New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life

Jürgen F.H. Strassert,[1] Mahwash Jamy,[1] Alexander P. Mylnikov,[2] Denis V. Tikhonenkov,[2] and Fabien Burki*,[1,3]

[1]Department of Organismal Biology, Program in Systematic Biology, Uppsala University, Uppsala, Sweden
[2]Institute for Biology of Inland Waters, Russian Academy of Sciences, Borok, Yaroslavl Region, Russia
[3]Science for Life Laboratory, Uppsala University, Uppsala, Sweden

*Corresponding author: E-mail: fabien.burki@ebc.uu.se.
Associate Editor: Beth Shapiro

Abstract

The resolution of the broad-scale tree of eukaryotes is constantly improving, but the evolutionary origin of several major groups remains unknown. Resolving the phylogenetic position of these "orphan" groups is important, especially those that originated early in evolution, because they represent missing evolutionary links between established groups. Telonemia is one such orphan taxon for which little is known. The group is composed of molecularly diverse biflagellated protists, often prevalent although not abundant in aquatic environments. Telonemia has been hypothesized to represent a deeply diverging eukaryotic phylum but no consensus exists as to where it is placed in the tree. Here, we established cultures and report the phylogenomic analyses of three new transcriptome data sets for divergent telonemid lineages. All our phylogenetic reconstructions, based on 248 genes and using site-heterogeneous mixture models, robustly resolve the evolutionary origin of Telonemia as sister to the Sar supergroup. This grouping remains well supported when as few as 60% of the genes are randomly subsampled, thus is not sensitive to the sets of genes used but requires a minimal alignment length to recover enough phylogenetic signal. Telonemia occupies a crucial position in the tree to examine the origin of Sar, one of the most lineage-rich eukaryote supergroups. We propose the moniker "TSAR" to accommodate this new mega-assemblage in the phylogeny of eukaryotes.

Key words: TSAR, Telonemia, phylogenomics, eukaryotes, tree of life, protists.

Introduction

The eukaryote tree of life is one of the oldest working hypotheses in biology. Over the past 15 years, this tree has been extensively reshaped based on phylogenomic studies of an ever-increasing diversity of protists (i.e., microbial eukaryotes) (Burki et al. 2007; Rodriguez-Ezpeleta, Brinkmann, Burger, et al. 2007; Brown et al. 2013; Yabuki et al. 2014; Katz and Grant 2015; Burki et al. 2016; Brown et al. 2018). The current tree places most lineages into a few redefined "supergroups"—a taxonomic rank-less concept that was devised as inclusive clades reasonably supported by molecular and/or morphological evidence (Simpson and Roger 2002, 2004; Keeling et al. 2005). As few as eight supergroups are recognized, sometimes arranged in even higher-order divisions assuming the root of the tree falls outside of these mega-clades: Diaphoretickes, including Sar, Haptista, Cryptista, Archaeplastida; Amorphea, including Amoebozoa and Obazoa; Excavata, including Discoba and Metamonada (Burki 2014; Simpson et al. 2017; Adl et al. 2019). Importantly, this view of the tree remains a working hypothesis and whilst most groups have been robustly and consistently supported, others have not and still await confirmation (this is mainly the case for Archaeplastida and Excavata) (Burki et al. 2016; Heiss et al. 2018).

The backbone of the tree also contains a series of long-known and newly discovered enigmatic protist taxa that do not fit within any established supergroup based on ultrastructure, ribosomal DNA phylogenies, or even protein-coding genes (Brugerolle et al. 2002; Walker et al. 2006; Heiss et al. 2011; Yabuki et al. 2012; see Adl et al. 2018 for a more exhaustive list of *incertae sedis* taxa within eukaryotes). Interestingly, phylogenomics suggested that many of these "orphan" taxa are themselves supergroup-level lineages, likely representing early diverging evolutionary intermediates between the established supergroups (Zhao et al. 2012; Brown et al. 2013; Cavalier-Smith et al. 2015; Janouškovec et al. 2017; Heiss et al. 2018; Lax et al. 2018). Individual orphan lineages can also turn out to be evolutionarily related, forming new larger deeply branching eukaryotic clade as in the case of the recently proposed "CRuMs" supergroup (Brown et al. 2018). Taken together, orphans can profoundly impact our understanding of ancient eukaryote evolution by defining new major branches in the tree and help establishing the branching pattern among all main eukaryotic groups.

The genus *Telonema* is one such orphan taxon that has proven difficult to place in eukaryote-wide phylogenies. Known for over a century, this genus contains only two named species, *Telonema subtile* (syn. *Telonema subtilis*) and *Telonema antarcticum* (Griessmann 1913;

Open Access

Klaveness et al. 2005). However, environmental sequencing of the 18S rDNA has shown that the genetic diversity of *Telonema* is relatively high in marine and freshwater environments and in fact likely represent several genera (Shalchian-Tabrizi et al. 2007; Bråte et al. 2010; Adl et al. 2018). *Telonema* and related sequences are frequently detected in these aquatic environments, and although they are usually not very abundant, they are thought to play significant ecological roles as biflagellated heterotrophic predators of bacteria and small-sized eukaryotes (Klaveness et al. 2005). The ultrastructure of *T. subtile* and *T. antarcticum* revealed unique morphological features, such as a multilayered subcortical lamina and complex cytoskeleton, that distinguish *Telonema* from all other eukaryotes (Klaveness et al. 2005; Shalchian-Tabrizi et al. 2006; Yabuki et al. 2013). Yet, *Telonema* also possesses structures that are characteristics of other groups: for example, the peripheral vacuoles that are similar to the cortical alveoli in Alveolata, tripartite tubular hairs on the long flagellum that are similar to the mastigonemes in Stramenopila, or wheel-like structures resembling extrusomes in some Rhizaria (Klaveness et al. 2005; Shalchian-Tabrizi et al. 2006; Yabuki et al. 2013). This combination of unique and shared ultrastructural features makes *Telonema* highly distinctive from other eukaryotic groups and has supported the notion that it forms a phylum-level lineage—Telonemia—without close relatives (Shalchian-Tabrizi et al. 2006). Accordingly, the first molecular phylogenies based on up to four genes provided only few clues as to the position of the group, other than it was not closely related to any known groups (Shalchian-Tabrizi et al. 2006). Later, more substantial analyses based on larger data sets containing 127–351 genes also failed to pinpoint with statistical significance the origin of Telonemia, but this was likely due at least in part to the small number of genes sampled from only one species (*T. subtile*) (Burki et al. 2009, 2012, 2016; Lax et al. 2018).

