MICROBIOLOGY

An excavate root for the eukaryote tree of life

Caesar Al Jewari* and Sandra L. Baldauf*

Much of the higher-order phylogeny of eukaryotes is well resolved, but the root remains elusive. We assembled a dataset of 183 eukaryotic proteins of archaeal ancestry to test this root. The resulting phylogeny identifies four lineages of eukaryotes currently classified as "Excavata" branching separately at the base of the tree. Thus, Parabasalia appear as the first major branch of eukaryotes followed sequentially by Fornicata, Preaxostyla, and Discoba. All four excavate branch points receive full statistical support from analyses with commonly used evolutionary models, a protein structure partition model that we introduce here, and various controls for deep phylogeny artifacts. The absence of aerobic mitochondria in Parabasalia, Fornicata, and Preaxostyla suggests that modern eukaryotes arose under anoxic conditions, probably much earlier than expected, and without the benefit of mitochondrial respiration.

INTRODUCTION

The major outlines of the eukaryote tree of life are beginning to coalesce, largely through a combination of multilocus phylogeny (phylogenomics) and the discovery of unique taxa filling in longstanding gaps in the tree (*1*). However, the root of the tree remains uncertain. One taxon set of particular interest is the so-called Excavata, a diverse collection of almost exclusively single-celled eukaryotes, many of which ingest their food via a deep (excavated) feeding groove (*2*). Taxa assigned to this group include Discoba, Fornicata, Preaxostyla, and Parabasalia (*3*). Discoba are mostly aerobic unicells, but with exceptionally diverse mitochondria and mitochondrial DNA (*4*). The remaining excavates are restricted to anaerobic or low-oxygen environments and have only what appear to be degenerate mitochondria-derived organelles [mitosomes or hydrogenosomes; (*5*)], if any at all (*6*). These "anaerobic excavates" or Metamonada (*1*, *3*) are strictly unicellular with diverse, often notable, morphologies and are likely to be major players in anoxic environments such as marine sediments, which are among the largest and least explored planetary ecosystems (*7*, *8*). The best studied anaerobic excavates are parasites, e.g., *Giardia* and *Trichomonas*, but the true diversity and distribution of the various excavate lineages are poorly understood (*9*). The phylogenetic position of these taxa is also poorly defined, and their presumed monophyly has not been tested in a rooted multigene tree.

Phylogenetic rooting most often relies on an outgroup. Because over half of the universal eukaryotic genes are derived from Bacteria or Archaea (*10*), the eukaryote tree can potentially be rooted with either archaeal or bacterial homologs. However, these genes are quite different. Universal eukaryotic genes of bacterial ancestry (euBacs) tend to be associated with mitochondria-related functions, while genes of archaeal ancestry (euArcs) tend to be involved in information processing such as replication, transcription, translation and protein modification, sorting, and degradation (*10*). In previous work, we showed that euBac phylogeny places Discoba as the sister group to all actively mitochondriate eukaryotes (*11*). However, these analyses excluded the anaerobic excavates, because they lack mitochondrial DNA and most mitochondrially targeted euBacs. By contrast, all eukaryotes have most euArcs, and therefore, we turned to euArc sequences to test the position of the anaerobic excavates in the eukaryote tree of life.

Evolutionary models are an integral part of molecular phylogeny. These models consist of parameters for substitution processes that are either estimated from the data or empirically estimated from highly curated external data [e.g., (*12*)]. One of the most influential recent advances in phylogenetic modeling is category profile mixture (CAT) models, which calculate the site likelihood for each alignment column as the weighted mean over all observed substitution patterns in the alignment. This has the attractive feature of showing a good fit to the data but at great expense in terms of demand on computation time and memory as well as involving a certain degree of circularity. Moreover, it has long been known that, at least for protein sequences, amino acid substitution patterns are primarily constrained by relatively simple patterns of protein secondary structure and solvent accessibility (*12*–*14*). These factors can be accounted for in phylogenetic analysis either with protein mixture models, which try to capture the structure based on site-wise amino acid composition, or with partition models, where site likelihood is calculated on the basis of a prior known structure. However, the latter is rarely used in practice because the relevant structural information is often unavailable or challenging to work with.

We have assembled a dataset of 183 euArc proteins and used these to explore the position of the eukaryote root. The data include broad and deep taxonomic sampling of eukaryotes and Archaea, including new and/or newly assembled public data for 31 excavate taxa. To facilitate analyses of the data, we developed a method to use predicted protein secondary structure and solvent accessibility to partition the data into the six main structure-solvent categories and then analyzed the data with substitution matrices specific for each site category. The model is simple and fast and provides a good fit to the data. Our analyses of a concatenation of the 183 euArc proteins using a variety of phylogenetic models, including the protein structure partition model, and various controls for deep phylogeny artifacts strongly and consistently support an excavate root for the eukaryote tree of life.

Program in Systematic Biology, Evolutionary Biology Center, Uppsala University, Uppsala SE-75236, Sweden.
*Corresponding author. Email: sandra.baldauf@ebc.uu.se (S.L.B.); cjewari@gmail.com (C.A.J.)

RESULTS

Eukaryotes can be roughly divided into six major divisions or supergroups. These are Amorphea (including animals and fungi), Diaphoretickes (including plants, most algae, and numerous and diverse eukaryotic microbes, saprobes, and parasites), one major lineage of aerobic excavates (Discoba), and three lineages of anaerobic excavates (Parabasalia, Preaxostyla, and Fornicata) (table S1) (*1*, *3*). Sequences for a broad taxon sampling of the six supergroups plus most major divisions of Archaea were extracted from public data for 183 euArc protein orthologs, identified by iterative screening of an intersection between the archaeal and eukaryotic clusters of orthologous genes (arCOG and KOD, respectively) databases (*15*). Because publicly available assembled data for excavates are sparse, we further augmented the excavate data with four in-house transcriptomes and 27 in-house assembled public sequence-read-archive (SRA) files (table S1).

