Title

Reducing long-branch effects in multi-protein data uncovers a close relationship between Alveolata and Rhizaria

Authors

Ding He[1*], Roberto Sierra[2], Jan Pawlowski[2] and Sandra L. Baldauf [1]

[1]Program in Systematic Biology, Department of Organismal Biology, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden.

[2]Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland.

*Author for correspondence

Ding He

Norbyvägen 18D, Evolutionary Biology Centre, 75236 Uppsala, Sweden

Tel: +46 (0)18-47-2777

Email: ding.he@ebc.uu.se

Abstract

Rhizaria is a major eukaryotic group of tremendous diversity, including amoebae with spectacular skeletons or tests (Radiolaria and Foraminifera), plasmodial parasites (Plasmodiophorida) and secondary endosymbionts (Chlorarachniophyta). Current phylogeny places Rhizaria in an unresolved trichotomy with Stramenopila and Alveolata (supergroup "SAR"). We assembled a 147-protein data set with extensive rhizarian coverage (M147), including the first transcriptomic data for a euglyphid amoeba. Phylogenetic pre-screening of individual proteins indicated potential problems with radically misplaced sequences due either to contamination of rhizarian sequences amplified from wild collected material and/or extremely long branches (xLBs). Therefore, two data subsets were extracted containing either all proteins consistently recovering rhizarian monophyly (M34) or excluding all proteins with ≥3 xLBs (defined as ≥2x the average terminal branch length for the tree). Phylogenetic analyses of M147 give conflicting results depending on the outgroup and method of analysis but strongly support an exclusive Rhizaria + Alveolata (R+A) clade with both data subsets (M34 and M37) regardless of phylogenetic method used. Support for an R+A clade is most consistent when a close outgroup is used and decreases with more distant outgroups, suggesting that support for alternative SAR topologies may reflect a long-branch attraction artefact. A survey of xLB distribution among taxa and protein functional category indicates that small "informational" proteins in particular have highly variable evolutionary rates with no consistent pattern among taxa.

Keywords

SAR; Rhizaria; Phylogenomics; Deep phylogeny; Eukaryote phylogeny; Long branch attraction.

1. Introduction

Rhizaria is a relatively recently recognized supergroup of eukaryotes consisting almost entirely of single-celled organisms (Nikolaev et al. 2004; Burki and Pawlowski 2006). These taxa inhabit diverse ecological niches and possess an apparently vast biological diversity that is still mostly unexplored (Burki and Keeling 2014). The group consists largely of heterotrophic naked and testate amoebae characterized by thread-like pseudopodia functioning in food acquisition and cell movement. Testate rhizarians include the Foraminifera with multi-chambered calcium carbonate shells and the Radiolaria with complex "ornate" skeletons of strontium sulfate or silica, famously illustrated by Ernst Haeckel. Both of these groups contribute substantial microfossils that are widely used for paleogeographic and paleoclimatic reconstruction (Boudagher-Fadel 2012). Most of the naked filose rhizarian amoebae comprise the Filosa (core Cercozoa), which includes the chlorarachniophyte algae, the euglyphid testate amoebae, and the multicellular sorocarpic *Guttulinopsis vulgaris* (Brown et al. 2012). Other rhizarian lineages include obligate parasites of plants (*Plasmodiophora*) and marine invertebrates (Haplosporidia).

Rhizaria is an ancient and diverse taxon, with so-far no recognized morphological synapomorphy. However, their pseudopodia (rhizopodia), where present, are highly distinct from those of the other two major amoeboid groups, the Amoebozoa (lobopodia) and Heterolobosea (eruptive pseudopodia) (Adl et al. 2012). Rhizopodia tend to be long and thin, sometimes anastomosing to form dynamic networks and primarily involved in passive prey capture, while the pseudopodia of Amoebozoa and Heterolobosea tend to be short and blunt allowing active pursuit and capture of prey (Adl et al. 2012). Cultivation of most rhizarians has proven challenging, and therefore their phylogeny has relied heavily on molecular sequences. Although the latter data remain scarce, recent multigene phylogenies strongly support rhizarian monophyly, after excluding two divisions of the former Heliozoa (Nikolaev

et al. 2004). More recent phylogenomic data have made it possible to begin to identify and resolve relationships among the major divisions of Rhizaria as well (Sierra et al. 2013; Sierra et al. 2016).

Nonetheless, the exact higher order placement of Rhizaria in the eukaryote tree of life remains uncertain. Multigene phylogeny strongly and consistently places Rhizaria together with the former core chromalveolates (Sierra et al. 2016). These are the Alveolata, including Ciliophora, Dinoflagellata and Apicomplexa, and the Stramenopila, including various heterotrophic lineages (*e.g.* Oomycetes, Bicosoecida, Labyrinthuloides, and Opalinida) as well as several major lineages of uni- and/or multicellular algae. The resulting superclade is referred to as SAR, based on the first letters of the three component taxa (Burki et al. 2007). However , the position of Rhizaria within SAR remains uncertain and all three possible topologies have been recovered (Parfrey et al. 2010; Burki et al. 2012; Sierra et al. 2013). The closest sister group to SAR appears almost certainly to be the Archaeplastida (formerly Plantae) and probably at least some components of the controversial supergroup "Hacrobia" (including Haptophyta and Cryptophyta; (Okamoto et al. 2009; Burki et al. 2012)). Together these five major lineages (Stramenopila, Alveolata, Rhizaria, Hacrobia and Archaeplastida) are now designated as suprakingdom Diaphoretickes (Adl et al. 2012).

