Supporting Information

Hampl et al. 10.1073/pnas.0807880106

SI Text

Assembling Sequence Data and Constructing Alignments. This work examines several EST datasets that have not previously been subjected to phylogenomic analysis: the diplomonad *Spironucleus barkhanus* (described in ref. 1), *Trimastix pyriformis* (described in ref. 2), the oxymonad *Monocercomonoides* sp. (described in ref. 3), the stramenopile *Blastocystis hominis* (described in ref. 4) and *Andalucia incarcerata* (description follows). Total RNA was extracted from dense cultures of *A. incarcerata* isolate MB1 (5) using standard Tri-Reagent-based extraction protocols. From this material, polyA$^+$ mRNA was purified and plasmid cDNA libraries were constructed by Amplicon Express (Pullman). Target genes were selected from among 3,456 ESTs that were obtained from the *A. incarcerata* library by Sanger sequencing technology.

Previously published alignments were generously provided by Dr. Hervé Phillipe [Université de Montréal (6)]. Taxon sampling was augmented through BLASTp and tBLASTn searches of National Center for Biotechnology Information nucleotide and protein databases (www.ncbi.nlm.nih.gov) and local laboratory databases of EST sequences (including those described above), with an initial cutoff of 10^{-10}, and a threshold of 10^{-40} for known gene families (sources of data are summarized in Table S1 and Table S2). The dataset assembly was completed in late 2006, when we started our analysis, and data that became available later could not be included. The work was performed both manually and automatically by an in-house bioinformatics tool, AutoBlast. The gene alignments were manually screened for aberrant sequences, including obvious paralogs and contamination (e.g., human sequences in the *Toxoplasma* genome database). The screen was performed by BLAST searches against GenBank databases and by manual inspection of gene trees. Gene alignments were generated in ProbCons (7) under default settings, edited by eye in MacClade version 4.06 (Sinauer Associates) and BioEdit (8) and ambiguously aligned positions were removed using GBLOCKS (9) with all gapped positions allowed, and the minimum number of sequences for a flank position set to 50% of the number of taxa, plus one. Single gene trees were generated using RAxML (10) under the PROTCAT-WAG model with 4 categories of rate variation. 143 genes were maintained in the dataset, based on length, taxon distribution, and level of sequence conservation. The entire concatenated alignment retained 35,584 sites. The proportion of missing data per taxon varied between 2% and 82% (average 44%). Recent simulation studies indicate that this proportion of missing data is not expected to compromise phylogenetic inference (6, 11, 12). The list of the genes and the information on the taxon representation across the genes is given in Table S2. The alignments are available at www.natur.cuni.cz/~vlada/phylogenomicanalysis.

Exhaustive Tree Searching Under the "Uniform" Model. In this section we give a more detailed description of the exhaustive search procedure that was briefly described in the *Results*.

Starting with the maximum likelihood tree estimated by RAxML, controversial deep nodes among the Excavata and eukaryotes were collapsed (nodes designated by asterisks in Fig. 1), and resolved in all possible ways (945 trees). To keep the number of trees examined in the exhaustive search manageable, nodes with lower support within well-supported clans (e.g., within Archaeplastida + Haptophyta) were not collapsed. For each of the 945 topologies, total likelihoods were calculated for the concatenated dataset. A single model and parameters were therefore used for all proteins in the concatenate; this we refer to as the uniform model. In addition, RELL bootstrap values (13) were determined for all relevant nodes by resampling, with replacement, the site likelihoods, and the sum of resampled log likelihoods was calculated for each tree. Using this method, 10,000 replicates were performed, and the ML tree was identified after each replicate. A consensus of these 10,000 trees was made.

Exhaustive Search with Protein-Specific Parameters ("Separate Analysis"). For separate analysis, we repeated the exhaustive search exactly as described above but used independent branch lengths and shape parameters for the gamma rates across sites distribution for each protein alignment partition–separate model. For proteins for which there were missing taxa, the relevant branches were "pruned" from trees before branch length estimation and likelihood calculation for that particular alignment. RELL bootstrap values were calculated as above except that resampling was carried out separately for each protein. Separate analysis allows for protein-specific evolutionary dynamics and may decrease the effect of LBA that is caused primarily by model misspecification. Although the number of parameters increased drastically in the separate analysis (from 94 to 13 442) its use was justified by the Akaike Information Criterion (AIC) (14) and likelihood ratio tests (see below).

Although the best topology under the separate model recovered *Malawimonas* and other excavates as 2 separate branches, the RELL bootstrap support for monophyly of Excavata increased from 7% to 30%. As in case of the uniform model, the likelihood difference between the best topology and several topologies with monophyletic Excavata was not statistically significant under the separate model (see below "Topology testing" and Table S3).

Comparison of Uniform and Separate Models. When nested models of differing complexity can be used, they should be evaluated to determine whether the additional parameters included in the more complex model are justified. The Akaike Information Criterion (14) (AIC) (Eq. **1**) and Bayesian Information Criterion (15) (BIC) (Eq. **2**) are frequently used to assess model fit. A second-order correction of the AIC (Eq. **3**) can also be used in cases where the number of parameters in the more complex model is large (16). These criteria are defined by the following equations.

$$\mathrm{AIC} = -2\,\log L + 2K \qquad [1]$$

$$\mathrm{AIC_c} = -2\,\log L + 2K\left(\frac{n}{n-K-1}\right) \qquad [2]$$

$$\mathrm{BIC} = -2\log L + 2K\,\log n \qquad [3]$$

where K is the number of parameters in the statistical model, and L is the likelihood function and n number of sites. In each case, the model with the smaller information criterion value is preferred.

Model fit can also be assessed by likelihood-ratio testing. The log of the ratio between the likelihood under the more complex model and the likelihood under the simpler model, multiplied by 2, is χ^2 distributed, with degrees of freedom equal to the number of additional parameters in the more complex model.

For our dataset we obtained conflicting results with different tests. Using the first-order AIC the separate model is preferred

over the uniform model (Δ AIC = 8,615.2). Using the second-order AIC or BIC, the uniform model is preferred (Δ AIC_C = −7,706.8, Δ BIC = −104,571.2). The separate model is preferred by the likelihood ratio test ($P < 0.0001$).

Note that regardless of whether or not the more complex model is selected by these criteria, the use of the separate model is consistent with standard phylogenetic practice of estimating protein-specific branch lengths when doing single-gene phylogenies. Furthermore, use of a more complex model than strictly necessary is expected to simply increase the variance of the tree estimate but is unlikely to bias this estimate.

Topology Testing. We tested the statistical significance of the likelihood difference between the best topology, in which *Malawimonas* branches separately from other Excavata (Fig. 1), and 10 plausible topologies with a monophyletic Excavata grouping. For these analyses we used the AU, KH and SH tests implemented in the program Consel (17) using the site-likelihoods associated with each topology provided by RAxML. The results for these 3 tests for the uniform and separate models are summarized in Table S3. The topologies that could not be rejected at a 5% level of significance were those in which *Malawimonas* formed the basal lineage within Excavata, or branched as the sister group to metamonads. Some tests also did not reject the topology with *Malawimonas* branching as the sister-group to Discoba.

