each clone (Snoeyenbos-West et al., 2002). *Hsp90* was amplified with CAC CTG ATG TCT YTN ATH ATH AAY and CTG GCG AGA NAN RTT NAR NGG, and reamplified with nested primers TCT CTG ATC ATC AAY RCN TTY TAY and AGA GAT GTT NAR NGG NAN RTC. Primers for actin, alpha-tubulin and beta-tubulin are from Tekle et al., 2008. Phusion DNA Polymerase (Finnzymes Inc), a strict proofreading enzyme, was used to amplify the genes of interest and Invitrogen Zero Blunt Topo cloning kits were used for cloning. Sequencing of cloned plasmid DNA was accomplished using vector- or gene-specific primers and the BigDye terminator kit (Applied Biosystems). Sequences were run on an ABI 3100 automated sequencer. We have fully sequenced 1-4 clones of each gene and surveyed up to 10 clones per taxon in order to detect paralogs. *Stephanopogon apogon* ssu-rDNA is extremely large and we were unable to amplify it using standard methods. Instead, we amplified three overlapping fragments that were then combined for use in our analyses. All new sequences, including any paralogs identified, have been deposited in GenBank (GQ377645-GQ377715 and HM244866-HM244878).

Cultures of microbial eukaryotes for EST sequencing were obtained from ATCC or the Culture Collection of Algae and Protozoa (CCAP; Table S1) and grown in Corning culture flasks according to supplier's recommended protocols. Cultures of *Heteromita* sp. were kindly provided by Linda Amaral Zettler and subsequently deposited at ATCC (ATCC PRA-74). Cultures were harvested and pooled as needed to obtain ~2x10^7 cells. Cells were pelleted and mRNA was extracted using the Qiagen Oligotex direct mRNA protocol. The resultant mRNA was quantitated by NanoDrop and/or Agilent Bioanalyzer RNA chip. Complementary DNA was generated using the ClonTech SMART cDNA construction protocol and ligated into the Lucigen pSMART vector (*Diplonema papillatum)* or the ClonTech pDNRlib vector (all others). Electrocompetent cells were transformed using the ligation products and plated on LB-kanamycin agar. Clones were grown in 96 well polypropylene 2.0 ml deep well growth blocks containing 1.2 ml superbroth (with 30ug/ml kanamycin) per well and plasmid DNA was prepared using a modified alkaline lysis procedure adapted for automation (GenomicSolutions RevPrep Orbit or Beckman BiomekFX). ~10,000 clones from each library were sequenced bidirectionally with vector primers using Sanger cycle sequencing (Applied Biosystems BigDye Terminator chemistry). Paired reads from the same clone were trimmed using custom Perl scripts and assembled based on sequence overlap using phrap (www.phrap.org). Clustering was done after assembly of paired reads, by TGICL (Pertea et al., 2003), and was used to group highly similar sequences that were extremely likely to be copies of the same gene. The size of a cluster thus reflects number of transcripts of a particular gene (gene copy number and expression level).

Dataset Assembly

Taxa and genes were selected to maximize taxonomic diversity and evenness given the availability of molecular data. This strategy was used to improve phylogenetic accuracy by breaking up long branches with dense sampling across the eukaryotic tree (Hillis, 1998). The classifications systems of Patterson 1999 and Adl et al., 2005 were used as guides as we aimed to sample eukaryotic diversity by including representatives of as many lineages defined by ultrastructural identities as possible (Table S2). These lineages have generally proven to be robust as they are well-supported in molecular analyses (e.g. Adl et al., 2005; Parfrey et al., 2006; Yoon et al., 2008), including the current study (Figs. 1-4 and Fig. S1), and they represent monophyletic groups that serve as a proxy for taxonomic diversity. Our dataset has representatives from 72 lineages, including 53 out of the 71 lineages plus 7 of 200 unplaced genera as defined in Patterson 1999. Additionally we include three unplaced lineages isolated more recently, *Malawimonas jakobiformis* (O'Kelly and Nerad, 1999), *Breviata anathema* (Walker et al., 2006), and American Type Culture Collection (ATCC) strain # 50646 (an isolate given the candidate name "*Soginia anisocystis*" that has yet to be described formally). We use an updated classification (Adl et al., 2005) to designate lineages in Amoebozoa and Rhizaria that belonged to the single unsupported clade ('Ramicristate') from Patterson 1999 (Table S2). In order to maximize taxon evenness along with breadth, we chose limited but diverse members from within lineages where possible (e.g. we included 15 phylogenetically distant animals).

To maximize gene sampling for diverse taxa, we include markers historically targeted by PCR-based analyses (e.g. ssu-rDNA, *actin*, *elongation factor 1α*; Table S3) plus commonly sequenced expressed sequence tags (ESTs; e.g. ribosomal proteins, *14-3-3*; Table S3). The comprehensively sampled ssu-rDNA and the historical markers facilitate inclusion of many additional taxa for which only these genes have been characterized (Table S4). The minimum sequence data required for inclusion was nearly full-length ssu-rDNA, which provided the core of information necessary for phylogenetic placement with large amounts of missing data (Wiens and Moen, 2008).

SSU-rDNA sequences were hand curated for target taxa by removing introns, unalignable regions, non-nuclear rDNAs, and misannotated sequences. This alignment was crucial to overall accuracy because nearly half of the target taxa are represented only by SSU, thus several alignment and masking methods were assessed to ensure the robustness of the ssu-rDNA alignment. SSU-rDNA sequences were aligned by HMMER (Eddy, 2001), version 2.1.4 with default settings, taking secondary structure into account. HMMER used a set of previously aligned sequences to model the secondary structure of a sequence. The training alignment for building the model, consisting of all available ssu-rDNA eukaryote sequences (as of December 2008) aligned according to their secondary structure, was downloaded from the European Ribosomal Database (Wuyts et al., 2002). An additional ssu-rDNA alignment was constructed in MAFFT 6 implemented in SeaView (Galtier et al., 1996) with the E-INS-i algorithm (Katoh and Toh, 2008). Both alignments were further edited manually in MacClade v4.05 (Maddison and Maddison, 2002). To assess the effect of rate heterogeneity on the ssu-rDNA topologies, we partitioned the data matrices into eight rate classes using the GTR model with invariable sites and rate variation among sites following a discrete gamma distribution, as implemented in HyPhy version .99b package (Kosakovsky Pond et al., 2005). We then ran analyses without the fastest and two fastest rate classes, resulting in 1019 characters. However, the reduced datasets resulted in less resolution in the backbone without improving apparent the long-branch attraction. Thus, we used the alignment generated in MAFFT and masked with GBlocks (Talavera and Castresana, 2007) and by eye in MacClade, resulting in 867 unambiguously aligned characters.

Analysis of protein coding genes relied on a custom-built pipeline and database that combined Perl and Python scripts to identify homologs from diverse eukaryotes. Our goal in developing this pipeline was to ensure that we captured the broadest possible set of sequences given the tremendous heterogeneity among microbial eukaryotes. All available protein and EST data from our target taxa (Table S4) were downloaded from GenBank in January 2009 and ESTs were analyzed in all six translated frames to identify correct sequences for our alignment. A fasta file of six sequences representing the six 'supergroups' was created for each target gene and used to query our database of target taxa by BLASTP. Results were limited by length, e-value and identity, and all sequences with greater than 1% divergence within each taxon were retained for assessment of paralogy. The resulting sequences were aligned with ClustalW (Thompson et al., 1994) and the resulting single gene alignments were assessed by eye to remove non-homologous sequences.

The inferred amino acid sequences for each of the 15 protein genes from our data pipeline were combined with the new sequences generated for this study and again aligned in Clustal W (Thompson et al., 1994). The alignment was adjusted by eye in MacClade (Maddison and Maddison, 2002). As these alignments included all paralogs extracted from the pipeline, individual gene trees were examined to choose appropriate orthologs. For example, in cases where paralogs formed a monophyletic group, the shortest branch sequence was retained. When paralogs fell into multiple locations on the tree, we aimed to maintain orthologous groups that included the greatest taxonomic representation. The individual gene alignments were then concatenated to build a 16 gene, 451-taxon matrix with 6578 unambiguously aligned characters, including ssu-rDNA. All other datasets were constructed by removing taxa and/or genes from this matrix. All data matrices are available at TreeBase (submission ID S10552).

Creation of Subdata Matrices

We created an array of data matrices by subsampling our full data matrix of 16 genes (15 protein-coding genes plus ssu-rDNA) and 451 taxa (denoted all:16) in order to assess the impact of taxon sampling, missing data, and gene sampling. First, seven datasets were created to assess the impact of missing data and taxon sampling (summarized in Table 1). The least inclusive of these contained 16 genes and all 88 taxa that had at least 10 of the 16 genes (10:16), which resulted in 17% missing data. Similarly, the 6:16 and 4:16 matrices include all taxa with at least six and four of the targeted 16 genes, respectively. SSU-rDNA is ubiquitously sampled in our dataset and many phylogenetic hypotheses are based on ssu-rDNA genealogies. To address the concern that ssu-rDNA was driving our results, we deleted it from each of the 16 gene datasets resulting in 9:15, 5:15, 3:15, and all:15 matrices.

To assess the relative importance of gene versus taxon sampling, we compared our full analysis to datasets with taxon sampling based a recent phylogenomic analysis (Hampl et al., 2009; Table S5, Hampl:16 gene) and phylogenetic analysis (Yoon et. al., 2008; Table S5, Yoon:

16 gene). We also analyzed a dataset of the four genes used by Yoon et al., 2008 (actin, alpha tubulin, beta tubulin, and ssu-rDNA) with our taxon sampling (Table S5, all: 4 gene). While a thorough test of the impact of gene sampling would require a large number of analyses of datasets with genes systematically deleted, we feel that this approach provides insight into the contributions of genes and taxa.