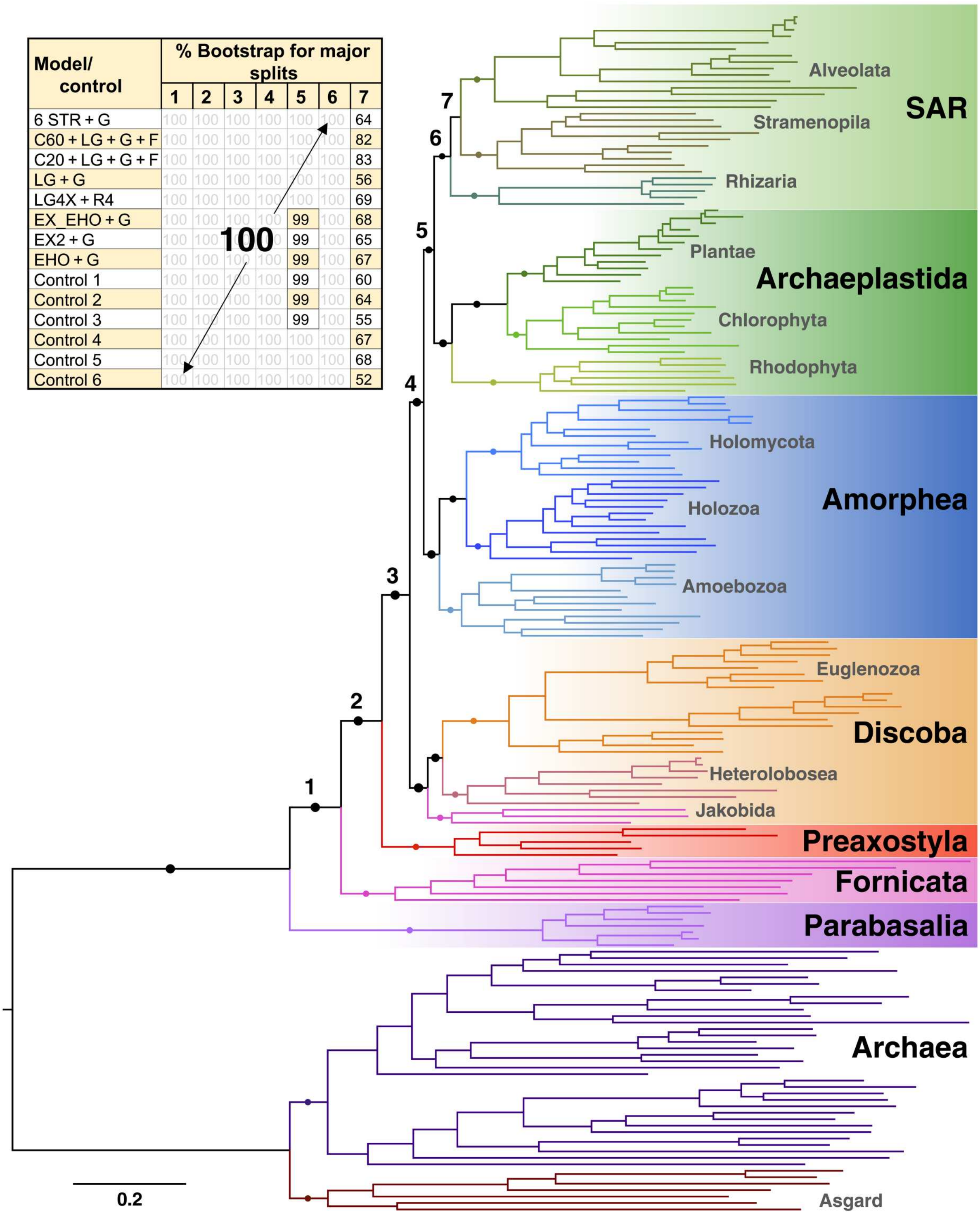
To facilitate the analysis of this large and complex dataset and to address a very deep evolutionary question, we developed a simplified procedure to use protein secondary structure to model the underlying evolutionary process. This procedure benefits from a recently developed deep learning method that predicts protein secondary structure and solvent accessibility with high speed and accuracy (*16*). In the procedure used here, individual protein structures are predicted based on full-length sequences for 10 taxa from across the dataset, with the resulting predicted structures mapped back to trimmed alignments. This allows the data to be partitioned, according to a majority consensus of the predicted structures, into six site categories corresponding to buried and exposed helices, sheets, and loops (data S1). The resulting partitioned matrix is then analyzed in a maximum likelihood framework using predetermined substitution matrices for each site category (*12*).

Phylogenetic analysis of a concatenation of the 183 euArc orthologs for 186 taxa (tables S1 to S3) produces a single well-resolved phylogeny of eukaryotes with 99 to 100% bootstrap support for all eukaryote supergroups [indicated by labels in Fig. 1, table S1, and (*3*)]. The rooted phylogeny places the four excavate lineages separately as the first four major branches in the eukaryote subtree, with 100% bootstrap support for each excavate branch point (Fig. 1 and figs. S1 to S3). The first of these branches is Parabasalia, which is then followed sequentially by Fornicata, Preaxostyla, and, lastly, Discoba as sister group to a clade of Amorphea + Diaphoretickes. This topology is fully reconstructed with a variety of evolutionary models, all with four gamma rate categories: the C20 and C60 profile mixture models (*17*), protein structure mixture models [EX2, EHO, and EX_EHO + G; (*12*)], and the protein structure–based partition model introduced here (Fig. 1, figs. S1 to S3, and data S1 and S2). In all cases, the trees show full support (100% bootstrap) for the four excavate branch points and all eukaryote supergroups with the exception of the internal branching order within Diaphoretickes subgroup SAR (Stramenopila + Alveolate + Rhizaria; Fig. 1).

A comparison of fit among the various evolutionary models shows that all models produce a comparable fit to these data, with structure-based models providing nearly as good a fit as the C20 and C60 profile mixture models. However, the structure models do this with a fraction of the complexity (number of categories) and run time, with the structure partition model introduced here being the simplest and fastest (Fig. 2). The relative simplicity of the partitioned structure models should also reduce the risk of overfitting (*18*), as well as having the obvious advantage of reducing the demand on analytical time and resources (Fig. 2). For example, full analyses of the 183-protein euArc data with the structure partition model required 46 Central Processing Unit (CPU) hours and 5 GB of memory. This is compared to more than 1000 hours and 106 GB of memory for analyses with the C20 mixture model and over 3200 hours and 338 GB for the C60 model on a 20-core supercomputer array with 500 GB of memory. The predicted structure partition model described here also obviates the need for solved protein structures, which are mostly available for only one or a few taxa per protein, if any at all. Thus, predicted structures can be generated for multiple taxa from across the data and then used to calculate a consensus predicted structure reflecting the full dataset. Moreover, the structural information only needs to be calculated once and then can be reused for additional analyses such as gene-wise or taxon-wise jackknifing or other controls.

