Bacterial proteins pinpoint a single eukaryotic root

Romain Derelle[a,b,1], Guifré Torruella[c], Vladimír Klimeš[d], Henner Brinkmann[e], Eunsoo Kim[f], Čestmír Vlček[g], B. Franz Lang[h], and Marek Eliáš[d]

[a]Centre for Genomic Regulation, 08003 Barcelona, Spain; [b]Universitat Pompeu Fabra, 08003 Barcelona, Spain; [c]Institut de Biologia Evolutiva, Consejo Superior de Investigaciones Científicas–Universitat Pompeu Fabra, 08003 Barcelona, Spain; [d]Faculty of Science, Department of Biology and Ecology, University of Ostrava, 710 00 Ostrava, Czech Republic; [e]Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, D-38124 Braunschweig, Germany; [f]Sackler Institute for Comparative Genomics and Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024; [g]Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, 142 20 Prague 4, Czech Republic; and [h]Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, Montreal, QC, Canada H3T 1J4

Edited by Thomas Martin Embley, University of Newcastle upon Tyne, Newcastle upon Tyne, United Kingdom, and accepted by the Editorial Board January 13, 2015 (received for review October 28, 2014)

The large phylogenetic distance separating eukaryotic genes and their archaeal orthologs has prevented identification of the position of the eukaryotic root in phylogenomic studies. Recently, an innovative approach has been proposed to circumvent this issue: the use as phylogenetic markers of proteins that have been transferred from bacterial donor sources to eukaryotes, after their emergence from Archaea. Using this approach, two recent independent studies have built phylogenomic datasets based on bacterial sequences, leading to different predictions of the eukaryotic root. Taking advantage of additional genome sequences from the jakobid *Andalucia godoyi* and the two known malawimonad species (*Malawimonas jakobiformis* and *Malawimonas californiana*), we reanalyzed these two phylogenomic datasets. We show that both datasets pinpoint the same phylogenetic position of the eukaryotic root that is between "Unikonta" and "Bikonta," with malawimonad and collodictyonid lineages on the Unikonta side of the root. Our results firmly indicate that (*i*) the supergroup Excavata is not monophyletic and (*ii*) the last common ancestor of eukaryotes was a biflagellate organism. Based on our results, we propose to rename the two major eukaryotic groups Unikonta and Bikonta as Opimoda and Diphoda, respectively.

eukaryote phylogeny | phylogenomics | Opimoda | Diphoda | LECA

The root of eukaryotes refers to the deepest node in the eukaryote crown group: i.e., to a node that separates the two monophyletic groups resulting from the first cladogenesis event of all extant eukaryotes. Knowing the position of the eukaryotic root is necessary for exploring and understanding the evolution of extant eukaryotes, with the root pointing to their last common ancestor. The presence of any character (e.g., cytological, molecular, or genomic) in eukaryotic lineages can be fully understood only by reconstructing its evolution from this ancestral time point. Therefore, the last two decades have witnessed intense efforts to identify the position of the eukaryotic root (1–4).

Because the "core" of the eukaryotic cell is most similar to an archaeon or an archaeon-related organism (5–7), the first and obvious way of rooting the eukaryotic tree has relied on performing phylogenetic analyses based on eukaryotic proteins of archaeal origin. More than a hundred of such phylogenetic markers (mostly proteins of the translational machinery) have been identified and used in eukaryotic phylogenetic studies (8, 9). However, the archaeal sequences differ substantially from their eukaryotic orthologs, resulting in extremely long phylogenetic distances between archaea and eukaryotes. The use of a distant outgroup in phylogenetic reconstruction is known to be highly problematic because (*i*) the remaining phylogenetic signal is very weak, and therefore, correct positioning of the root is even weaker, and (*ii*) it creates a nonphylogenetic signal that is often stronger than the phylogenetic signal itself, thereby favoring long-branch attraction artifacts (LBAs) causing a basal position of fast evolving species in the ingroup. Indeed, phylogenetic inferences using archaeal sequences as an outgroup constantly find fast evolving eukaryotes at the base of all other eukaryotes (9–12).

In the absence of a close outgroup, rare cytological and genomic changes specific to some eukaryotic lineages have also been considered for rooting of the eukaryotic tree. In this context, the leading hypothesis used to be the Unikonta–Bikonta dichotomy, in which unikonts and bikonts are ancestrally characterized by (arguably) either a single or two flagella, respectively. This subdivision seemed to be supported by the distribution of certain gene fusions (13), and a specific myosin paralog (14), but both characters later proved to have a more complex evolutionary history (2). Furthermore, the idea of the "uniflagellate" ancestry for unikonts became untenable (2). For this reason, the concept of Unikonta has been recently superseded by proposing a "megagroup" Amorphea, which embraces unikonts as well as some previous bikont lineages (15). Other root positions were suggested more recently by assuming the most parsimonious explanation of the phylogenetic distribution of particular characters (16, 17), but secondary losses and lineage-specific modifications make such ad hoc inferences questionable. Instead of considering a priori selected characters supposed to reflect the deep history of eukaryotes, an alternative and more rational approach would consist of analyzing the evolution of an entire class of characters. Such analyses have been conducted using rare replacements and indels of amino acids within highly conserved regions (18), and gene duplication events (19), inferring alternative eukaryotic roots lying between Archaeplastida and all remaining eukaryotes, and between Opisthokonta and all

Significance

The root of eukaryote phylogeny formally represents the last eukaryotic common ancestor (LECA), but its position has remained controversial. Using new genome sequences, we revised and expanded two datasets of eukaryotic proteins of bacterial origin, which previously yielded conflicting views on the eukaryotic root. Analyses using state-of-the-art phylogenomic methodology revealed that both expanded datasets now support the same root position. Our results justify a new nomenclature for the two main eukaryotic groups and provide a robust phylogenetic framework to investigate the early evolution of the eukaryotic cell.

Author contributions: R.D. designed research; R.D. performed research; R.D., G.T., V.K., E.K., Č.V., B.F.L., and M.E. contributed new reagents/analytic tools; R.D. and H.B. analyzed data; and R.D., B.F.L., and M.E. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. T.M.E. is a guest editor invited by the Editorial Board.