In this study, we tested whether better gene coverage and taxon sampling for Telonemia using a phylogenomic approach can help to infer with confidence the phylogenetic position of this important group of eukaryotes. To this effect, we established cultures for three telonemid species, one of which corresponded to the described *T. subtile*, and generated high-quality transcriptomes for all. The availability of different data sets from closely related species allowed us to robustly identify the correct copies from contaminant sequences, a task that is otherwise not easy with single data sets for orphan lineages. Our analyses unambiguously place Telonemia in a sister position to the supergroup Sar, forming a new mega-assemblage that further resolves the tree of eukaryotes.

Results

Improved Data Set and Taxon Selection

To place Telonemia in the tree of eukaryotes, we assembled a 248-genes data set with a broad taxon sampling. All genes were carefully inspected by reconstructing gene phylogenies to evaluate orthology and detect contamination (see Materials and Methods). Given the taxonomic distribution

of available genomic data sets and the tentative placement of Telonemia in previous phylogenomic works (Burki et al. 2009, 2012, 2016), special attention was made to include representatives of Diaphoretickes lineages with sufficient data. For the telonemids, at least two sequences from different data sets were required to group together in gene trees in order to be retained. This requirement became possible by the addition of the new telonemid data sets and greatly lowered the risk that unidentified contaminants were wrongly assigned to Telonemia. Overall, the sequence coverage for the three telonemid operational taxonomic units (OTUs) ranged between 58% and 99% (supplementary table 1, Supplementary Material online), which constitutes a large improvement over previous studies (e.g., Burki et al. 2016). Sequences corresponding to the bodonid prey *Procryptobia sorokini*, easily identifiable in gene trees from the presence of other related taxa (e.g., *Neobodo designis*, *Bodo saltans*), were reassigned accordingly. In total, 293 taxa were initially analyzed (supplementary fig. S1, Supplementary Material online), but this number was reduced to 109 OTUs to allow computational-heavy analyses. This final taxon sampling was selected based on the taxonomic affiliation and data coverage of the OTUs, excluding faster-evolving lineages when the selection of shorter branches was possible (fig. 1, Supplementary fig. S1, Supplementary Material online, and supplementary table 1, Supplementary Material online).