The combined speed and accuracy of the predicted structure model allowed us to run a series of controls for two important artifacts potentially affecting deep phylogeny. The possibility of artifactual attraction of exceptionally long ingroup branches to a distant outgroup [long-branch attraction (LBA)] is an especially important consideration for a rooted tree. On the basis of simple inspection, the multiple-excavate root does not appear to reflect a series of LBA artifacts for the four excavate groups, as most excavates do not have especially long branches relative to other eukaryotes or the outgroup (Fig. 1). We tested this further by deleting the euArc proteins with the largest ingroup-outgroup distances in their individual trees [single-gene trees (SGTs)] using increasingly stringent cutoffs of <0.9, <0.8, and < 0.7 substitutions per site (controls 1 to 3, table S4). All three "long-branch–depleted" datasets yield essentially the same topology and support values for major nodes as the full dataset —maximal bootstrap support for the four excavate splits (Fig. 1 and fig. S2). Deleting the outgroup entirely leads to no change in ingroup topology or support, and outgroup-free rooting using the nonreversible amino acid model NONREV (*19*) in IQTREE2 (*20*) produces a multiexcavate root but without statistical support for the exact placement of the root within excavates (fig. S4) (*19*).

Another possible source of artifact would be the incorrect detection of paralogs, because early eukaryote evolution involved rampant gene duplication, and many euArcs are multicopy in eukaryotes (*21*). Thus, the early excavate branches seen here could reflect uniquely retained paralogs that were lost in other lineages, rather than species phylogeny. Paralogy was rigorously assessed throughout the initial orthology assignment phase by iterative phylogenetic screening with stepwise pruning of multigene families. These "potential orthologs" were then further screened for any signs of paralogy and other potential artifacts to arrive at the 183 euArc orthologs used to construct the rooted tree (Fig. 1 and Materials and Methods). Moreover, deep paralogs, if any remain, should appear as strongly supported deep branches in SGTs, but we found that strongly supported deep excavate branches are rare in the euArc SGTs (table S4 and data S3). Nonetheless, as a further control, we removed all proteins for which any excavate taxon, group or subgroup, appears as even a moderately supported deep branch in their SGT, using increasingly stringent cutoff values of >70, >60, and >50% bootstrap support (controls 4 to 6, table S4). This also controls for the possibility of a few proteins with a very strong individual phylogenetic signal overwhelming a widespread but weaker



Model/ control	% Bootstrap for major splits						
	1	2	3	4	5	6	7
6 STR + G	100	100	100	100	100	100	64
C60 + LG + G + F	100	100	100	100	100	100	82
C20 + LG + G + F	100	100	100	100	100	100	83
LG + G	100	100	100	100	100	100	56
LG4X + R4	100	100	100	100	100	100	69
EX_EHO + G	100	100	100	100	99	100	68
EX2 + G	100	100	100	100	99	100	65
EHO + G	100	100	100	100	99	100	67
Control 1	100	100	100	100	99	100	60
Control 2	100	100	100	100	99	100	64
Control 3	100	100	100	100	99	100	55
Control 4	100	100	100	100	100	100	67
Control 5	100	100	100	100	100	100	68
Control 6	100	100	100	100	100	100	52

Fig. 1. A rooted phylogeny of eukaryotes based on eukaryotic proteins of archaeal descent. A concatenated alignment of 183 proteins with 45,443 aligned positions and 85% overall data occupancy (tables S1 and S2 and data S5) was analyzed by maximum likelihood with eight different evolutionary models. The tree shown was derived using the deduced protein structure partitioned model (6 STR + G). Solid circles indicate nodes with 100% bootstrap support from all models and controls, and branch lengths are drawn to scale as indicated by the scale bar. Bootstrap support values for the major nodes are shown in the table at the top. Controls 1 to 3 used stepwise reduction in ingroup-outgroup distances, while controls 4 to 6 used stepwise reduction in individual tree support for early branching excavates (table S4). Taxon names and all bootstrap values for the 6 STR + G tree are shown in fig. S1.

