Single Cell Transcriptomics, Mega-Phylogeny, and the Genetic Basis of Morphological Innovations in Rhizaria

Anders K. Krabberød,[1] Russell J.S. Orr,[1] Jon Bråte,[1] Tom Kristensen,[1] Kjell R. Bjørklund,[2] and Kamran Shalchian-Tabrizi*[,1]

[1]Department of Biosciences, Centre for Integrative Microbial Evolution (CIME) and Centre for Epigenetics Development and Evolution (CEDE), University of Oslo, Oslo, Norway

[2]Department of Research and Collections, Natural History Museum, University of Oslo, Oslo, Norway

*Corresponding author: E-mail: kamran@ibv.uio.no.

Associate editor: Nicole Perna

Abstract

The innovation of the eukaryote cytoskeleton enabled phagocytosis, intracellular transport, and cytokinesis, and is largely responsible for the diversity of morphologies among eukaryotes. Still, the relationship between phenotypic innovations in the cytoskeleton and their underlying genotype is poorly understood. To explore the genetic mechanism of morphological evolution of the eukaryotic cytoskeleton, we provide the first single cell transcriptomes from uncultured, free-living unicellular eukaryotes: the polycystine radiolarian *Lithomelissa setosa* (Nassellaria) and *Sticholonche zanclea* (Taxopodida). A phylogenomic approach using 255 genes finds Radiolaria and Foraminifera as separate monophyletic groups (together as Retaria), while Cercozoa is shown to be paraphyletic where Endomyxa is sister to Retaria. Analysis of the genetic components of the cytoskeleton and mapping of the evolution of these on the revised phylogeny of Rhizaria reveal lineage-specific gene duplications and neofunctionalization of α and β tubulin in Retaria, actin in Retaria and Endomyxa, and Arp2/3 complex genes in Chlorarachniophyta. We show how genetic innovations have shaped cytoskeletal structures in Rhizaria, and how single cell transcriptomics can be applied for resolving deep phylogenies and studying gene evolution in uncultured protist species.

Key words: **cytoskeleton, phylogeny, protists, Radiolaria, Rhizaria, SAR, single-cell transcriptomics.**

Introduction

One of the major eukaryotic innovations is the cytoskeleton, consisting of microtubules, filaments, and motor proteins. Together these structures regulate the internal milieu of the cell, and aid in movement, cytokinesis, phagocytosis, and predation (Grain 1986; Vale 2003; Wickstead and Gull 2011). Of essential importance, and the main focus of this work, the cytoskeleton of unicellular eukaryotes determines the morphological design of the cell.

The evolution of the eukaryote cytoskeleton is an intriguing story of gene evolution. Homologs to actin and tubulin genes can be found in Eubacteria and Archaea, whereas the origin of motor proteins is unclear, as they lack distinct homologs in prokaryotes (Vale 2003; Wickstead and Gull 2011). Early in the evolution of eukaryotes the cytoskeletal filaments of prokaryotes evolved new functions and new motor proteins were invented, in addition to a large repertoire of molecules that modify and interact with both the cytoskeleton and these motor proteins (Wickstead and Gull 2011; Cavalier-Smith et al. 2014).

Most of what we know about the eukaryote cytoskeleton comes from studies of humans, plants, and fungi (Jékely 2007; Wickstead and Gull 2011), whereas less is known about the genetic machinery and the molecular architecture of the cytoskeleton in nonmodel single-celled eukaryotes (protists). Our current knowledge about the evolution of cytoskeletal genes in protists stems from human pathogens, for example, *Plasmodium*, *Toxoplasma* and *Cryptosporidium* (Wickstead and Gull 2011; Burki and Keeling 2014), but virtually nothing is known about how the evolution of these genes has shaped cytoskeletal morphology in other protists.

In this paper, we trace the evolution of key cytoskeletal genes in a major group of eukaryotes, Rhizaria, consisting predominantly of understudied single-celled protists (Burki and Keeling 2014). Although no clearly defined phenotypic synapomorphies for Rhizaria have been described (Pawlowski 2008), there is a common theme to many rhizarians: well-developed pseudopodia which are often reticulose or filose. The different groups of rhizarians use their pseudopodia in different ways: Some form complicated reticulose networks, such as chlorarachniophytes, whereas others use pseudopodia stiffened by microtubules to capture prey, for example, Radiolaria, to move molecules and organelles, as in Foraminifera, or even as oars in Taxopodida (Cachon et al. 1977; Anderson 1978; Sugiyama et al. 2008; Bass et al. 2009). How these widely different applications of pseudopodia have evolved and how the morphological evolution is reflected in changes to cytoskeletal genes are unknown. In the formation of pseudopods in eukaryotes, actin and myosin interact in order to make a protrusion in the plasma membrane

Article

creating the leading edge of the pseudopod. Nucleators anchor actin to the cell membrane, and actin-related proteins (i.e., the Arp2/3-complex) recruit additional actin filaments to form the branching network that supports the pseudopod (Giannone et al. 2007; Mogilner and Keren 2009; Ura et al. 2012). Rigid pseudopods stiffened with additional bundles of microtubules can be found in Radiolaria and Foraminifera (Anderson 1983; Lee and Anderson 1991).

To understand the evolution of the cytoskeleton and pseudopodia in Rhizaria a fully resolved phylogenetic tree is vital, but getting a stable phylogeny for the entire group has proven problematic. Some lineages have apparently evolved extremely fast (such as Foraminifera), making them unstable in molecular phylogenies. The biggest problem in reconstructing the rhizarian phylogeny is, however, the lack of molecular data from key groups (Burki and Keeling 2014). The main reason for this is that we currently are not able to culture more than a handful of species. To overcome this problem, we have used transcriptomes from single cells of two uncultured Rhizaria species (*Sticholonche zanclea* and *Lithomelissa setosa*) to build multi-gene phylogenies and investigated the genetic basis of cytoskeletal differences in Rhizaria.