Given large amounts of new rhizarian EST data including the first transcriptome from a rhizarian testate amoeba (*Euglypha rotunda*, Filosa), we sought to re-test the SAR topology using far broader taxon sampling than previously possible. Two pre-existing multi-gene data sets were used: a 119-protein data set rich in informational proteins (Burki et al. 2013) and a 37-protein set rich in mitochondrial proteins (He et al. 2014). The data sets were expanded to include the new sequences and then systematically screened for two important artefacts - potentially contaminant sequences derived from wild-collected material and extraordinarily long phylogenetic branches to which some SAR taxa are particularly prone. Combining the

data and filtering for these artefacts uncovered a single strong signal that places Rhizaria as more closely related to Alveolata than to Stramenopila. A survey of extreme long-branch distribution by taxon and gene showed surprising patterns, particularly among informational proteins.

2. Materials and Methods

2.1. The Euglypha transcriptome

Euglypha rotunda CCAP 1530/1 was cultured in NCL/0.01% NPA media. The cells were collected and total RNA extraction was performed using the NucleoSpin XS kit (Macherey-Nagel). An Illumina RNA-seq library was constructed using the TruSeq kit (Illumina) and sequenced in a 200-cycle run using a HiSeq 2500. The sequencing yielded 151,851,442 raw reads that were trimmed and filtered based on quality scores using in-house scripts. The high-quality reads were assembled into contigs using velvet/oases software (Schulz et al. 2012) using the cDNA read assembly parameters.

2.2. Initial data set assembly and phylogenetic screening

A previously developed rhizarian-enriched data set of 119 proteins (Burki et al. 2013) was augmented with a the 37 euBac protein data set previously developed for deep eukaryote phylogeny (euBac proteins; (He et al. 2014)). Rhizarian sequences to be added to the euBac data were identified by tBLASTn search against 14 rhizarian translated EST databases (Sierra et al. 2013) and the *E. rotunda* RNAseq contigs (see above). All tBLASTn hits with e-value < 1e-5 were translated and added to their corresponding single protein sequence alignments, which were then realigned using MUSCLE (Edgar 2004). Ambiguously aligned regions were selected and trimmed using Gblocks (Castresana 2000) using the less stringent criteria implemented in SeaView (Gouy et al. 2010).

Individual proteins were screened using single gene trees (SGTs) constructed by RAxML with the PROTGAMMALGF model and 100 rapid bootstrap replicates (Stamatakis 2014). SGTs were examined by eye to identify potential non-orthologous or contaminant sequences. These were defined as rhizarian sequences strongly grouping with non-rhizarian taxa (mlBP > 60%) through several rounds of progressive taxon elimination (see below). Four euBac proteins were then excluded due to lack of reliable rhizarian sequences, and four proteins included in both data sets were reduced to single copies. The result was a full data set of 147 protein partitions (M147). Non-rhizarian taxa were selected with an aim toward a balanced sampling of other eukaryote major groups (supplementary table S1).

To minimize missing data within Rhizaria, five chimeric rhizarian taxa were constructed using sequences from multiple representatives of robust rhizarian clades (Sierra et al. 2013) as follows: *Foraminifera* (*Ammonia sp*, *Brizalina sp*, *Bulimina marginata*, *Globobulimina turgida* and *Nonionellina sp*.), *Acantharea* (*Astrolonche serrata*, *Amphilonche elongata* and *Phyllostaurus siculus*), *Polycystinea* (*Collozoum sp* and *Spongosphaera streptacantha*), *Cercozoa* (*Aulacantha scolymantha* and *Paracercomonas marina*), *Plasmodiophorida* (*Plasmodiophora brassicae* and *Spongospora subterranea* (supplementary table S1).

2.2.1 Screening for potential contamination in rhizarian data

There are two major sources of phylogenetic artefact in rhizarian molecular data, contaminant sequences from non-axenic substrates and the tendency of rhizarian sequences to form extremely long branches in evolutionary trees. Contaminant sequences were defined here as those failing to group with Rhizaria in SGTs. Eighteen proteins (~12%) showed strictly monophyletic Rhizaria, while an additional 16 proteins produced a "nearly-monophyletic" Rhizaria with one misplaced taxon as follows: 1) twelve partitions included

one rhizarian sequence grouping with other eukaryotes, and 2) four partitions included one non-rhizarian grouping with an otherwise monophyletic Rhizaria. All potentially contaminant sequences in the 16 partitions (1-2 per partition) were then masked and new SGTs constructed, all of which now produced strict rhizarian monophyly. These 16 masked protein sets were then combined with the 18 other partitions producing rhizarian monophyly without masking to give a set of 34 "mono-rhizarian" partitions (M34; supplementary table S2).