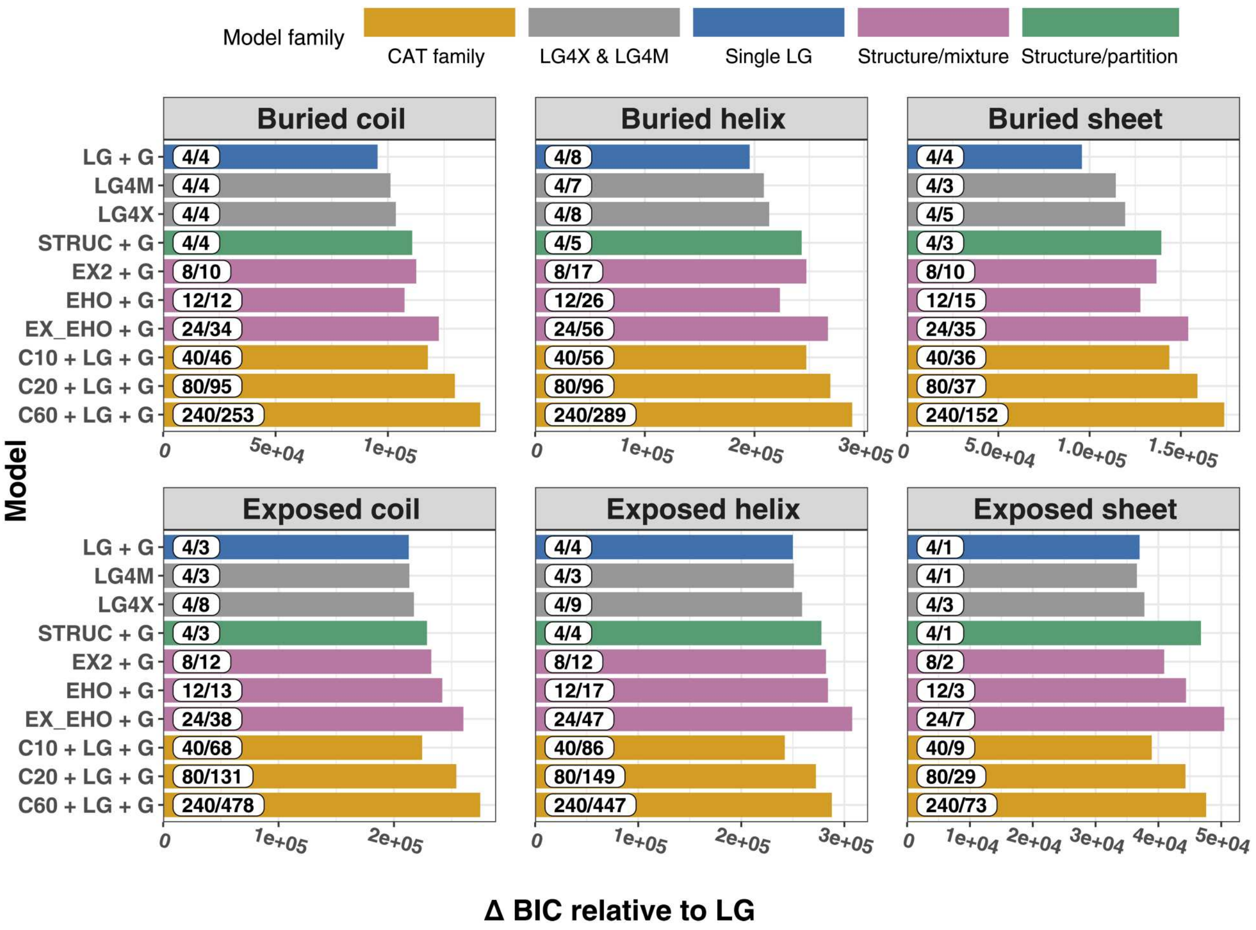


Fig. 2. Comparative fit of evolutionary models for the six main protein structure elements. Goodness of fit based on Bayesian information criteria [BIC; (*40*)] was calculated using model fit [IQTREE.v1.6.12; (*41*)] on a 183-protein alignment partitioned into six predicted structural elements using NetSurfP-3.0 (*16*) and evaluated on the best global tree (Fig. 1). Bars show improvement in BIC scores for various models relative to the LG model (without gamma) and are colored according to their general model type as indicated by the key to the right. The analyses were run on a 20-core CPU and include optimizing branch lengths and model parameters. The numbers of categorical mixture components for each model and CPU time in minutes are shown at the far left of each bar, to the left and right of a slash, respectively. All phylogenetic analyses of the 183 euArc data with these models produce the same tree (Fig. 1 and data S2). Raw values are provided in table S5.

opposing signal, which can occur even with large multilocus datasets (*22*). Again, all three controls reconstruct all major eukaryote groups with 99 to 100% boostrap support including 100% support for the four excavate splits (Fig. 1 and fig. S3).

DISCUSSION

Phylogenomic analyses using diverse evolutionary models, including a fast and well-fit protein structure partition model (Fig. 2), place the four major lineages of excavate eukaryotes as the first four branches of the eukaryote tree of life (Fig. 1 and fig. S1). Together, these proteins represent the core of eukaryotic and archaeal information processing plus diverse other cellular processes (tables S2 and S3). Controls for deep phylogeny artifacts, particularly LBA and deep paralogy, produce no decrease in support for this topology (figs. S1 to S3). Thus, rooted phylogenies based on proteins of both archaeal (Fig. 1) and bacterial (*11*) descent identify excavates as the earliest branches of eukaryotes. The fact that the first three of these branches are essentially modern cells but lacking aerobic mitochondria suggests, among other things, that eukaryotic complexity arose without the benefit of mitochondrial respiration.

The euArc phylogeny is notably robust with strong resolution of nearly all well-established major divisions of eukaryotes (Fig. 1), consistent with a variety of analyses based on different genes, taxa, methods, and models (*1*, *3*). The tree is also unusually symmetrical, with the major groups showing roughly similar distances from the root. This is especially true if one disregards the two *Giardia* species (Fornicata), which have the long terminal branches typical of parasites. Although the tree lacks an outgroup for Archaea, the topology is also consistent with the Asgard archaea as sister to eukaryotes (*23*). The one poorly resolved lineage is the Diaphoretickes assemblage referred to as SAR (*24*), which we found was also unstable in euBac phylogeny (*11*). The branching order within SAR is an important question, given that the group includes most of the marine algae and major parasites of animals and crop plants. The reason for the group's instability in these trees is not clear but may be better addressed in analyses without a distant outgroup.

All the models that we use in this study reconstruct the same phylogeny and with nearly identical support values for all major nodes (Fig. 1). Nonetheless, the partitioned structure model has a number of advantages. First, it provides model fitness close to that of more complex models without exploding the parameter space or the mixture of components. This makes the model more efficient and avoiding the issue of circularity. Second, the current dataset is very robust, so that any reasonably accurate model reproduces the same tree (Fig. 1). However, this would not necessarily be the case with smaller datasets (fewer genes or taxa) or more challenging phylogenetic problems, e.g., very distantly or very closely related sequences or taxa, where site-wise sequence information may be insufficient to accurately model protein structure. Thus, it remains to be determined whether the structure mixture models and structure partition model perform similarly with various quality data. There is also the question of the carbon footprint of highly demanding computational analyses, which are hard to justify if simpler models of similar accuracy are available. Perhaps simpler models can also be more readily combined with other models such as corrections for variation among lineages, either in terms of substitution rates (*25*) or other factors. These are likely major impediments to accurate phylogenetic reconstruction at many levels. Further refinement of the structure partition method might be gained by including weaker physical factors such as aromaticity, charge, and rotational isomer state (*26*–*28*). However, no fast and easy method that integrates all these factors is now available, and the possible added value would need to be evaluated.