Results

Single Cell Transcriptomics of Two Uncultured Protists

We generated cDNA libraries from two rhizarian specimens: *Lithomelissa setosa* and *Sticholonche zanclea* (supplementary fig. S1, Supplementary Material online). The cDNA was sequenced on the Illumina MiSeq platform, 300 bp paired end. This resulted in 19,894,654 reads for *S. zanclea* and 11,590,658 for *L. setosa*, which were *de novo* assembled using the Trinity platform (Haas et al. 2013). Assembly resulted in two single cell transcriptomes (SCT) with 4,749 predicted genes for *S. zanclea and* 2,122 predicted genes for *L. setosa* (table 1). Subsampling and re-assembly of reads showed that the sequencing threshold for both libraries was close to maximum (supplementary fig. S2, Supplementary Material online). We assessed the suitability of the data for phylogenomic reconstruction using the BIR pipeline (Kumar et al. 2015). Using 255 seed alignments covering the eukaryote Tree of Life (Burki et al. 2012) we identified 54 and 16 corresponding orthologous gene sequences from *S. zanclea* and *L. setosa*, respectively. In addition BIR extracted 3,534 gene sequences from Marine Microbial Eukaryote Transcriptome Sequencing Project, MMETSP (Keeling et al. 2014) and 793 proteins from GenBank with TaxID 543769 (Rhizaria) and added these to their corresponding gene alignments (supplementary table S1, Supplementary Material online). After concatenation of all gene alignments we had a super-matrix consisting of 91 taxa and 54,898 amino acids (255 genes).

Table 1. Single Cell Transcriptome Statistics.

Species Name	Raw Reads	Contigs[a]	GC-content (%)	Predicted Genes[b]
Sticholonche zanclea	19,894,654	19,509	53.5	4,749
Lithomelissa setosa	11,590,658	12,212	48.8	2,122

[a]Number of contigs assembled by Trinity (Haas et al. 2013).
[b]The number of genes predicted by TransDecoder in the Trinity platform.

Bayesian CATGTR Trees Show Congruent Phylogeny for SAR and Subgroups

In the Bayesian analysis of the full dataset (255 genes 54,898 AA, 91 taxa, fig. 1) using the CATGTR model in PhyloBayes (Lartillot et al. 2013), Stramenopiles and Alveolates formed a clade, with Rhizaria as sister, with maximum statistical support [1.00 posterior probability (pp)]. The relationship and support values did not change for SAR (Stramenopiles, Alveolata, and Rhizaria) after four categories of fast evolving sites had been removed (supplementary table S2, Supplementary Material online). Within Rhizaria, each of the three groups Foraminifera, Radiolaria, and Taxopodida were all monophyletic, and together formed a cluster (i.e., Retaria) with maximum support even when fast evolving sites were removed (i.e., always 1.00 pp). Radiolaria and Foraminifera were placed together as a monophyletic group (0.71 pp), with *S. zanclea* branching off as sister to both. This topology remained constant after removal of fast evolving sites (supplementary table S2, Supplementary Material online). The posterior probability for the monophyly of Radiolaria together with Foraminifera, i.e., excluding *S. zanclea*, increased to 0.97 when fast evolving sites were removed (supplementary table S2, Supplementary Material online). Endomyxa was monophyletic (1.00 pp) and always sister to Retaria with full support (1.00 pp), rendering Cercozoa paraphyletic. Filosa was monophyletic in all analyses (1.00 pp).

ML Trees Converged towards Bayesian Topology after Removal of Fast Evolving Sites

In contrast, the maximum likelihood (ML) analysis of the full dataset using the LG model (255 genes, 54,898 AA, 91 taxa; supplementary fig. S3, Supplementary Material online), grouped Alveolata with Rhizaria instead of the Stramenopiles [96% bootstrap support (bs)]. Retaria was recovered with high support (88% bs) as in the Bayesian tree, but *S. zanclea* was no longer placed ancestrally to Radiolaria and Foraminifera. Instead *S. zanclea* was sister to Radiolaria (88% bs). Importantly, however, *S. zanclea* changed to a basal position in Retaria after four categories of fast evolving sites were removed, consistent with all the CATGTR Bayesian trees (supplementary table S2, Supplementary Material online).

Removal of four categories of fast evolving sites did not change the monophyly of Foraminifera and Radiolaria (excluding *S. zanclea*) or the sister relation between alveolates and Rhizaria, but the support values were reduced in both instances to 50% bs for Radiolaria together with Foraminifera and 67% bs for Alveolata together with Rhizaria