Data deposition: The phylogenetic matrices and trees have been deposited in the TreeBase database, treebase.org (accession no. 16424).

[1]To whom correspondence should be addressed. Email: romain.derelle@crg.es.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1420657112/-/DCSupplemental.

remaining eukaryotes, respectively. However, the interpretation of these rare genomic changes is problematic because they are prone to homoplasy (20–23).

Recently, an innovative phylogenomic approach has been developed to root the eukaryotic tree by using a closer outgroup to eukaryotes, eubacterial lineages (4). Indeed, eukaryotes have been, and still are, the receptacles of massive lateral gene transfers from diverse prokaryotic sources leading to the chimeric nature of extant eukaryotic genomes (7, 24). The most important of these gene transfers occurred during the acquisition of mitochondria by incorporating an alpha-proteobacterial endosymbiont, when hundreds of alpha-proteobacterial genes have been transferred to the nucleus during the reduction of the endosymbiont (5, 7, 25, 26). We have previously proposed the use of 42 eukaryotic proteins with a mitochondrial function (encoded by the nuclear or mitochondrial genomes) as phylogenetic markers to root the eukaryotic tree with their orthologous alpha-proteobacterial sequences (27). The close outgroup obtained from this "ALPHA-PROT dataset" suggested the eukaryotic root between Unikonta/Amorphea plus malawimonads and Bikonta, although only with moderate support.

A similar, but less restrictive approach has then been used by He et al. (28), in which 37 eukaryotic genes acquired by ancient lateral gene transfers from different eubacterial sources are combined into a single dataset. Here, eubacterial lineages are grouped into a so-called "composite outgroup." While sharing a significant number (13 out of 37) of genes with the ALPHA-PROT dataset, phylogenetic analyses from this "EUBAC dataset" yielded an alternative eukaryotic root between Discoba and remaining eukaryotes, with maximal support. Evidently, the incongruence of results obtained from these two datasets suggests that further studies are needed to understand the source of disagreement and to arrive at a robust rooting hypothesis for the eukaryote phylogeny.

In this study, we reinvestigated the phylogenetic signal contained in the two eukaryotic datasets based on bacterial proteins, by increasing their eukaryotic taxonomic sampling with newly available data. The newly sequenced nuclear genomes include two malawimonads, the jakobid *Andalucia godoyi* and the cryptomonad *Goniomonas avonlea*. Particular attention was paid to the alternative positions of the eukaryotic root, whose support values were quantified by the metric "node support" (27), and to the phylogenetic positions of two eukaryotic lineages with uncertain affinity, malawimonads and collodictyonids (represented in our analysis by *Collodictyon tricilliatum*).

Results

Improving the ALPHA-PROT and EUBAC Datasets. Although the gene composition of the EUBAC dataset reanalyzed is exactly as published in He et al. (28) (i.e., 37 proteins), we considerably updated the gene composition of the ALPHA-PROT dataset. From the 42 proteins analyzed in Derelle and Lang (27), we removed the 10 proteins encoded by the mitochondrial genome of all eukaryotes because these markers mainly bear a non-phylogenetic signal (27). We also removed four additional proteins (Cytc2, Mtrf1, Nad9, and Sco1) due to their shortness and low level of sequence conservation across eukaryotes, for which eukaryote–eukaryote lateral gene transfer or contamination was difficult to detect. Instead, 11 new phylogenetic markers were identified from eukaryotic proteomes using the same criteria as in Derelle and Lang (27) (*Material and Methods*). All of these modifications resulted in a final ALPHA-PROT dataset of 39 proteins, of which 15 are shared with the EUBAC dataset. We then added sequences from a wide range of eukaryotic taxa representing most of the known eukaryotic lineages, including sequences from the two malawimonad species (*Malawimonas jakobiformis* and the still formally undescribed distantly related "*Malawimonas californiana*") and from *Andalucia godoyi*. We finally verified the orthology status of the sequences for all single-gene alignments, and all detected outliers (i.e., contaminants, lateral gene transfers, and paralogs) were excluded from the phylogenomic analyses.

Single-protein alignments from both datasets were trimmed and concatenated using the same methods, and saturated positions were eliminated using a tree-independent method that includes an auto-stopping criterion (29). The two cleaned phylogenetic matrices have a similar number of conserved positions (9,261 and 9,555 positions for the ALPHA-PROT and EUBAC datasets, respectively), identical eukaryotic taxonomic sampling (only the composition of the outgroup differs between the two datasets), similar levels of missing data (about 20% in both datasets) (*SI Appendix*, *Supplementary Data S1*), and similar saturation levels (*SI Appendix*, *Supplementary Data S1*). As shown by saturation-plot analyses, a significant improvement in terms of the saturation level has been obtained in the EUBAC dataset compared with the study by He et al. (28) (*SI Appendix*, *Supplementary Data S1*).

Both Datasets Converge on the Same Position of the Eukaryote Root. Cross-comparison tests implemented in PhyloBayes and performed on both datasets revealed that the CAT-GTR model has a better fit to both datasets than the CAT model. Because the CAT model has a better fit to the ALPHA-PROT and EUBAC datasets than the empirical LG model (27, 28), we can safely assume that the CAT-GTR model is the best fitting model of both datasets. Therefore, we decided to analyze the two datasets in a Bayesian framework under the CAT-GTR and CAT models (using posterior probabilities as support values). In addition, a search for the best Maximum Likelihood (ML) tree combined with an ML bootstrap analysis under the GTR model (referred to hereafter as "ML GTR" model/analysis) was conducted. All phylogenetic trees are shown in *SI Appendix*, *Supplementary Data S3*.