Amino Acid Recoding. We recoded the amino acids in the alignment into functional categories using 2 different recoding schemes (1). Using a model based on amino acid biochemical properties (Dayhoff model), amino acids were grouped into 4 categories: YWFVILM, STGP, NEQD, KHR, with cysteines replaced by gaps (2). Following a JTT substitution matrix for assigning distances between amino acids based on observed rates of substitutions, the 20 residues were reassigned into 4 categories: ANGTPS, RDEQK, ILMFV, and HWYC. Phylogenies and 100 bootstrap replicates were generated from the recoded alignments using RAxML under the GTRGAMMA model with 4 categories of rate variation. In both cases the resulting topologies were very similar to the tree in Fig. 1, but the RAxML bootstrap support for the branch separating *Malawimonas* from the other excavates was lower at −54% and 70%, respectively (Figs. S1 and S2).

Removal of Fast Evolving Sites. We used 2 different methods to identify fast evolving sites: (1) a parsimony-based method (18) and (2) a likelihood method (a modification of the approach used in refs. 19 and 20).

Parsimony-Based Fast Site Removal. The sites were divided into rate categories using a parsimony-based method described by Brinkmann and Phillippe (18). The sites were assessed into 20 rate categories using the tree from the Fig. 1 and clades that received RAxML bootstrap support 95% and higher. The fastest categories of sites were progressively removed from the alignment. Assessment of rate categories and alignment trimming was performed by the program Slowfaster (21). The 6 most trimmed alignments containing sites in rate categories 0 + 1+2 + 3+4 + 5+6 and less were analyzed by ML with bootstrapping (100 replicates) using RAxML. The bootstrap values for bipartitions of interest were plotted against the number of sites (Fig. S3). The bootstrap support of the unikont clan was used as a control.

Maximum Likelihood-Based Fast Site Removal. Conditional mean rates at sites were calculated for all sites using CODEML (model WAG+Γ with 8 categories) for each of the 100 trees produced by RAxML from bootstrapping our main dataset. For each site, the average rate was calculated over all 100 conditional mean rates. The 1,000 sites with the highest rates were iteratively removed to produce nested alignments with between 8,584 and 35,584 sites (a total of 28 alignments; no datasets with <8,584 sites were produced). ML trees with bootstrap support (100 replicates) were estimated using RAxML for each of these alignments and the bootstrap values for bipartitions of interest were plotted against the size of the alignment (Fig. S4). The bootstrap support of the unikont clan was used as a control for overall resolution.

For both methods the support for the *Malawimonas* + unikont grouping increased up to 98% at some level of exclusion and then dropped concomitantly with the collapse of the overall tree topology (as indicated by the drop of support for unikonts, plotted as a "control"). The support for the common clade of Excavata and *Malawimonas* remained very low.

Removal of Long Branching (LB) Taxa Using Alternative Root Positions. In addition to placing the root between "unikonts" and "bikonts" (see Results), 2 other positions for the root were used for experiments with removal of LB taxa: (1) a root in the middle of the common ancestral branch of diplomonads and parabasalids and (2) the "corrected midpoint root". The corrected midpoint root was estimated as the midpoint of a tree from which 10% of taxa with the longest terminal branches (*Trichomonas*, *Spironucleus*, *Giardia*, *Blastocystis*, and *Sawyeria*) were trimmed as outliers. The corrected midpoint root lies on the branch between Discoba and the rest of eukaryotes.

Root positions were used to calculate root-to-tip distances for taxa. Up to 26 taxa were sequentially removed according to this measure and trees were constructed and bootstrapped from the truncated alignments using RAxML exactly as in the main analysis described in the main text. The changes in bootstrap support for the bipartitions of interest are plotted against the number of removed taxa in Figs. S7 and S8. The changes in support were broadly similar to those obtained using the unikonts/bikonts rooting. The support for monophyly of Excavata as a whole and for the Excavata + unikonts bipartition increased concomitantly with the decrease of support for the bipartition of *Malawimonas* and Unikonts. This trend reversed temporarily after exclusion of 14 and more taxa and then reverted again. As in case of the main experiment described in the text, the dip in the support for Excavata and Excavata + unikonts was probably caused by artificial attraction between newly emerging long branches of *Cercomonas* (after exclusion of *Bigelowiella*) and jakobids or *Trimastix* (after exclusion of other Excavata).

Comparison of RAxML and PhyloBayes Results Under the CAT Model. To determine whether the tree recovered by PhyloBayes was an artifact of the tree-search heuristic or was truly the tree with highest posterior probability, PhyloBayes was run with the CAT model, but with fixed trees. Two chains were run for each of the following analyses: one analysis had the tree fixed to be the ML tree recovered by RAxML and a second was fixed to the tree recovered by PhyloBayes. This procedure was repeated for the datasets generated by the removal of 14 LB taxa, and the removal of 1,750 LB sequences.

A standard method in Bayesian analysis to compare the fit of 2 models is to compute the Bayes factor:

$$\mathrm{BF} = \frac{\int P(\theta|M_1)P(D|\theta, M_1)d\theta}{\int P(\theta|M_2)P(D|\theta, M_2)d\theta}. \qquad [4]$$

In this case the 2 models (M_1 and M_2) correspond to the 2 different tree topologies. Evaluating the integrals in the Bayes

factor calculation is impossible analytically, and can be very difficult to approximate using numerical methods. One such approximation to the integrals is to use the harmonic mean of the post-burn-in likelihoods from the posterior distribution (22). However, this estimate is extremely unstable because of infinite variance. In section 7 of ref. 22, Newton and Raftery suggest an importance sampling method that generates a more stable smoothed estimate of the marginal likelihoods. Using this method we calculated marginal likelihoods for the chains constrained to have the "PhyloBayes tree" and the chains constrained to have the "RAxML tree." The Bayes factor was calculated (or $\log_{10}$BF, see Eq. **5**) to compare the fit of the 2 trees by merging the results of 2 independent chains for each set of constraints.

$$\text{log BF in decibans} = 0.1 \log_{10}\text{BF} = 0.1 \log_{10}\left[\frac{\mu_{\text{PhyloBayes}}}{\mu_{\text{RAxML}}}\right]. \quad [5]$$

For the 14 LB dataset, log BF = −5.462 decibans and for the 1,750 LB dataset, log BF = −12.430 decibans. These negative values in both cases indicate support for the RAxML topology over the PhyloBayes topology when evaluated under the CAT+ Γ model. While Bayes factors <5 are not considered significant, it is customary to regard a 5 < log BF <10 in magnitude as "substantial" evidence in support of the preferred model and log BF > 10 as "strong" evidence. Thus, it is especially puzzling that the topology of highest posterior probability in the independent PhyloBayes runs on both of these datasets is not supported by these Bayes factor calculations. The cause of these discrepancies is unknown at this time, but their existence is sufficient, in our view, to call into question the PhyloBayes results. Thus, the high posterior probabilities supporting the non-monophyly of excavates in the PhyloBayes analyses are misleading.

Analysis of Excavata-Only Dataset. To examine relationships within the Excavata and avoid systematic biases resulting from interactions with outgroup taxa, we also estimated the unrooted tree of the Excavata (Fig. S10). The relationships among Excavata corresponded exactly with those shown in Fig. 3*B*.

Analyses Including Cryptophyta. When the analyses for this article were underway, phylogenomic data from the cryptophytes *Guillardia theta* and *Rhodomonas salina* became publicly available. Cryptophyta represent probably the most important taxon that was absent in our original analyses. To examine the position of Cryptophyta in the phylogenetic tree constructed from our dataset we added available Cryptophyta data to our gene alignments. Cryptophyta sequences were carefully examined by BLAST and by inspection of gene trees, and sequences that likely originated from the nucleomorph genome were excluded. After this cleaning, *Guillardia theta* was represented in 42 gene alignments and sequences of *Rhodomonas salina* were available for 8 genes for which *Guillardia theta* data were missing. The sequences of *Guillardia theta* and *Rhodomonas salina* were then amalgamated to form a composite taxon "Cryptophyta" that was represented by 50 out of 143 genes. The new dataset was analyzed in RAxML using both uniform and separate models (WAG+Γ, PROTGAMMACAT setting, 4 categories of rate variation). Support for bipartitions was assessed by 100 bootstrap replicates also under both uniform and separate models. The tree is shown in Fig. S9.