It is important to distinguish the secondary structure partitioning model that we present here from partitioning by locus (gene or protein). The latter is essentially partitioning by tertiary structure, which is a fairly rough level of partitioning and does not appear to contribute strongly to phylogenetic accuracy or may even hinder it (*29*). Nor is our protocol similar to the automated partitioning method of PartitionFinder (*30*), which works by stepwise merging of loci where such mergers lead to higher Bayesian information criterion (BIC) score. Partitioning by locus or even combined loci fundamentally differs from the fine-level secondary structure partitioning used here. Moreover, protein secondary structure and solubility exposure have been explicitly shown to strongly constrain amino acid substitution patterns throughout the protein, unlike tertiary structure (*12*, *13*).

Many potentially interesting taxa are still missing from the current euArc tree. This is primarily because they lacked sufficient data for adequate prescreening at the time of these analyses. Screening is particularly problematic for intriguing but newly described species where genomic data tend to be limited and/or of mixed quality. This makes it difficult to identify orthologs with confidence, which is critical here. We have also omitted phylogenetically challenging taxa such as Cryptophytes and Haptophytes, which are probable members of Diaphoretickes, possibly closely related to Archaeplastida. While these taxa are interesting in their own right, they are not directly relevant to the question addressed here (*11*). We also do not include the phylogenetically problematic Malawimonads, for which there is a lack of sufficient quality data. In future, it would be especially interesting to include these taxa, as they are often, if sporadically, assigned to Excavata.

An anaerobic excavate root raises interesting questions regarding the nature of the last eukaryote common ancestor (LECA) and the origin of mitochondria. If LECA had a respiratory-competent mitochondrion, as is widely held, then an early ancestor of each of the three anaerobic excavate lineages would have had to migrate independently to a low-oxygen environment. Meanwhile, each lineage would also have had to retain at least one fully mitochondriate branch that remained extant long enough to give rise to the next surviving split in the tree. However, there is now no evidence of any aerobic branch in any of the three anaerobic excavate groups. Each anaerobic excavate lineage would also have had to independently reduce their mitochondrion to nonrespiring hydrogenosomes or mitosomes [mitochondria-related organelles (MROs); (*31*)]. The latter at least is not, together, unlikely as multiple examples of such reductions have been documented in other eukaryotes (*5*, *32*). However, a theoretically simpler explanation would be that the LECA simply had an MRO, most likely a hydrogenosome (*33*), and that mitochondrial respiration arose later, sometime after the divergence of Preaxostyla and before the emergence of Discoba (Fig. 3). This would suggest that aerobic mitochondria arose by a separate endosymbiosis from that which gave rise to hydrogenosomes.

Such a "serial endosymbiotic" scenario is consistent with the mixed ancestry of euBacs in general and especially mitochondrial proteins, only a small fraction of which trace to α-proteobacteria (*34*). A late advent of respiration is also consistent with the often much higher sequence conservation of eukaryotic genes of α-proteobacterial descent compared to genes tracing to other bacteria (*11*, *32*). A relatively late origin of respiration would also help explain the unique presence of notably α-proteobacterial like mitochondria genomes in the Jakobida [Discoba; (*35*)]. Given that the most common donors of euBacs are α-proteobacteria followed by δ-proteobacteria (*32*), while the second most common donor of mitochondrial proteins is γ-proteobacteria (*11*, *36*), the simplest scenario would be a γ- or δ-proteobacterial endosymbiosis followed by an α-proteobacterial one (Fig. 3). Alternatively, there may have been multiple endosymbioses of varying success both before and

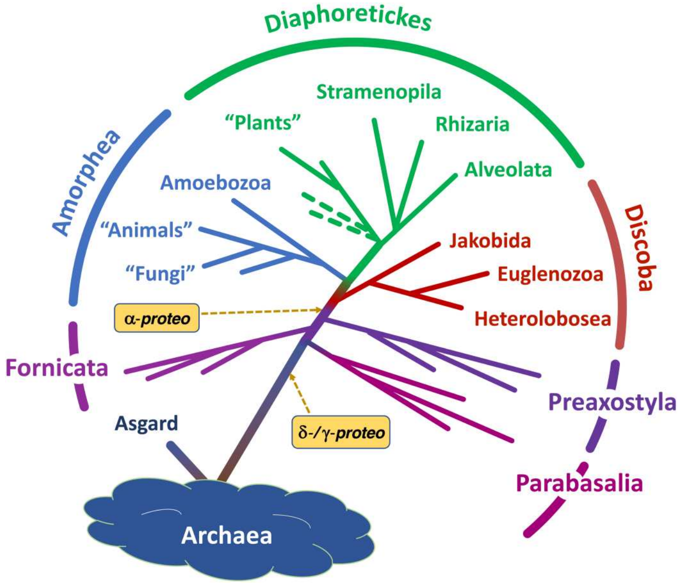


Fig. 3. A proposed stepwise scenario for the origin of mitochondria and mitochondria-like organelles. A schematic version of the rooted euArc phylogeny in Fig. 1 is shown with two proposed endosymbiotic events, the earlier most likely involving a γ- and/or δ-proteobacterium (γ-/δ-*proteo*) and the second an α-proteobacterium (α-*proteo*).

after LECA, as previously suggested (*32*). It should be noted that these scenarios share similarities with the archaezoa hypothesis of Cavalier-Smith (*37*), although that was later abandoned after the discovery of MROs and tree reconstruction artifacts (*38*).

Excluding taxa from the euArc tree means that we cannot rule out the possibility that one or more of these taxa, or other yet-undiscovered taxa, may represent earlier branches, given that much eukaryote diversity remains unknown (*3*, *7*). However, no addition of taxa will change the fundamental relationship described here, i.e., that the earliest branches of extant eukaryotes include multiple anaerobic lineages with predominantly excavate morphology. The implications of this are profound. For example, eukaryotic cellular and molecular complexity probably predate mitochondrial respiration. Modern eukaryotes could also have arisen before the great oxygen event (*4*), which is consistent with recent molecular dating (*5*, *39*). Eukaryotes probably also had an excavate morphology for much of their early history, and this morphology may have formed the basis for other morphological innovations. However, it is important to note that an excavate morphology is so far unknown for Parabasalia, which were primarily assigned to Excavata based on unrooted trees. Thus, this enigmatic taxon may be a key to understanding eukaryote origins and the nature of LECA and the forces that shaped it.