Under the CAT-GTR model, the two datasets converged to a eukaryotic root lying between two principal eukaryotic clades. One is composed of the taxa classified as Amorphea plus malawimonads and collodictyonids. The other embraces Discoba (Jakobida, Heterolobosea, and Euglenozoa) and the recently introduced megagroup Diaphoretickes (15). These two clades will hereafter be referred to as Opimoda and Diphoda, respectively (Fig. 1). The rational and formal definitions of these two new names are given in *Discussion*. This topology is supported by node support values of 1 and 0.94 for the ALPHA-PROT and EUBAC datasets, respectively (Fig. 2*A*). The two topologies are identical, with the only exception being the positions of malawimonads and *Collodictyon* within Opimoda (sister group to Amoebozoa or sister group to all other Opimoda in the ALPHA-PROT and EUBAC trees, respectively). These results are therefore in agreement with those obtained from an earlier variant of the ALPHA-PROT dataset (27, 30) and contradict the rooting hypothesis previously obtained from the EUBAC dataset by He et al. (28).

Under the less-fitting CAT and ML GTR models, the ALPHA-PROT dataset yielded a topology identical to the one obtained under the CAT-GTR model. Therefore, both the position of the eukaryotic root and the position of malawimonads and *Collodictyon* within Amorphea were congruent under the three evolutionary models used in this study (Fig. 2*A*). On the other hand, the EUBAC dataset showed different topologies depending on the model used to analyze it: the eukaryotic root lay between Discoba and remaining eukaryotes in ML under the GTR model and between *Collodictyon* and remaining eukaryotes in the Bayesian analysis with the CAT model, but in both cases with very low node supports (Fig. 2*A*). These results indicate that only the CAT-GTR model, the best fitting model to the dataset, has enough statistical power to infer the eukaryotic root from the EUBAC dataset whereas the simpler models tend to produce unresolved topologies.

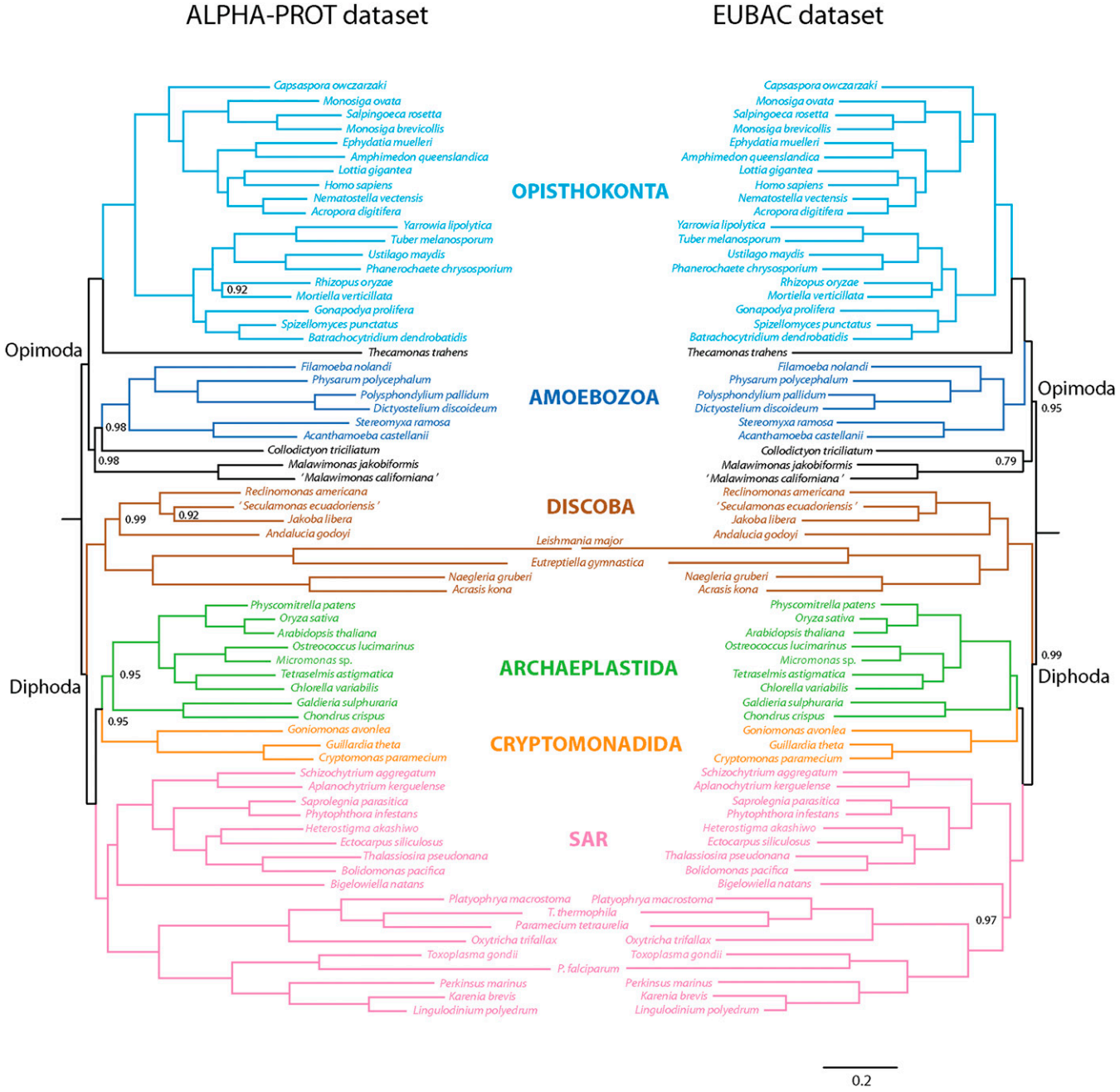

Fig. 1. Bayesian consensus trees. Bayesian consensus trees obtained from the ALPHA-PROT (left trees) and EUBAC (right trees) datasets under the CAT-GTR + Γ4 model. Posterior probabilities equal to 1 are not displayed. The two outgroups (Alpha-Proteobacteria and Eubacteria, respectively) are not shown for design reasons (gain of space).