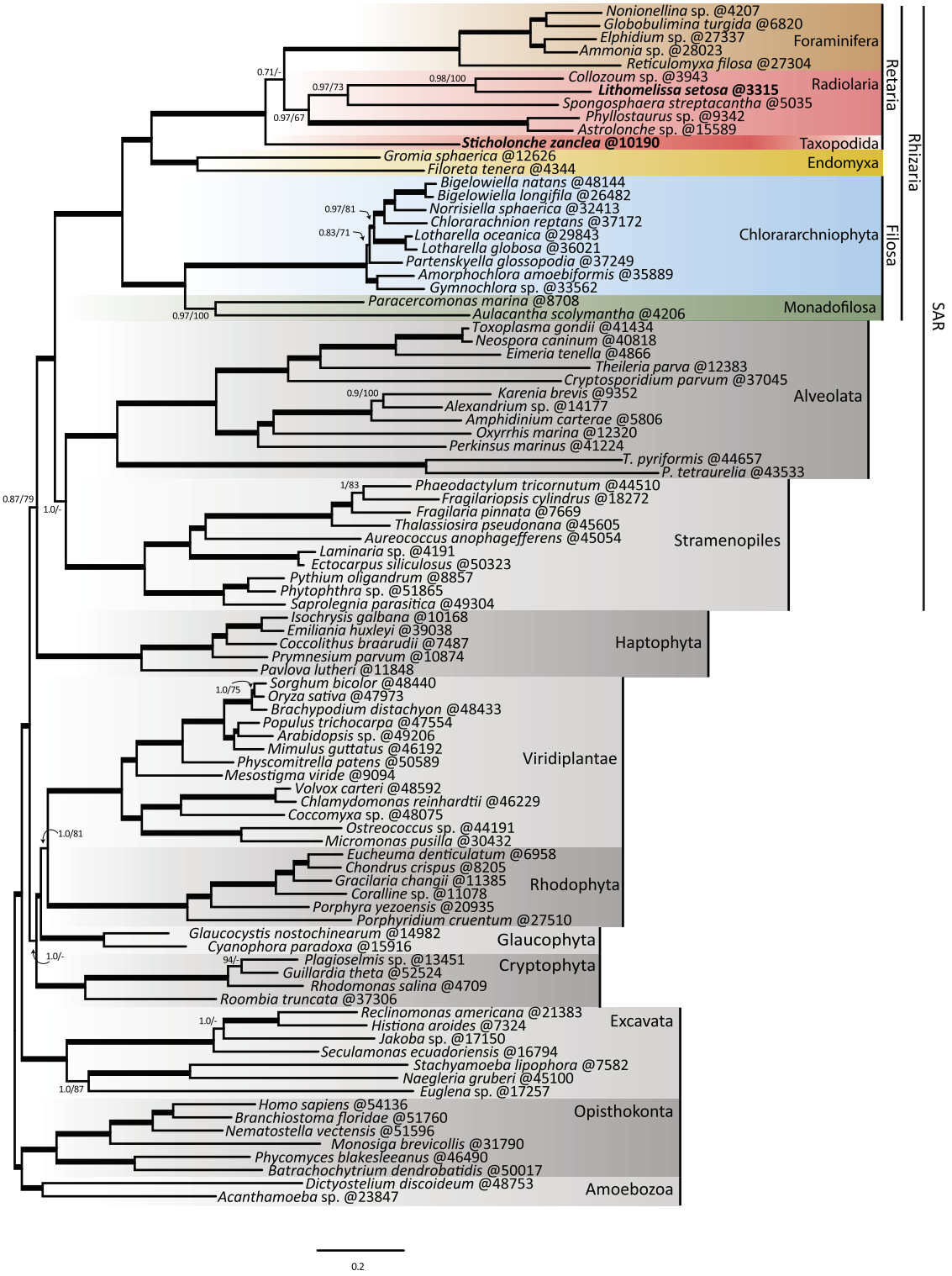

Fig. 1. Bayesian phylogeny of eukaryotes (CATGTR model, 255 genes, 54,898 AA, and 91 taxa, maxdiff 0.2666) with bootstrap values from maximum likelihood analysis added. Thick branches represent maximal support (posterior probability = 1, and bootstrap support = 100%). Radiolarian species sequenced for this paper are shown in bold. Number after "@" is the concatenated sequence length for each taxon. Important clades in Rhizaria are colored for easier identification: brown = Foraminifera, dark red = Taxopodida, red = Radiolaria, yellow = Endomyxa, blue= Chlorarachniophyta (Filosa), and green = Mondaofilosa (Filosa). The scale bar represents 0.1 substitutions per site.

(supplementary fig. S3 and table S2, Supplementary Material online). Endomyxa and Retaria group together with full support, as in the Bayesian analysis (100% bs), making Cercozoa paraphyletic. As in the Bayesian phylogeny haptophytes appeared as sister to SAR (77% bs) and changed the position basal to the plants, glaucophytes, and cryptomonads (73% bs) after removal of four categories of fast evolving sites. Species with more than 10% missing data in the final concatenated

data matrix were placed on the ML phylogeny using the Evolutionary Placement Algorithm (supplementary table S3, Supplementary Material online; Berger and Stamatakis 2011). Five species were placed in Endomyxa, five in Filosa, two in Radiolaria, and finally ten species in Foraminifera.

Influence of Fast Evolving Sites and the Choice of Model on the Phylogeny

The discrepant topologies of the Bayesian (CATGTR) and ML (LG) trees could be due to the different models implemented in these two approaches. We assessed the influence of these two models by running Bayesian inferences using the LG model (the opposite: running ML with a CATGTR model is currently not possible). This was done on a smaller alignment to reduce the computational burden (146 genes, 33,081 AA, 91 taxa, see Materials and Methods for further explanation). The resulting Bayesian LG tree grouped *S. zanclea* with Radiolaria (0.67 pp, Supplementary table S2, Supplementary Material online) as in the ML (LG) tree, and not as sister to Foraminifera and Radiolaria, as in all Bayesian trees with the CATGTR model. Other branching patterns in the Rhizaria phylogeny were unaffected.

We repeated the ML (LG) analyses after removing four categories of fast evolving sites on the full dataset as well as the reduced dataset. While Alveolata and Rhizaria formed a clade in the full and small datasets (85% bs, supplementary table S2, Supplementary Material online), removal of four categories of fast evolving sites moved alveolates to the Stramenopiles in the dataset with 146 genes (74% bs, supplementary table S2, Supplementary Material online), a result congruent with the Bayesian topology. The support for Alveolata together with Rhizaria was also weakened in the dataset with 255 genes when four categories of fast evolving sites were removed, from 96% bs to 67% bs. When Foraminifera was excluded from the 255 genes dataset with four categories of fast evolving sites removed, Stramenopiles and Alveolata formed a group with Rhizaria as sister (57% bs; supplementary table S2, Supplementary Material online).

Actin Radiation in Rhizaria and Unique Duplications in Retaria and Endomyxa

We identified 6 actin sequences in our SCTs. From MMETSP, we identified 18 foraminiferan actin and 18 chlorarachniophyte actin sequences. Phylogenetic analysis of these and other available actin sequences retrieved from GenBank and Pfam revealed that Retaria (including *S. zanclea*) have two distinct paralogs of actin—actin1 and 2—where actin2 is fully supported (fig. 2). Actin1 is supported in the Bayesian analysis (0.87 pp) but not in by ML analysis. Endomyxaean actins form a weakly supported monophyletic group together with retarian actin1 (14% bs/0.68 pp). This clade, in turn, groups with retarian actin2 (59% bs/0.97 pp), and is a synapomorphy for Retaria and Endomyxa.