Tree Inference and Testing

The 248 genes/109 OTUs supermatrix was analyzed by maximum likelihood (ML) and Bayesian tree reconstructions using site-heterogeneous mixture models of evolution, specifically LG+C60+F+Γ4 with posterior mean site frequency (PMSF) profiles and CAT+GTR+Γ4, respectively (see Materials and Methods). These models have been shown to be more robust against heterogeneity in the amino acid substitution process across sites, which are important to take into consideration especially along the most ancient branches of the eukaryotic tree (Lartillot and Philippe 2008; Quang et al. 2008; Wang et al. 2018). As shown in figure 1, both models recovered a similar overall topology, with most currently accepted supergroups receiving maximal support (e.g., Sar, Cryptista, Haptista, Discoba, Obazoa, and Amoebozoa). Most interestingly, in both analyses the position of Telonemia was identical and unequivocal, placed within Diaphoretickes as sister to the Sar group (bootstrap [BP] = 100%, posterior probability [PP] = 1.0). Cryptista was robustly inferred in a clade including the members of Archaeplastida, but the branching pattern in this clade was unclear: whilst the LG+C60+F+Γ4+PMSF tree remained unresolved (BP = 56% for the grouping of Cryptista with Rhodophyta), the CAT+GTR+Γ4 model recovered a monophyletic Archaeplastida with 1.0 PP (fig. 1). The position of Haptista (and its possible sister taxon, *Ancoracysta twista*; see Janouškovec et al. 2017) also remained ambiguous showing a sister relationship to the Archaeplastida + Cryptista clade with the LG+C60+F+Γ4+PMSF model (BP = 100%) but corresponding to an unconverged node in the CAT+GTR+Γ4 analysis (two chains supporting the named
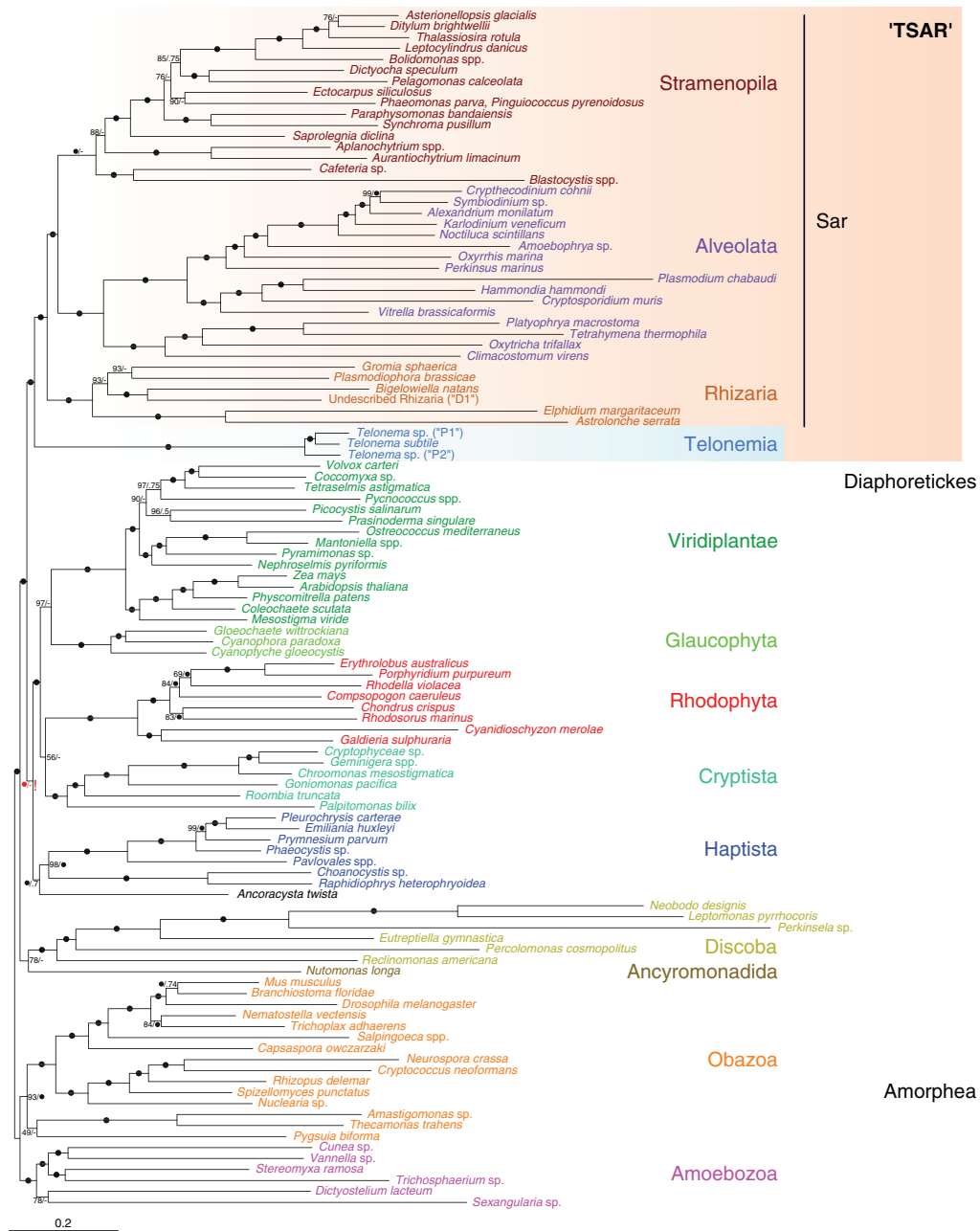
FIG. 1. ML tree showing the phylogenetic position of Telonemia within eukaryotes. The tree corresponds to the best ML tree inferred from a concatenated alignment of 248 protein-coding genes using the LG+C60+G+F-PMSF model. Support at nodes is derived from 100 BP replicates and Bayesian PPs. Black circles represent maximum support in both analyses. A number appended to a black circle indicates full support in only one of the two analyses (i.e., BP/PP). Nodes with dashes (–) indicate an alternative branching in the Phylobayes tree (see supplementary figs. S2 and S3, Supplementary Material online). The ambiguous position of Haptista with this data set is highlighted with an exclamation mark (see text for more information).

relationship, and two chains supporting a sister relationship to Telonemia + Sar; for details, see supplementary figs. S2 and S3, Supplementary Material online).

The relationships within Diaphoretickes show two strongly supported blocks, Telonemia + Sar and Cryptista + Archaeplastida (whose monophyly remains ambiguous), as well as a third group, Haptista, whose position is unclear (fig. 1). In order to test for the consistency of these results, we performed two alterations of the main 248 genes/109 OTUs supermatrix. First, we evaluated the persistence of

the phylogenetic signal for the Telonemia + Sar and Cryptista + Archaeplastida groupings when progressively decreasing the number of genes used to infer the trees (fig. 2A). Random subsamples of the total 248 genes were generated without replacement in 20% decrements (80% to 20%), with a replication scheme that ensured a 95% probability of sampling every gene in each subset (Brown et al. 2018). Each replicate was then subjected to ML tree reconstruction under the LG+C60+F+Γ4 model and the group support monitored with ultrafast bootstrap (UFBoot) approximation
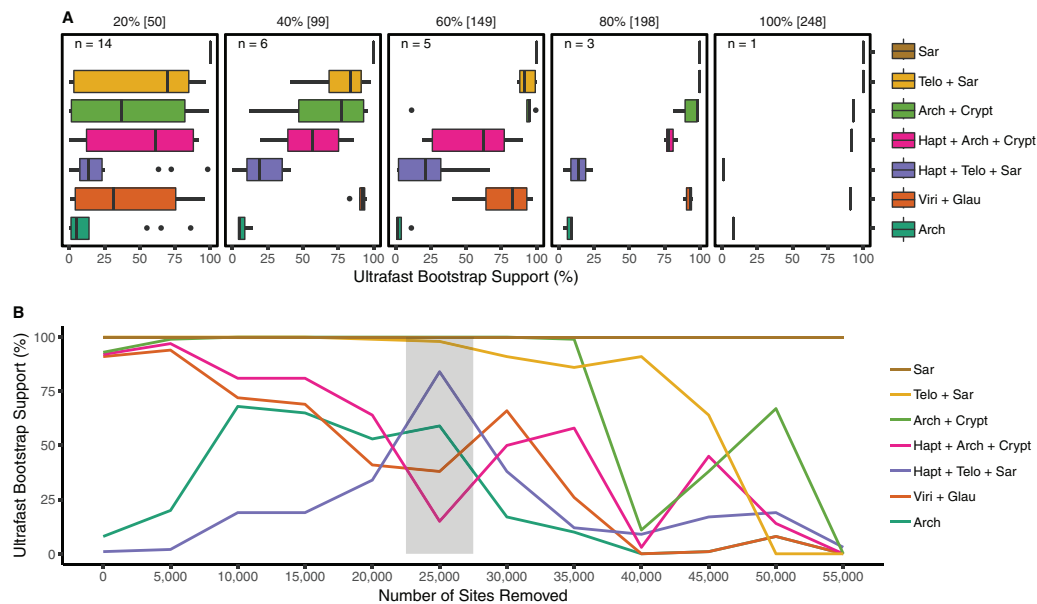
FIG. 2. Gene subsampling and fast-evolving sites removal analyses. (*A*) UFBoot approximation for selected nodes of interest using the full supermatrix (100%) and subsets of randomly sampled genes (20–80%; see Materials and Methods). The variability of support values is shown by Box-and-Whisker plots. (*B*) UFBoot approximation UFBoot for the same nodes as in (*A*) using the full supermatrix (0 sites removed) and subsets from which fast-evolving sites were removed in 5,000 increments (see Materials and Methods). The gray box indicates the reduced data set which recovered a different position for Haptista. Arch, Archaeplastida; Crypt, Cryptista; Glau, Glaucophyta; Hapt, Haptista; Telo, Telonemia; Viri, Viridiplantae.