Outparalogs Detected in the He et al. (28) Study Are Responsible for the "Alternative Root." Surprisingly, a significant number of distant paralogs, also called outparalogs (31) (i.e., sequences belonging to different eukaryotic orthologous groups that originated by gene duplication before the radiation of extant eukaryotes), were detected in the EUBAC dataset analyzed in He et al. (28), most of them in Discoba (five out of the six markers in which outparalogs have been detected). The list of these sequences and their corresponding single-gene trees are provided in *SI Appendix, Supplementary Data S2*. When these outparalogs were removed from the original matrix (i.e., outparalogs replaced by "X" in the filtered matrix "M20845"), the EUBAC dataset recovered the Opimoda–Diphoda root under the CAT-GTR and CAT models with node supports equal to 0.95 and 0.97, respectively, whereas the less reliable ML analysis under the GTR model showed the alternative rooting obtained in He et al. (28) (i.e., between Discoba and other eukaryotes), although only with a moderate node support of 83% (*SI Appendix, Supplementary Data S3*). These results indicate that the "alternative rooting hypothesis" obtained by He et al. (28) is the consequence of distant paralogs from Discoba species that are present in the dataset.

Addressing Possible Phylogenetic Artifacts and Shortcomings. Biases in amino acid compositions are a frequent source of artifacts in phylogenetic reconstruction (see, for instance, ref. 32). Although principal component analyses of amino acid compositions

A

dataset		model	Opimoda, Diphoda	Collodictyon, other euk.	Malawimonas, other euk.	Discoba, other euk.
ALPHA-PROT		CAT-GTR	1	0	0	0
		CAT	1	0	0	0
		ML GTR	78	4	9	1
EUBAC	all genes	CAT-GTR	0.94	0.02	0.01	0
		CAT	0.05	0.87	0.04	0
		ML GTR	7	1	36	50
	reduced	CAT-GTR	0.33	0.24	0.2	0
		CAT	0.05	0.8	0.1	0
		ML GTR	4	0	58	31

B

dataset		ML GTR	CAT	CAT-GTR
ALPHA-PROT		0.77	0.83	0.82
EUBAC	all genes	0.99	1.57	1.38
	reduced	0.94	1.42	1.27

C

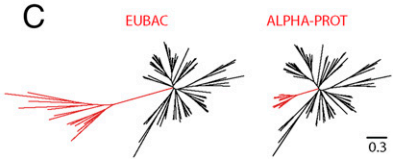

Fig. 2. Summary of phylogenetic results. (*A*) Node supports for the alternative rooting hypothesis. Shaded boxes indicate the topology obtained in the best ML tree or the Bayesian consensus tree. "EUBAC reduced" refers to the EUBAC dataset without the 10 distant markers identified in He et al. (28). (*B*) Relative distances outgroup-eukaryotes. Shaded boxes indicate the lowest distances outgroup-eukaryotes for each of the three evolutionary models. (*C*) Unrooted Bayesian consensus trees obtained from the ALPHA-PROT and EUBAC datasets under the CAT-GTR + Γ4 model (eukaryotic relationships shown in Fig.1) with their outgroup highlighted in red.

performed on both datasets did not show any reason to suspect such a bias (*SI Appendix*, *Supplementary Data S1*), we still addressed this question by performing CAT-GTR analyses after recoding the two datasets with the Dayhoff6 recoding scheme. In both cases, the topologies recovered the Opimoda–Diphoda root, although with low posterior probabilities (due to the loss of information after the six-states recoding) (*SI Appendix*, *Supplementary Data S2*).

The main argument of He et al. (28) against the results produced in Derelle and Lang (27) was a substantial level of incongruence of phylogenetic signal between markers used in that study. We repeated the congruence tests using Conclustador (33) and could not detect any incongruence in datasets built by both He et al. (28) and Derelle and Lang (27) (*SI Appendix*, *Supplementary Data S4*). Conclustador detected incongruence of phylogenetic signal between genes in both the ALPHA-PROT and EUBAC datasets built in the present study, but we argue that this incongruence of phylogenetic signal was quantitative (i.e., different amounts of phylogenetic signal) rather than qualitative (i.e., conflicting phylogenetic signals) (*SI Appendix*, *Supplementary Data S4*). This conclusion is further supported by the absence of outlier sequence in both of these datasets as revealed by our Phylo-MCOA (34) analyses (*Material and Methods*). According to He et al. (28), the putative incongruence between genes in the earlier version of the ALPHA-PROT dataset was responsible for the unorthodox groupings of Jakobida with Viridiplantae and of discicristates (e.g., *Naegleria* and *Leishmania*) with Amoebozoa, observed in the absence of discicristates and Jakobida, respectively, and would explain why the Discoba root was not obtained when analyzing the ALPHA-PROT dataset. We repeated the same analyses with the modified ALPHA-PROT dataset (i.e., alternatively lacking the Jakobida and Discicristata groups) using CAT and ML GTR models, but we did not observe these groupings (*SI Appendix*, *Supplementary Data S2*). All together, these results demonstrate that the Opimoda–Diphoda root obtained here in this study is not the result of conflicting phylogenetic signals.

Finally, a remaining point is the incongruence of the results obtained from the EUBAC dataset in this study between the three evolutionary models whereas results obtained from the ALPHA-PROT dataset were all congruent. It is important to notice that all eukaryotic relationships obtained under the three models were virtually identical, and that only the position of the outgroup differed between the three topologies. This observation is symptomatic for the presence of a distant outgroup that generates, due to its large distance to the ingroup, incorrect positions of the root via LBA artifacts. We tested this hypothesis by measuring the average ingroup–outgroup distance (normalized by the average intraingroup distances) for each phylogenetic tree obtained in this study (Fig. 2*B*). Although distances obtained from both datasets under the ML GTR model were similar, they seemed to be significantly different when calculated from the CAT topologies: the ingroup–outgroup distance of the EUBAC dataset was almost twice as large as the distance calculated from the ALPHA-PROT dataset. Most likely, the longer distances inferred by the site heterogeneous models (CAT and CAT-GTR) were the consequence of the ability of these models to detect multiple substitutions per site. Therefore, we posit that the incongruence of the results observed between the three evolutionary models was the direct consequence of the rather distant eubacterial outgroup.