Photosynthetic lineages have chimeric genomes that are composed of genes originating both from the host eukaryote, the endosymbiotic plastid (through EGT), and, in cases of secondary or greater endosymbiosis, from the symbiont nucleus. If genes of multiple origins are retained in our concatenated dataset, the resulting conflicting signal between host, symbiont and plastid can mislead phylogenetic reconstruction. This chimerism may contribute to the instability observed for photosynthetic lineages without clear sister groups (red algae, green algae, glaucocystophytes, cryptomonads, and haptophytes). Thus, we used two methods to detect discordance among loci that could signal EGT. First, the 16 genes from representatives of each of these photosynthetic lineages were analyzed by top BLASTp hit. We scored the first two lineages hit, with red algae, green algae and plants, or glaucophytes taken as evidence for EGT. Nine genes showed some evidence of EGT, and these were removed to create non-EGT datasets (5:non-EGT and 3:non-EGT; Table S6). The second approach was to use Concaterpillar to identify protein-coding genes with discordant histories (Leigh et al, 2008), which could be caused by EGT or LGT. Repeated runs yielded difference results, indicating an absence of supported discordances. Nevertheless, we analyzed several gene sets identified by Concaterpillar as concordant, including 1) the largest set of concordant genes plus SSU (3: cater 7 gene; Table S6), 2) a 13-gene dataset that excluded the three genes that were not concordant with any others (5: cater 13 gene; Table S6). To target discordance caused by EGT, we ran

Concaterpillar on photosynthetic lineages alone and analyzed the largest concordant gene set (5: cater 9 gene; Table S6).

Phylogenetic Analyses

Genealogies for this study were constructed almost exclusively in RaxML. The MPI version of RaxML 7.0.4 with rapid bootstrapping was used (Stamatakis et al., 2008). The ssu-rDNA partition was analyzed with GTR+gamma as this was the best fitting model available in RAxML, according to MrModelTest (Nylander 2004). ProtTest (Abascal et al., 2005) was used to select the appropriate model of sequence evolution for the amino acid data using the 9:15 dataset. WAG was found to be the best fitting model for the concatenated data, but rtREV was the best for some of the individual partitions and both WAG and rtREV were among the top three models for all but one gene (and with similar likelihood scores). We ran our data under both WAG and rtREV models and found consistent results, indicating that our interpretations are robust to at least this level of model choice. The results presented are from the WAG analyses and the rtREV analyses differed only in level of bootstrap support for key nodes (usually ±5 points). In initial analyses, the appropriate number of independent bootstrap replicates was determined for each data set using bootstopping criteria in RAxML 7.0.4 as implemented on CIPres portal 2 (Miller et al., 2009). All analyses stopped after 200 or fewer replicates, except all:16, which stopped after 400 replicates. In later analyses, using the MPI version of RAxML, which does not implement a bootstopping criterion, 200 rapid bootstrap replicates followed by a full maximum likelihood search was used for all analyses except all:16, for which 600 bootstrap replicates were run. Because of the computational cost of the all:16 analysis, this was run as six separate analyses: one hundred bootstraps followed by a full ML search and five other runs of

100 bootstraps each. These data were combined in RAxML to complete the analysis. We found no significant difference in comparisons between fast and slow RAxML bootstrap methods (Fig S1i), which we tested because the fast bootstrapping method in RAxML can produce misleading results particularly for long-branch taxa (Leigh, 2008). The results of rapid bootstrapping are shown.

To investigate the stability of our tree topology under different analytic methods, select datasets were analyzed with Bayesian approaches and Parsimony (Fig. S1s-v). Parsimony analysis of 10:16, implemented in Paup* (Swofford, 2002), yielded a less resolved version of the RAxML topology (i.e. Excavata as a polytomy) that is generally concordant with the more resolved tree obtained by maximum likelihood methods. The one exception was the misplacement of some rapidly evolving lineages (including *Giardia*, Microsporidia, Foraminifera, and *Entamoeba*). PhyloBayes was run on the 9:15 dataset using the CAT model with recoded amino acids. The amino acids were recoded using the Dayhoff (6) model, based on the chemical properties of the amino acids. PhyloBayes was stopped after building two chains of > 13000 trees with a maxdiff of 0.26, which indicates weak convergence, but that the chains disagreed on at least one clade 26% of the time. A burn-in of 100 trees was removed and the posterior probabilities were calculated after sampling every other tree. The topology of the consensus tree is consistent with, though less well resolved than the results from RAxML. The parallel version of MrBayes 3.1.4 was used to analyze the 10:16 data matrix using the GTR+I+ γ (for nucleotide partition) and WAG (for amino acid partition) models of sequence evolution (Ronquist and Huelsenbeck, 2003). Six simultaneous MCMCMC chains were run for 5,600,000 generations, sampling every 1000 generations. An average standard deviation of split frequencies of <0.1 indicated weak convergence. Stationarity was determined by plotting the maximum likelihood values of the two runs, and 10,756 trees were retained. The resulting topology is the same as shown in Figure 2, except that *Breviata* nests within Amoebozoa sister to *Mastigamoeba* + *Entamoeba.* Most nodes are strongly supported: posterior probability equals 1.00 for Amoebozoa, Opisthokonta, Rhizaria, and SAR, and 0.66 for Excavata and 'Unikonta'.

Topology Testing

We performed the approximately unbiased (AU) test (Shimodaira, 2002) as well as the more conventional Kishono-Hasegawa (KH) and Shimodaira-Hasegawa (SH) tests, as implemented in Consel 0.1j (Shimodaira and Hasegawa, 2001) to test the monophyly of 'Chromalveolata', 'Archaeplastida' and 'Chromista'. The most likely trees with these groups constrained to be monophyletic were built, and the site likelihood values for each constrained topology and the unconstrained topology were estimated using RAxML 7.0.4 (Table S7). In addition, we explored in Paup* v4.08b (Swofford, 2002) the number of Bayesian trees that were consistent with these hypotheses (Table S7).

RESULTS AND DISCUSSION

Robust Topology of the Eukaryotic Tree of Life

Many major clades were consistently recovered across our analyses (Table 1). These stable groups receive moderate to strong support in analyses with limited missing data (Fig. 2) and less support as missing data increases. The Opisthokonta, which includes animals and fungi, and the heterogeneous clade Rhizaria are recovered in all analyses with strong support (Fig. 1 and Table 1). Excavata are recovered in all analyses with moderate support (Fig. 1 and Table 1). Amoebozoa receives low to moderate support in all but our most inclusive analysis (all:16) where Mastigamoebidae + *Entamoeba* form a separate clade with *Breviata* and Centroheliozoa (Fig. 1 and Table 1). Both Rhizaria and Amoebozoa are heterogeneous assemblages of organisms with diverse body plans (Pawlowski and Burki, 2009; Tekle et al., 2009) that were created based on molecular analyses (Parfrey et al., 2006). There are no defining morphological features or molecular signatures for Rhizaria, which now encompasses nearly 30 of the 75 lineages with ultrastructural identities (Pawlowski and Burki, 2009). Excavata was hypothesized in part on the basis of ultrastructural characters associated with the ventral feeding groove (Simpson, 2003), but is generally polyphyletic in phylogenetic (Parfrey et al., 2006; Simpson et al., 2006) and phylogenomic analyses unless rapidly evolving taxa and characters are removed from the analyses (Rodríguez-Ezpeleta et al., 2007a; Hampl et al., 2009). We also find strong support for the clade of stramenopiles, alveolates, plus Rhizaria (SAR; Burki et al., 2007; Hackett et al., 2007; Burki et al., 2008) and a sister relationship between stramenopiles and Rhizaria (Fig. 2 and Table 1). This latter finding is at odds with many phylogenomic analyses (Rodríguez-Ezpeleta et al., 2007a; Burki et al., 2008; Hampl et al., 2009) that find stramenopiles and alveolates are sister to one another.

In contrast, the relationships among photosynthetic lineages and the position of most orphan lineages (e.g. *Breviata* and Centroheliozoa) remain unresolved, as discussed below. Further, the root of the eukaryotic tree of life has been hypothesized to be between a clade containing Amoebozoa and Opisthokonta ('Unikonta') and all remaining eukaryotes (Stechmann and Cavalier-Smith, 2003), although there is conflict among evidence (reviewed in Roger and Simpson, 2009; Tekle et al., 2009). In our analyses, we find at best moderate support for 'Unikonta' (Table 1), but concatenated analyses such as these cannot resolve the root.

In exploring the tradeoffs between increasing taxonomic sampling and decreasing missing data, we analyzed varying combinations of genes and taxa using almost exclusively a maximum likelihood approach implemented in the software RAxML 7.0.4 (Stamatakis et al., 2005). Node support was highest when we included taxa with ten or more of our targeted 16 genes (10:16, with 17% missing data and 88 taxa: Fig. 2 and Table 1). As taxa are added, node support decreases (Table 1, bootstrap support in Fig. 1) due to the diminishing amount of character data available to estimate a growing number of relationships (i.e. 211 of 451 taxa are represented by ssu-rDNA only). Put another way, stochastic error increases with increasing missing data because the signal to noise ratio is decreasing. The mosaic structure of missing data in phylogenomic studies using ESTs is known to decrease phylogenetic accuracy (Hartmann and Vision, 2008). However, Wiens and Moen (2008) found that taxa with large amounts of missing data (up to 90%) could be accurately placed so long as there is a shared core of informative data. The ubiquitous ssu-rDNA plus a few well-sampled protein genes likely provide such a core of informativeness in this study.

In addition to allowing assessment of the phylogenetic diversity of eukaryotes, a strength of this taxon-rich analysis is that it enables us to assess clade stability by comparing tree topology across analyses that vary in numbers of taxa and genes included. Much of the topology remains consistent across all analyses: supported clades (Table 1) and most clades with ultrastructural identities (bold lineages Fig. 1; Table S2) are recovered regardless of the number of genes/level of missing data included. We argue that this is strong evidence that these clades are accurately reconstructed—they reflect true relationships. The ability to accurately place so many lineages that are represented only by ssu-rDNA demonstrates the robustness of these analyses.