Arp2/3 Complex Gene Duplication in Chlorarachniophyta

Of the seven genes encoding components of the Arp2/3 complex, which is responsible for branching of actin filaments and recruitment of new actin, we identified *Arp2*, *Arp3*, *ARPC2*, and *ARPC5* from *S. zanclea*, but only *Arp2* from *L. setosa*. From MMETSP, we identified sequences of all seven genes from both Chlorarachniophyta and Foraminifera. Phylogenetic analysis of these genes revealed that all chlorarachniophytes have two distinct paralogs of both *Arp2* (fig. 3A) and *APRC1* (fig. 3B), recovered with maximum support (100% bs/1.0 pp).

Neofunctionalization of Arp2/3 in Chlorarachniophytes

Comparative evolutionary analyses of the duplicated Arp2/3 complex genes (*Arp2* and *ARPC1*) were performed by examining the evolutionary rates for each paralog, and then mapping the genes to structural models using Consurf (Ashkenazy et al. 2010; Celniker et al. 2013). The analysis showed that the two different forms of Arp2 (Arp2a and Arp2b, supplementary fig. S4, Supplementary Material online) and the two different forms of ARPC1 (ARPC1a and ARPC1b, supplementary fig. S5, Supplementary Material online) follow a pattern where the most conserved sites are localized inside the protein structure. Comparison between the Arp2 a and b proteins shows shared conserved residues in contact surfaces against other proteins in the Arp2/3 complex (fig. 3C). Similarly, the two different paralogs of ARPC1 show shared conserved sites localized inside the complex (fig. 3D). In contrast, the surfaces of the two Arp2 and ARPC1 copies show more variable substitution rates, and all paralogs have patches with mutually exclusive conserved residues.

Myosin Evolution in Rhizaria

We extracted 133 myosin transcripts with rhizarian origin from MMETSP. A phylogenetic reconstruction of the newly identified rhizarian myosins together with already published myosin classes spanning a broad taxonomical range of eukaryotes (Richards and Cavalier-Smith 2005; Sebé-Pedrós et al. 2014) revealed two known classes (If and IV) and three previously unknown classes of myosin in Rhizaria (XXXV, XXXVI, and XXXVII, following the naming scheme of Sebé-Pedrós et al. (2014); fig. 4; supplementary fig. S8, Supplementary Material online). Myosin XXXVII is currently the only known synapomorphy for Rhizaria. It is highly supported (100% bs and 1.0 pp) and has a molecular signal distinct from other described myosins (Richards and Cavalier-Smith 2005; Sebé-Pedrós et al. 2014). The insertion in ribosomal protein 10a reported in Burki et al. (2010) is not present in all Rhizaria species and should not be regarded as synapomorphic for the group. For instance, *Minchinia chitonis* in the MMETSP lacks this insertion. In the same manner, the insertion in polyubiquitin is lacking in several radiolarian species (i.e., *Collozoum inerme*, *Larcopyle butschlii*, and *Sphaerozoum punctatum*, GenBank accessions CAI77900, BAK61738, and BAK61751) and is therefore not a synapomorphy for Rhizaria. In the rhizarian-specific myosin class XXXVII, there has been an additional radiation within the chlorarachniophytes into three separate paralogs, all fully supported (100% bs and 1.0 pp, fig. 4). Rhizarians have also gained a large repertoire of myosin IV variants, with six paralogs in Chlorarachniophyta and two