MATERIALS AND METHODS

Initial euArc protein selection and dataset assembly

A dataset of universal eukaryotic proteins of archaeal descent (euArcs) was assembled beginning with 719 proteins identified as shared by Archaea and Eukarya based on a cross section of the arCOG and KOD databases (*15*). Individual datasets were assembled for each of the 719 proteins by BLASTp search of the GenBank nr data for a taxonomically broad set of Archaea, Eukarya, and Bacteria. Datasets were further augmented by BLASTp search of in-house data for four additional excavate taxa and in-house assemblies of an additional 27 excavates from publicly available SRA files (table S1) using Trinity version 2.13.2 (*40*). The four in-house excavate transcriptomes were sequenced from mono-eukaryote cultures as previously described (*41*, *42*). Taxonomic identifiers were prepended to sequence names as shown in table S1. Sequence searches used BLASTp with a relaxed *e* value cutoff of e-05 to extract all possible paralogs. The resulting sequence sets were then ranked on the basis of their taxonomic coverage across Archaea and Eukarya, and 456 protein sets were identified as having wide taxon coverage in both domains.

Ortholog detection

Each of the 456 candidate proteins was aligned individually using Mafft with FFT-NS-i algorithm and --maxiterate 1000 (*43*) and then trimmed using TrimAl auto1 to remove regions of ambiguous alignment (*44*). Each protein set was then screened for multigene families (paralogs) with maximum likelihood and the Shimandaro-Hasegawa (SH) test as implemented in FastTree 2.1 (*45*). Proteins showing non- or weakly monophyletic eukaryotes were deleted, while proteins showing multiple, well-separated, strongly supported eukaryote-wide clades were broken down into single candidate orthologs. Candidate orthologs were defined as highly distinct clades (long subtending branch) that were also out-paralog-free and showed strong and consistent support for eukaryote monophyly (SH > 0.95). This resulted in 441 candidate orthologs. Bitscores were then recalculated for each sequence for each candidate ortholog by BLASTp search using two eukaryote query sequences from within each ortholog candidate against the rest of its sequences with the highest score retained. Each candidate ortholog was further screened for in-paralogs by iterative phylogenetic analysis using IQTREE (*46*), with command arguments as indicated in data S4. In-paralogs, i.e., multiple copies in the same or closely related taxa (same genus), were reduced to a single sequence according to the following criteria: (i) lowest recalculated bitscore, (ii) fewest gaps after trimming, and/or (iii) shortest terminal branch in their IQTREE inference.

Ortholog assessment

Candidate orthologs were further subjected to multiple rounds of screening and filtering by phylogenetic inference using IQTREE with LG + G4 and ultrafast bootstraps. This resulted in the elimination of proteins that failed one or more of the following criteria: (i) failure to support eukaryote monophyly with ≥90% bootstrap support, (ii) Eukarya closer to Bacteria than Archaea, (iii) strong disagreement (≥70% bootstrap) with accepted eukaryote phylogeny, i.e., one or more taxa falling within a noncanonical major clade [listed as kingdoms in table S1 and (*3*)], (iv) lack of taxa for more than two major subgroups of Archaea, (v) lack of taxa for any of the eukaryote superkingdoms [table S1 and (*3*)], or (vi) one or more major groups showing overall substantially lower bitscores than the remaining eukaryotes groups.

The final rounds of assessment were conducted using IQTREE as above, after removing the bacterial sequences. Any remaining eukaryotic within-kingdom paralogs (in-paralogs) in terminal or near-terminal clades in the resulting trees were reduced to the single shortest-branched sequence. In the case of any remaining nonterminal paralogs, the protein in question was deemed as a suspect of hidden paralogy and discarded. Sequences with apparent xenologs, contaminant data, or sequences producing excessively long branches were also removed. This was followed by a repeat round of screening for the five criteria listed above. The final result is a set of 183 individual ortholog datasets.

A final set of alignments was produced for each of the 183 orthologs using mafft with L-INS-i algorithm --maxiterate 1000 (*43*) and trimmed using TrimAl auto1 to remove regions of ambiguous alignment (*44*). The alignments were then further trimmed to remove alignment columns with more than 80% missing data in either Eukarya or Archaea and subjected to maximum likelihood and rapid bootstraps analysis with RAxML (*47*) under the LGPROTCAT model. The resulting final set of individual trees (SGTs; data S3) was used as the basis for six control analyses (table S4). Command arguments for tree inferences are given in data S4.

Phylogenetic inference, models, and model fitness

Maximum likelihood trees were constructed for the full concatenated 183 protein alignment (supermatrix) with and without data partitioning. Unpartitioned supermatrix analyses were conducted using five method-model combinations as follows: (i) LG model with four gamma rate categories using RAxML with rapid bootstraps (*47*) and (ii) IQTREE (*46*) with LG4X model (*48*) and ultrafast bootstraps (*49*); (iii) C20 and (iv) C60 profile mixture models (*17*) using IQTREE with the LG + C20 + G4 + F and LG + C60 + G4 models, respectively, and ultrafast bootstraps; and (v) protein structure mixture models with EX_EHO + G4, EX2 + G4, EHO + G4 (*12*) using IQTREE and ultrafast bootstraps. Trees are summarized in Fig. 1 and provided in newick format in data S2. All IQTREE ultrafast bootstraps used the optimization flag –bnni for a more thorough search of tree space.

Protein structure partition model

Protein structure partitioning consists of five steps: (i) structure prediction, (ii) structure consensus calculation, (iii) consensus mapping onto trimmed alignments, (iv) alignment partitioning by consensus structure, and (v) phylogenetic analysis. Plurality consensus structures are determined for each protein with NetSurfP-3.0 (*16*) using full sequences for 10 species from across the tree. The 10 predicted structures are then used to calculate a plurality consensus predicted structure site-wise for each protein, which is then mapped onto the full trimmed alignment for that protein. Last, each protein alignment is partitioned into the six predicted structure categories —buried and exposed helices, sheets, and loops—and subjected to phylogenetic analysis using corresponding structure-based exchangeability matrices (*12*).