Formal Definition of Discoba. Discoba A.G.B. Simpson, new clade name.

Definition. Node-based: The least inclusive clade containing *Jakoba libera* (Ruinen, 1938) Patterson, 1990; *Andalucia godoyi,* Lara et al., 2006; *Euglena gracilis* Klebs 1883; and *Naegleria gruberi* (Schardinger, 1899) Alexeieff, 1912. The taxon does not apply if this clade includes *Homo sapiens* Linnaeus 1758 or *Arabidopsis thaliana* (Linnaeus) Heynhold, 1842 (Abbreviated definition: < *Jakoba libera* and *Andalucia godoyi* and *Euglena gracilis* and *Naegleria gruberi*).

Etymology. A portmanteau of Discicristata and *Jakoba*, representing the 2 major clades that comprise this taxon.

Composition. Discicristata (i.e., Heterolobosea Page and Blanton, 1985 and Euglenozoa Cavalier-Smith, 1981 emend. Simpson, 1997) and Jakobida Cavalier-Smith, 1993 emend. Adl et al., 2005.

Reference Phylogeny. Fig. 1, this article. Note that *Andalucia godoyi* and *Andalucia incarcerata* are related to the exclusion of other described species studied to date (5).

Diagnostic Apomorphies and Synonyms. None known.

1. Andersson JO, et al. (2007) A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics* 8:51.
2. Hampl V, et al. (2008) Genetic evidence for a mitochondriate ancestry in the ''amitochondriate'' flagellate *Trimastix pyriformis*. *PLoS ONE* 3:e1383.
3. Hampl V, et al. (2005) Inference of the phylogenetic position of oxymonads based on 9 genes: Support for Metamonada and Excavata. *Mol Biol Evol* 22:2508–2518.
4. Stechmann A, et al. (2008) Organelles in *Blastocystis* that blur the distinction between mitochondria and hydrogenosomes. *Curr Biol* 18:580–585.
5. Simpson AGB, Perley TA, Lara E (2008) Lateral transfer of the gene for a widely used marker, alpha-tubulin, indicated by a multi-protein study of the phylogenetic position of *Andalucia* (Excavata). *Mol Phylogenet Evol* 47:366–77.
6. Philippe H, et al. (2004) Phylogenomics of eukaryotes: Impact of missing data on large alignments. *Mol Biol Evol* 21:1740–1752.
7. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res* 15:330–340.
8. Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp* Ser:95–98.
9. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.
10. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
11. Wiens JJ (2005) Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? *Syst Biol* 54:731–742.
12. Wiens JJ (2006) Missing data and the design of phylogenetic analyses. *J Biomed Inform* 39:34–42.
13. Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol* 31:151–160.
14. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Contr ACM* 19:716–723.
15. Schwarz G (1978) Estimating the dimension of the model. *Ann Statist* 6:461–464.
16. Hurvich CM, Tsai C.-L (1989) Regression and time series model selection in small samples. *Biometrika* 76:297–307.
17. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.
18. Brinkmann H, Philippe H (1999) Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol Biol Evol* 16:817–825.
19. Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguna J (1999) Acoel flatworms: Earliest extant bilaterian Metazoans, not members of Platyhelminthes. *Science* 283:1919–1923.
20. Rodriguez-Ezpeleta N, et al. (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol* 56:389–399.
21. Kostka M, Uzlikova M, Cepicka I, Flegr J (2008) SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of *Blastocystis*. *BMC Bioinformatics* 15:341.
22. Newton MA, Raftery AE (1994) Approximate bayesian inference with the weighted likelihood bootstrap. *J Roy Stat Soc Ser B* 56:3–48.

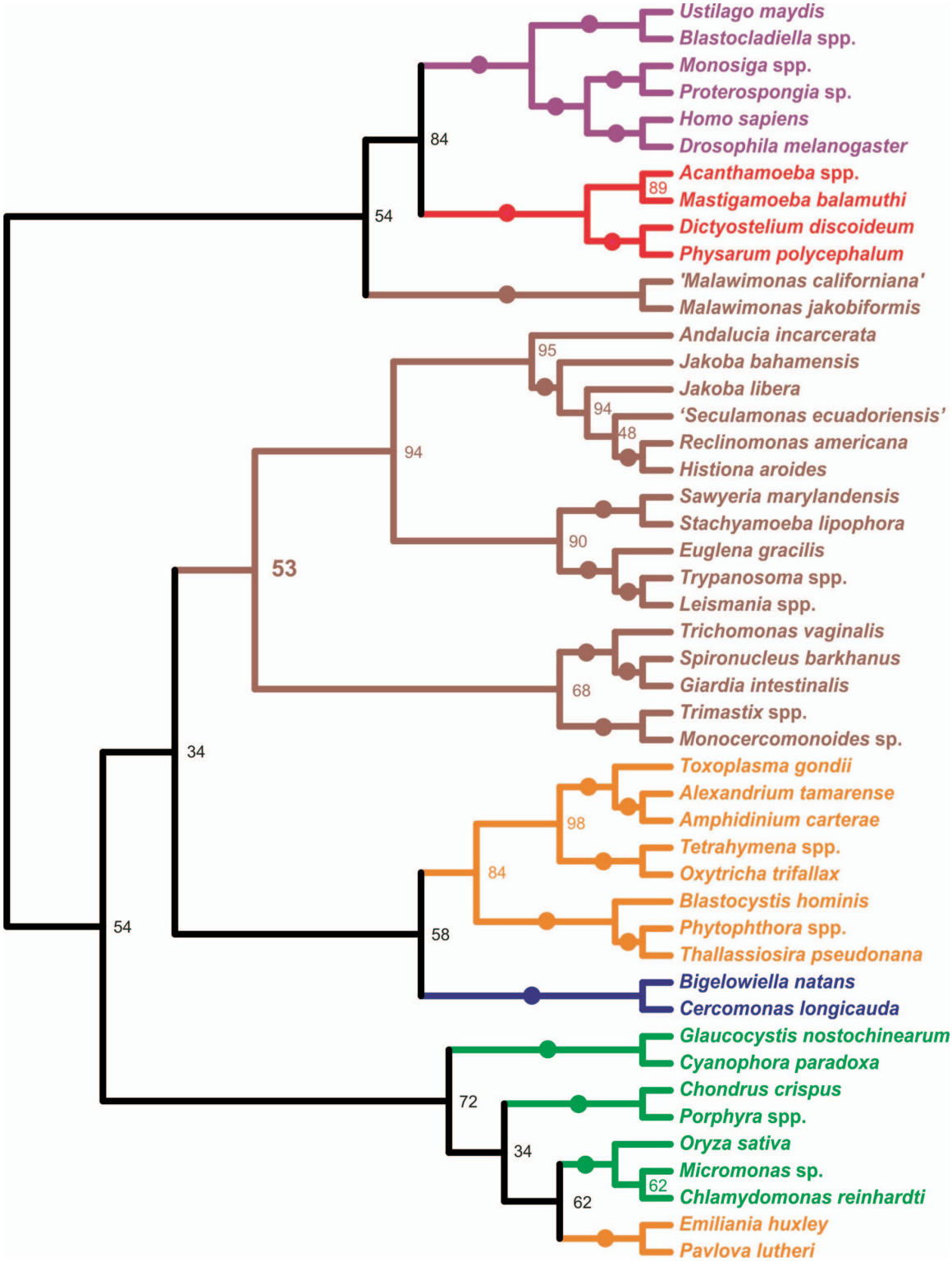

Fig. S1. A phylogenetic tree constructed from the dataset after recoding the amino acids into functional classes according to the Dayhoff model. The tree was constructed by RAxML using the GTRGAMMA model. The representatives of the 6 supergroups are color-coded. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping, branches that received maximum support are indicated by full circles.