He et al. (28) proposed to reduce the Eubacteria–Eukaryota distance by removing from the EUBAC dataset the 10 markers displaying the lowest level of similarity between eubacterial and eukaryotic sequences. Although this operation led to the removal of a significant fraction of the dataset (about 25%), the gain in decreasing patristic distances was rather limited: the ALPHA-PROT displayed by far the shortest distances between eukaryotes and the outgroup under the three models considered (Fig. 2*B*). Phylogenetic analyses based on this reduced EUBAC dataset gave similar results as those obtained from the full EUBAC dataset, with the exception of the CAT-GTR consensus tree that resulted in an unresolved polytomy malawimonads—*C. triciliatum*—other eukaryotes (Fig. 2*A*). Therefore, these analyses show that

the procedure for reducing the Eubacteria–Eukaryota distance as proposed by He et al. (28) seems to be inefficient.

Discussion

ALPHA-PROT and EUBAC Datasets Contain a Congruent Phylogenetic Signal. The analyses presented here demonstrate that both eukaryotic datasets based on proteins of bacterial origin bear a congruent phylogenetic signal, which is demonstrated by nearly identical topologies and the same position of the eukaryotic root in most analyses. The relationships within the eukaryotic subtree are fully consistent with the results of recent phylogenomic analyses based on independent sequence datasets (35–40), indicating that the eukaryotic genes of bacterial origin confer a phylogenetic signal useful for inferring the evolutionary history of eukaryotic lineages. In contradiction to the "alternative eukaryotic root" of He et al. (28) placed between Discoba and the remaining eukaryotes, our analyses of the EUBAC dataset place the root between Opimoda and Diphoda when the most realistic substitution models are used. The explanation for this difference lies primarily in the phylogenetic matrix of He et al. (28) that includes outparalogs, mostly from jakobids. Outparalogs carry an erroneous, strong phylogenetic signal that inevitably supplants the correct but comparatively weaker phylogenetic signal contained in the dataset.

In addition, the composite outgroup built with the EUBAC dataset tends to create a rather long distance between eukaryotes and the outgroup, making the phylogenetic signal difficult to extract by simple evolutionary models. This issue applies particularly when early eukaryotic lineages with uncertain affinity, such as malawimonads and collodictyonids, are included in analyses. The EUBAC dataset has been designed to allow for more characters and eventually to replace the more restricted ALPHA-PROT dataset by combining eukaryotic proteins originating from different bacterial sources into a single phylogenetic matrix. However, the relatively distant outgroup created by this approach does not seem to be appropriate for inferring the root of the eukaryotic tree. Finally, it is likely that the composite outgroup that combines markers for which phylogenetic relationships between eukaryotes and eubacterial lineages are by definition incongruent creates a noisy (nonphylogenetic) signal not suitable to infer deep eukaryotic relationships (41).

The Eukaryotic Supergroup Excavata Is Not Monophyletic. Both the updated ALPHA-PROT and EUBAC datasets place the two enigmatic lineages malawimonads and collodictyonids within, or as sister to, the Amorphea group, supporting previous results obtained with the earlier version of the ALPHA-PROT dataset (27, 30). Collodictyonids exhibit a suite of unique cellular features, leaving this lineage without clear affinity (15, 42), so its position close to or within Amorphea was not anticipated before phylogenomic analyses. The phylogenetic position of malawimonads indicated by our analyses is more striking. The suspension-feeding groove and the organization of the flagellar apparatus were interpreted as evidence for a specific relationship of malawimonads with other taxa of the supergroup Excavata: i.e., Discoba and Metamonada (43, 44). However, the monophyly of these three groups has never been convincingly demonstrated by phylogenetic analyses with molecular characters. Particularly, malawimonads do not branch together with the Discoba clade in any of the recent phylogenomic analyses with rich taxon and gene sampling and more realistic substitution models (30, 35, 36, 38). Instead, malawimonads form a clan with Amorphea whereas Discoba form a clan with Diaphoretickes in these analyses, which is consistent with our results obtained using largely nonoverlapping datasets. We posit that these results recover a genuine phylogenetic signal, thus indicating that the supergroup Excavata is at least diphyletic.

An open question is the position of the anaerobic group Metamonada. In phylogenomic analyses they are placed either as a sister group of Discoba, especially when long-branch representatives are included (39, 43), or they may be a sister group of malawimonads, as suggested by analyses where metamonads are represented only by the relatively slowly evolving *Trimastix* (36, 39, 43). This group, composed exclusively of anaerobic species, was not included in our analyses due to poor representation of the genes in both the ALPHA-PROT and EUBAC datasets—because most of these genes are functionally associated with a conventional, aerobic mitochondrion. Therefore, their phylogenetic position with respect to the eukaryotic root advocated here remains to be determined.

The phylogenetic position of malawimonads and Discoba on the opposite sides of the eukaryotic root open a fundamental question relative to the early evolution of eukaryotes: (*i*) Was the last eukaryotic common ancestor (LECA) an excavate-like organism as proposed by Cavalier-Smith (16, 45) and as suggested by the complex ultrastructure of the flagellar apparatus shared by the different excavate lineages (46) or (*ii*) does this topology represent another example of convergent acquisitions, so well-known from the complex evolutionary history across eukaryotes (e.g., convergent acquisitions of multicellularity, amoeboid stage or photosynthetic metabolism)?

New Names for New Clades. Corroborating with previous studies (27, 30), the traditional nomenclature Unikonta/Bikonta is challenged in this study by the deep branching position on the unikont side of the root of species that have two or more flagella: the apusomonads, malawimonads, and collodictyonids. This result, in addition to the phylogenetic position of the biflagellate breviate *Pygsuia biforma* as sister-group to Opisthokonta and apusomonads (36), and the fact that the supergroup Amoebozoa had ancestrally two flagella (2) led to the conclusion that the last common ancestor of unikonts was a biflagellate organism. Therefore, as acknowledged by Adl et al. (15), the term Unikonta should no longer be used. In addition, the biflagellate state also corresponds to the ancestral state of the last common ancestor of eukaryotes. That means that the name Bikonta is as invalid because it now reflects the ancestral state of all eukaryotes, calling for an adequate naming of the two principal eukaryotic clades resolved in this study.