We tested the hypothesis that ssu-rDNA was driving our results, as this gene is ubiquitously sampled but is not present in phylogenomic analyses. However, the 15-protein datasets yielded similar topologies that were again robust to varying taxonomic representation (Table 1). We also looked for supported incongruences among loci using Concaterpillar (Leigh et al., 2008) on the 15 protein-coding genes. Repeated runs yielded varying gene sets, suggesting there are no well-supported incongruences. Analyses of several of these gene sets yielded a topology consistent with that depicted in Figures 1 and 2, although support was low in analyses with few genes (Table S6). Here again, the placement of photosynthetic lineages was unstable, suggesting that they may be responsible for discordance among loci.

We also assessed the extent to which choice of these particular 16 genes versus the breadth of our taxon sampling impacted the generation of stable topologies by comparing to previously published studies. Using our 16 genes and a taxon set comparable to Hampl et al. 2009 that included only 48 taxa representing 19 lineages, we generated a highly supported tree similar to what we find using broader taxon sampling (Table S5). Indeed, with our 16 genes and this Hampl-like data set, we recover monophyletic Excavata with 82% BS while this clade is only monophyletic after removal of rapidly evolving lineages in the phylogenomic analysis (Hampl et al. 2009). In contrast, using the broader taxon set of Yoon et al. 2008 (101 taxa representing 26 lineages) generates a topology that is less well supported at many nodes, and Excavata is polyphyletic (Table S5). Finally, using all our taxa and the 4 genes from Yoon et al. generates poorly supported topologies (Table S5). Together, these analyses demonstrate that it is an interaction of gene choice and taxon sampling that yields well-resolved trees.

The ability of our taxon rich approach to place lineages known to be problematic for phylogenetic reconstruction into correct territories, including Microsporidia, *Giardia* and ciliates (e.g. Hirt et al. 1999; Zufall et al. 2006; Yoon et al. 2008; Hampl et al. 2009), is a testament to the role of sufficient gene and taxon sampling in accurately reconstructing relationships. Other analyses with fewer taxa and/or genes routinely remove rapidly evolving taxa and/or sites so that these clades "behave" (Hackett et al. 2007; Rodríguez-Ezpeleta et al. 2007a; Burki et al. 2008; Yoon et al. 2008; Hampl et al. 2009). However, removal of taxa weakens the credibility of the process and support for taxonomic hypotheses while also decreasing the power of interpretation of the resulting phylogenetic trees (Hillis, 1998).

Orphan Lineages

Our taxon-rich analyses enable inclusion of numerous unplaced lineages that have only limited molecular data. Some of these remain orphans (i.e. without clear sister taxa) including *Breviata*, Centroheliozoa, *Ancyromonas,* and *Micronuclearia*, as their position is unstable and support values are very low (Table S8). These taxa may be either independent lineages or that their sister taxa have yet to be sequenced. Consistent with other analyses, we find support for the sister relationships of Apusomonadida with Opisthokonta (Cavalier-Smith and Chao, 2003; 85-100%; Table S8), and the non-photosynthetic kathablepharids with cryptomonads (Okamoto and Inouye, 2005; 65-88%; Table S8). *Telonema* is consistently basal to green algae (including plants), albeit with low support (Table S8), which is in contrast to the hypothesis that this lineage is sister to cryptomonads (Shalchian-Tabrizi et al. 2006). Several unplaced lineages represented only by SSU are placed within robust groups, but often on long branches and with low support (Paramyxea, *Mikrocytos*; Table S8). We believe that their placement is artifactual, either due to long-branch attraction, or the lack of a sequenced sister lineage. In support of this hypothesis, these taxa also bounce around in analyses of ssu-rDNA alone with and without rapidly evolving sites (as described in methods).

Photosynthetic Lineages

Our analyses do not resolve the placement of many lineages with photosynthetic ancestry including the green algae, red algae (rhodophytes), glaucocystophytes, haptophytes and cryptomonads. Notably, there is no support in any analysis for ‘Archaeplastida’ (‘Plantae’) or ‘Chromalveolata’ (Tables 1 and S6) or the nested hypothesis ‘Chromista’ (stramenopiles, cryptomonads, and haptophytes). These hypothesized clades rest on the assertion that plastid acquisition is a rare event, happening once in the ‘Archaeplastida’ (primary acquisition of a cyanobacterium in the ancestor of red algae, green algae and glaucocystophtes; Cavalier-Smith, 1981) and once in ‘Chromalveolata’ (secondary acquisition of a red algal plastid in the ancestor of stramenopiles, alveolates, haptophytes, and cryptomonads; Cavalier-Smith, 1999) or ‘Chromista’ (Cavalier-Smith, 2004). We hypothesize that the lack of resolution among the photosynthetic lineages (cryptomonads, haptophytes, glaucocystophytes, rhodophytes and green algae) is due to conflicting signal following endosymbiotic gene transfer from plastid genomes or from the nuclei of secondary (or tertiary) eukaryotic endosymbionts (Martin and Schnarrenberger 1997; Lane and Archibald 2008; Tekle et al. 2009). We discuss this hypothesis and alternatives below.

Our analyses, like many others (Cavalier-Smith 2004; Parfrey et al. 2006; Rodríguez-Ezpeleta et al. 2007a; Yoon et al. 2008; Kim and Graham 2008; Hampl et al. 2009), find polyphyletic ‘Chromalveolata’ and thus falsify the chromalveolate hypothesis as it was originally proposed. Further, ‘Chromalveolata’ and the nested hypothesis ‘Chromista’ (stramenopiles, cryptomonads, and haptophytes) are rejected by the approximately unbiased test (p = .007 and p < 0.001 respectively) and other statistical methods, and this topology was not found among the 10,756 trees in Bayesian analyses (Table S7). A single endosymbiotic event at the base of the chromalveoate lineages necessitates that the descendant lineages be monophyletic, although not everyone agrees with this interpretation (Keeling, 2009). Instead, our analyses are consistent with alternative hypotheses that postulate multiple secondary endosymbioses of red algal plastids (Grzebyk et al. 2003; Howe et al. 2008; Bodyl et al. 2009) in the ancestors of former chromalveoate lineages such as stramenopiles, dinoflagellates, haptophytes and cryptomonads.

Recent findings indicate that plastid acquisition is not as rare as once assumed, challenging the central tenet that plastid acquisition is much more difficult than loss. Two independent primary endosymbioses that may be first steps toward organelles have been detailed in the testate amoeba *Paulinella chromatophora* (Nakayama and Ishida, 2009) and the diatom *Rhopalodia gibba* (Kneip et al. 2008). Numerous secondary endosymbiotic events are also known in lineages such as euglenids, chlorarachniophytes, and kathablepharids (Archibald, 2009) and there is evidence for tertiary endosymbiosis in diatoms (Moustafa et al. 2009) and dinoflagellates (Archibald, 2009). Thus, plastid acquisition is more common across the eukaryotic tree of life than previously believed. The possibility that plastid acquisition may have occurred multiple times will make a stable resolution of photosynthetic lineages difficult (Lane and Archibald 2008; Bodyl et al. 2009).

As the stramenopiles and alveolates (two putative members of the 'Chromalveolata') form a well-supported clade including Rhizaria (SAR), we suggest it is time to abandon the chromalveolate hypothesis. While some argue for expanding the chromalveolate concept to include Rhizaria and other heterotrophic assemblages of eukaryotes as descendants of an ancestor with a red algal symbiont (Hackett et al. 2007; Burki et al. 2009; Keeling 2009) we do not think this revision is warranted due to the large number of losses and replacement of plastids that would have to had occurred. Instead, multiple endosymbioses are a much more parsimonious scenario, and are consistent with the monophyly of former 'chromalveolate' lineages in analyses of plastid genes (Yoon et al. 2002; Bodyl 2005; Parfrey et al. 2006). Similarly, the mere handful of genes that are potentially of photosynthetic origin in heterotrophic lineages such as ciliates (16 genes from a total of 27,446 in the complete genome; Reyes-Prieto et al. 2008) or the basal dinoflagellate *Oxyrrhis marina* (8 genes from 9,876 ESTs; Slamovits and Keeling, 2008) are more consistent with the "you are what you eat" hypothesis (Doolittle, 1998) than the 'Chromalveolata' hypothesis.

A single primary plastid acquisition at the base of 'Archaeplastida' is the prevailing view (Gould et al. 2008; Archibald 2009; Keeling 2009). The 'Archaeplastida' hypothesis is supported by many shared features plastids and their integration into the host cell, including plastid protein import machinery, conserved gene order and metabolic pathways (Mcfadden 2001; Larkum et al. 2007; Gould et al. 2008). While analyses of few genes do not generally support 'Archaeplastida' (Parfrey et al. 2006; Kim and Graham, 2008), support is strong in some phylogenomic analyses (Rodríguez-Ezpeleta et al. 2005; Rodríguez-Ezpeleta et al. 2007a Burki et al. 2008, though see Hampl et al. 2009). It has been suggested that 100+ genes are necessary to recover 'Archaeplastida' with strong support (Rodríguez-Ezpeleta et al. 2005).