(Hoang et al. 2018). Both groups (i.e., Telonemia + Sar and Cryptista + Archaeplastida) showed remarkable stability until as few as 60% of the genes were subsampled, corresponding to an alignment length of 149 genes (in average 34,588 AA positions), but started to crumble with fewer genes (fig. 2*A*). None of the other tested groups were as robust against the decreasing number of genes, with the exception of Sar (used as control) that remained maximally supported throughout the experiment (fig. 2*A*).

The second alteration was the removal of the fastest-evolving sites from the full supermatrix, which are more likely to be homoplasic and thus leading to systematic errors due to model misspecification (Rodriguez-Ezpeleta, Brinkmann, Roure, et al. 2007). Sites were sorted from fastest to slowest and progressively removed in 5,000 sites increments until only 3,469 sites were left. Each sub-data set was again used in ML tree reconstruction (LG+C60+F+Γ4 model) with UFBoot approximation (Hoang et al. 2018). The Telonemia + Sar relationship remained strongly supported until 25,000 sites were removed, after which the support was progressively lost but not replaced by an alternative supported relationship (fig. 2*B*). Similarly, the Cryptista + Archaeplastida group retained strong support until the removal of 35,000 sites with no other supported grouping emerging when more sites were removed. In contrast, the position of Haptista was markedly altered by the removal of fast-evolving sites. Starting from a maximally supported sister relationship to the Cryptista + Archaeplastida group (figs. 1 and 2*B*), Haptista shifted to a sister position to the Telonemia + Sar group after the removal of 25,000 sites (82% UFBoot), before reverting again to its original position with further site removal albeit with no real statistical significance (fig. 2*B*). In order to further

assess the strength of the Telonemia + Sar + Haptista relationship, we ran a Bayesian analysis with the CAT+F81 model (Lartillot et al. 2013) using the 25,000 sites-removed data set. Interestingly, the two converged chains were fully consistent with the ML tree and the position of Haptista as sister to Telonemia + Sar received maximal PP (fig. 3 and supplementary fig. S4, Supplementary Material online).

Discussion

The position of Telonemia in the tree of eukaryotes has been debated ever since the group was first described (Vørs 1992; Klaveness et al. 2005; Shalchian-Tabrizi et al. 2006). The difficulty in placing Telonemia is due to a combination of several factors, including the ancient origin of the group, the sparse sampling of molecular data for group members, and the lack of closely related outgroup lineages. Earlier phylogenomic studies often placed Telonemia in the vicinity of Sar or close to cryptist or haptist lineages but never convincingly (Burki et al. 2009, 2012, 2016; Katz and Grant 2015; Lax et al. 2018). Here, we tested whether improving the gene coverage in our phylogenomic supermatrix and adding more telonemid representatives help to infer the origin of this group with more confidence. The results of the phylogenetic analyses show that this is the case: for the first time, the placement of Telonemia in the global phylogeny of eukaryotes receives strong and consistent support across multiple analyses (ML and Bayesian analyses using site-heterogeneous models, fig. 1). Telonemia was inferred as the most closely related lineage to the Sar supergroup. This position is robust against variations in the gene sets although it requires a substantial amount of data to be recovered (typically above 150 genes), and is stable
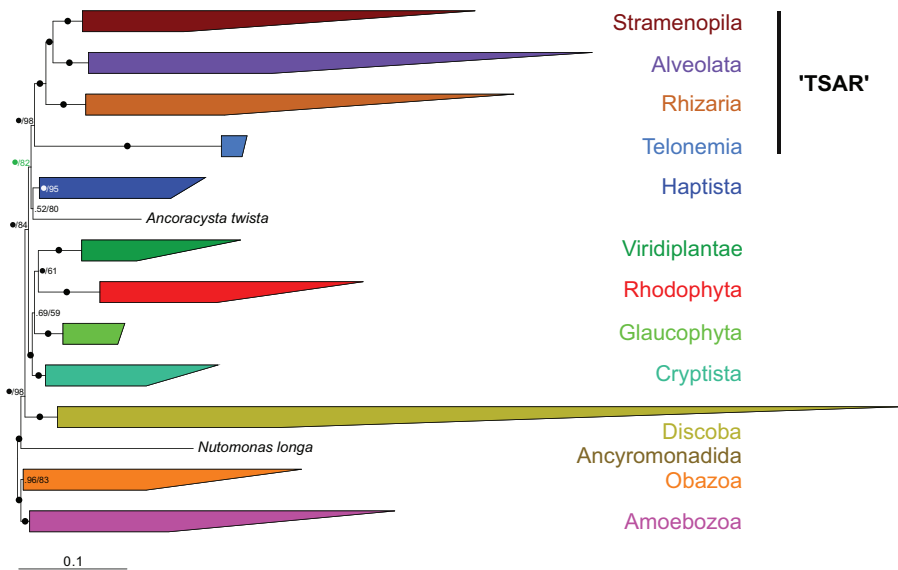
FIG. 3. Simplified Phylobayes consensus tree inferred from a reduced concatenated alignment (25,000 fast-evolving sites removed) of the 248 protein-coding genes under the CAT+F81+Γ4 model (for a full version showing all taxa, see supplementary fig. S4, Supplementary Material online). Node support is given by Bayesian PPs and UFBoot approximation inferred from a ML analysis (LG+C60+G+F model) using the same subalignment (i.e., 25,000 sites removed; see text and fig. 2). Nodes with black circles are fully supported in both analyses. A number appended to a black circle indicates full support in only one of the two analyses (i.e., PP/UFBoot). Note the complete congruence between the Bayesian and ML analyses (all nodes are recovered by both methods), including the position of Haptista.