2.2.2 Systematic screening for extreme long-branches

Extremely long branches (xLBs) are defined here as branches ≥2 times the average terminal branch length (AtBL) for an individual partition (protein). AtBL was calculated from SGT branch lengths and defined as the sum of all terminal branches divided by the number of sequences (taxa) in the partition. AtBL and xLB values were calculated with a custom python script (supplementary file 1) utilizing the ETE python toolkit (Huerta-Cepas et al. 2010). The number of xLBs per partition was then tabulated by hand for each of the three major subdivisions of SAR. Twenty-seven partitions were found to have ≤2 xLB. These 27 were concatenated to give the M27 or "xLB-depleted" data set. The overall number of xLBs per partitions ranged from 0-9, with a mean of 3.8.

2.3. Phylogenetic analyses

Three super matrices were assembled using SequenceMatrix (Vaidya et al. 2011): the full data set (M147, 33136 amino-acid positions), the "mono-rhizarian" data set (M34, 9393 amino-acid positions), and the "xLB-depleted" data set (M27, 8737 amino acid positions). Maximum-likelihood phylogenetic analyses with rapid bootstrapping (mlBP) were conducted using RAxML version 8 with the PROTGAMMALGF model and 100 rapid bootstrap

replicates (Stamatakis 2014). Bayesian phylogenetic inference with posterior probabilities (biPP) was conducted with constant sites removed using two independent MCMC chains under CAT-GTR, CAT-Poisson, CAT-covarion or CAT-dayhoff6, all with four discrete gamma rate categories to model among site variation with PhyloBayes 3.3 (Lartillot et al. 2009) or its 1.5a MPI version (Lartillot et al. 2013). Topological convergence of chains was determined (maxdiff < 0.1) after discarding ~50% of cycles as the burnin. Both mlBP and biPP analyses were run either on local machines or using the resources of the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) or the Cyberinfrastructure for Phylogenetic Research (CIPRES) (Miller et al. 2010).

2.4. Approximately unbiased (AU) test

Alternative phylogenetic hypotheses were evaluated using the AU test (Shimodaira 2002). All model parameters were re-estimated for each tree topology and per-site log-likelihoods were calculated using RAxML (Stamatakis 2014). The resulting p-values were calculated from multi-scale RELL bootstrap replicates with default settings using CONSEL (Shimodaira and Hasegawa 2001).

3. Results

3.1. Phylogenomic data sets

Rhizaria is one of the most poorly sequenced major eukaryotic groups, *e.g.* complete genome data are available for only two taxa (the secondarily endosymbiotic green alga, *Bigelowiella natans*, and the naked foraminiferan, *Reticulomyxa filosa*) and, until recently, very little molecular data were available for any other taxon. This lack of sequence and taxon sampling has hindered resolution of the phylogenetic position of Rhizaria within the SAR clade. To take advantage of a recent large influx of rhizarian sequences, we combined two

previously developed multi-protein data sets to address this question. These data set consist of a set of 37 eukaryotic proteins of bacterial ancestry augmented with EST data from 15 Rhizaria (euBacs; (He et al. 2014)) and 119 largely informational proteins with extensive rhizarian coverage that still failed to resolve the position of Rhizaria within SAR (Burki et al. 2013). New transcriptomic data from the Euglyphid amoeba, *E. rotunda*, were then added, thus giving a data set with representatives of nearly every known major division of Rhizaria (supplementary table S1). To reduce potential missing data effects, chimeric taxa were constructed with sequences from sets of closely related species within the rhizarian taxa Foraminifera, Acantharea, Polycystinea, Cercozoa and Plasmodiophorida (Supplementary table S1). The full non-redundant data set consists of 147 proteins with a taxonomically broad and balanced set of 44 Diaphoretickes taxa and 33136 aligned amino-acid positions (M147; supplementary table S1).

Preliminary analysis of the M147 data failed to consistently resolve the branching order within SAR. However, examination of the 147 SGTs indicated two potential problems with the data: many of the proteins failed to reconstruct a monophyletic Rhizaria, and/or showed extremely long Rhizarian branches. Since few rhizarian species can be cultured, most of the EST data were prepared directly from wild collected material (Sierra et al. 2013). Even with careful preparation of material there remains a potentially high risk of contamination from frequent symbionts or other protists trapped within rhizarian tests or partly digested food within the rhizarian cells. However, contamination can be difficult to assess in poorly resolve SGTs. Therefore we took a strict approach of deleting all partitions with multiple taxa violating rhizarian monophyly. In fact, only 18 SGTs showed a monophyletic rhizarian clade, while an additional 16 proteins produced a “nearly-monophyletic” Rhizaria with a single “misplaced” sequence that could be easily masked (supplementary table S2). After masking of these apparently illegitimate sequences, these 34 “mono-rhizarian” partitions were

combined giving a data set of 9393 aligned amino acid positions (M34; supplementary tables S1 and S2).