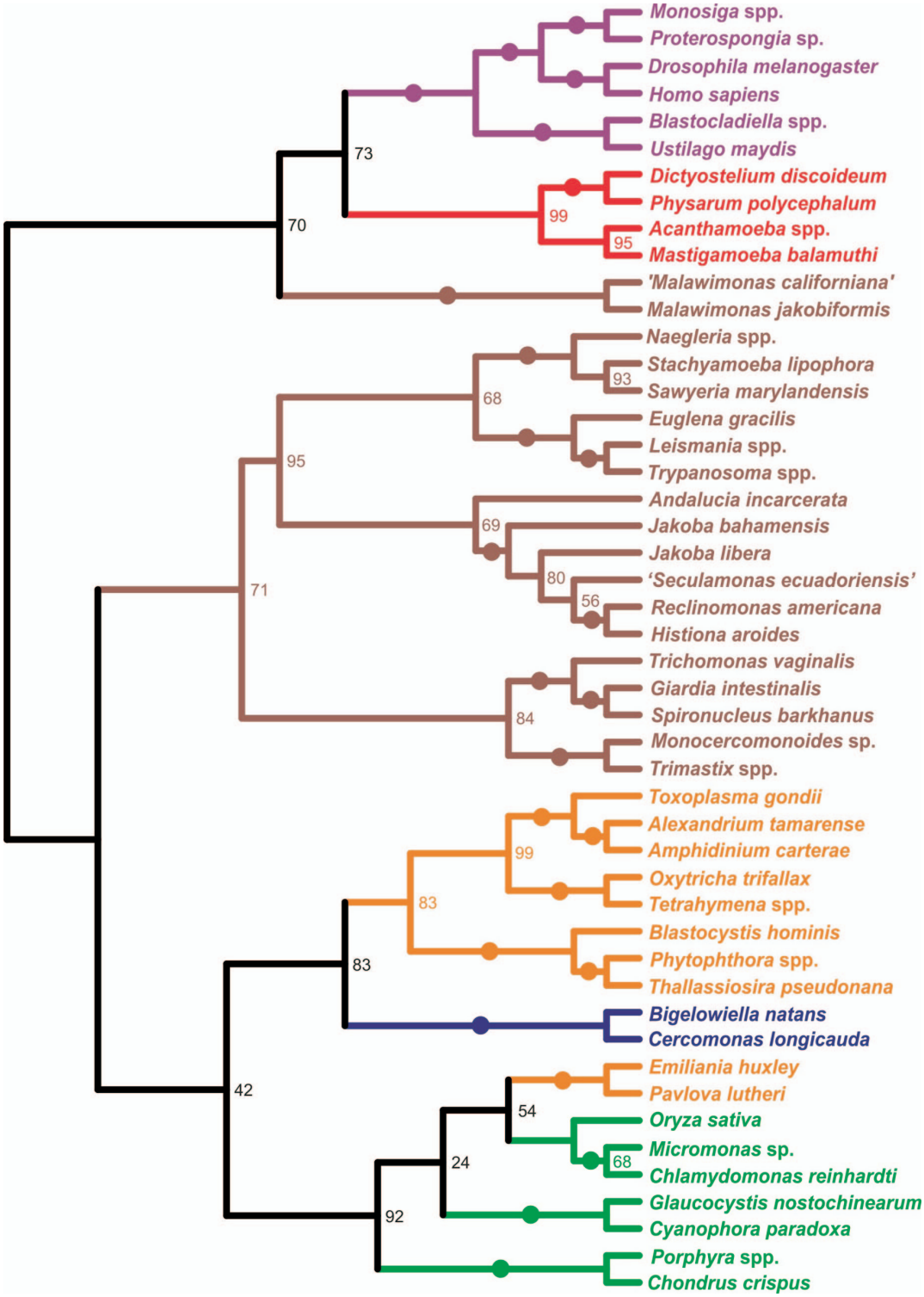

Fig. S2. A phylogenetic tree constructed from the dataset after recoding the amino acids into functional classes according to the JTT model. The tree was constructed by RAxML using the GTRGAMMA model. The representatives of the 6 supergroups are color-coded. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping, branches that received maximum support are indicated by full circles.

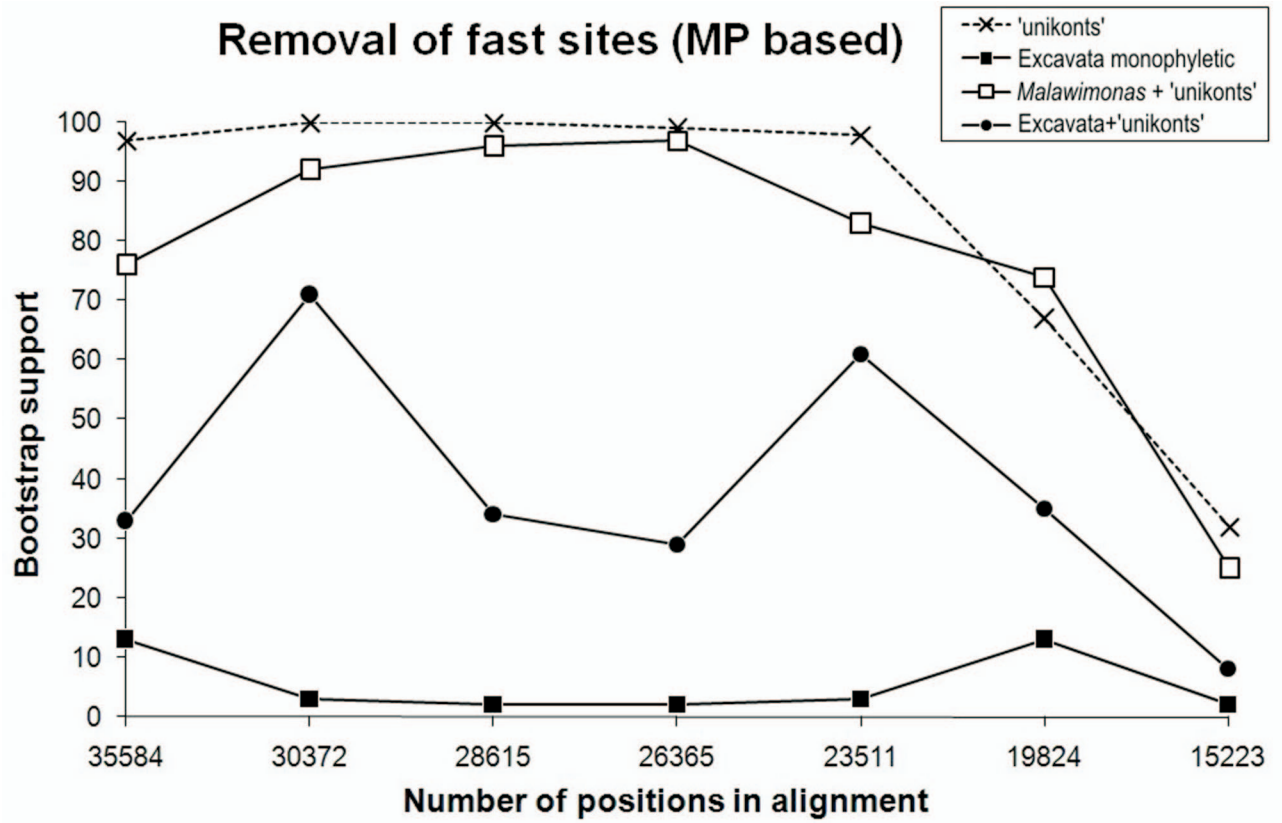

Fig. S3. A graph summarizing the results of the parsimony-based fast site removal experiment. The support for the nodes of interest calculated by RAxML bootstrapping is plotted against the number of fastest sites that were removed from the concatenated alignment. The support for unikonts (X) is used as a control.