For the 183 euArc proteins, secondary structures were predicted for four Archaea and six eukaryotes, with taxa selected on the basis of least amount of gap positions in the trimmed alignments (data S1). In terms of predicted structure, the resulting consensus alignments show 95% match of the individual predicted structures to the consensus, and 80% of the alignment columns have strict consensus. For solubility exposure, 93% of residues match the consensus and 74% of columns have strict consensus (data S1). Phylogenetic analyses of the partitioned data were run using six structure-based exchangeability matrices and frequencies (*12*) in IQTREE 1.6.12 with the -spp option, which assigns separate gamma rate categories to each partition. Ultrafast bootstrap analysis (-bb 1000) was optimized as above. The model was used for analyses of the full dataset and six control analyses (table S4). The resulting trees are provided in newick format in data S2.

Comparison of model fitness, complexity, and run times

Model fitness comparison tests were performed for each of the six protein structure categories using IQTREE model finder (*50*) and the BIC scores. Run times are the CPU time in minutes required to optimize the likelihood and branch lengths of the best global tree (Fig. 1) under each model as reported in the IQTREE log file for each run. Raw fitness scores for BIC, Akaike information criterion (AIC), AIC corrected for small sample size (AICc), and the log-likelihood values and CPU time are provided in table S5.

Supplementary Materials

This PDF file includes:
Figs. S1 to S4
Legends for tables S1 to S5
Legends for data S1 to S6

Other Supplementary Material for this manuscript includes the following:
Tables S1 to S5
Data S1 to S6

REFERENCES AND NOTES

1. F. Burki, A. J. Roger, M. W. Brown, A. G. B. Simpson, The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).
2. A. G. B. Simpson, Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *Int. J. Syst. Evol. Microbiol.* **53**, 1759–1777 (2003).
3. S. M. Adl, D. Bass, C. E. Lane, J. Lukeš, C. L. Schoch, A. Smirnov, S. Agatha, C. Berney, M. W. Brown, F. Burki, P. Cárdenas, I. Čepička, L. Chistyakova, J. del Campo, M. Dunthorn, B. Edvardsen, Y. Eglit, L. Guillou, V. Hampl, A. A. Heiss, M. Hoppenrath, T. Y. James, A. Karnkowska, S. Karpov, E. Kim, M. Kolisko, A. Kudryavtsev, D. J. G. Lahr, E. Lara, L. Le Gall, D. H. Lynn, D. G. Mann, R. Massana, E. A. D. Mitchell, C. Morrow, J. S. Park, J. W. Pawlowski, M. J. Powell, D. J. Richter, S. Rueckert, L. Shadwick, S. Shimano, F. W. Spiegel, G. Torruella, N. Youssef, V. Zlatogursky, Q. Zhang, Revisions to the classification, nomenclature, and diversity of eukaryotes. *J. Eukaryot. Microbiol.* **66**, 4–119 (2019).
4. G. Burger, M. Valach, Perfection of eccentricity: Mitochondrial genomes of diplonemids. *IUBMB Life* **70**, 1197–1206 (2018).
5. M. Van Der Giezen, Hydrogenosomes and Mitosomes: Conservation and evolution of functions. *J. Eukaryot. Microbiol.* **56**, 221–231 (2009).
6. A. Karnkowska, V. Vacek, Z. Zubáčová, S. C. Treitli, R. Petrželková, L. Eme, L. Novák, V. Žárský, L. D. Barlow, E. K. Herman, P. Soukal, M. Hroudová, P. Doležal, C. W. Stairs, A. J. Roger, M. Eliáš, J. B. Dacks, Č. Vlček, V. Hampl, A eukaryote without a mitochondrial organelle. *Curr. Biol.* **26**, 1274–1284 (2016).
7. T. Cordier, I. B. Angeles, N. Henry, F. Lejzerowicz, C. Berney, R. Morard, A. Brandt, M.-A. Cambon-Bonavita, L. Guidi, F. Lombard, P. M. Arbizu, R. Massana, C. Orejas, J. Poulain, C. R. Smith, P. Wincker, S. Arnaud-Haond, A. J. Gooday, C. de Vargas, J. Pawlowski, Patterns of eukaryotic diversity from the surface to the deep-ocean sediment. *Sci. Adv.* **8**, eabj9309 (2022).
8. T. Hoshino, H. Doi, G.-I. Uramoto, L. Wörmer, R. R. Adhikari, N. Xiao, Y. Morono, S. D'Hondt, K.-U. Hinrichs, F. Inagaki, Global diversity of microbial communities in marine sediment. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 27587–27597 (2020).
9. M. Kolisko, J. D. Silberman, I. Cepicka, N. Yubuki, K. Takishita, A. Yabuki, B. S. Leander, I. Inouye, Y. Inagaki, A. J. Roger, A. G. B. Simpson, A wide diversity of previously undetected free-living relatives of diplomonads isolated from marine/saline habitats. *Environ. Microbiol.* **12**, 2700–2710 (2010).
10. J. A. Cotton, J. O. McInerney, Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17252–17255 (2010).
11. C. Al Jewari, S. L. Baldauf, Conflict over the eukaryote root resides in strong outliers, mosaics and missing data sensitivity of site-specific (CAT) mixture models. *Syst. Biol.*, syac029 (2022).
12. S. Q. Le, O. Gascuel, Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. *Syst. Biol.* **59**, 277–287 (2010).
13. J. L. Thorne, N. Goldman, D. T. Jones, Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13**, 666–673 (1996).
14. N. Goldman, J. L. Thorne, D. T. Jones, Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458 (1998).
15. N. Yutin, K. S. Makarova, S. L. Mekhedov, Y. I. Wolf, E. V. Koonin, The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **25**, 1619–1630 (2008).
16. M. H. Høie, E. N. Kiehl, B. Petersen, M. Nielsen, O. Winther, H. Nielsen, J. Hallgren, P. Marcatili, NetSurfP-3.0: Accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Res.* **50**, W510–W515 (2022).
17. S. Q. Le, O. Gascuel, N. Lartillot, Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
18. E. S. Allman, J. A. Rhodes, S. Sullivant, When do phylogenetic mixture models mimic other phylogenetic models? *Syst. Biol.* **61**, 1049–1059 (2012).
19. S. Naser-Khdour, B. Quang Minh, R. Lanfear, Assessing confidence in root placement on phylogenies: An empirical study using nonreversible models for mammals. *Syst. Biol.* **71**, 959–972 (2022).
20. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
21. J. Vosseberg, J. J. E. van Hooff, M. Marcet-Houben, A. van Vlimmeren, L. M. van Wijk, T. Gabaldón, B. Snel, Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat. Ecol. Evol.* **5**, 92–100 (2021).
22. X. X. Shen, C. T. Hittinger, A. Rokas, Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat. Ecol. Evol.* **1**, 1–10 (2017).
23. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. Bäckström, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