when the most saturated sites are removed. It is thus unlikely to be due to convergence from undetected multiple substitutions (fig. 2). To acknowledge the strength of this relationship and facilitate further discussions, we introduce the constructed name "TSAR" to refer to the Telonemia + Sar grouping. We propose that TSAR forms a provisional supergroup of eukaryotes.

The hypothesis of TSAR has multiple implications. Most importantly, it confirms that Telonemia is a *bona fide* eukaryotic phylum, at the same phylogenetic level as other major phyla such as the Sar members (Stramenopila, Alveolata, and Rhizaria) (Burki et al. 2007; Hackett et al. 2007; Rodriguez-Ezpeleta, Brinkmann, Burger, et al. 2007). A unique combination of ultrastructural cellular characteristics suggested that Telonemia represents a deeply branching lineage without clear evolutionary affinities (Shalchian-Tabrizi et al. 2006; Yabuki et al. 2013). However, several morphological features, for example the peripheral vacuoles, extrusomes, or tubular tripartite hairs, were also suggested to resemble cell structures found in specific lineages such as Stramenopila, Alveolata, or Rhizaria (Shalchian-Tabrizi et al. 2006; Yabuki et al. 2013). Here, we show that Telonemia is indeed closely related to the Sar group, indicating that these morphological similarities might be homologous. If true, this could explain the origin of important synapomorphies in members of Sar from a common ancestor of the group, namely the alveoli in Alveolata from peripheral vacuoles and the mastigonemes in Stramenopila from similar flagellar hairs. Telonemia thus appears crucial for understanding the early evolution of Sar, a globally distributed and highly diverse assemblage of eukaryotes likely containing the majority of protist lineages

(del Campo et al. 2014; de Vargas et al. 2015; Grattepanche et al. 2018). More generally, our study identifies Telonemia as crucial to infer the evolution of other important characteristics of eukaryotes, for example their cytoskeleton that underlies cell morphology and motility (Dacks et al. 2016). With its unique, highly complex multilayered cytoskeleton (Yabuki et al. 2013), it is possible that Telonemia has retained characteristics of an ancestral complex organization before the cytoskeleton was differentially reduced in other eukaryotic groups. A better understanding of the cytoskeleton evolution and other cell systems will emerge as more data are gathered from diverse eukaryotes, especially from deep-branching lineages like Telonemia.

In addition to TSAR, our analyses strongly recovered another major clade in Diaphoretickes, including Archaeplastida members (not consistently monophyletic) and Cryptista (AC clade; fig. 1). This association has been recovered before in taxon-rich phylogenomics, and its implication for eukaryote evolution discussed (Burki et al. 2016; Cenci et al. 2018). However, the AC clade still awaits confirmation because other recent studies displayed alternative relationships, albeit with no conclusive support and poorer taxon sampling for the relevant groups (Yabuki et al. 2014; Cavalier-Smith et al. 2015, 2018; Katz and Grant 2015). Our analyses also failed to consistently place Haptista in the tree. Based on the full data set, ML and some unconverged Bayesian chains favored a close relationship to the AC clade (fig. 1). But after the removal of the 25,000 fastest-evolving sites, both ML and converged Bayesian analyses supported the placement of Haptista and TSAR together (fig. 3). The position of Haptista has varied dramatically with conflicting confidence

in support (e.g., Yabuki et al. 2014; Katz and Grant 2015; Burki et al. 2016; Cavalier-Smith et al. 2018; Cenci et al. 2018). Hence, moving forward to improve the deep structure of Diaphoretickes, it will be crucial to specifically address the causes for the lack of consensus for the position of Cryptista, Haptista, and the monophyly of Archaeplastida.

Despite these lasting uncertainties surrounding the relationships within Diaphoretickes, we present here consistent results that bring us one step closer to a fully resolved eukaryotic tree of life. Telonemia was once a notorious rogue lineage, but using new transcriptome data for this important but neglected phylum, we show that it now occupies a robust position in the tree as part of an expanded supergroup (TSAR). We thus advise that Telonemia should be routinely included in future works addressing the broad-scale evolution of eukaryotes, in the same way as other major lineages are included. More generally, this study follows a series of recent phylogenomic investigations that successfully placed in the tree several orphan taxa (Brown et al. 2013, 2018; Burki et al. 2016; Janouškovec et al. 2017; Lax et al. 2018). This demonstrates that obtaining adequate genomic data sets (including transcriptomics) for orphans can go a long way towards filling important evolutionary gaps. With the maturation of culture-free genomic solutions, we expect that inaccuracy due to poor taxon sampling will continue to reduce in phylogenomic investigations. Thus, even uncultured long-standing mystery taxa will soon fill the tree of life and contribute to a better understanding of ancient eukaryote evolution.

Materials and Methods

RNA Extraction, Sequencing, and Assembling

Clonal cultures of three telonemid species were isolated from Arctic surface waters of the Kara Sea, N 75.888, E 89.508, total depth: 52 m, water temperature: 2.3°C. Isolates were propagated on the bodonid *Procryptobia sorokini* Frolov, Karpov, and Mylnikov, 2001 strain B-69 grown in Schmalz-Pratt's medium by using the bacterium *Pseudomonas fluorescens* as food (Tikhonenkov et al. 2014).