Rhizarian phylogeny is further confounded by the highly divergent nature of some rhizarian sequences, which cause them to form long branches in phylogenetic trees. This raises the possibility of artifactual clustering of long branches (long branch attraction or LBA). LBA is particularly worrying here, as many Alveolata also tend to have extremely long branches. Elimination of long branches to improve phylogenetic signal is common practice, although this is often done in an *ad hoc* manner (Hampl et al. 2009). To avoid this, and since the branches in question here are not just long but extremely long (≥2 times the length of most other branches in the tree), we developed a simple metric to allow us to systematically survey the full M147 data set for extreme long branches.(xLBs). We defined an xLB as a sequence with a terminal branch length (tBL) two or more times the average terminal branch length (avTBL) for its SGT (tBL/avTBL ≥ 2; details in Materials and Methods). This metric was used to count the number of xLBs per taxon per protein partition, as well as by protein functional category (tables 1 and 2).

A total of 676 xLBs were found in the full M147 data set, including 58 branches >4 times the average terminal branch length for their respective partition (table 2). Rhizaria account for slightly more xLBs overall than alveolates (29% and 24%, respectively), while the overall xLB contribution of stramenopiles is much lower (0.9%; table 2). However in the most extreme category (>4 times average branch length), Rhizaria make a much larger contribution (40%) than either Alveolata (21%) or Stramenopila (14%) (table 2). Thus Rhizaria have strikingly long terminal branches for a substantial number of the proteins comprising the M147 data set (22 proteins; Table 2).

Examination of the frequency of xLBs by protein functional category shows that the tendency to form long branches is not evenly spread across the M147 data set. Proteins

involved in information processing (primarily ribosomal proteins) are among the most highly xLB-prone partitions in the data set. For example, the small and large ribosomal subunit proteins have on average 4.8 and 4.5 xLBs per protein, respectively (table 1). The most extreme case, however, is β-tubulin with nine xLBs in a 49-taxon tree (table 1). Among the most conservative of the M147 proteins are the euBac proteins, which have on average 2.2 xLBs per protein for this taxon set (table 1). The contribution of Rhizaria to the xLB content of the various partitions is also uneven, indicating that the xLB variation among the proteins is not attributable to a specific subset of taxa (table 1).

3.2. The phylogenetic position of Rhizaria

Phylogenetic analyses of the full (M147), mono-Rhizaria (M34), xLB-depleted (M27) data sets using maximum likelihood (mlBP; (Stamatakis 2014)) and Bayesian inference (biPP; (Lartillot et al. 2009)) produce nearly identical trees overall, with all examined eukaryote supergroups reconstructed as monophyletic with full support. This includes full support for the SAR clade as a whole, as well as its three major divisions (Stramenopila, Alveolata and Rhizaria; fig. 1). Phylogenetic relationships within these three divisions are also consistent with previous results (fig. 1; (Burki et al. 2013; Sierra et al. 2013)). Most importantly, maximum likelihood analyses of all three data sets place Rhizaria as more closely related to Alveolata than to Stramenopila (85-99% mlBP; fig. 1). Support for this hypothesis (H1) is particularly strong when the data are cleaned of partitions with potential contaminant sequences (M34: 94% mlBP) or large numbers of xLBs (M27: 99% mlBP) (fig. 1). M34 and M27 also fully support H1 with biPP (1.0 biPP; fig. 1). However, Bayesian analysis of M147 rejects H1 in favor of an Alveolata + Stramenopila clade (hypothesis H2; 1.0 biPP, fig. 1). Thus, M34 and M27 consistently and strongly support H1, while M147 gives weaker results with mlBP and contradictory results with biPP.

To test if two common types of model misspecification - heterotachy and compositional bias - could be effecting biPP results, we conducted Bayesian analyses with a co-varion model to account for heterotachous processes and a with dayhoff6 amino-acid recoding scheme to reduce potential compositional bias (Lartillot et al. 2009). The covarion model had no effect on the results with M34 or M27 (0.99 biPP and 1.00 biPP for H1, respectively), while M147 still recovered topology H2 but with markedly reduced support (0.92 biPP) (supplementary figures S1-S3). The dayhoff6 recoding also had no effect on M147 (0.99 biPP for H2), while M34 and M27 are unresolved, possibly due to the substantial reduction in phylogenetic signal caused by dayhoff recoding (Lartillot et al. 2009) (supplementary figures S4-S6).

3.3. Impact of outgroup selection

Empirical and theoretical studies have shown that using outgroups that are closer to the ingroup (i.e. with a shorter edge connecting to the ingroup) and composed of taxa with shorter terminal branches can counter statistical artifacts such as LBA and random rooting (de la Torre-Bárcena et al. 2009; Schneider and Cannarozzi 2009; Torruella et al. 2012). All current analyses of multigene data support Archaeplastida and Hacrobia (PLH) as close sister taxa to SAR, with Amorphea (AMR) and Discoba (DSC) as more distant relatives (Burki et al. 2013; He et al. 2014; Yabuki et al. 2014). Therefore we tested three alternative hypotheses of SAR - Rhizaria + Alveolata (H1), Alveolata + Stramenopila (H2), and Rhizaria + Stramenopila (H3) - with these close and more distant outgroups. Analyses were conducted using mlBP, biPP and the approximately unbiased (AU) test (table 3; supplementary fig. S8) (Shimodaira 2002).