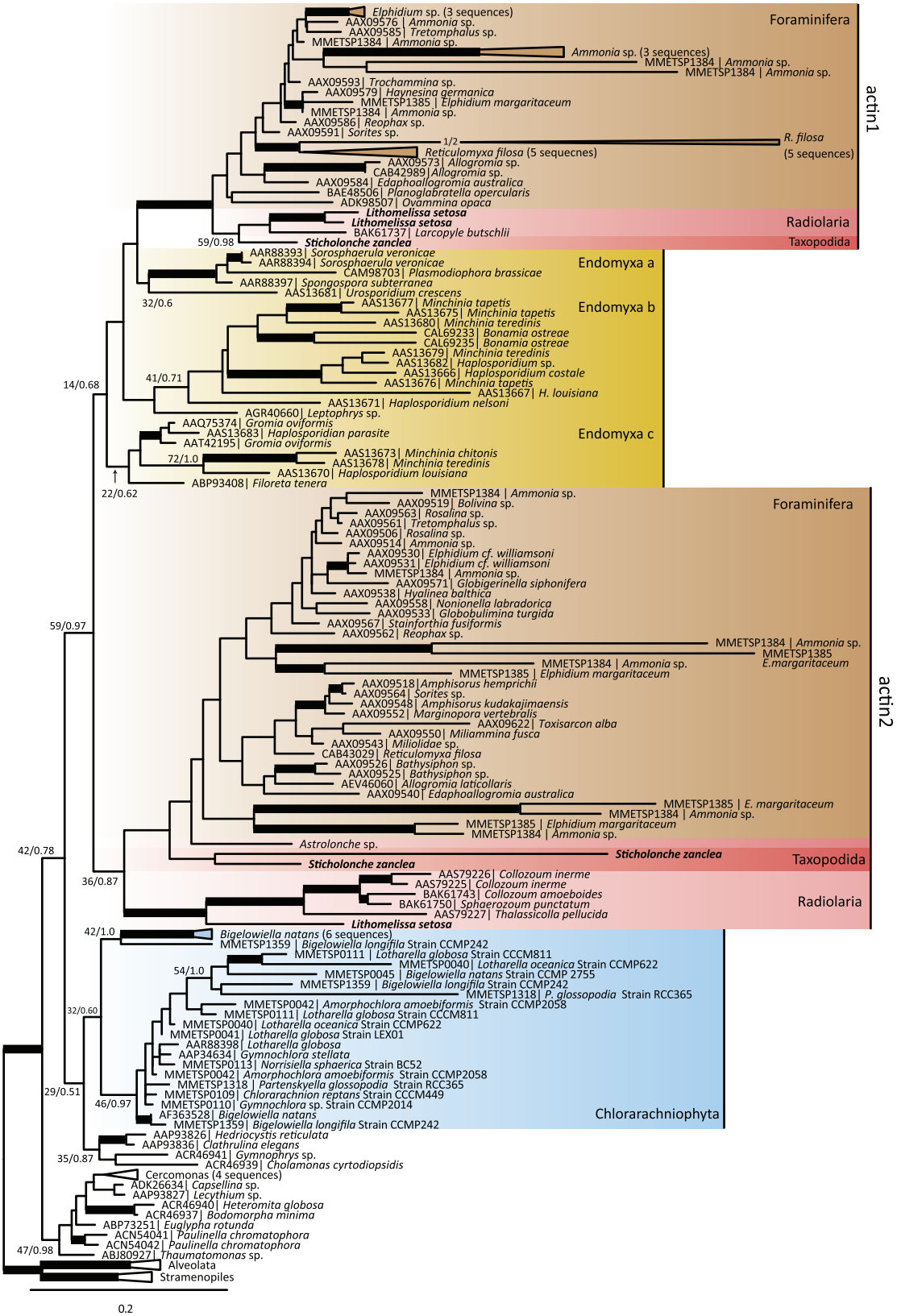

Fig. 2. Actin phylogeny (229 taxa, 374 AA). Thick branches represent bootstrap support >75% and posterior probability >0.9. Some branches are collapsed to save space. Support values for selected nodes discussed in the text are added for clarity. The scale bar represents 0.2 substitutions per site. The coloring scheme is the same as in figure 1.