It may seem straightforward to simply expand the meaning of the taxon Amorphea to embrace collodictyonids and malawimonads, which would thus cover one of the two principal clades. However, the taxon Amorphea was established with a node-based phylogenetic definition stating that it corresponds to the least inclusive clade containing *Homo sapiens*, *Neurospora crassa*, and *Dictyostelium discoideum* (15): i.e., the least inclusive clade containing Opisthokonta and Amoebozoa. Under this definition, malawimonads and collodictyonids branch either within Amorphea (ALPHA-PROT dataset) or as sister group to Amorphea (EUBAC dataset). The latter represents a position favored by phylogenetic analyses, based on conventional phylogenomic matrices without an outgroup (36, 38–40), and therefore a taxon including Amorphea, collodictyonids, and malawimonads has never been defined. Therefore, we propose that the original definition of the Amorphea be kept and a new, more inclusive taxon embracing Amorphea, malawimonads, and collodictyonids be established.

For the reasons mentioned above, we propose two newly named formal taxa using branch-based phylogenetic definitions in which all specifiers are extant. These two taxa are defined by the position of the eukaryotic root obtained in this study as follows:

Opimoda: The most inclusive clade containing *Homo sapiens*, Linnaeus (1758) (Opisthokonta); *Dictyostelium discoideum*, Raper (1935) (Amoebozoa); and *Malawimonas jakobiformis*, O'Kelly and Nerad (1999); but not *Arabidopsis thaliana*, (Linnaeus) Heynhold (1842) (Archaeplastida); *Bigelowiella natans*, Moestrup and Sengco (2001) (Rhizaria); *Goniomonas avonlea*,

Kim and Archibald (2013), and *Jakoba libera*, (Ruinen, 1938) Patterson (1990).

Diphoda: The most inclusive clade containing *Arabidopsis thaliana*, (Linnaeus) Heynhold (1842) (Archaeplastida); *Bigelowiella natans*, Moestrup and Sengco (2001) (Rhizaria); *Goniomonas avonlea*, Kim and Archibald (2013); and *Jakoba libera*, (Ruinen, 1938) Patterson (1990); but not *Homo sapiens* Linnaeus (1758) (Opisthokonta); *Dictyostelium discoideum*, Raper (1935) (Amoebozoa); and *Malawomonas jakobiformis*, O'Kelly and Nerad (1999).

In the absence of obvious morphological synapomorphies, the chosen names Opimoda and Diphoda are two acronyms that stand for OPIsthokonta and aMOebozoa and for DIscoba and diaPHOretickes, respectively. We believe that the nomenclature proposed here will offer a neutral framework (i.e., one that does not reflect any presumed ancestral state), suitable for further phylogenetic investigations and studies of eukaryotic evolution. At the present stage, deep phylogenetic relationships of the group Opimoda, which most likely include "other" enigmatic eukaryotic lineages (e.g., ancyromonads and *Mantamonas*) (45, 47, 48), represent the most challenging task in our understanding of the early stages of the evolution of eukaryotes and the precise nature of LECA.

Materials and Methods

Genome Sequencing. The nuclear genomes *A. godoyi* of and *M. californiana* were sequenced using the 454 method on the Titanium platform according to GS FLX Library Preparation Method protocols (Roche). Shotgun and paired-end libraries were prepared and run to get over 50-fold read coverage. Reads were assembled using Newbler 2.6 (Roche), and bacterial sequences were recognized and removed by blast, yielding draft assemblies of the nuclear genomes (20.2 Mb of unique sequence contained in 174 scaffolds for *A. godoyi* and 46.5 Mb of unique sequence contained in 1,123 scaffolds for *M. californiana*). The nuclear genome of *M. jakobiformis* was sequenced using one run of Illumina HiSEq. 2000 from a paired-end library. Reads were assembled using Abyss (49), and bacterial sequences were recognized and removed by blast, yielding a draft assembly of 71.1Mb (3,491 contigs; N50 = 87 kb).

The draft genome assembly of the cryptomonad *G. avonlea* was based on data generated from two short-insert and two mate-pair (2 kbp, 6 kbp) libraries on an Illumina HiSEq. 2000 sequencer. Reads from short-insert libraries were error corrected using ALLPATHS-LG (50) before being assembled using Abyss over a range of k-mer values. The assembly used in this study had the total length of 227.9 Mb (143,882 contigs; N50 = 25 kb). Finally, gene predictions were obtained from the genome assemblies using Augustus (51). The protein sequences used in this study are available in *SI Appendix, Supplementary Data S5*.

Dataset Preparation. For the purpose of identifying new ALPHA-PROT phylogenetic markers, all proteins from *Phytophthora infestans* and *Amphimedon queenslandica* that have a predicted mitochondrial localization were retrieved from the Ensembl database. These sequences were used as initial reference datasets for blasting locally a large collection of prokaryotic and eukaryotic predicted proteomes downloaded from the National Center for Biotechnology Information RefSeq database. Only those alignments were retained for which (*i*) eukaryotic proteins have an alpha-proteobacterial origin, (*ii*) orthologous sequence relationships were assessed with confidence, and (*iii*) the genes are encoded by the nuclear genome in most of eukaryotic lineages.

Phylogenetic matrices used in He et al. (28) were downloaded from TreeBase (www.treebase.org). The matrix M20844 was divided into single-gene alignments to rebuild the EUBAC dataset: for each species, the complete protein sequences were retrieved by blast in replacement of the trimmed sequences.

A wide range of eukaryotic species were added by blast to both the ALPHA-PROT and EUBAC datasets (*SI Appendix, Supplementary Data S1*). This set of species was selected to represent most of the eukaryotic lineages for which sequences are available, with the exception of anaerobic eukaryotes (e.g., breviates, metamonads) and lineages known to be extremely unstable to avoid converge issues in Bayesian analyses (e.g., we kept only two Archaeplastida and one "Hacrobia" lineages) (3).

Single-gene alignments were aligned with T-coffee (52) by masking in the alignments all characters that had a consistency index lower than 9 (which corresponds to the highest value). To check orthologous relationships, alignments were then trimmed by trimAl (53) to remove positions with more than 50% of gaps and blocks of length lower than four positions. A search for the best RAxML tree under the PROTGAMMALG model combined with 100 ML bootstraps was then performed from each alignment, and trees were screened manually to detect and remove outliers. These cases were detected by searching for splits in individual protein trees that were supported by ML bootstrap values ≥70% and that conflicted with well-accepted eukaryotic supergroups.