For cDNA preparation, cells were harvested following peak abundance after eating most of the prey. Cells were collected by centrifugation (2,000 × g, room temperature) on the 0.8 μm membrane of Vivaclear Mini columns (Sartorium Stedim Biotech Gmng, Germany, Cat. No. VK01P042). Total RNA was extracted using the RNAqueous-Micro Kit (Invitrogen, Cat. No. AM1931) and was converted into cDNA prior to sequencing using the SMART-Seq v4 Ultra Low Input RNA Kit (Takara Bio USA, Inc., Cat. No. 634888). Sequencing was performed on the Illumina MiSeq platform with read lengths of 300 bp using the NexteraXT protocol (Illumina, Inc., Cat. No. FC-131-1024) to construct paired-end libraries.

Reads were trimmed with Trimmomatic v. 0.36 (Bolger et al. 2014) using the following parameters: LEADING: 5, TRAILING: 5, SLIDINGWINDOW: 5: 16, MINLEN: 80. Potential vector sequences were detected (and removed if applicable) by BLAST searches against the UniVec database (e-value: 1e-4). Cleaned reads were assembled and translated into amino acids with Trinity v. 2.4.0 and TransDecoder,

respectively (default settings Haas et al. 2013). In addition to the reads obtained in our study, publicly available reads of other taxa were trimmed, assembled, and translated as described above but with slightly different settings for the trimming procedure (LEADING: 3, TRAILING: 3, SLIDINGWINDOW: 4: 15, MINLEN: 36).

Phylogenomic Data Set Construction

A previously published data set comprising 263 genes and 234 taxa was used as starting point for this study (Burki et al. 2016). A broad range of new taxa were added by the following successive steps (sources of data: ensemblgenomes.org, imicrobe.us/#/projects/104, ncbi.nlm.nih.gov, onekp.com): 1) Protein sequences for each taxon were clustered with CD-HIT (Fu et al. 2012) using an identity threshold of 85%; 2) Homologous copies were retrieved by BLASTP searches using the 263 genes as queries (e-value: 1e-20; coverage cutoff: 0.5); 3) Phylogenetic trees were constructed and carefully inspected in order to detect and remove contaminants and select orthologous copies. Sequences were removed if their positions in gene trees conflicted with an initial species tree based on an intermediate concatenated alignment (i.e., after first round of cleaning, see below); this initial species tree globally corresponded to a consensus phylogeny of eukaryotes (e.g., Burki 2014). An automated coloring scheme was developed for easier visual inspection of the gene trees in FigTree v. 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/; Last accessed February 6, 2019). Three rounds of inspection were performed, first to detect deep paralogs and then progressively refined to remove contamination and select orthologs. Sequences were aligned with MAFFT v. 7.310 (Katoh and Standley 2013) using the -auto option (first round) and with MAFFT L-INS-i using the default settings (second and third round), and filtered with trimAL v. 1.4 (Capella-Gutiérrez et al. 2009) using a gap threshold of 0.8 (all three rounds). FastTree v. 2.1.10 (Price et al. 2010) employing -lg -gamma plus options for more accurate performances (first round) and RAxML v. 8.2.10 (Stamatakis 2014) employing PROTGAMMALGF and 100 rapid BP searches (second and third round) were used to infer single gene ML trees. Three genes (FTSJ1—RNA methyltransferase 1, tubb—beta-tubulin, and eif5A—eukaryotic translation initiation factor 5A) were excluded from the base data set due to suspicious clustering of major groups that appeared after the inclusion of new taxa: in FTSJ1, a duplication of the entire Sar group was detected; in tubb, ambiguous relationships were observed due to duplications within Stramenopila and Opisthokonta; in eif5A, an unexpected clustering of Cryptista and Rhizaria was recovered. Twelve other genes were removed because they lacked telonemid sequences.

The initial data set comprised 248 genes and 293 taxa. An unaligned version of these genes was subjected to PREQUAL v. 1.01 (Whelan et al. 2018), a new software to remove or mask stretches of nonhomologous characters in individual sequences. A PP threshold of 0.95 (-filterthresh) was used. MAFFT G-INS-i with the VSM option (−unalignlevel 0.6) was then used for alignment, and ambiguously aligned sites were filtered out with BMGE v.

1.12 (Criscuolo and Gribaldo 2010) with the parameters -g 0.2, -b 5, -m, and BLOSUM75. Partial sequences belonging to the same taxon that did not show evidence for paralogy or contamination on the gene trees were merged. Of the initial 293 taxa, a reduced set of 148 taxa was selected to allow computationally intensive analyses based on 1) knowledge of sequence divergence (slower-evolving species were retained when possible), 2) to represent all major eukaryotic groups with sufficient data (amino acid positions in supermatrix >55%); of all major eukaryotic lineages with genomic or transcriptomic data, only Picozoa was deemed of insufficient coverage, and 3) sequence completeness among the genes. Furthermore, in order to reduce missing data, several strains and species complexes were combined when monophyly inferred from a preliminary concatenated alignment was unambiguous (supplementary fig. S1, Supplementary Material online; see also supplementary table 1, Supplementary Material online for a list of 109 taxa and chimera that were used as OTUs. Finally, all 248 genes were concatenated into a supermatrix (58,469 amino acid positions) with SCaFos v. 1.25 (Roure et al. 2007).