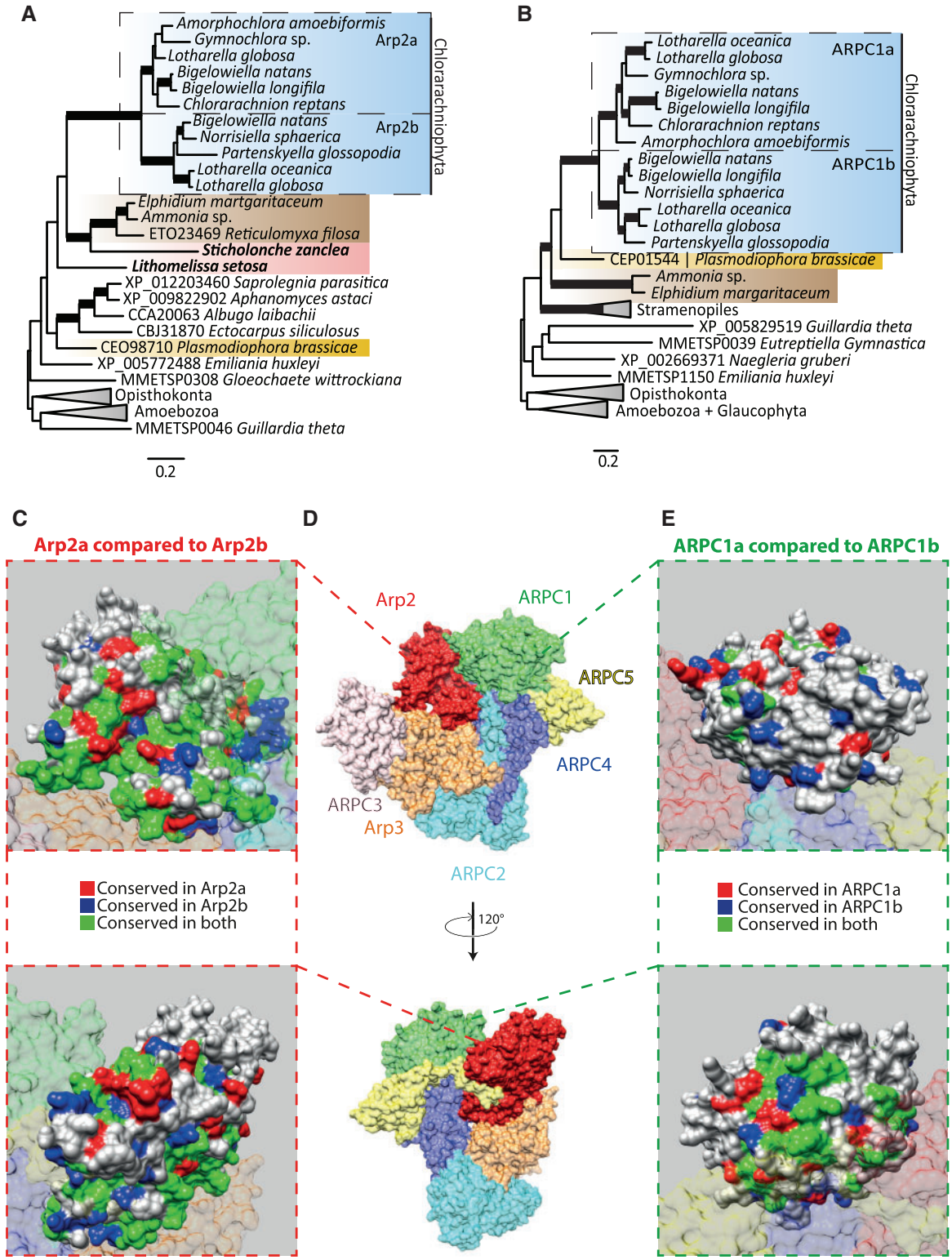

Fig. 3. Two genes from the Arp2/3 complex with a recent gene duplication in Chlorarachniophyta. (*A*) Phylogeny of Arp2 (39 taxa, 373 AA), and (*B*) phylogeny of APRC1 (34 taxa, 328 AA). Coloring of groups in both phylogenies as in figure 1 (Brown = Foraminifera, red= Radiolaria, yellow = Endomyxa, and blue= Filosa). Thick branches represent bootstrap support > 75% and posterior probability > 0.9 and the scale bar equals 0.2 substitutions per site. (*C*) Molecular model and **c**omparison of conserved residues between the two paralogs of Arp2 a and b superimposed on PDB accession 4JD2. (*D*) The Arp2/3 complex (PDB accession 4JD2) showing the position of arp2 (red) and ARPC1 (green) relative to the other molecules in the complex. (*E*) Molecular model and comparison of conserved residues between the two paralogs of ARPC1 a and b superimposed on PDB accession 4JD2. For figures C and E coloring display areas uniquely conserved in either paralog a (red) or paralog b (blue) or areas that are shared between the paralogs (green). Both Arp2 and ARPC1 are shown from two different sides (displayed in upper and lower panels).

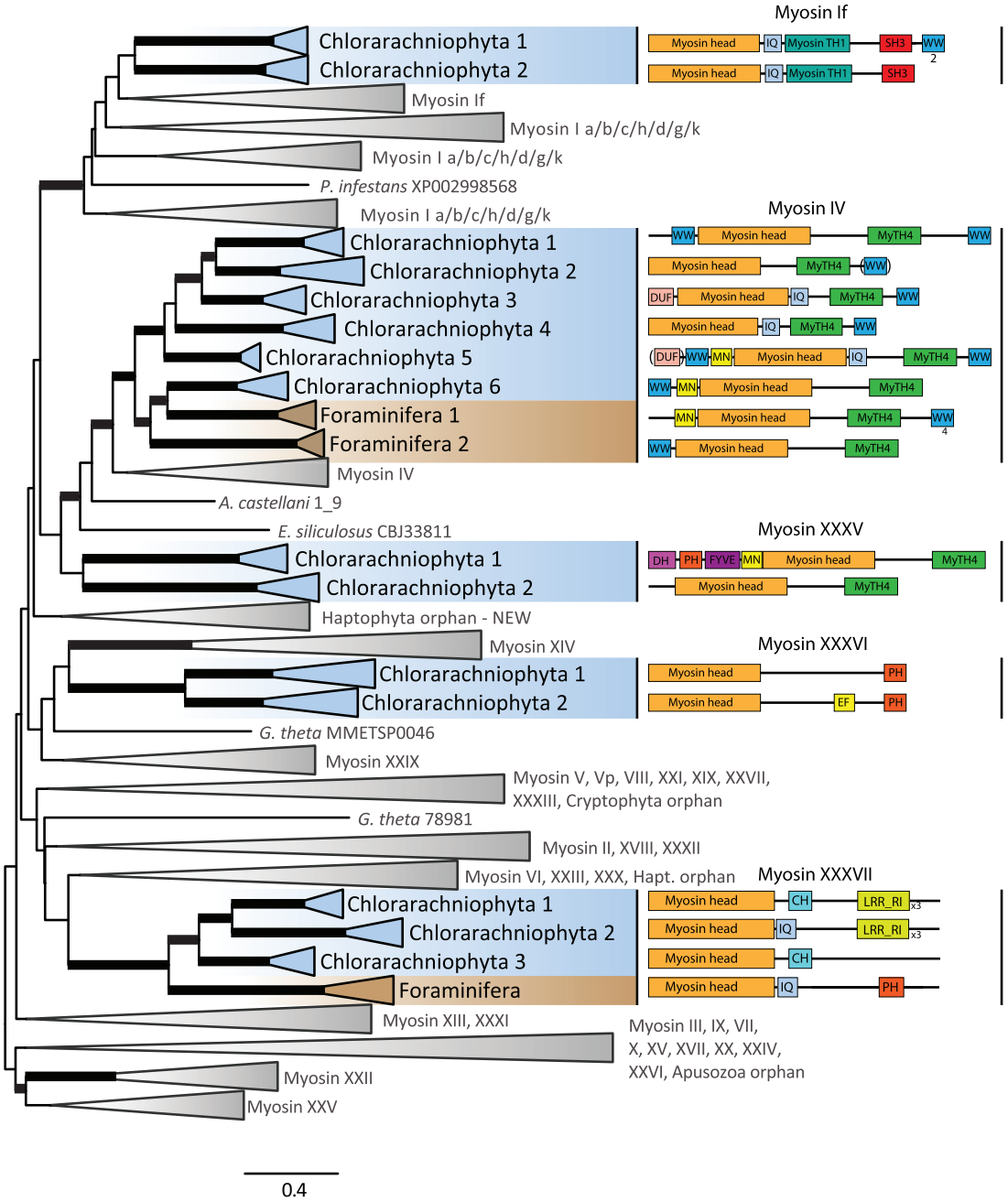

Fig. 4. Myosin maximum likelihood phylogeny (830 taxa, 754 AA). Groups colored according to taxonomy as in figure 1 (blue= Chlorarachniophyta, brown= Foraminifera). Branches are collapsed according to myosin class affiliation and following the nomenclature of Sebé-Pedrós et al. (2014). The tree is midpoint-rooted and thick branches represent bootstrap support > 75%, and Bayesian support > 0.8 pp. The domain architectures for each class with representatives from Rhizaria are shown. A complete ML tree without collapsed branches as well as IPR annotation of functional domains is listed in supplementary figure S8, Supplementary Material online.

in Foraminifera. All paralogs were well supported phylogenetically (bs > 90% and pp >0.9) and differed from each other in functional domains (fig. 4). Two paralogs from Chlorarachniophyta resembled myosin IV by having a MYTH4 domain at the C-terminal, but with additional domains at the N-terminal usually not present in myosin IV (Richards and Cavalier-Smith 2005; Sebé-Pedrós et al. 2014). We have chosen to assign them to a new class (myosin XXXV). Finally, there was a group unique to Chlorarachniophyta with two paralogs, named myosin XXXVI (fig. 4).