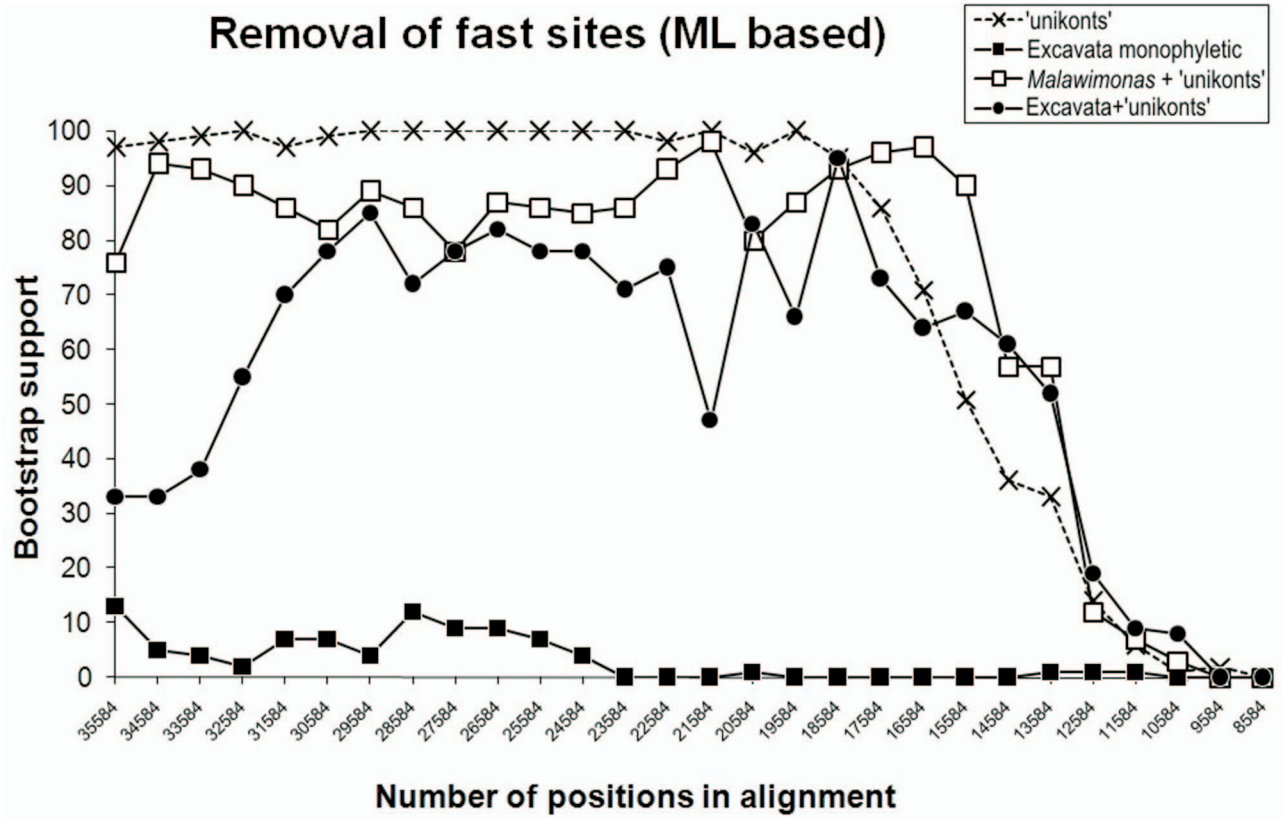

Fig. S4. A graph summarizing the results of the likelihood-based fast site removal experiment. The support for the nodes of interest calculated by RAxML bootstrapping is plotted against the number of fastest sites that were removed from the concatenated alignment. The support for unikonts (X) is used as a control.