Both the M34 and M27 data sets support H1 (98-99% mlBP, 0.98-1.00 biPP) and reject the two alternative hypotheses ($p<0.05$) with all outgroups except Discoba, which is

possibly the most distantly related outgroup here and one that includes some notoriously long branches with currently available data (table 3) (He et al. 2014). In fact M27 rejects both alternative hypotheses with $p<0.05$ with all outgroups except when Discoba is used as the sole outgroup, in which case there is no significant support for any hypothesis with any method (table 3). M34 also shows consistent support for H1 with mlBP and all outgroups except Discoba, although with biPP M34 supports H2 or H3 with Amorphea or Discoba, respectively (table 3). In contrast, results with M147 are highly inconsistent with mlBP and the AU test, alternatively supporting all three hypotheses with different outgroup combinations. Nonetheless, this data set consistently supports H2 with biPP (table 3). Among the outgroups, Discoba clearly gives the most inconsistent results, and is the only outgroup that allows the recovery of H3, an hypothesis that is significantly rejected by all data sets with all other outgroups and methods (table 3).

4. Discussion

A 147-protein data set was enriched for rhizarian sequences to give the broadest taxonomic representation of the group to date, including the first deep sequencing data from Euglyphida. Analyses of these data support Rhizaria as more closely related to Alveolata than either is to Stramenopila (H1, fig. 1), particularly when the data are masked for two striking signatures of artefact, contaminant sequences (M34; (mono-rhizarian) and extremely long branches (M27; xLB-depleted). While analyses of the full M147 data give conflicting results depending on method and outgroup, M27 and M34 consistently support H1 with all methods and all but the most distant outgroup (fig. 1; table 3). The fact that support for H1 is highest with xLB-depleted data (M27; table 3) suggests that this topology is not an LBA artefact. The fact that H2 and H3 are mostly supported in analyses using distant outgroups (AMR and DSC;

table 3) suggests that support for competing SAR topologies may reflect LBA of Rhizaria and/or Alveolata to long outgroup branches.

Altogether the M34 and M27 data sets share 11 proteins (supplementary table S2). This is partly because the selection criteria for "xLB-depleted" and "mono-rhizaria" are not entirely independent; highly divergent but legitimate sequences will also give xLBs that may cause them to be radically misplaced. Thus, some of the misplaced sequences in our control trees may reflect LBA rather than contamination. However, xLBs are problematic regardless of their cause, so xLB but taxonomically correct sequences can still cause LBA and excluding them will still reduce the xLB content of the data.. This suggests that the results obtained for M34 and M27 may both reflect decreased xLB content. In fact we found very few well-supported examples of contamination in these data (>60% mlBP), although this is difficult to assess reliably in short informational proteins.

The distribution of xLBs showed no clear pattern in either broad protein functional category or species taxonomy (table 1). Even the presumably conservative "informational" proteins do not appear to be immune. For example, the 43 ribosomal proteins (r-proteins) that make up a large component of the 119-protein phylogenomic data set seem to be surprisingly xLB-prone. In fact r-proteins as a whole have some of the highest xLB content in the M147 data set with an average of 4.5 and 4.8 xLBs per protein for the 24 large and 19 small subunit r-proteins, respectively (Table 1). It should be noted that this does not necessary challenge the theory that informational genes have slower evolutionary rates than "operational" (metabolic) genes, since most r-proteins have a fairly peripheral role in protein synthesis (Lecompte 2002; Noller 2005). In addition, r-proteins may be involved in a variety of extra ribosomal functions that tend to be taxon specific, which could lead to accelerated and taxon-specific evolutionary rates (Warner and McIntosh 2009).

Some contradictory results are seen here with different phylogenetic methods and data sets. Specifically, M147 supports all possible hypotheses with mlBP depending on the outgroup but consistently supports H2 with biPP (table 3). Meanwhile, H2 is consistently rejected by both M27 and M34 except with distant outgroups (table 3). It is tempting to attribute such conflicting results to model misspecification in RAxML, particularly across-site substitution heterogeneity, a major contributor to phylogenetic inaccuracy (Lartillot et al. 2007). Theoretically this should be modelled somewhat more accurately by the site-heterogeneous CAT model used in PhyloBayes (Lartillot and Philippe 2004). However, the fact that PhyloBayes only supports H2 with distant outgroups or highly xLB-rich data (97 of the M147 proteins have ≥3 xLBs and/or ≥2 radically misplaced sequences) suggests that this method or its implementation may be sensitive to misleading signal associated with LBA or data contamination.

Rhizaria have long been one of the most poorly molecularly sampled major taxa within eukaryotes (Burki and Keeling 2014). This has undoubtedly contributed to the difficulty in resolving its higher-level phylogenetic affinities. For example, the group was only recognized as a major taxon in 2004 (Nikolaev et al. 2004) and even more recently assigned to supergroup SAR (Burki et al. 2007; Burki 2014). Within SAR, recent analyses have tended to favor the long-held Stramenopila+Alveolata grouping ("halvaria") (Burki and Keeling 2014), supported by many phylogenetic analyses of both nuclear and chloroplast genes over the past 10 years (Keeling 2013). However, these analyses mostly either lack substantial data for Rhizaria (Zhao et al. 2012; Yabuki et al. 2014) and/or have only extremely long rhizarian branches to work with (Brown et al. 2012; Burki et al. 2013).