Phylogenomic Analyses

ML trees were inferred from the supermatrix using IQ-TREE v. 1.6.3 (Nguyen et al. 2015). Following the Akaike information criterion and the Bayesian information criterion, the best fitting model was the LG+C60 mixture model (Quang et al. 2008; Minh et al. 2013) with stationary amino acid frequencies optimized from the data set (F) and four discrete gamma (Γ) categories for rate heterogeneity (LG+C60+F+Γ4). The tested models were site-homogeneous LG +/− F, I, Γ4; site-heterogeneous LG +/− F, Γ4 and C10–C60. The best ML tree for the full data set was used as guide tree under the same model to estimate the PMSFs (Wang et al. 2018) model, which was used for further tree reconstructions (LG+C60+F + Γ4+PMSF). Support values for the best LG+C60+F + Γ4+PMSF tree were obtained using nonparametric BP with 100 replicates.

Bayesian analyses were inferred using PHYLOBAYES-MPI v. 1.8 (Lartillot et al. 2013) with the CAT+GTR+Γ4 model (Lartillot and Philippe 2004). Constant sites were removed to decrease running time (-dc). Four independent Markov Chain Monte Carlo (MCMC) chains were run for >6,300 generations (all sampled). For each chain, the burnin period was chosen after monitoring the evolution of the log-likelihood (Lnl) at each sampled point, removing the generations before stabilization (generally 1,250) prior to computing a 50% majority-rule consensus tree of all chains using bpcomp. Global convergence between chains was assessed by the maxdiff statistics measuring the discrepancy in PPs. Unfortunately, global convergence was never achieved in any combination of two chains; this is a known issue of PHYLOBAYES, which is especially acute with taxon-rich data sets and has been reported on many occasions (e.g., Katz and Grant 2015; Burki et al. 2016; Kang et al. 2017). The discrepancies between chains concerning nodes under scrutiny in this study were labeled as unresolved.

Gene Subsampling and Fast-Evolving Sites Removal

From the 248 gene data set, random subsampling without replacement of 80% to 20% (in 20% increments) of genes were performed with replications (3, 5, 6, and 14 replicates in each subset, respectively); this allowed a >95% probability that every gene is represented at least once (following an approach described in Brown et al. 2018). The genes in each subset were then concatenated with SCaFoS. Fast-evolving sites were estimated with IQ-TREE v. 1.6.5 (option -wsr) giving the PMSF tree as fixed tree topology and using the LG+C60+F + Γ4 model. Sites were then removed in 5,000 increments using a Perl script. For all the subsamples with reduced number of genes or sites, phylogenetic trees with UFBoot (1,000 replicates) were inferred using IQ-TREE (LG+C60+F+Γ4) with the -bb, -wbtl, and -bnni flags employed. One sub-alignment (25,000 sites removed) was used in a Bayesian tree inference using the site-heterogeneous CAT+F81+Γ4 model as implemented in PHYLOBAYES-MPI v. 1.8 (Lartillot et al. 2013). We did not use here the CAT+GTR + Γ4 model for computational tractability. Two independent MCMC chains were run for >6,300 cycles. Before computing the majority-rule consensus tree from the two chains combined, the initial 2,500 trees in each MCMC run were discarded as burnin after checking for convergence in likelihood and in clade PPs (maxdiff <0.3).

Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

References

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al. 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. J Eukaryot Microbiol 66:4–119.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Bråte J, klaveness D, Rygh T, Jakobsen KS, Shalchian-Tabrizi K. 2010. Telonemia-specific environmental 18S rDNA PCR reveals unknown

diversity and multiple marine-freshwater colonizations. *BMC Microbiol.* 10:168.

Brown MW, Heiss AA, Kamikawa R, INAGAKI Y, Yabuki A, Tice AK, Shiratori T, Ishida K-I, Hashimoto T, Simpson AGB, et al. 2018. Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biol Evol.* 10(2): 427–433.

Brown MW, Sharpe SC, Silberman JD, Heiss AA, Lang BF, Simpson AGB, Roger AJ. 2013. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proc R Soc B.* 280(1769):20131755.

Brugerolle G, Bricheux G, Philippe H, Coffe 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.

Burki F, Inagaki Y, Bråte J, Archibald JM, Keeling PJ, Cavalier-Smith T, Sakaguchi M, Hashimoto T, Horak A, Kumar S, et al. 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, telonemia and centroheliozoa, are related to photosynthetic chromalveolates. *Genome Biol Evol.* 1:231–238.

Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, Smirnov A, Mylnikov AP, Keeling PJ. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc R Soc B* 283:

Burki F, Okamoto N, Pombert J-F, Keeling PJ. 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proc R Soc B.* 279(1736):2246–2254.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* 2(8):e790.

Burki F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb Perspect Biol.* 6(5):a016147.

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.

Cavalier-Smith T, Chao EE, Lewis R. 2015. Multiple origins of Heliozoa from flagellate ancestors: new cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. *Mol Phylogenet Evol.* 93:331–362.

Cavalier-Smith T, Chao EE, Lewis R. 2018. Multigene phylogeny and cell evolution of chromist infrakingdom Rhizaria: contrasting cell organisation of sister phyla Cercozoa and Retaria. *Protoplasma* 59:429.

Cenci U, Sibbald SJ, Curtis BA, Kamikawa R, Eme L, Moog D, Henrissat B, Maréchal E, Chabi M, Djemiel C, et al. 2018. Nuclear genome sequence of the plastid-lacking cryptomonad Goniomonas avonlea provides insights into the evolution of secondary plastids. *BMC Biol.* 16(1):137.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol.* 10:210.

Dacks JB, Field MC, Buick R, Eme L, Gribaldo S, Roger AJ, Brochier-Armanet C, Devos DP. 2016. The changing view of eukaryogenesis - fossils, cells, lineages and how they all come together. *J Cell Sci.* 129(20):3695–3703.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, et al. 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348(6237):1261605–1261605.

del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. 2014. The others: our biased perspective of eukaryotic genomes. *Trends Ecol Evol.* 29(5): 252–259.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152.

Grattepanche J-D, Walker LM, Ott BM, Paim Pinto DL, Delwiche CF, Lane CE, Katz LA. 2018. Microbial Diversity in the eukaryotic SAR clade: illuminating the darkness between morphology and molecular data. *Bioessays* 40(4):e1700198.