The 'Archaeplastida' hypothesis is not supported in our analyses (Table 1,2) or those of others (Parfrey et al. 2006; Kim and Graham 2008; Yoon et al. 2008; Hampl et al. 2009). Here, the 'Archaeplastida' lineages red algae, green algae, and glaucocystophytes are never monophyletic, but instead generally form a poorly supported cluster with the secondarily photosynthetic haptophytes and cryptomonads plus other non-photosynthetic lineages (Table 1 and Fig. 2). This lack of resolution is not simply a byproduct of our overall approach as the analyses yield relatively well-supported nodes for much of the rest of the tree (Table 1 and Figs. 1,2), and recover groups with ultrastructural identities with strong support, including photosynthetic lineages (e.g. green algae including land plants; Fig. 2). The confounding effects of EGT (from plastid or nucleus of secondary enodsymbiont) may explain the lack of resolution and failure to recover 'Archaeplastida'. Being aware of these issues, we attempted to identify conflicting signal and remove genes impacted by EGT both by inspection of individual genes using BLAST analyses and by assessing concordant datasets identified by Concaterpillar (Table S6 and Fig. S1m-r). These approaches failed to yield robust placement of the problematic photosynthetic lineages (Table S6). For example, we hypothesized that the secondarily photosynthetic haptophytes and cryptomonads were branching within 'Archaeplastida' due to EGT; however, 'Archaeplastida' remains polyphyletic in analyses without haptophytes and cryptomonads (Table S6). In contrast to the 'Archaeplastida', other lineages with photosynthetic ancestry are robustly placed in clades containing both photosynthetic and heterotrophic lineages (e.g. dinoflagellates within alveolates, diatoms within Stramenopiles, and euglenids as sister to kinetoplastids). This may reflect differential timing of endosymbiotic events as ancient events will be more difficult to reconstruct than recent secondary transfers because 1) more genes in the plastid were available for transfer early and 2) more time for subsequent confounding events will have elapsed.

Alternatively, non-monophyly of 'Archaeplastida' may be reflective of the true host histories if there were multiple endosymbiotic events in the ancestors of red algae, green algae and glaucocystophytes. Many scenarios are consistent with both the non-monophyly of

‘Archaeplastida’ and the similarities of the plastids of these lineages (Palmer 2003; Stiller 2003; Larkum et al. 2007). Two of these are 1) multiple primary endosymbioses of closely related cyanobacteria followed by a convergent path of plastid reduction plus extinction of intervening bacterial lineages and 2) a single primary endosymbiosis into one lineage followed by ancient secondary endosymbioses into the remaining ‘Archaeplastida’ lineages. Monophyly of plastids would also be expected if there were multiple acquisitions of related cyanobacteria (Larkum et al. 2007) followed by extinction of cyanobacterial lineages. Such scenarios, as well as a single primary acquisition, is also consistent with the well-supported monophyly of plastids with respect to cyanobacteria (Rodríguez-Ezpeleta et al. 2005; Parfrey et al. 2006) plus possibly the confounding data on the divergent Rubisco genes in red and green algae (Delwiche and Palmer, 1996). Further, the phylogenetic position of ‘Archaeplastida’ lineages may be difficult to resolve because their sister groups have not yet been sequenced, or are extinct. The unstable position of these lineages across our analyses mimics the patterns observed in orphan lineages (Table S8) in support of this hypothesis. Under these scenarios, phylogenomic analyses that recover ‘Archaeplastida’ may be picking up misleading EGT signal of genes independently transferred into the host nucleus of these three lineages.

We suspect that resolving relationships among photosynthetic groups will require more intensive taxon and more careful gene sampling to disentangle signals from host and symbiont genomes, coupled with the recognition that plastid genes may be derived from several sources (Larkum et al. 2007). These data, combined with methods that distinguish between conflicting phylogenetic signal (Ahmadinejad et al. 2007; Leigh et al. 2008) or gene-tree species-tree reconciliation (Wehe et al. 2008; Akerborg et al. 2009), are likely required to elucidate the history of photosynthetic lineages.

Relationships within the well sampled Rhizaria and Excavata

We subsampled the data set to estimate relationships within two diverse clades, Excavata and Rhizaria, for which we had large numbers of taxa. We analyzed a 97-taxon dataset of Rhizaria that included all lineages with previously published data plus additional multigene data for 12 taxa added for this study (Table S1). Three major clades are strongly supported, though the relationships among them are unresolved: 1) Cercozoa, 2) Foraminifera plus Polycystinea and Acantharea (formerly classified with Phaeodarea as radiolarians), and 3) the parasitic Haplosporidia and Plasmodiophorida with *Gromia* and vampyrellids (Fig. 3; Bass et al. 2009). We show that *Theratromyxa,* a nematode-eating soil amoeba, is related to vampyrellid amoebae (Fig. 3; 100% bootstrap support; BS), and together they are sister to the plant parasites plasmodiophorids (100% BS). The ssu-rDNA sequence for *Theratromyxa* is identical to an amoeba isolated from Siberia where it was misidentified as *Arachnula impatiens* (EU567294; Bass et al. 2009).

The topology within the Excavata is consistent with previous hypotheses and clades with ultrastructural identities (Simpson, 2003, Fig. 4), when contaminant EST data originally mislabeled as *Streblomastix strix* is excluded (Slamovits and Keeling, 2006). Excavata is often polyphyletic in other analyses because *Malawimonas* branches outside the other clades of Excavata (Rodríguez-Ezpeleta et al. 2007a; Hampl et al. 2009), while in analyses of fewer genes Excavata members fall into two or three clades (Parfrey et al. 2006; Simpson et al. 2006). While *Malawimonas* nests robustly within Excavata in our analyses, it does not have a stable sister group and may represent an independent lineage (Fig. 4). Our analyses confirm that *Stephanopogon* (unplaced in Patterson 1999) branches within Heterolobosea (Cavalier-Smith and Nikolaev 2008; Yubuki and Leander 2008) and suggests that another enigmatic flagellate, ATCC 50646 ("*Soginia anisocystis*") is a basal member of Heterolobosea.

Conclusions

The robust tree of life emerging from this study demonstrates the benefits of improved taxon sampling for reconstructing deep phylogeny as our analyses produce stable topologies that include a broad representation of eukaryotes. The current study, combined with insights from other studies referenced herein, has refined the eukaryotic tree of life from over 70 major lineages (Patterson, 1999) to ~16 major groups (Fig. 5, eutree.lifedesks.org). Most significantly, we attribute the stability of major clades (e.g. Excavata, Amoebozoa, Opisthokonta, and SAR) to broader taxonomic sampling combined with analyses of sufficient characters (16 genes, or 6578 characters). In our view, inclusion of more taxa coupled with carefully chosen genes is necessary to further resolve the 16 or so major lineages of microbial eukaryotes for which sister group relationships remain uncertain.

Supplementary Material

Supplementary material can be found at: http://www.sysbio.oxfordjournals.org.

Funding

This work was made possible by the US National Science Foundation Assembling the Tree of Life grant to L.A.K. and D.J.P. (043115) and US National Institutes of Health

5R01AI058054-05 to M.L.S. Funding to collect Foraminifera was provided by a Society of Systematic Biologists MiniPEET grant to L.W.P.

ACKNOWLEDGEMENTS

We are grateful to Robert Molestina at ATCC who provided DNAs through a collaborative NSF grant. We acknowledge the assistance of Kasia Hammar, Leslie Murphy and Jillian Ward in preparation and sequencing of EST libraries. Our manuscript was improved following detailed comments from the editors, Alastair Simpson, and one anonymous reviewer. Thanks to David Hillis for conversations on early versions of the manuscript. Many thanks also to Daniel JG Lahr for comments and discussions, and to Wayne Pfeiffer and Mark Miller at CIPRES plus Tony Caldanaro at Smith College for technical help running analyses.

REFERENCES

Adl, S. M., A. G. B. Simpson, M. A. Farmer, R. A. Andersen, O. R. Anderson, J. R. Barta, S. S. Bowser, G. Brugerolle, R. A. Fensome, S. Fredericq, T. Y. James, S. Karpov, P. Kugrens, J. Krug, C. E. Lane, L. A. Lewis, J. Lodge, D. H. Lynn, D. G. Mann, R. M. McCourt, L. Mendoza, O. Moestrup, S. E. Mozley-Standridge, T. A. Nerad, C. A. Shearer, A. V. Smirnov, F. W. Spiegel, and M. Taylor. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J. Euk. Microbiol. 52:399-451.

Ahmadinejad, N., T. Dagan, and W. Martin. 2007. Genome history in the symbiotic hybrid *Euglena gracilis*. Gene 402:35-39.

Akerborg, O., B. Sennblad, L. Arvestad, and J. Lagergren. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci, USA. 106:5714-5719.

Archibald, J. M. 2009. The puzzle of plastid evolution. Curr Biol 19:R81-R88.

Bass, D., E. E. Y. Chao, S. Nikolaev, A. Yabuki, K. I. Ishida, C. Berney, U. Pakzad, C. Wylezich, and T. Cavalier-Smith. 2009. Phylogeny of novel naked filose and reticulose Cercozoa: Granofilosea cl. n. and Proteomyxidea revised. Protist 160:75-109.

Brinkmann, H., M. Van der Giezen, Y. Zhou, G. P. De Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743-757.

Cavalier-Smith, T. 1981. Eukaryote kingdoms - seven or nine. Biosystems 14:461-481.

Cavalier-Smith, T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Euk. Microbiol. 46:347-366.

Cavalier-Smith, T. 2004. Only six kingdoms of life. Proc. R. Soc. B Biol. Sci. 271:1251-1262.

Cavalier-Smith, T., and E. E. Y. Chao. 2003. Phylogeny of Choanozoa, Apusozoa, and other Protozoa and early eukaryote megaevolution. J. Mol. Evol. 56:540-563.

Delwiche, C. F., and J. D. Palmer. 1996. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. Mol. Biol. Evol. 13:873-882.

Doolittle, W. F. 1998. You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 14:307-311.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401-410.

Galtier, N., M. Gouy, and C. Gautier. 1996. Seaview and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. 12:543-548.

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

Hartmann, S., and T. J. Vision. 2008. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evol. Biol. 8:95.

Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239-257.

Hedtke, S. M., T. M. Townsend, and D. M. Hillis. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55:522-529.

Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47:3-8.

Katoh, K., and H. Toh. 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics 9:286-298.

Keeling, P. J. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. J. Euk. Microbiol. 56:1-8.