In cases where several sequences of a given species were present in the alignment, the slowest evolving one was selected (according to the branch lengths in RAxML trees). Given the large diversity of eubacterial lineages used in the EUBAC dataset, we did not check their orthologous relationships and simply used those published in He et al. (28).

Assembly of Sequences into the Phylogenetic Matrices. Single-gene alignments cleaned from outliers were then concatenated into phylogenetic matrices. We aligned them with T-coffee (same parameters as mentioned above) and trimmed them using Gblocks (54) under the following parameters: maximum proportion of gaps equal to 20%, minimum size of a block equal to 5, and maximum number of contiguous nonconserved positions equal to 3. Trimmed alignments were concatenated into two phylogenetic matrices (called ALPHA-PROT and EUBAC) using a custom-made script. Finally, we removed fast evolving sites from both matrices using a method described in *SI Appendix, Supplementary Data S1*. The phylogenetic matrices have been deposited in the TreeBASE database (accession number 16424), and single-gene alignments are available upon request.

Phylogenetic Analyses. We performed statistical comparisons of the CAT-GTR and CAT models from both datasets by using a cross-validation test implemented in PhyloBayes (55), based on the topology of Fig. 1 without *Malawimonas* species, *C. triciliatum*, and the two outgroups. Ten replicates were performed: 9/10 for the learning set and 1/10 for the test set. Markov chain Monte Carlo (MCMC) chains were run for 3,000 cycles with a burn-in of 1,500 cycles for the CAT model and 1,500 cycles with a burn-in of 100 cycles for the CAT-GTR model. For both datasets, the CAT-GTR model was found to have a much better statistical fit than CAT (a likelihood score of 347.7 ± 42.7703 and 292.88 ± 47.6114 in favor of CAT-GTR for the ALPHA-PROT and EUBAC datasets, respectively).

Bayesian inferences were performed with the CAT-GTR and CAT models, using the "-dc" option, by which constants sites are removed, implemented in the program PhyloBayes. For the plain posterior estimation, two independent runs were performed with a total length of 8,000 and 15,000 cycles under the CAT-GTR and CAT models, respectively. Convergence between the two chains was ascertained by calculating the difference in frequency for all their bipartitions using a threshold maxdiff <0.3 (bipartitions of eukaryotic relationships were <0.1 in all analyses). The first 3,000 and 6,000 points were discarded as burn-in in the CAT-GTR and CAT analyses, respectively, and the posterior consensus was computed by selecting 1 tree every 10 over both chains. The recoding of amino acids into the six Dayhoff functional categories was performed using the "-recode" command implemented in PhyloBayes, and runs of 15,000 cycles under the CAT-GTR model were performed from these recoded datasets (using a burn-in of 6,000 cycles).

ML analyses were performed using RAxML; in each case, the search for the best ML tree was conducted under the PROTGAMMAGTR model starting from three random trees, and 400 ML bootstraps were analyzed with the rapid BS algorithm under the same model.

Miscellaneous. Congruence of phylogenetic signal between genes was tested using Conclustador version 0.4a (33) using default parameters. For these tests, trimmed aliments used to build the multigene matrices were analyzed by RAxML: 100 bootstraps were generated and combined with a search for the best ML tree using the fast algorithm under the PROTGAMMALG model. The detection of outliers was performed using Phylo-MCOA (34) using default parameters from this set of best ML trees. Phylo-MCOA could not detect any outlier sequence in both the ALPHA-PROT and EUBAC datasets.

Principal component analyses were computed using the R package ade4.

Distances used to build saturation plots were obtained as follows: uncorrected distances were calculated using a custom-made script, and patristic distances were retrieved from the best RAxML tree (obtained under the PROTGAMMAGTR model) using the ETE package (56). Node supports for the alternative rooting hypotheses were calculated using the ETE package.

ACKNOWLEDGMENTS. We thank A. Narechania for assistance with genome assembly of *G. avonlea*. The work of R.D. was supported by Howard Hughes Medical Institute International Early Career Scientist Program Grant 55007424, Spanish Ministry of Economy and Competitiveness Grant BFU2012-31329 as part of the European Molecular Biology Organization Young Investigator Program, and two grants from the Spanish Ministry of Economy and Competitiveness, "Centro de Excelencia Severo Ochoa 2013-2017" Grant Sev-2012-0208 and Grant BES-2013-064004 funded by the European Regional Development Fund. This study was further supported by a Czech Science Foundation Grant 13-24983S and Project CZ.1.05/2.1.00/03.0100 (Institute of Environmental Technology) financed by Structural Funds of the European Union (to M.E.), American Museum of Natural History start-up grants (to E.K.), the Canadian Research Chair program (B.F.L.), and the Natural Sciences and Engineering Research Council of Canada (B.F.L.).