α- and β-Tubulin Gene Duplications in Retaria

We report 16 new α-tubulin and 19 new β-tubulin sequences from our two transcriptomes: 4 α-tubulin and 9 β-tubulin sequences from *L. setosa*, 12 α-tubulin and 10 β-tubulin sequences from *S. zanclea*. Additionally, we identified 26 α-tubulin and 42 β-tubulin sequences from other rhizarian species in the MMETSP data (i.e., 12 Chlorarachniophyta and 14 Foraminifera α-tubulins, 10 Chlorarachniophyta and 32 Foraminifera β-tubulins). The phylogenetic tree of rhizarian α-tubulin revealed two different versions of the gene: the canonical version of the α-tubulin gene (α1-tubulin; α1)

and a novel variant (α2-tubulin; α2) found only in Retaria (fig. 5A). The split separating the two versions received maximal support (100% bs/1.0 pp). We identified α2-tubulin in both *L. setosa* and *S. zanclea*. Together with the available data from other Rhizaria, the phylogeny shows that this paralog is unique for Retaria. In Foraminifera there were several paralogs of α2, with most copies in *Reticulomyxa filosa* (25 copies). Foraminifera α2 was divided into two groups, but the bootstrap and posterior probability values for dividing these (70% bs/0.97 pp) were variable (fig. 5A).

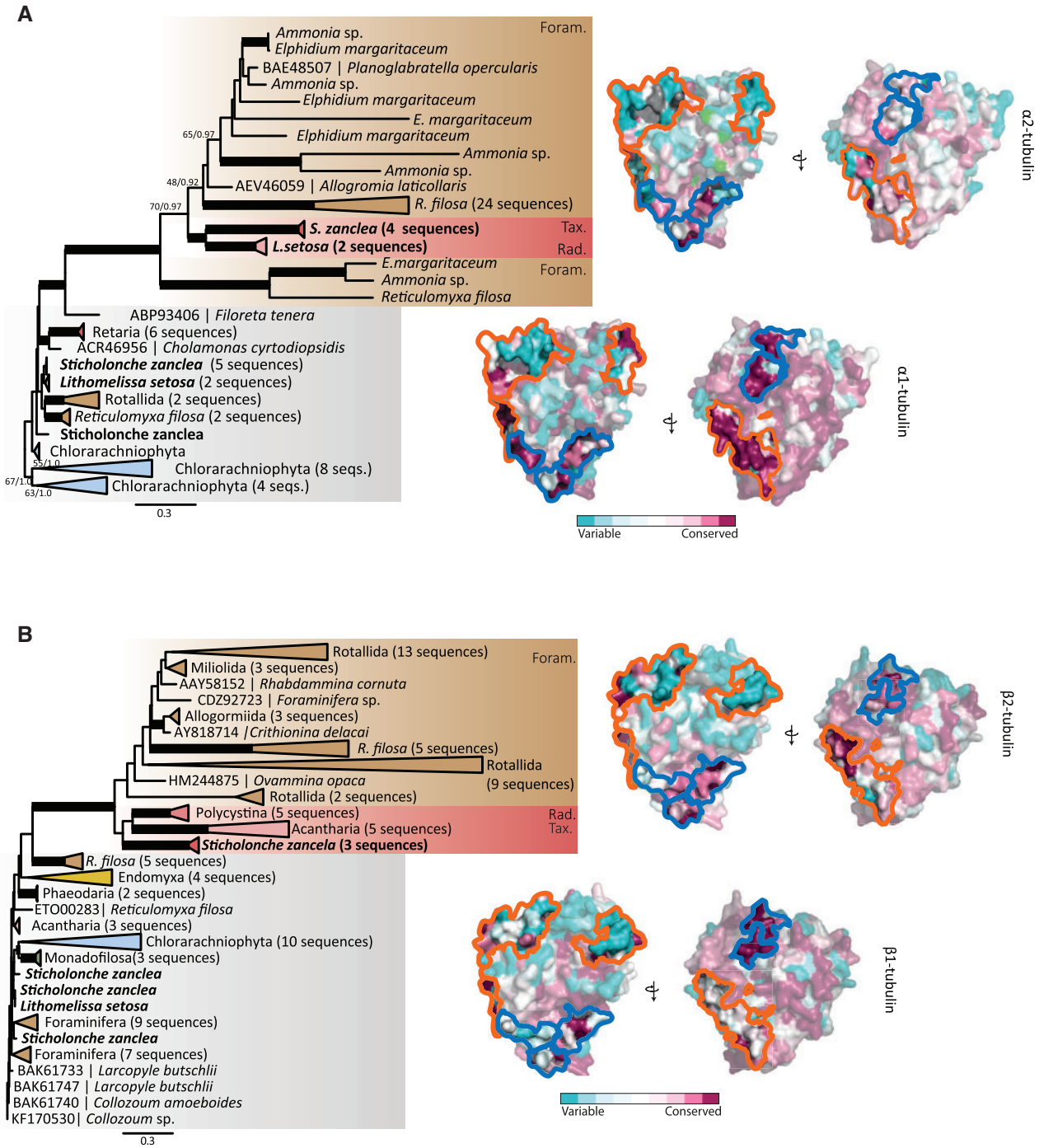

Fig. 5. Tubulin phylogenies and molecular models of two paralogs of α- and β-tubulin. (*A*) α-tubulin phylogeny and molecular model (75 taxa, 453 AA). (*B*) Phylogeny and molecular model β-tubulin (104 taxa, 456 AA). Thick branches in the phylogeny represent bootstrap values >75% and posterior probability >0.9. Some branches are collapsed to save space. Grey areas mark the canonical tubulins (α1- and β1 tubulins) while the colors for branches of the novel variants (α2- and β2 tubulins) are the same as in figure 1 (i.e., brown = Foraminifera, red = Radiolaria, yellow = Endomyxa, and blue = Filosa). Residues on the 3D models are colored according to the evolutionary rates calculated by Consurf and modeled using PDB accession 3du7 as template. Turquoise residues are highly variable and maroon means conserved residues. Functional important areas of the α- and β-tubulin molecules are outlined: blue marks areas important for protofilament assembly and disassembly while orange areas represent lateral interaction sites between protofilaments.