Kim, E., and L. E. Graham. 2008. EEF2 analysis challenges the monophyly of Archaeplastida and Chromalveolata. PLoS ONE 3:e2621.

Kneip, C., C. Voss, P. J. Lockhart, and U. G. Maier. 2008. The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. BMC Evol. Biol. 8:30.

Knoll, A. H., E. J. Javaux, D. Hewitt, and P. Cohen. 2006. Eukaryotic organisms in proterozoic oceans. Philos. Trans. R. Soc. B Biol. Sci. 361:1023-1038.

Larkum, A. W. D., P. J. Lockhart, and C. J. Howe. 2007. Shopping for plastids. Trends Plant Sci. 12:189-195.

Leigh, J. W. 2008. Congruence in phylogenomic data: Exploring artifacts in deep eukaryotic phylogeny *in* Biochemistry Dalhousie, Halifax, NS.

Leigh, J. W., E. Susko, M. Baumgartner, and A. J. Roger. 2008. Testing congruence in phylogenomic analysis. Syst. Biol. 57:104-115.

Maddison, D. R., and W. P. Maddison. 2002. MacClade 4.05. Sinauer Assoc.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. 46:523-536.

Medlin, L., H. J. Elwood, S. Stickel, and M. L. Sogin. 1988. The characterization of enzymatically amplified eukaryotes 16s-like ribosomal RNA coding regions. Gene 71:491-500.

Miller, M. A., M. T. Holder, R. Vos, P. E. Midford, T. Liebowitz, L. Chan, P. Hoover, and T. Warnow. 2009. The CIPRES portals.

Moustafa, A., B. Beszteri, U. G. Maier, C. Bowler, K. Valentin, and D. Bhattacharya. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science 324:1724-1726.

Nakayama, T., and K. Ishida. 2009. Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. Curr. Biol. 19:R284-R285.

Nylander, J.A. 2004. MrModelTest. Uppsala. Distributed by the author. Evolutionary Biology Centre, Uppsala University.

O'Kelly, C. J., and T. A. Nerad. 1999. *Malawimonas jakobiformis* n. Gen., n. Sp (Malawimonadidae n. Fam.): A *Jakoba*-like heterotrophic nanoflagellate with discoidal mitochondrial cristae. J. Euk. Microbiol. 46:522-531.

Okamoto, N., and I. Inouye. 2005. The katablepharids are a distant sister group of the Cryptophyta: A proposal for Katablepharidophyta divisio nova/Kathablepharida phylum novum based on ssu-rDNA and beta-tubulin phylogeny. Protist 156:163-179.

Parfrey, L. W., E. Barbero, E. Lasser, M. Dunthorn, D. Bhattacharya, D. J. Patterson, and L. A. Katz. 2006. Evaluating support for the current classification of eukaryotic diversity. PLoS Genetics 2:2062-2073.

Patterson, D. J. 1999. The diversity of eukaryotes. Am. Nat. 154:S96-S124.

Pawlowski, J., and F. Burki. 2009. Untangling the phylogeny of amoeboid protists. J. Euk. Microbiol. 56:16-25.

Pertea, G., X. Q. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. 2003. TIGR gene indices clustering tools (tgicl): A software system for fast clustering of large EST datasets. Bioinformatics 19:651-652.

Rannala, B., and Z. H. Yang. 2008. Phylogenetic inference using whole genomes. Annu. Rev. Genomics Hum. Genet. 9:217-231.

Reyes-Prieto, A., A. Moustafa, and D. Bhattacharya. 2008. Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. Curr. Biol. 18:956-962.

Rodríguez-Ezpeleta, N., H. Brinkmann, S. C. Burey, B. Roure, G. Burger, W. Loffelhardt, H. J. Bohnert, H. Philippe, and B. F. Lang. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. Curr. Biol. 15:1325-1330.

Rodríguez-Ezpeleta, N., H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang, and H. Philippe. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389-399.

Roger, A. J., and L. A. Hug. 2006. The origin and diversification of eukaryotes: Problems with molecular phylogenetics and molecular clock estimation. Phil. Trans. R. Soc. B Biol. Sci. 361:1039-1054.

Roger, A. J., and A. G. B. Simpson. 2009. Evolution: Revisiting the root of the eukaryote tree. Curr. Biol. 19:R165-R167.

Rokas, A., and S. B. Carroll. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337-1344.

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

Shalchian-Tabrizi, K., W. Eikrem, D. Klaveness, D. Vaulot, M. A. Minge, F. Le Gall, K. Romari, J. Throndsen, A. Botnen, R. Massana, H. A. Thomsen, and K. S. Jakobsen. 2006. Telonemia, a new protist phylum with affinity to chromist lineages. Proc. R. Soc. B Biol. Sci. 273:1833-1842.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492-508.

Shimodaira, H., and M. Hasegawa. 2001. Consel: For assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246-1247.

Simpson, A. G. B. 2003. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon excavata (eukaryota). Int. J. Syst. Evol. Micr. 53:1759-1777.

Simpson, A. G. B., Y. Inagaki, and A. J. Roger. 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. Mol. Biol. Evol. 23:615-625.

Slamovits, C. H., and P. J. Keeling. 2006. A high density of ancient spliceosomal introns in oxymonad excavates. BMC Evol. Biol. 6:8.

Slamovits, C. H., and P. J. Keeling. 2008. Plastid-derived genes in the nonphotosynthetic alveolate *oxyrrhis marina*. Mol. Biol. Evol. 25:1297-1306.

Snoeyenbos-West, O. L. O., T. Salcedo, G. B. McManus, and L. A. Katz. 2002. Insights into the diversity of choreotrich and oligotrich ciliates (class: Spirotrichea) based on genealogical analyses of multiple loci. Int. J. Syst. Evol. Micr. 52:1901-1913.

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the RAxML web-servers. Syst. Biol. 57(5): 758-771.

Stamatakis, A., T. Ludwig, and H. Meier. 2005. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456-463.

Stechmann, A., and T. Cavalier-Smith. 2003. The root of the eukaryote tree pinpointed. Curr. Biol. 13:R665-R666.

Swofford, D. 2002. Paup*. Phylogenetic analysis using parsimony (*and other methods). version 4.0b8. Sinauer Assoc.

Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. 407-514 *in* Molecular systematics (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Assoc., Sunderland MA.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564-577.

Tekle, Y. I., L. W. Parfrey, and L. A. Katz. 2009. Molecular data are transforming hypotheses on the origin and diversification of eukaryotes. Bioscience 59:471-481.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Walker, G., J. B. Dacks, and T. M. Embley. 2006. Ultrastructural description of *Breviata anathema*, n. Gen., n. Sp., the organism previously studied as "*Mastigamoeba invertens*". J. Euk. Microbiol. 53:65-78.

Wehe, A., M. S. Bansal, J. G. Burleigh, and O. Eulenstein. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics 24:1540-1541.

Wiens, J. J., and D. S. Moen. 2008. Missing data and the accuracy of Bayesian phylogenetics. J Syst. Evol. 46:307-314.

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

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51:588-598.

TABLE 1. Support for major clades of eukaryotes in analyses containing varying levels of taxon inclusion and missing data.

Supported clades	10: 16	6: 16	4: 16	all: 16	9: 15	5: 15	3: 15	all: 15
Opisthokonta	99	97	97	69	100	99	85	19
Rhizaria	100	99	94	82	100	100	47	29
SAR	97	98	63	22	100	100	32	19
Rhizaria + stramenopiles	94	94	57	26	92	96	29	18
Excavata	83	77	65	6	84	76	44	19
Amoebozoa	59	46	49	nm	68	56	44	5
‘Unikonta'	63	39	21	nm	54	50	15	3
Weak/unsupported hypotheses								
‘Archaeplastida’	nm	nm	nm	nm	nm	nm	nm	nm
‘Chromalveloata'	nm	nm	nm	nm	nm	nm	nm	nm
Cryptomonads + haptophytes	33	50	nm	29	38	56	22	25
Haptophytes + SAR	nm	nm	15	nm	nm	nm	nm	nm
Alveolates + stramenopiles	nm	nm	nm	nm	nm	nm	nm	nm
Red algae + green algae	nm	nm	nm	nm	nm	nm	nm	nm
Red, Green, Glauco, Hapto, Crypt	47	32	nm	9	39	27	16	8
Dataset statistics								
Number of taxa	88	111	160	451	88	111	160	240
Number of lineages	26	30	45	72	26	30	45	54
% Missing data (characters)	17	25	38	69	19	28	43	59

Supported clades are stable across analyses, albeit with decreasing support as the percentage of missing data increases. Varying bootstrap support values are represented with warmer colors corresponding to higher levels of support in RAxML analyses. nm= not monophyletic. Column headings describe the data sets. For example, ‘10: 16’ includes all taxa that have at least ten of the 16 genes, with a total of 88 taxa representing 26 lineages and containing 17% missing data. The ‘all: 15’ includes the protein coding genes from all taxa and contains 59% missing data. See Table S2 for lineages and Figure S1a-h for individual trees.