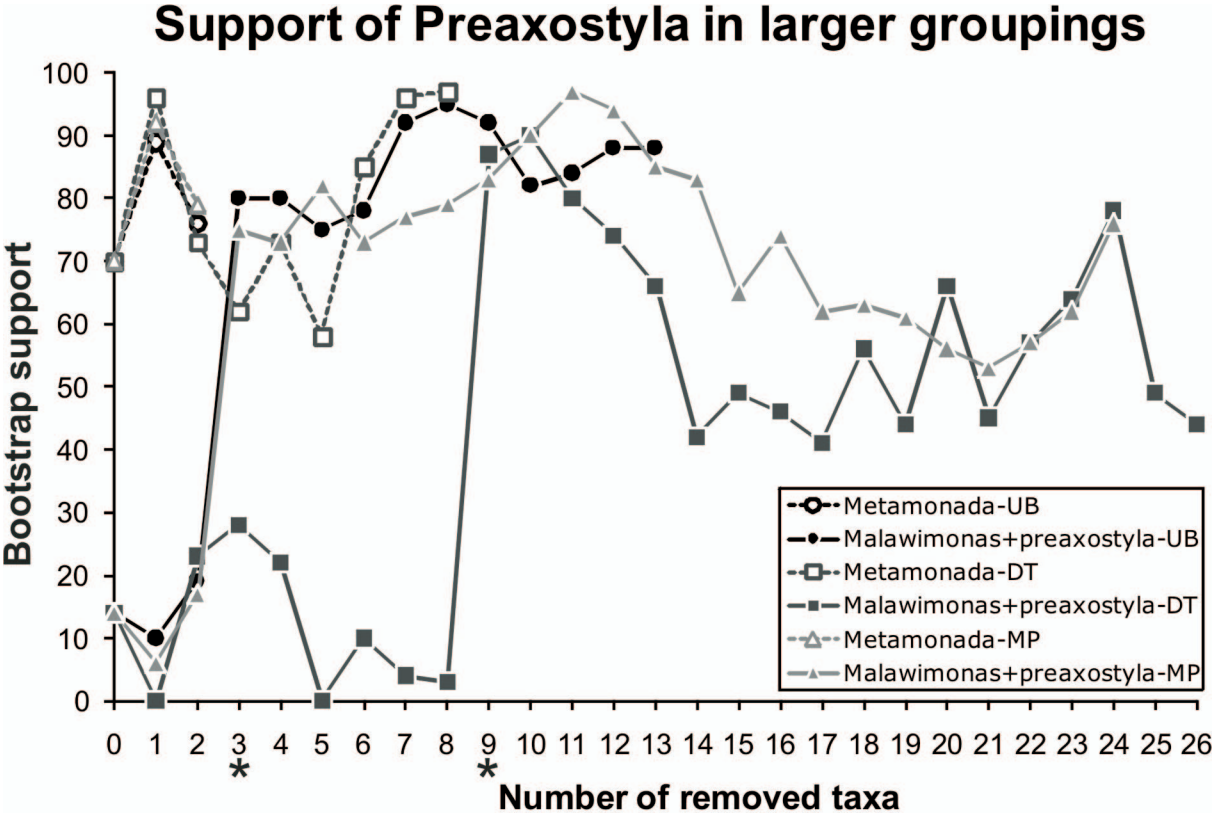

Fig. S5. A graph summarizing the results of LB taxon removal with respect to the relationships of Preaxostyla and other excavates. The support for a clade of Preaxostyla (*Monocercomonoides* plus *Trimastix*) with parabasalids and diplomonads to form Metamonada, and the support for the Preaxostyla with *Malawimonas* are shown as long-branch taxa are removed. These trends are tracked for 3 separate rooting positions and thus 3 different orders of taxon removal. Note that the support for *Malawimonas* plus Preaxostyla rises dramatically as soon as the final representative of the parabasalid/diplomonad group is removed (denoted with an asterisk), irrespective of which rooting position is used. This suggests that the long-branch nature of the parabasalid/diplomonad group is obscuring any relationship of *Malawimonas* with the Preaxostyla. Conceptually, however, it is important to note that these 2 relationships are not mutually exclusive since once the last representative of the diplomonad/parabasalid group is removed from the dataset, the Preaxostyla become synonymous with the grouping of Metamonada. UB, DT and MP in the legend designate the LB gene sequence removal analysis based on unikont, diplomonad + trichomonad and midpoint root, respectively.

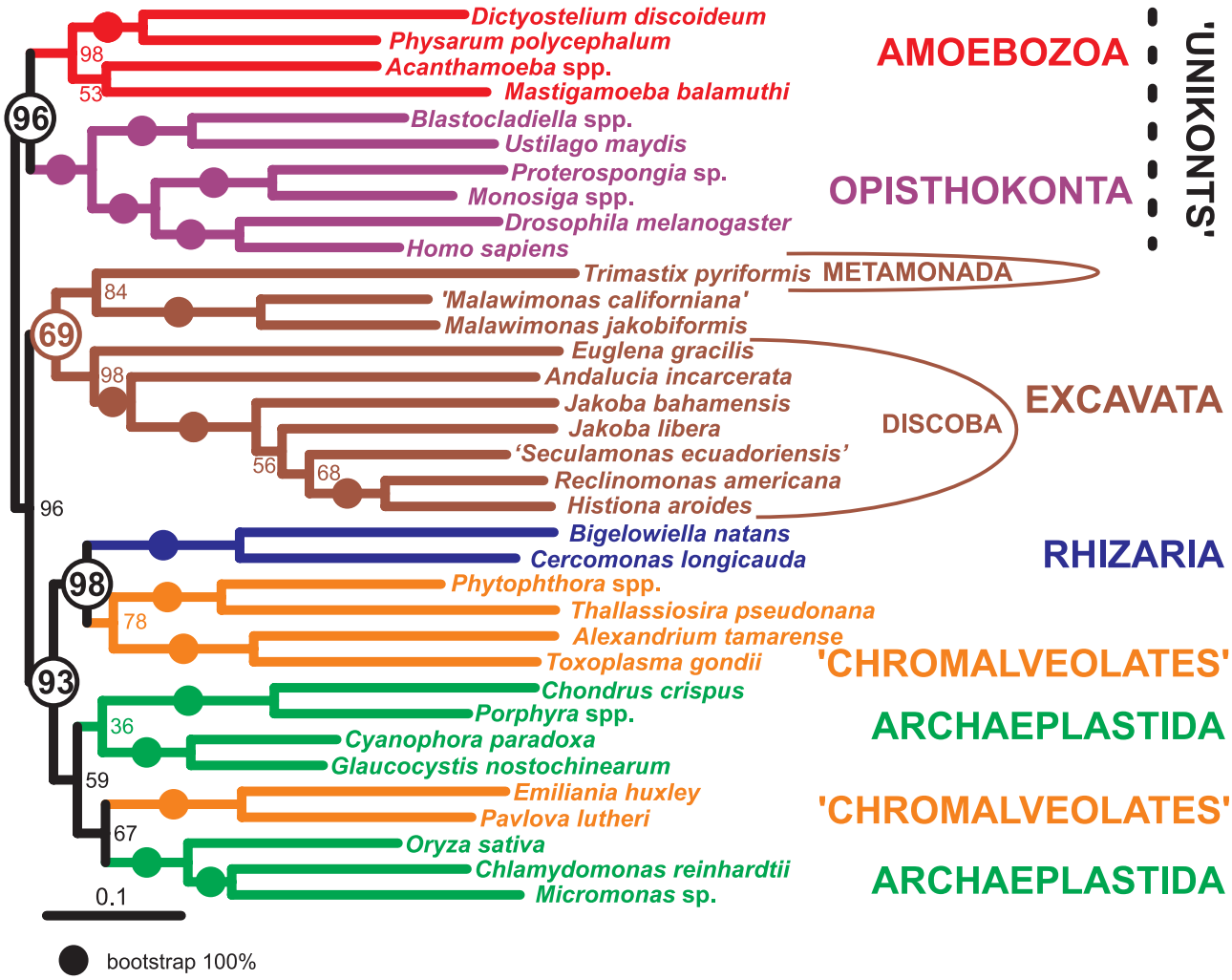

Fig. S6. Phylogenetic tree constructed from the dataset after removal of 13 of the longest-branch taxa (using Unikonta/Bikonta rooting point). The tree was constructed by RAxML under the WAG+ Γ model. The representatives of the 6 supergroups are color-coded. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping. Branches that received maximum support by all methods are indicated by full circles.