24. J. F. H. Strassert, M. Jamy, A. P. Mylnikov, D. V. Tikhonenkov, F. Burki, New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* **36**, 757–765 (2019).
25. S. M. Crotty, B. Q. Minh, N. G. Bean, B. R. Holland, J. Tuke, L. S. Jermiin, A. V. Haeseler, GHOST: Recovering historical signal from heterotachously evolved sequence alignments. *Syst. Biol.* **69**, 249–264 (2020).
26. O. Soyer, M. W. Dimmic, R. R. Neubig, R. A. Goldstein, Using evolutionary methods to study G-protein coupled receptors. *Pac. Symp. Biocomput.*, 625–636 (2002).
27. M. W. Dimmic, D. P. Mindell, R. A. Goldstein, Modeling evolution at the protein level using an adjustable amino acid fitness model, in *Biocomputing 2000* (World Scientific, 1999), pp. 18–29.
28. U. Perron, A. M. Kozlov, A. Stamatakis, N. Goldman, I. H. Moal, Modeling structural constraints on protein evolution via side-chain conformational States. *Mol. Biol. Evol.* **36**, 2086–2103 (2019).
29. E. Susko, A. J. Roger, Long branch attraction biases in phylogenetics. *Syst. Biol.* **70**, 838–843 (2021).
30. R. Lanfear, B. Calcott, S. Y. W. Ho, S. Guindon, PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701 (2012).
31. P. Peña-Diaz, J. Lukeš, Fe–S cluster assembly in the supergroup Excavata. *J. Biol. Inorg. Chem.* **23**, 521–541 (2018).
32. T. M. Embley, Multiple secondary origins of the anaerobic lifestyle in eukaryotes. *Philos. Trans. R Soc. B Biol. Sci.* **361**, 1055–1067 (2006).
33. M. M. Leger, M. Kolisko, R. Kamikawa, C. W. Stairs, K. Kume, I. Čepička, J. D. Silberman, J. O. Andersson, F. Xu, A. Yabuki, L. Eme, Q. Zhang, K. Takishita, Y. Inagaki, A. G. B. Simpson, T. Hashimoto, A. J. Roger, Organelles that illuminate the origins of Trichomonas hydrogenosomes and Giardia mitosomes. *Nat. Ecol. Evol.* **1**, 1–7 (2017).
34. M. W. Gray, Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 10133–10138 (2015).
35. G. Burger, M. W. Gray, L. Forget, B. F. Lang, Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol. Evol.* **5**, 418–438 (2013).
36. A. A. Pittis, T. Gabaldón, Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).
37. T. Cavalier-Smith, Eukaryotes with no mitochondria. *Nature* **326**, 332–333 (1987).
38. J. Tovar, G. León-Avila, L. B. Sánchez, R. Sutak, J. Tachezy, M. van der Giezen, M. Hernández, M. Müller, J. M. Lucocq, Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. *Nature* **426**, 172–176 (2003).
39. J. F. H. Strassert, I. Irisarri, T. A. Williams, F. Burki, A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).
40. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
41. C. J. Fu, S. Sheikh, W. Miao, S. G. E. Andersson, S. L. Baldauf, Missing genes, multiple ORFs, and C-to-U type RNA editing in Acrasis kona (Heterolobosea, Excavata) mitochondrial DNA. *Genome Biol. Evol.* **6**, 2240–2257 (2014).
42. D. He, C. J. Fu, S. L. Baldauf, Multiple origins of eukaryotic cox15 suggest horizontal gene transfer from bacteria to Jakobid mitochondrial DNA. *Mol. Biol. Evol.* **33**, 122–133 (2016).
43. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
44. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
45. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010).
46. L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
47. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
48. S. Q. Le, C. C. Dang, O. Gascuel, Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **29**, 2921–2936 (2012).
49. D. T. Hoang, O. Chernomor, A. von Haeseler, B. Q. Minh, L. S. Vinh, UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
50. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

Acknowledgments: We thank C. Fu for help with excavate cell culture and sequencing, U. Perron and N. Goldman for helpful discussion of protein models, J. Silberman for the *Paratrimastix* culture, A. Simpson for the *Andalucia* and *Seculamonas* cultures, M. Brown for the *Acrasis* culture, and the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for computational resources (project snic2021-5-353). **Funding:** This work was supported by the Swedish Research Council (Vetenskapsrådet) under grant VR 2017-04351 (S.L.B.) and by the Department of Organismal Biology, Uppsala University (C.A.J.). **Author contributions:** Conceptualization: S.L.B. Methodology: C.A.J. and S.L.B. Formal analysis: C.A.J. Data curation: C.A.J. Validation: C.A.J. and S.L.B. Resources: S.L.B. Investigation: S.L.B. and C.A.J. Visualization: S.L.B. and C.A.J. Funding acquisition: S.L.B. Project administration: S.L.B. Supervision: S.L.B. Writing—original draft: S.L.B. and C.A.J. Writing—review and editing: S.L.B. and C.A.J. Competing interests: The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials, which include the supermatrix, control alignments, the resulting trees in newick format and the full set of raw sequences for each euArc in fasta format (data S1 to S6). New sequences generated for this study have been submitted to GenBank as follows: *Andalucia godoyi* (accession nos. OP718981 to OP719152), *Seculamonas equadoriensis* (accession nos. OP760301 to OP760474), and *Paratrimastix eleionoma* (accession nos. OP718805 to OP718980) (table S1).

Submitted 19 August 2022
Accepted 29 March 2023
Published 28 April 2023
10.1126/sciadv.ade4973