FIGURE LEGENDS

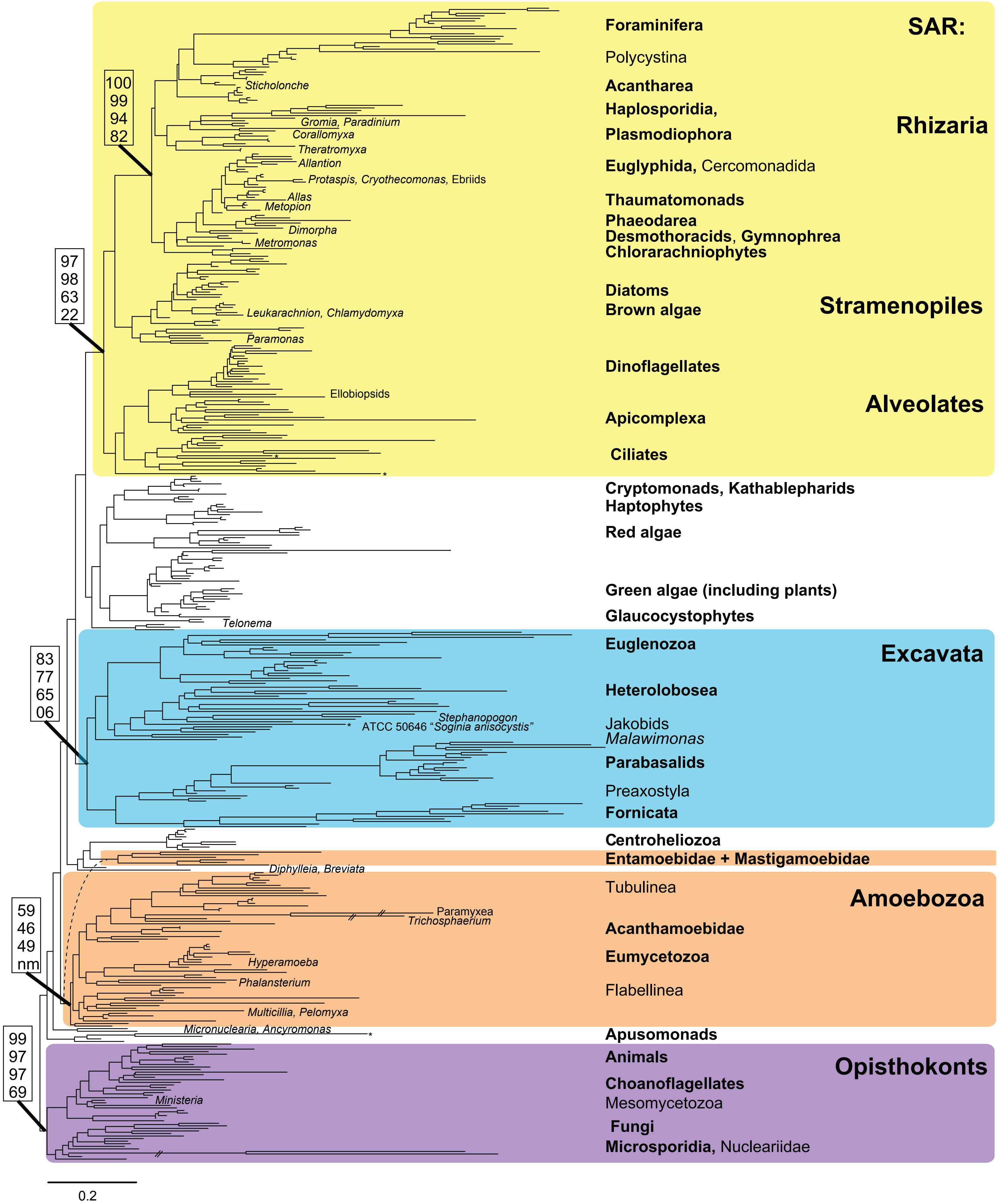
FIGURE 1. Most likely eukaryotic tree of life reconstructed using all 451 taxa and all 16 genes (ssu-rDNA plus 15 protein genes). Major nodes in this topology are robust to analyses of subsets of taxa and genes, which include varying levels of missing data (Table 1). Clades in bold are monophyletic in analyses with 2 or more members except in all:15 in which taxa represented by a single gene were sometimes misplaced. Numbers in boxes represent support at key nodes in analyses with increasing amounts of missing data (10:16, 6:16, 4:16, and all:16 analyses; see Table 1 for more detail). Given uncertainties around the root of the eukaryotic tree of life (see text), we have chosen to draw the tree rooted with the well-supported clade Opisthokonta. Dashed line indicates alternate branching pattern seen for Amoebozoa in other analyses. Long branches, indicated by //, have been reduced by half. †Preaxostyla is not monophyletic likely because of contaminating *Streblomastix strix* ESTs. The six lineages labeled by * represent taxa that are misplaced, probably due to LBA, listed from top to bottom with expected clade in parentheses. These are *Protoopalina japonica* (Stramenopiles), *Aggregata octopiana* (Apicomplexa), *Mikrocytos mackini* (Haplosporidia), *Centropyxis laevigata* (Tubulinea), *Marteilioides chungmuensis* (unplaced), and *Cochliopodium spiniferum* (Amoebozoa).

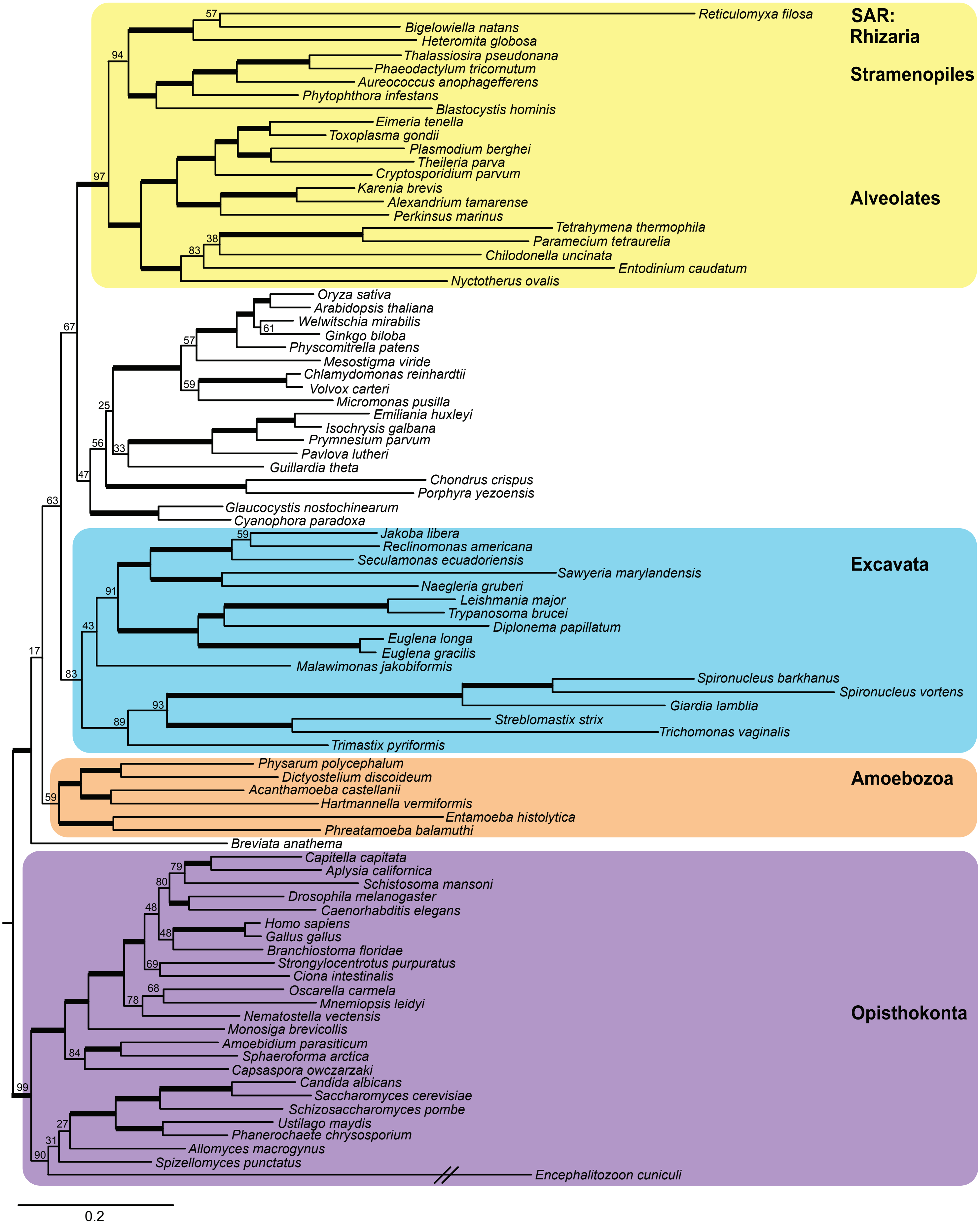
FIGURE 2. Most likely eukaryotic tree of life reconstructed with 10:16, which includes 88 taxa and 16 genes (ssu-rDNA plus 15 protein genes). Thickened lines receive >95% bootstrap support. Other notes as in Methods and Figure 1.