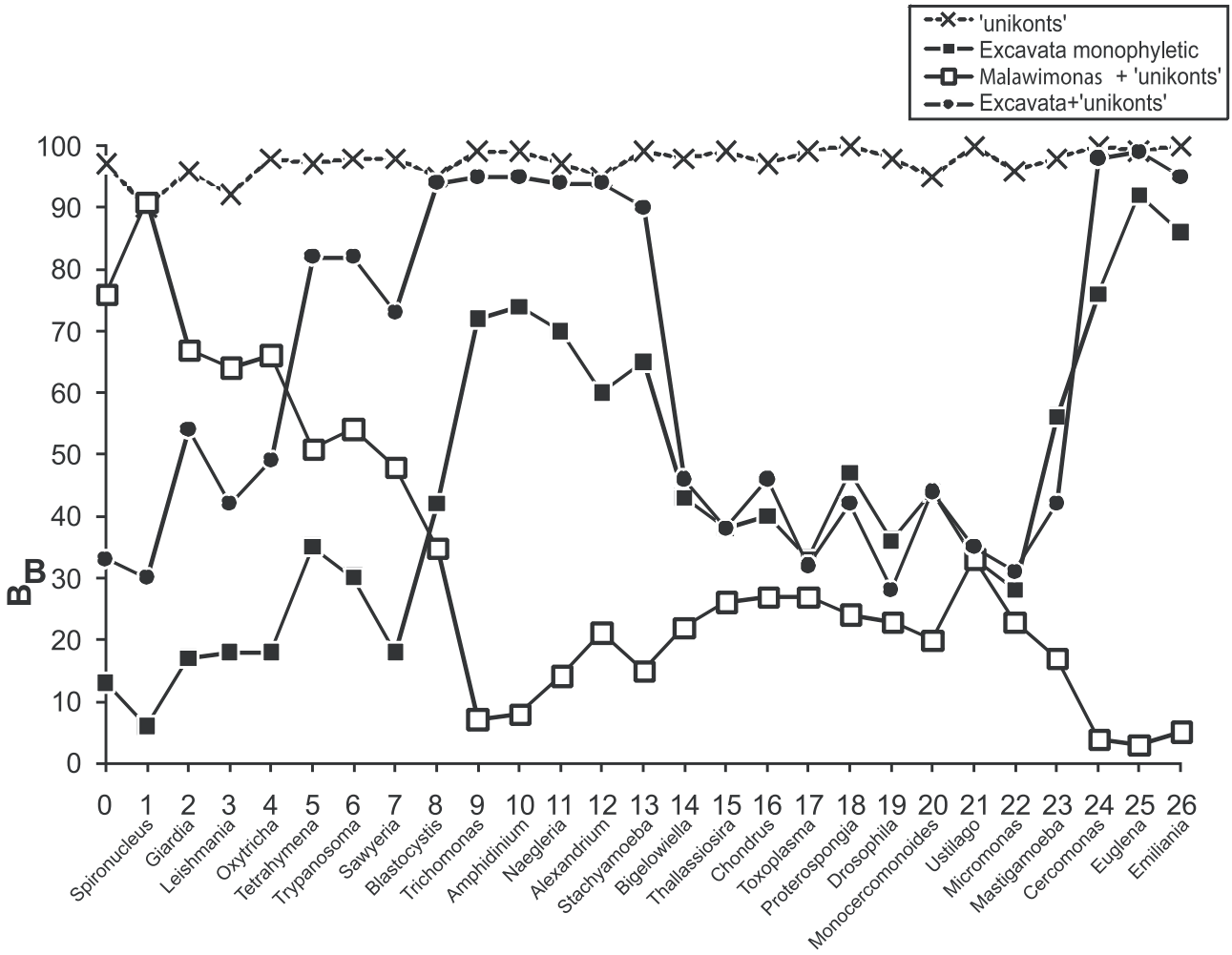

Fig. S7. A graph summarizing the results of LB taxa removal experiment using a diplomonad/trichomonad root. The support for the nodes of interest calculated by RAxML bootstrapping is plotted against the number of longest-branch taxa that were removed from the concatenate. The support for Unikonta (X) is used as a control. A root position at the midpoint of the branch connecting common branch of diplomonads and trichomonads with the rest of eukaryotes was used to calculate root-to-tip distances of taxa. The order of the removed taxa is given on the *x* axis.

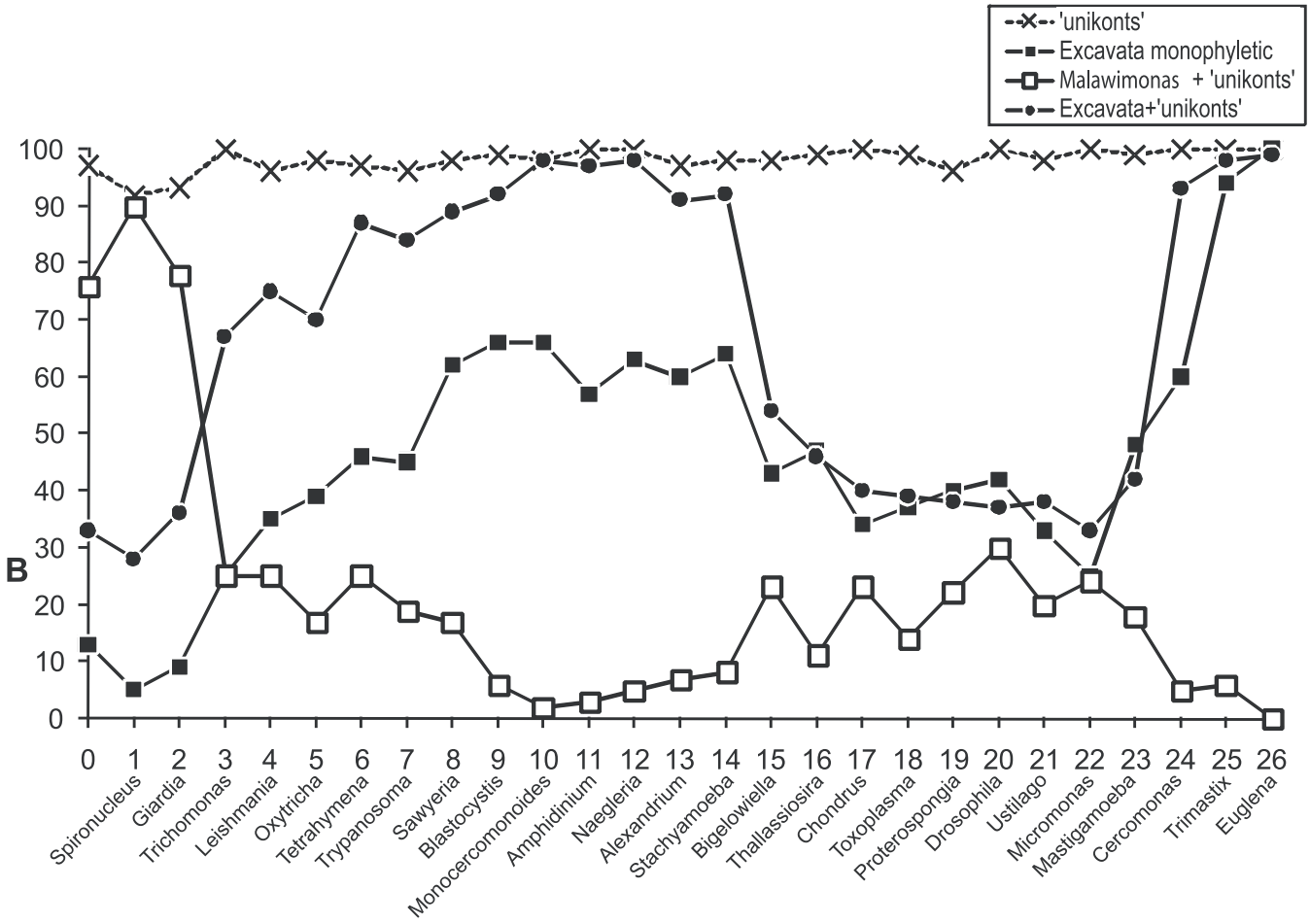

Fig. S8. A graph summarizing the results of LB taxa removal experiment using a corrected midpoint root. The support for the nodes of interest calculated by RAxML bootstrapping is plotted against the number of longest-branch taxa that were removed from the concatenate. The support for Unikonta (X) is used as a control. A root position at the corrected midpoint of the tree was used to calculate root-to-tip distances of taxa. The order of the removed taxa is given on the *x* axis.