In contrast, the results presented here strongly support Alveolata as the sister group to Rhizaria, particularly when the data are masked for potential contaminant sequences and/or xLB sequences (Fig. 1; table 3). Thus, our analyses suggest that the inconsistent

resolution of rhizarian higher-order affinities has not only been due to a lack of rhizarian sequence data but also to the tendency of rhizarian sequences to form xLBs in molecular trees (Table 2). These xLBs were found in many genes (Table 1) as well as nearly all examined rhizarian taxa but also different subsets of taxa for different genes (data not shown), which make this problem particularly challenging to deal with. These results further underscore the importance of monitoring long branches particularly for analyses focused on resolving deep branches (Philippe et al. 2011; He et al. 2014).

Acknowledgments

We thank F. Burki for kindly providing the 119-protein alignment. For the use of computational resources we thank Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX, www.uppmax.uu.se), Cyberinfrastructure for Phylogenetic Research (CIPRES, www.phylo.org) and Vital-IT Center for high-performance computing of the SIB Swiss Institute of Bioinformatics (www.vital-it.ch). This work was supported by grants from the Swedish Research Council (Vetenskaprådet) and the Swiss National Science Foundation grant No. 31003A_140766 (to R.S. and J.P.). The *Euglypha rotunda* transcriptome is available at the Sequence Read Archive (SRA) under the SRS927575 identifier.

References

Adl SM, Simpson AGB, Lane CE, et al. 2012. The revised classification of eukaryotes. J. Eukaryot. Microbiol. 59:429–514.

Boudagher-Fadel MK. 2012. Biostratigraphic and Geological Significance of Planktonic Foraminifera. Elsevier

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative Multicellularity Evolved Independently in the Eukaryotic Supergroup Rhizaria. Curr. Biol. 22:1123–1127.

Burki F, Corradi N, Sierra R, Pawlowski J, Meyer GR, Abbott CL, Keeling PJ. 2013. Phylogenomics of the intracellular parasite *Mikrocytos mackini* reveals evidence for a mitosome in rhizaria. Curr. Biol. 23:1541–1547.
Burki F, Kaplan M, Tikhonenkov D V, Zlatogursky V, Minh B Q, Radaykina L V, Smirnov A, Mylnikov A P, Keeling P J. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc. R. Soc. B. Sci. 283:1823.
Burki F, Keeling PJ. 2014. Rhizaria. Curr. Biol. 24:R103–R107.
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 Biol. Sci. 279:2246–2254.
Burki F, Pawlowski J. 2006. Monophyly of Rhizaria and multigene phylogeny of unicellular bikonts. Mol Biol Evol 23:1922–1930.
Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One 2:e790.
Burki F. 2014. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb. Perspect. Biol. 6:a016147–a016147.
Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.
Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–224.
He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SL. 2014. An Alternative Root for the Eukaryote Tree of Life. Curr. Biol. 24:465–470.
Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G. B., & Roger, A. J. (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc. Natl. Acad. Sci. U. S. A., 106 (10), 3859-3864.
Keeling PJ. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu. Rev. Plant Biol. 64:583–607.
De la Torre-Bárcena JE, Kolokotronis S-O, Lee EK, Stevenson DW, Brenner ED, Katari MS, Coruzzi GM, DeSalle R. 2009. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. PLoS One 4:e5764.
Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7 Suppl 1:S4.
Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109.
Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst. Biol. 62:611–615.
Lecompte O. 2002. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res. 30:5382–5390.
Miller MA, Pfeiffer W, Schwartz T. 2010. 2010 Gateway Computing Environments Workshop (GCE). In: 2010 Gateway Computing Environments Workshop (GCE). IEEE. p. 1–8.
Nikolaev SI, Berney C, Fahrni JF, Bolivar I, Polet S, Mylnikov AP, Aleshin V V, Petrov NB, Pawlowski J. 2004. The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. Proc. Natl. Acad. Sci. U. S. A. 101:8066–8071.
Noller HF. 2005. RNA structure: reading the ribosome. Science 309:1508–1514.

Okamoto N, Chantangsi C, Horák A, Leander BS, Keeling PJ. 2009. Molecular phylogeny and description of the novel katablepharid *Roombia truncata* gen. et sp. nov., and establishment of the Hacrobia taxon nov. PLoS One 4:e7080.
Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, Patterson DJ, Katz LA. 2010. Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life. Syst. Biol. 59:518–533.
Philippe H, Brinkmann H, Lavrov D V, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Schneider A, Cannarozzi GM. 2009. Support patterns from different outgroups provide a strong phylogenetic signal. Mol. Biol. Evol. 26:1259–1272.
Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.
Sierra R, Cañas-Duarte S J, Burki F, Schwelm A, Fogelqvist J, Dixelius C, González-García L N, Gile G H, Slamovits C H, Klopp C, Restrepo S, Arzul I, Pawlowski J. 2015. Evolutionary Origins of Rhizarian Parasites. Mol. Biol. Evol. doi: 10.1093/molbev/msv340
Sierra R, Matz M V, Aglyamova G, Pillet L, Decelle J, Not F, de Vargas C, Pawlowski J. 2013. Deep relationships of Rhizaria revealed by phylogenomics: A farewell to Haeckel's Radiolaria. Mol. Phylogenet. Evol. 67:53–59.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Torruella G, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I. 2012. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol. Biol. Evol. 29:531–544.
Vaidya G, Lohman DJ, Meier R. 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27:171–180.
Warner JR, McIntosh KB. 2009. How common are extraribosomal functions of ribosomal proteins? Mol. Cell 34:3–11.
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 29:1557–1568.