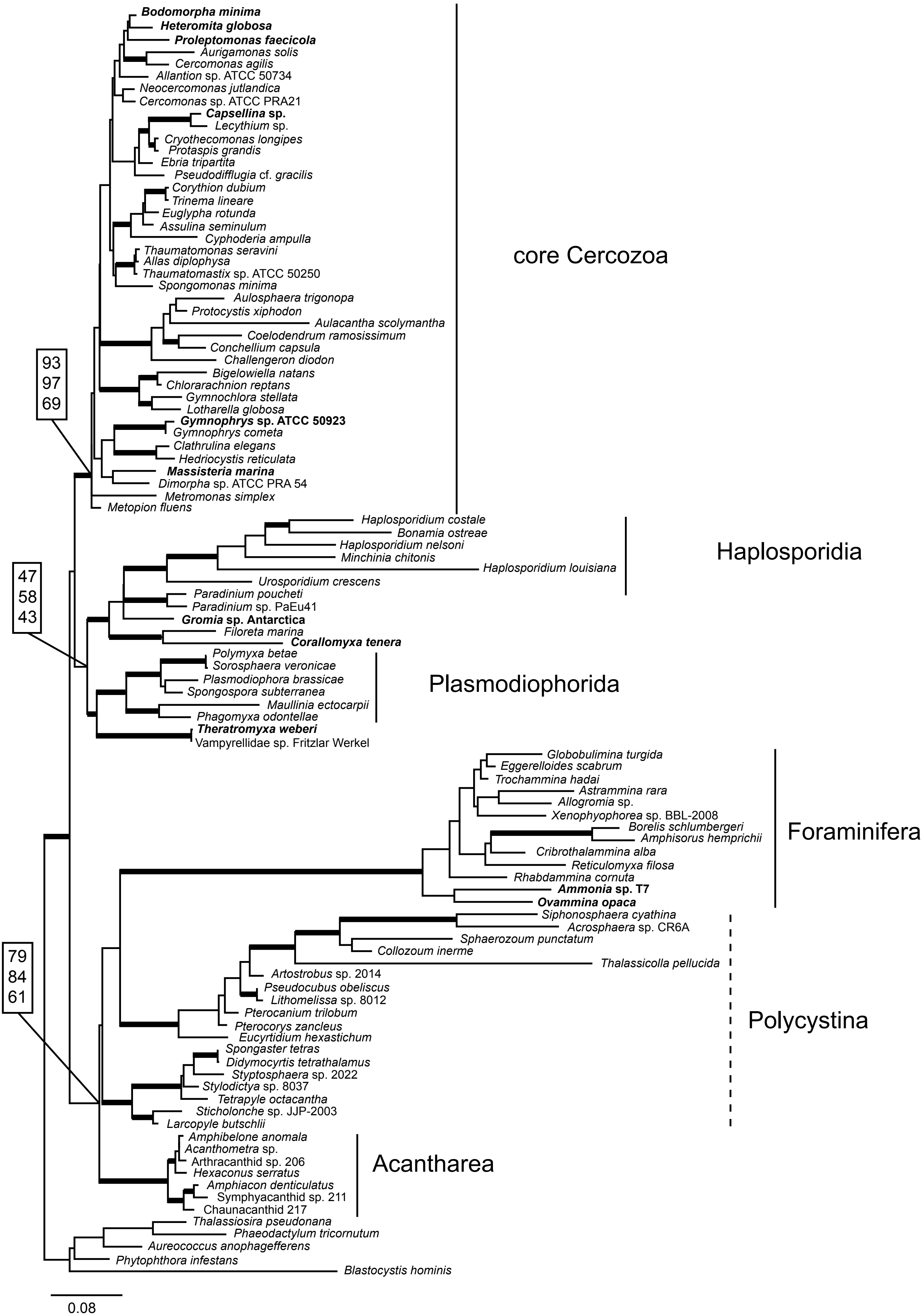
FIGURE 3. Maximum likelihood tree of Rhizaria reconstructed with 103 Rhizaria taxa and 16 genes. The ssu-rDNA partition was analyzed with GTR+gamma and proteins with rtREV. Thickened lines receive >80% bootstrap support in all analyses. Node support in boxes from Rhizaria:4-gene, Rhizaria:16-gene, all:16 analyses. Taxa with new data are bold. Dash lines indicate non-monophyly.

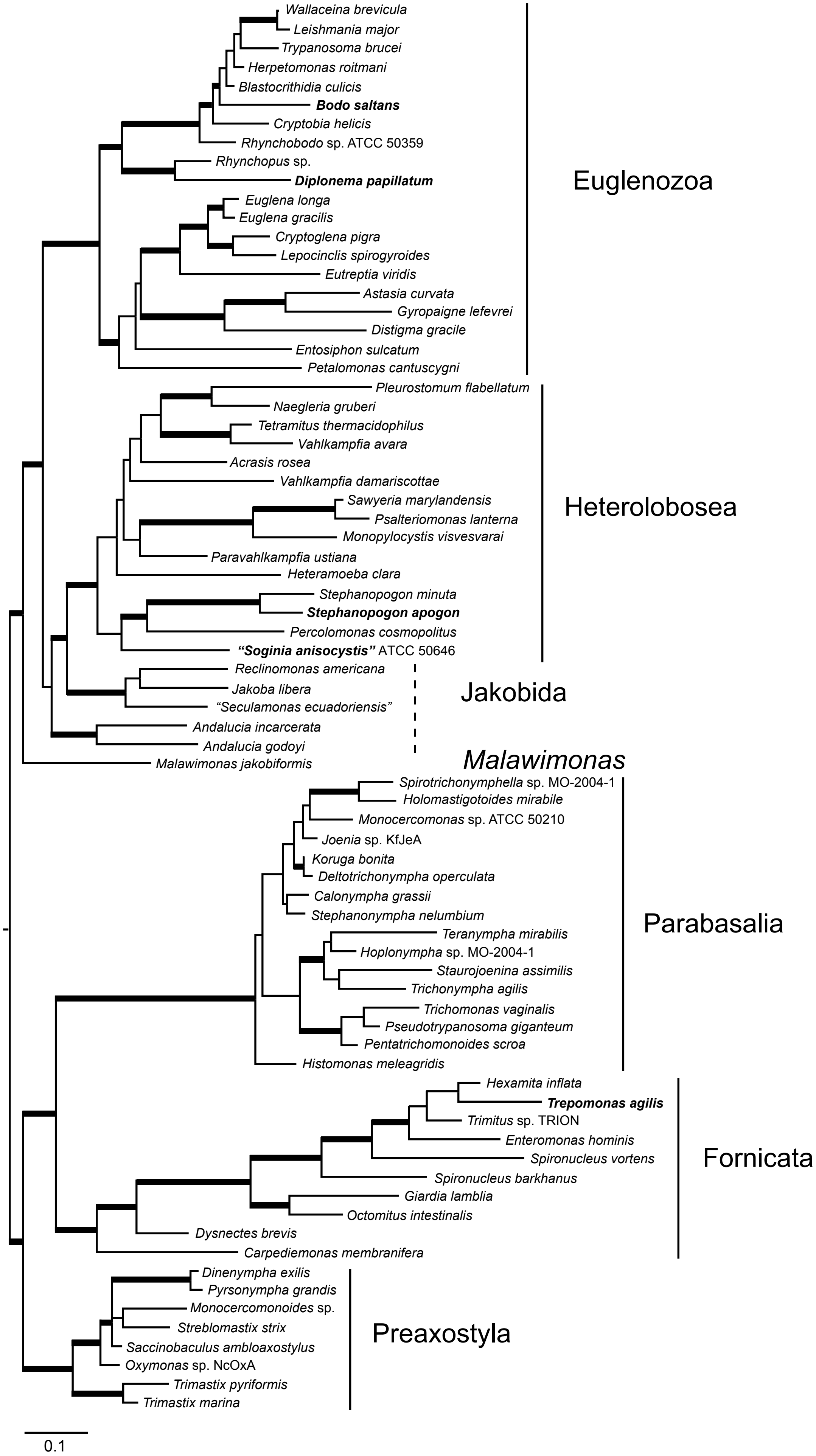
FIGURE 4. Maximum likelihood tree of Excavata with 75 taxa and 16 genes. The ssu-rDNA partition was analyzed with GTR+gamma and proteins with rtREV. See Figure 3 for other notes.

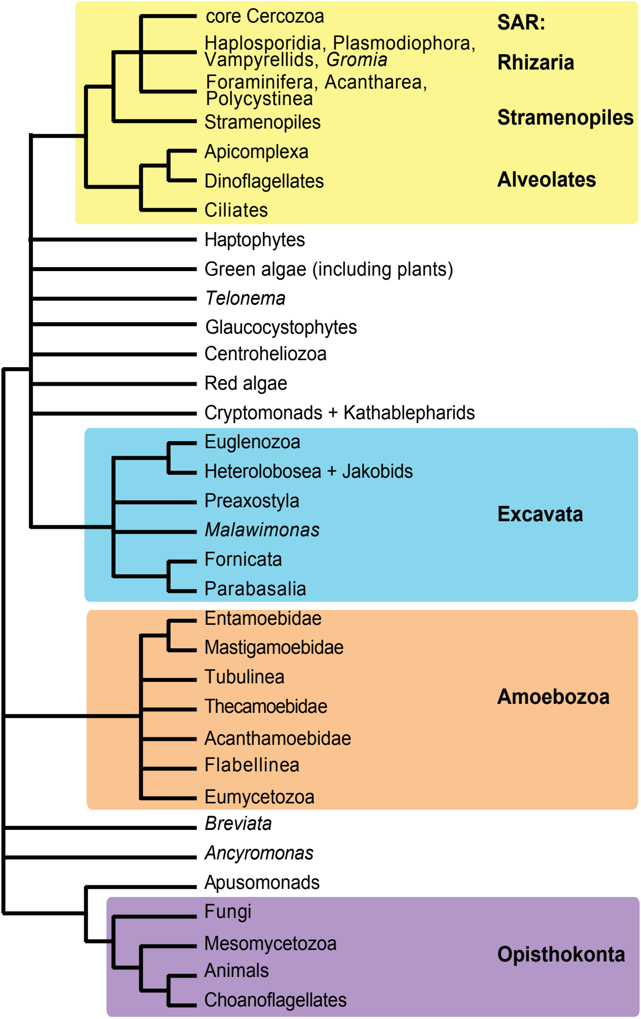
FIGURE 5. Summary of major findings—the evolutionary relationships among major lineages of eukaryotes. Clades have been collapsed into those that we view to be strongly supported. The many polytomies represent uncertainties that remain.