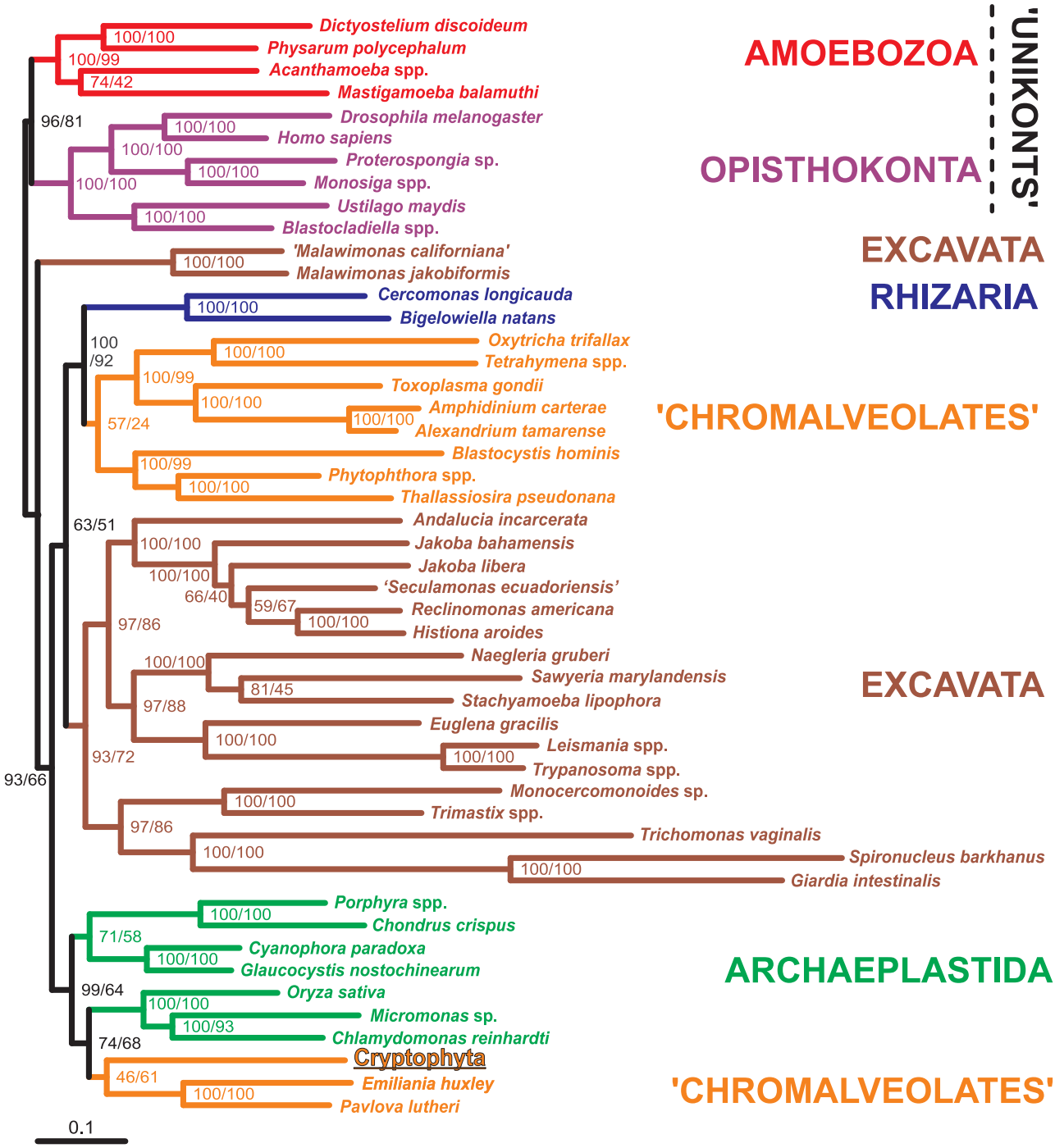

Fig. S9. A phylogenetic tree constructed from the data in which the sequences of Cryptophyta were included. The tree was constructed by RAxML under the WAG+ Γ model. The representatives of the 6 supergroups are color-coded. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping (uniform/separate model), branches that received maximum support are indicated by full circles.

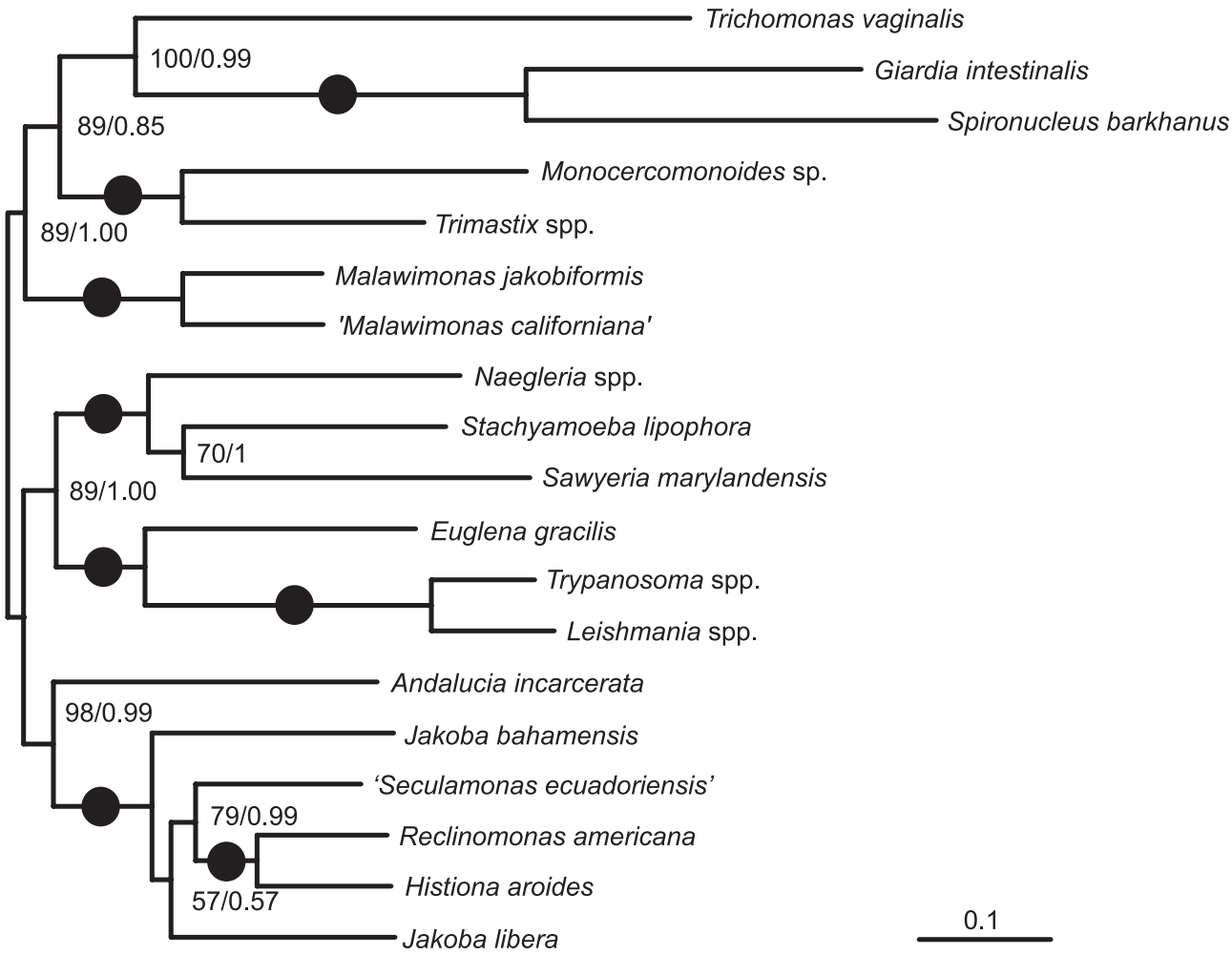

Fig. S10. Phylogenetic tree constructed from the set with only Excavata taxa. The tree was constructed by RAxML PROTCATWAG model. The numbers at the nodes indicate bootstrap support calculated by RAxML bootstrapping/PhyloBayes posterior probability. Branches that received maximum support by all methods are indicated by full circles.

Other Supporting Information Files

Table S1
Table S2
Table S3