FIGURE LEGENDS

Fig. 1. Phylogenomic analyses reveal strong support for Rhizaria + Alveolata. The tree shown was derived by Bayesian analysis of the M27 (xLB-depleted) data set using PhyloBayes and supports hypothesis H1 (Rhizaria + Alveolata). Identical tree topologies were obtained with the M147 (full) and M34 (mono-rhizaria) data sets using RAxML and

Bayesian inference, except where indicated by a large red circle or small empty circles. Filled black circles indicate branches receiving 100% mlBP and 1.00 biPP from all data sets; otherwise only the highest support values with mlBP and biPP are shown above and below the branches, respectively. Exact support values of two alternative SAR topologies are shown in the top left panel, with red background indicating support for H1 and green background support for H2. A schematic of H2 is shown in the green-dashed box below the panel. Branches are drawn to scale as indicated by the scale bar at the bottom.

TABLES

Table 1. Extreme long branches by protein functional category. The total number of proteins (# partitions) in the M147 data set for various function categories are given for: initiation factors (eif's), euBac proteins (He et al. 2014), psm subunits b-d (psmb-d), large (rpLs) and small (rpSs) ribosomal subunit proteins, tcp subunits (tcp's), and beta-tubulin (tubb). For each protein category, the total number of extreme long branches is given for all taxa (xLB_{ttl}) and for Rhizaria only (xLB_R) along with the average number of xLBs per protein (average xLB/part) for all taxa and the percentage of these xLBs that are attributable to Rhizaria (% R/ttl = xLB_R/xLB_{ttl})

	eif's	euBacs	psmb-d	rpL's	rpS's	tcp's	tubb	other	total
# partitions	3	35	11	24	19	5	1	62	150
xLB_{ttl}	13	77	54	107	91	18	9	199	568
xLB_R	5	26	16	38	42	4	1	60	192
average xLB/part	4.3	2.2	4.9	4.5	4.8	3.6	9	3.2	3.8
% R/ttl	38.5%	33.8%	29.6%	35.5%	46.2%	22.2%	11.1%	30.2%	33.8%

Table 2. Number of extreme long branches (xLBs) by taxon (full M147 data set). The total number of extremely long terminal branches (total xLBs) that were 2-3 times (2-3x), 3-4 times (3-4x), or greater than 4 times (>4x) the average terminal branch length in their respective single protein control trees are shown for all taxa (total xLBs), which includes the four outgroup taxa (Archaeplastida, Hacrobia, Amorphea and Excavata), and then separately for Rhizaria, Stramenopila and Alveolata.

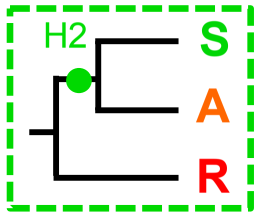
		Extreme long branches (xLBs) (times average branch length)			
		2-3x	3-4x	>4x	all xLBs (≥2x)
all taxa	total xLBs	479	139	58	676
	affected proteins	141	92	43	
Rhizaria	number of xLBs	135	36	23	194
	fraction of total	0.28	0.26	0.40	0.29
Stramenopila	number of xLBs	36	15	8	59
	fraction of total	0.075	0.11	0.14	0.09
Alveolata	number of xLBs	117	33	12	162
	fraction of total	0.24	0.24	0.21	0.24