SAR:
Foraminifera
Polycystina
Acantharea
Sticholonche
Haplosporidia,
Gromia, Paradinium
Plasmodiophora
Corallomyxa
Rhizaria
Theratromyxa
100
99
94
82
Allantion
Euglyphida, Cercomonadida
Protaspis, Cryothecomonas, Ebriids
Thaumatomonads
Allas
Metopion
Phaeodarea
Desmothoracids, Gymnophrea
Dimorpha
Metromonas
Chlorarachniophytes
97
98
63
22
Diatoms
Brown algae
Stramenopiles
Leukarachnion, Chlamydomyxa
Paramonas
Dinoflagellates
Ellobiopsids
Alveolates
Apicomplexa
*
Ciliates
*
Cryptomonads, Kathablepharids
Haptophytes
Red algae
Green algae (including plants)
Glaucocystophytes
Telonema
Euglenozoa
Excavata
Heterolobosea
83
77
65
06
Stephanopogon
* ATCC 50646 "Soginia anisocystis"
Jakobids
Malawimonas
Parabasalids
Preaxostyla
Fornicata
Centroheliozoa
Entamoebidae + Mastigamoebidae
Diphylleia, Breviata
Tubulinea
Amoebozoa
Paramyxea
Trichosphaerium
Acanthamoebidae
59
46
49
nm
Eumycetozoa
Hyperamoeba
Phalansterium
Flabellinea
Multicillia, Pelomyxa
Micronuclearia, Ancyromonas
*
Apusomonads
Animals
Opisthokonts
99
97
97
69
Choanoflagellates
Mesomycetozoa
Ministeria
Fungi
Microsporidia, Nucleariidae
0.2


SAR:
Rhizaria
Stramenopiles
Alveolates
Excavata
Amoebozoa
Opisthokonta
Reticulomyxa filosa
Bigelowiella natans
Heteromita globosa
Thalassiosira pseudonana
Phaeodactylum tricornutum
Aureococcus anophagefferens
Phytophthora infestans
Blastocystis hominis
Eimeria tenella
Toxoplasma gondii
Plasmodium berghei
Theileria parva
Cryptosporidium parvum
Karenia brevis
Alexandrium tamarense
Perkinsus marinus
Tetrahymena thermophila
Paramecium tetraurelia
Chilodonella uncinata
Entodinium caudatum
Nyctotherus ovalis
Oryza sativa
Arabidopsis thaliana
Welwitschia mirabilis
Ginkgo biloba
Physcomitrella patens
Mesostigma viride
Chlamydomonas reinhardtii
Volvox carteri
Micromonas pusilla
Emiliania huxleyi
Isochrysis galbana
Prymnesium parvum
Pavlova lutheri
Guillardia theta
Chondrus crispus
Porphyra yezoensis
Glaucocystis nostochinearum
Cyanophora paradoxa
Jakoba libera
Reclinomonas americana
Seculamonas ecuadoriensis
Sawyeria marylandensis
Naegleria gruberi
Leishmania major
Trypanosoma brucei
Diplonema papillatum
Euglena longa
Euglena gracilis
Malawimonas jakobiformis
Spironucleus barkhanus
Spironucleus vortens
Giardia lamblia
Streblomastix strix
Trichomonas vaginalis
Trimastix pyriformis
Physarum polycephalum
Dictyostelium discoideum
Acanthamoeba castellanii
Hartmannella vermiformis
Entamoeba histolytica
Phreatamoeba balamuthi
Breviata anathema
Capitella capitata
Aplysia californica
Schistosoma mansoni
Drosophila melanogaster
Caenorhabditis elegans
Homo sapiens
Gallus gallus
Branchiostoma floridae
Strongylocentrotus purpuratus
Ciona intestinalis
Oscarella carmela
Mnemiopsis leidyi
Nematostella vectensis
Monosiga brevicollis
Amoebidium parasiticum
Sphaeroforma arctica
Capsaspora owczarzaki
Candida albicans
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Ustilago maydis
Phanerochaete chrysosporium
Allomyces macrogynus
Spizellomyces punctatus
Encephalitozoon cuniculi
57
94
97
83
38
67
61
57
59
25
56
33
47
63
59
91
43
83
93
89
17
59
79
80
48
48
69
68
78
84
99
27
31
90
0.2


Bodomorpha minima
Heteromita globosa
Proleptomonas faecicola
Aurigamonas solis
Cercomonas agilis
Allantion sp. ATCC 50734
Neocercomonas jutlandica
Cercomonas sp. ATCC PRA21
Capsellina sp.
Lecythium sp.
Cryothecomonas longipes
Protaspis grandis
Ebria tripartita
Pseudodifflugia cf. gracilis
Corythion dubium
Trinema lineare
Euglypha rotunda
Assulina seminulum
Cyphoderia ampulla
Thaumatomonas seravini
Allas diplophysa
Thaumatomastix sp. ATCC 50250
Spongomonas minima
Aulosphaera trigonopa
Protocystis xiphodon
Aulacantha scolymantha
Coelodendrum ramosissimum
Conchellium capsula
Challengeron diodon
Bigelowiella natans
Chlorarachnion reptans
Gymnochlora stellata
Lotharella globosa
Gymnophrys sp. ATCC 50923
Gymnophrys cometa
Clathrulina elegans
Hedriocystis reticulata
Massisteria marina
Dimorpha sp. ATCC PRA 54
Metromonas simplex
Metopion fluens
core Cercozoa
93
97
69
Haplosporidium costale
Bonamia ostreae
Haplosporidium nelsoni
Minchinia chitonis
Haplosporidium louisiana
Urosporidium crescens
Paradinium poucheti
Paradinium sp. PaEu41
Gromia sp. Antarctica
Filoreta marina
Corallomyxa tenera
Haplosporidia
47
58
43
Polymyxa betae
Sorosphaera veronicae
Plasmodiophora brassicae
Spongospora subterranea
Maullinia ectocarpii
Phagomyxa odontellae
Theratromyxa weberi
Vampyrellidae sp. Fritzlar Werkel
Plasmodiophorida
Globobulimina turgida
Eggerelloides scabrum
Trochammina hadai
Astrammina rara
Allogromia sp.
Xenophyophorea sp. BBL-2008
Borelis schlumbergeri
Amphisorus hemprichii
Cribrothalammina alba
Reticulomyxa filosa
Rhabdammina cornuta
Ammonia sp. T7
Ovammina opaca
Foraminifera
Siphonosphaera cyathina
Acrosphaera sp. CR6A
Sphaerozoum punctatum
Collozoum inerme
Thalassicolla pellucida
Artostrobus sp. 2014
Pseudocubus obeliscus
Lithomelissa sp. 8012
Pterocanium trilobum
Pterocorys zancleus
Eucyrtidium hexastichum
Spongaster tetras
Didymocyrtis tetrathalamus
Styptosphaera sp. 2022
Stylodictya sp. 8037
Tetrapyle octacantha
Sticholonche sp. JJP-2003
Larcopyle butschlii
Polycystina
79
84
61
Amphibelone anomala
Acanthometra sp.
Arthracanthid sp. 206
Hexaconus serratus
Amphiacon denticulatus
Symphyacanthid sp. 211
Chaunacanthid 217
Acantharea
Thalassiosira pseudonana
Phaeodactylum tricornutum
Aureococcus anophagefferens
Phytophthora infestans
Blastocystis hominis
0.08


Wallaceina brevicula
Leishmania major
Trypanosoma brucei
Herpetomonas roitmani
Blastocrithidia culicis
Bodo saltans
Cryptobia helicis
Rhynchobodo sp. ATCC 50359
Rhynchopus sp.
Diplonema papillatum
Euglena longa
Euglena gracilis
Cryptoglena pigra
Lepocinclis spirogyroides
Eutreptia viridis
Astasia curvata
Gyropaigne lefevrei
Distigma gracile
Entosiphon sulcatum
Petalomonas cantuscygni
Euglenozoa
Pleurostomum flabellatum
Naegleria gruberi
Tetramitus thermacidophilus
Vahlkampfia avara
Acrasis rosea
Vahlkampfia damariscottae
Sawyeria marylandensis
Psalteriomonas lanterna
Monopylocystis visvesvarai
Paravahlkampfia ustiana
Heteramoeba clara
Stephanopogon minuta
Stephanopogon apogon
Percolomonas cosmopolitus
"Soginia anisocystis" ATCC 50646
Heterolobosea
Reclinomonas americana
Jakoba libera
"Seculamonas ecuadoriensis"
Andalucia incarcerata
Andalucia godoyi
Jakobida
Malawimonas jakobiformis
Malawimonas
Spirotrichonymphella sp. MO-2004-1
Holomastigotoides mirabile
Monocercomonas sp. ATCC 50210
Joenia sp. KfJeA
Koruga bonita
Deltotrichonympha operculata
Calonympha grassii
Stephanonympha nelumbium
Teranympha mirabilis
Hoplonympha sp. MO-2004-1
Staurojoenina assimilis
Trichonympha agilis
Trichomonas vaginalis
Pseudotrypanosoma giganteum
Pentatrichomonoides scroa
Histomonas meleagridis
Parabasalia
Hexamita inflata
Trepomonas agilis
Trimitus sp. TRION
Enteromonas hominis
Spironucleus vortens
Spironucleus barkhanus
Giardia lamblia
Octomitus intestinalis
Dysnectes brevis
Carpediemonas membranifera
Fornicata
Dinenympha exilis
Pyrsonympha grandis
Monocercomonoides sp.
Streblomastix strix
Saccinobaculus ambloaxostylus
Oxymonas sp. NcOxA
Trimastix pyriformis
Trimastix marina
Preaxostyla
0.1


core Cercozoa
Haplosporidia, Plasmodiophora, Vampyrellids, *Gromia*
Foraminifera, Acantharea, Polycystinea
Stramenopiles
Apicomplexa
Dinoflagellates
Ciliates
SAR:
Rhizaria
Stramenopiles
Alveolates
Haptophytes
Green algae (including plants)
Telonema
Glaucocystophytes
Centroheliozoa
Red algae
Cryptomonads + Kathablepharids
Euglenozoa
Heterolobosea + Jakobids
Preaxostyla
Malawimonas
Fornicata
Parabasalia
Excavata
Entamoebidae
Mastigamoebidae
Tubulinea
Thecamoebidae
Acanthamoebidae
Flabellinea
Eumycetozoa
Amoebozoa
Breviata
Ancyromonas
Apusomonads
Fungi
Mesomycetozoa
Animals
Choanoflagellates
Opisthokonta