1. Brinkmann H, Philippe H (2007) The diversity of eukaryotes and the root of the eukaryotic tree. *Adv Exp Med Biol* 607:20–37.
2. Roger AJ, Simpson AG (2009) Evolution: Revisiting the root of the eukaryote tree. *Curr Biol* 19(4):R165–R167.
3. Burki F (2014) The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb Perspect Biol* 6(5):a016147.
4. Williams TA (2014) Evolution: Rooting the eukaryotic tree of life. *Curr Biol* 24(4): R151–R152.
5. Koonin EV (2010) The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol* 11(5):209.
6. Guy L, Saw JH, Ettema TJ (2014) The archaeal legacy of eukaryotes: A phylogenomic perspective. *Cold Spring Harb Perspect Biol* 6(10):a016022.
7. Rochette NC, Brochier-Armanet C, Gouy M (2014) Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. *Mol Biol Evol* 31(4):832–845.
8. Bapteste E, et al. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci USA* 99(3):1414–1419.
9. Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54(5):743–757.
10. Arisue N, Hasegawa M, Hashimoto T (2005) Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. *Mol Biol Evol* 22(3):409–420.
11. Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311(5765):1283–1287.
12. Williams TA, Embley TM (2014) Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol Evol* 6(3):474–481.
13. Stechmann A, Cavalier-Smith T (2003) The root of the eukaryote tree pinpointed. *Burr Biol* 13(17):R665–R666.
14. Richards TA, Cavalier-Smith T (2005) Myosin domain evolution and the primary divergence of eukaryotes. *Nature* 436(7054):1113–1118.
15. Adl SM, et al. (2012) The revised classification of eukaryotes. *J Eukaryot Microbiol* 59(5):429–493.
16. Cavalier-Smith T (2010) Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol Lett* 6(3):342–345.
17. Wideman JG, Gawryluk RM, Gray MW, Dacks JB (2013) The ancient and widespread nature of the ER-mitochondria encounter structure. *Mol Biol Evol* 30(9):2044–2049.
18. Rogozin IB, Basu MK, Csürös M, Koonin EV (2009) Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome Biol Evol* 1:99–113.
19. Katz LA, Grant JR, Parfrey LW, Burleigh JG (2012) Turning the crown upside down: Gene tree parsimony roots the eukaryotic tree of life. *Syst Biol* 61(4):653–660.
20. Bapteste E, Philippe H (2002) The potential value of indels as phylogenetic markers: Position of trichomonads as a case study. *Mol Biol Evol* 19(6):972–977.
21. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6(5):361–375.
22. Rodríguez-Ezpeleta N, et al. (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol* 56(3):389–399.
23. Leonard G, Richards TA (2012) Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *Proc Natl Acad Sci USA* 109(52):21402–21407.
24. Andersson JO (2009) Gene transfer and diversification of microbial eukaryotes. *Annu Rev Microbiol* 63:177–193.
25. Andersson SG, Karlberg O, Canback B, Kurland CG (2003) On the origin of mitochondria: A genomics perspective. *Philos Trans R Soc Lond B Biol Sci* 358(1429): 165–177. discussion 177–169.
26. Gray MW, Burger G, Lang BF (2001) The origin and early evolution of mitochondria. *Genome Biol* 2(6), reviews1018.1–reviews1018.5.
27. Derelle R, Lang BF (2012) Rooting the eukaryotic tree with mitochondrial and bacterial proteins. *Mol Biol Evol* 29(4):1277–1289.
28. He D, et al. (2014) An alternative root for the eukaryote tree of life. *Curr Biol* 24(4): 465–470.
29. Goremykin VV, Nikiforova SV, Bininda-Emonds OR (2010) Automated removal of noisy data in phylogenomic analyses. *J Mol Evol* 71(5-6):319–331.
30. Zhao S, Shalchian-Tabrizi K, Klaveness D (2013) Sulcozoa revealed as a paraphyletic group in mitochondrial phylogenomics. *Mol Phylogenet Evol* 69(3):462–468.
31. Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 18(12):619–620.
32. Foster PG, Cox CJ, Embley TM (2009) The primary divisions of life: A phylogenomic approach employing composition-heterogeneous methods. *Philos Trans R Soc Lond B Biol Sci* 364(1527):2197–2207.
33. Leigh JW, Schliep K, Lopez P, Bapteste E (2011) Let them fall where they may: Congruence analysis in massive phylogenetically messy data sets. *Mol Biol Evol* 28(10): 2773–2785.
34. de Vienne DM, Ollier S, Aguileta G (2012) Phylo-MCOA: A fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. *Mol Biol Evol* 29(6):1587–1598.
35. Brown MW, Kolisko M, Silberman JD, Roger AJ (2012) Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. *Burr Biol* 22(12): 1123–1127.
36. Brown MW, et al. (2013) Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proc Biol Sci* 280(1769):20131755.
37. Burki F, Okamoto N, Pombert JF, Keeling PJ (2012) The evolutionary history of haptophytes and cryptophytes: Phylogenomic evidence for separate origins. *Proc Biol Sci* 279(1736):2246–2254.
38. Zhao S, et al. (2012) Collodictyon: An ancient lineage in the tree of eukaryotes. *Mol Biol Evol* 29(6):1557–1568.
39. Kamikawa R, et al. (2014) Gene content evolution in Discobid mitochondria deduced from the phylogenetic position and complete mitochondrial genome of Tsukubamonas globosa. *Genome Biol Evol* 6(2):306–315.
40. Yabuki A, et al. (2014) Palpitomonas bilix represents a basal cryptist lineage: Insight into the character evolution in Cryptista. *Sci Rep* 4:4641.
41. Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497(7449):327–331.
42. Brugerolle G, Bricheux G, Philippe H, Coffea G (2002) Collodictyon triciliatum and Diphylleia rotans (=Aulacomonas submarina) form a new family of flagellates (Collodictyonidae) with tubular mitochondrial cristae that is phylogenetically distant from other flagellate groups. *Protist* 153(1):59–70.
43. Hampl V, et al. (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci USA* 106(10):3859–3864.
44. Simpson AG (2003) Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *Int J Syst Evol Microbiol* 53(Pt 6): 1759–1777.
45. Cavalier-Smith T (2014) The neomuran revolution and phagotrophic origin of eukaryotes and cilia in the light of intracellular coevolution and a revised tree of life. *Cold Spring Harb Perspect Biol* 6(9):a016006.
46. Yubuki N, Leander BS (2013) Evolution of microtubule organizing centers across the tree of eukaryotes. *Plant J* 75(2):230–244.
47. Cavalier-Smith T (2013) Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. *Eur J Protistol* 49(2):115–178.
48. Paps J, Medina-Chacón LA, Marshall W, Suga H, Ruiz-Trillo I (2013) Molecular phylogeny of unikonts: New insights into the position of apusomonads and ancyromonads and the internal relationships of opisthokonts. *Protist* 164(1):2–12.
49. Simpson JT, et al. (2009) ABySS: A parallel assembler for short read sequence data. *Genome Res* 19(6):1117–1123.
50. Butler J, et al. (2008) ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res* 18(5):810–820.
51. Hoff KJ, Stanke M (2013) WebAUGUSTUS: A web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res* 41(web server issue):W123–W128.
52. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1):205–217.
53. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973.
54. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540–552.
55. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25(17): 2286–2288.
56. Huerta-Cepas J, Dopazo J, Gabaldón T (2010) ETE: A python Environment for Tree Exploration. *BMC Bioinformatics* 11:24.