Griessmann K. 1913. Über marine flagellaten. *Arch Protistenk* 32:1–78.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 8(8):1494–1512.

Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rümmele SE, Bhattacharya D. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. *Mol Biol Evol.* 24(8):1702–1713.

Heiss AA, Kolisko M, Ekelund F, Brown MW, Roger AJ, Simpson AGB. 2018. Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *R Soc Open Sci.* 5(4):171707.

Heiss AA, Walker G, Simpson AGB. 2011. The Ultrastructure of ancyromonas, a eukaryote without supergroup affinities. *Protist* 162(3):373–393.

Hoang DT, Chernomor O, Haeseler von A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* 35(2): 518–522.

Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Rohwer FL, Mylnikov AP, Keeling PJ. 2017. A New Lineage of eukaryotes illuminates early mitochondrial genome reduction. *Curr Biol.* 27(23):3717–3724.e5.

Kang S, Tice AK, Spiegel FW, Silberman JD, Pánek T, Čepička I, Kostka M, Kosakyan A, Alcântara DM, Roger AJ, et al. 2017. Between a pod and a hard test: the deep evolution of amoebae. *Mol Biol Evol.* 34(9):2258–2270.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30(4): 772–780.

Katz LA, Grant JR. 2015. Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. *Syst Biol.* 64(3): 406–415.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW. 2005. The tree of eukaryotes. *Trends Ecol Evol.* 20(12): 670–676.

Klaveness D, Shalchian-Tabrizi K, Thomsen HA, Eikrem W, Jakobsen KS. 2005. Telonema antarcticum sp. nov., a common marine phagotrophic flagellate. *Int J Syst Evol Microbiol.* 55(6):2595–2604.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21(6):1095–1109.

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Phil Trans R Soc B* 363(1496):1463–1472.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol.* 62(4):611–615.

Lax G, Eglit Y, Eme L, Bertrand EM, Roger AJ, Simpson AGB. 2018. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* 564(7736):410–414.

Minh BQ, Nguyen MAT, Haeseler von A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol.* 30(5):1188–1195.

Nguyen L-T, Schmidt HA, Haeseler von A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1): 268–274.

Price MNM, Dehal PSP, Arkin APA. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.

Quang LS, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24:2317–2323.

Rodriguez-Ezpeleta N, Brinkmann H, Burger G, Roger AJ, Gray MW, Philippe H, Lang BF. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Curr Biol.* 17(16): 1420–1425.

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

Roure B, Rodriguez-Ezpeleta N, Philippe H. 2007. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol.* 7(Suppl 1):S2.

Shalchian-Tabrizi K, Eikrem W, Klaveness D, Vaulot D, Minge MA, Le Gall F, Romari K, Throndsen J, Botnen A, Massana R, et al. 2006. Telonemia, a new protist phylum with affinity to chromist lineages. *Pro. Biol Sci.* 273(1595):1833–1842.

Shalchian-Tabrizi K, Kauserud H, Massana R, Klaveness D, Jakobsen KS. 2007. Analysis of environmental 18S ribosomal RNA sequences reveals unknown diversity of the cosmopolitan phylum Telonemia. *Protist* 158(2):173–180.

Simpson AG, Slamovits CH, Archibald JM. 2017. Protist diversity and eukaryote phylogeny. In: Archibald JM, Simpson AGB, Slamovits CH, editors. Springer International Publishing. p. 1657. https://link.springer.com/referenceworkentry/10.1007/978-3-319-28149-0_45

Simpson AGB, Roger AJ. 2002. Eukaryotic evolution: getting to the root of the problem. *Curr Biol.* 12(20):R691–R693.

Simpson AGB, Roger AJ. 2004. The real "kingdoms" of eukaryotes. *Curr Biol.* 14(17):R693–R696.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Tikhonenkov DV, Janouškovec J, Mylnikov AP, Mikhailov KV, Simdyanov TG, Aleoshin VV, Keeling PJ. 2014. Description of Colponema vietnamica sp.n. and Acavomonas peruviana n. gen. n. sp., two new alveolate phyla (Colponemidia nom. nov. and Acavomonidia nom. nov.) and their contributions to reconstructing the ancestral state of alveolates and eukaryotes. *PLoS One* 9(4):e95467.

Vørs N. 1992. Heterotrophic Amoebae, Flagellates and Heliozoa from the Tvärminne Area, Gulf of Finland, in 1988–1990. *Ophelia* 36(1):1–109.

Walker G, Dacks JB, Martin Embley T. 2006. Ultrastructural description of Breviata anathema, n. gen., n. sp., the organism previously studied as "Mastigamoeba invertens". *J Eukaryot Microbiol.* 53(2):65–78.

Wang H-C, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst Biol.* 67(2):216–235.

Whelan S, Irisarri I, Burki F. 2018. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* 496:311–312.

Yabuki A, Chao EE, Ishida K-I, Cavalier-Smith T. 2012. Microheliella maris (Microhelida ord. n.), an ultrastructurally highly distinctive new axopodial protist species and genus, and the unity of phylum Heliozoa. *Protist* 163(3):356–388.

Yabuki A, Eikrem W, Takishita K, Patterson DJ. 2013. Fine structure of Telonema subtilis Griessmann, 1913: a flagellate with a unique cytoskeletal structure among eukaryotes. *Protist* 164(4):556–569.

Yabuki A, Kamikawa R, Ishikawa SA, Kolisko M, Kim E, Tanabe AS, Kume K, Ishida K-I, Inagki Y. 2014. Palpitomonas bilix represents a basal cryptist lineage: insight into the character evolution in Cryptista. *Sci Rep.* 4:4641.

Zhao S, Burki F, Brate J, Keeling PJ, klaveness D, Shalchian-Tabrizi K. 2012. Collodictyon–an ancient lineage in the tree of eukaryotes. *Mol Biol Evol.* 29(6):1557–1568.