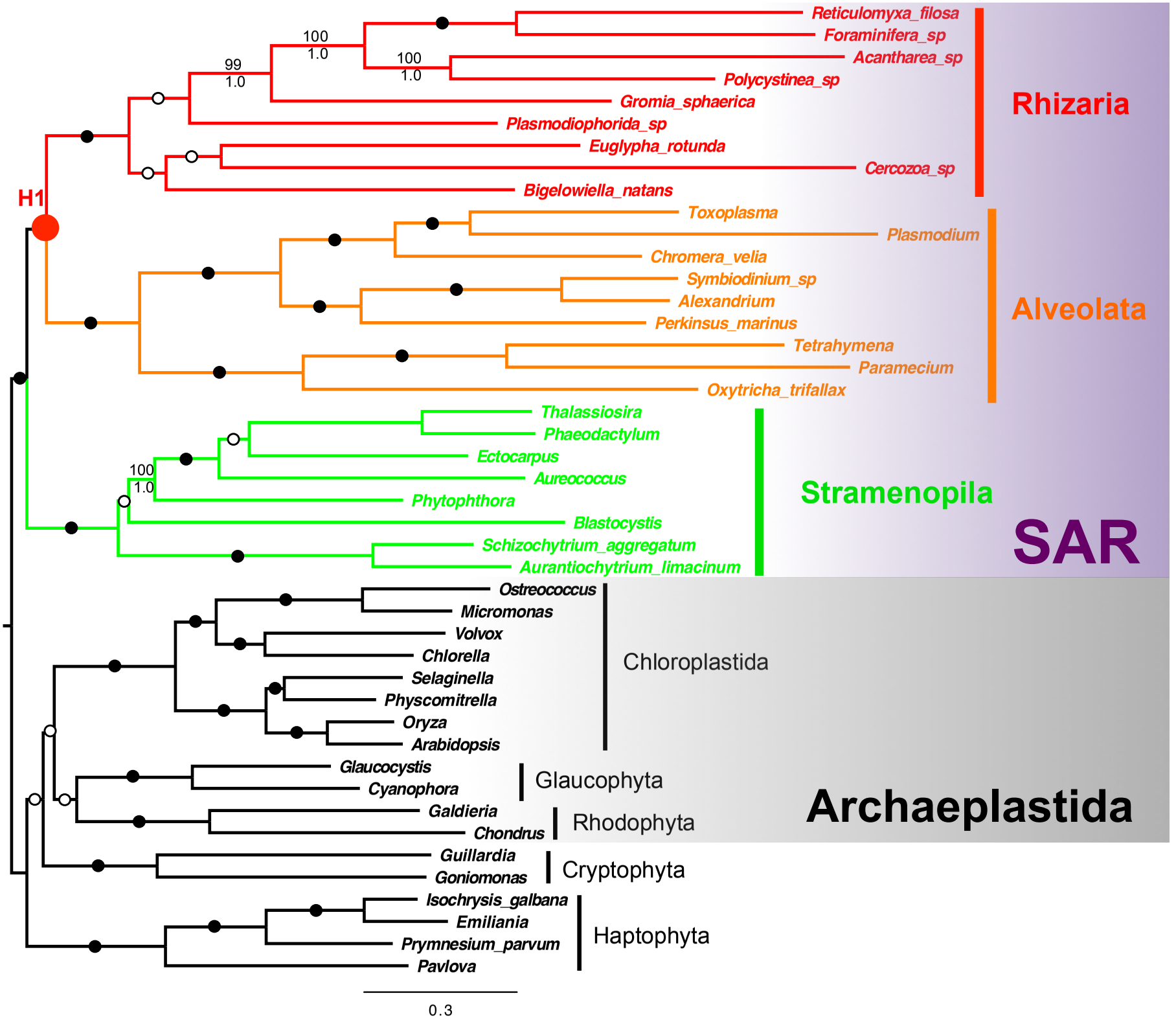
Table 3. Support for three alternative hypotheses of SAR topology by different data sets and with different outgroups. Numbers in parenthesis show support from mlBP and biPP for three alternative topologies for supergroup SAR: H1 (Rhizaria + Alveolata), H2 (Alveolata + Stramenopila) and H3 (Rhizaria + Stramenopila). Analyses were conducted with three different outgroups: PLH (Plants + Hacrobia), AMR (Amorphia) and DSC (Discoba). A single star (*) indicates no convergence; a double star (**) indicates CAT-Poisson model used because no convergence with CAT-GTR (biPP<0.6). P-values of the AU test results are shown in the order H1/H2/H3, with the favored hypothesis shown in parenthesis.

	M147			M34			M27		
	mlBP	biPP	pAU	mlBP	biPP	pAU	mlBP	biPP	pAU
All taxa	H2 (79)	H2 (1.00)*	(H2) 0.269/ 0.711/ 0.019	H1 (98)	H1 (0.98)**	(H1) 0.920/ 0.068/ 0.013	H1 (99)	H1 (1.00)	(H1) 0.996/ 0.006/ 0.007
PLH	H1 (85)	H2 (1.00)	(H1) 0.934/ 0.118/ 0.050	H1 (94)	H1 (1.00)	(H1) 0.976/ 0.039/ 0.020	H1 (99)	H1 (1.00)	(H1) 0.998/ 0.002/ 0.004
AMR	H2 (100)	H2 (1.00)	(H2) 0.032/ 0.971/ 0.003	H1 (82)	H2 (1.00)	(H1) 0.822/ 0.193/ 0.001	H1 (96)	H1 (1.00)	(H1) 0.979/ 0.013/ 0.037
DSC	H3 (81)	H2 (1.00)	(H3) 0.012/ 0.223/ 0.816	H3 (44)	H2 (0.89)	(H3) 0.380/ 0.433/ 0.619	H3 (54)	H2 (0.68)	(H1) 0.667/ 0.251/ 0.448

Figure 1

	mlBP	biPP
M27	99	1.0
M34	94	1.0
M147	85	1.0





***Supplementary Material / Phylogenetic tree data**

Click here to download Supplementary Material / Phylogenetic tree data: Heetal_suppl_final1.docx