Similarly, the β-tubulin trees contained a clearly divergent clade (i.e., β2-tubulins) with several copies for each retarian group (fig. 5B). All β2-tubulin copies were grouped together with high support (100% bs/1.0 pp) in agreement with earlier studies (Hou et al. 2013). We also found that the β2 copies were present in Taxopodida as well as in Foraminifera and Radiolaria.

Neofunctionalization of Tubulin Genes in Retaria

Comparative evolutionary analyses of the two tubulin paralogs were done to identify patterns of functional change. Evolutionary rates were estimated and site rates mapped to tubulin structural models using Consurf (Ashkenazy et al. 2010; Celniker et al. 2013). Highly conserved amino acid residues were assumed to be functionally important and variable residues of less importance for function. We compared separately α1 with α2, and β1 with β2, identifying sites conserved in one paralog and variable in the other. Such sites were believed to have undergone functional shifts, and were therefore considered important for evolution of microtubules of Retaria. We also examined regions of the α- and β-tubulin structures known to be important for microtubule function and dynamics. Evolutionary changes in these areas are likely to affect the function of the cytoskeleton.

Tracing evolutionary rates on the molecular structures of α- and β-tubulin (fig. 5; supplementary figs. S6 and S7, Supplementary Material online) revealed two patterns of functional change between the conventional and new tubulin genes: first, areas considered as functionally important and conserved in α- and β-tubulin are generally conserved for α1 and β1, though highly variable for α2 and β2. This pattern was observed for both longitudinal interactions important for protofilament assembly and disassembly (e.g., T7-loop and the 8H helix; blue areas in fig. 5; supplementary figs. S6 and S7, Supplementary Material online), as well as lateral interactions between protofilaments (e.g., the M-loop and the H12 helix; orange areas in fig. 5; supplementary figs. S6 and S7, Supplementary Material online). Second, several residues outside of the conventional longitudinal and lateral binding sites are highly conserved in both α2 and β2 while highly variable in the original α1 and β1 genes. Many of these residues are exposed on the surface of the monomers and could represent new sites for other tubulin interactions or surfaces for motor protein attachment and movement.

Discussion

The last common ancestor of Rhizaria was most likely a naked, heterotrophic flagellate that relied extensively on its pseudopodia to explore the environment and to catch prey (Cavalier-Smith 2009). Its pseudopods were supported by actin and at least one group of myosins unique to Rhizaria (fig. 6). Rhizarian cytoskeletons have since undergone evolutionary changes, and their diversification follows a pattern where the major groups have their own favored filament: the chlorarachniophytes have relied on actin to support their reticulose pseudopodia, whereas the axopodia and reticulopodia in Retaria have been stiffened by microtubules composed of tubulin. Although some structural differences between lineages are known, little is established regarding the genetic basis of these phenotypes.

Resolving Rhizarian Relationships

Within Rhizaria, it has been suspected for some time that Foraminifera and Radiolaria are closely related (Cavalier-Smith 2002; Krabberød et al. 2011). Recent phylogenomic analyses place Foraminifera either within Radiolaria, implying Radiolaria to be a paraphyletic group (Burki et al. 2013; Sierra et al. 2013; Sierra et al. 2016) or as sister to Radiolaria (Cavalier-Smith et al. 2015; Burki et al. 2016; He et al. 2016). However, these analyses lack two crucial pieces in the puzzle: representatives from Nassellaria, one of the major polycystine radiolarian orders, and *S. zanclea*, the only species of Taxopodida. We have generated transcriptome data and protein sequences from both the missing Radiolaria groups. In addition, we have reduced the impact of missing data in earlier phylogenomic analyses (Sierra et al. 2013; Cavalier-Smith et al. 2015; Burki et al. 2016; Sierra et al. 2016) by adding genes to Foraminifera and a substantially larger sampling of other Rhizaria species.

Using these data, our analyses always cluster Radiolaria, Foraminifera and Taxopodida into Retaria. We find Radiolaria (excluding Taxopodida) and Foraminifera as two distinct clusters (congruent with Cavalier-Smith et al. 2015). Endomyxa and Retaria form a monophyletic group, revealing Cercozoa as paraphyletic. In our multi-gene alignments, data from two important endomyxean clades (i.e., Haplosporida and Vampyrellida) are absent. However, we included representatives from the two clades on the ML tree with the Evolutionary Placement Algorithm (Berger and Stamatakis 2011), and they fall inside the endomyxean clade, strengthening the monophyly of Retaria and Endomyxa (fig. 6). Rhizaria is always placed as sister to Alveolata and Stramenopiles in Bayesian inferences (and in ML analyses corrected for phylogenetic noise).

Taxopodida and Endomyxa as Sister Lineages to Foraminifera and Radiolaria

Taxopodida has previously been placed within Radiolaria (Nikolaev et al. 2004; Krabberød et al. 2011), but has two different positions in our trees dependent on the analysis. The Bayesian CATGTR trees show Taxopodida as sister to Radiolaria and Foraminifera, while ML LG places the species as sister to Radiolaria. We assessed the basis for this discrepancy by running Phylobayes with the LG model used in the ML analyses. The resulting Bayesian tree placed Taxopodida as sister to Radiolaria, similar to the ML tree, clearly demonstrating inconsistency caused by the less favored LG model. It should also be noted that removing fast evolving sites in the ML LG analysis changed the tree correspondingly towards the Bayesian topology by placing Taxopodida at the base of Retaria (supplementary table S2, Supplementary Material online). While the Bayesian inferences with CATGTR model were congruent, the ML topologies were less stable and converged towards the Bayesian tree with removal of fast evolving sites. The stability of the Bayesian results may be due to the use of the CATGTR model, which more realistically

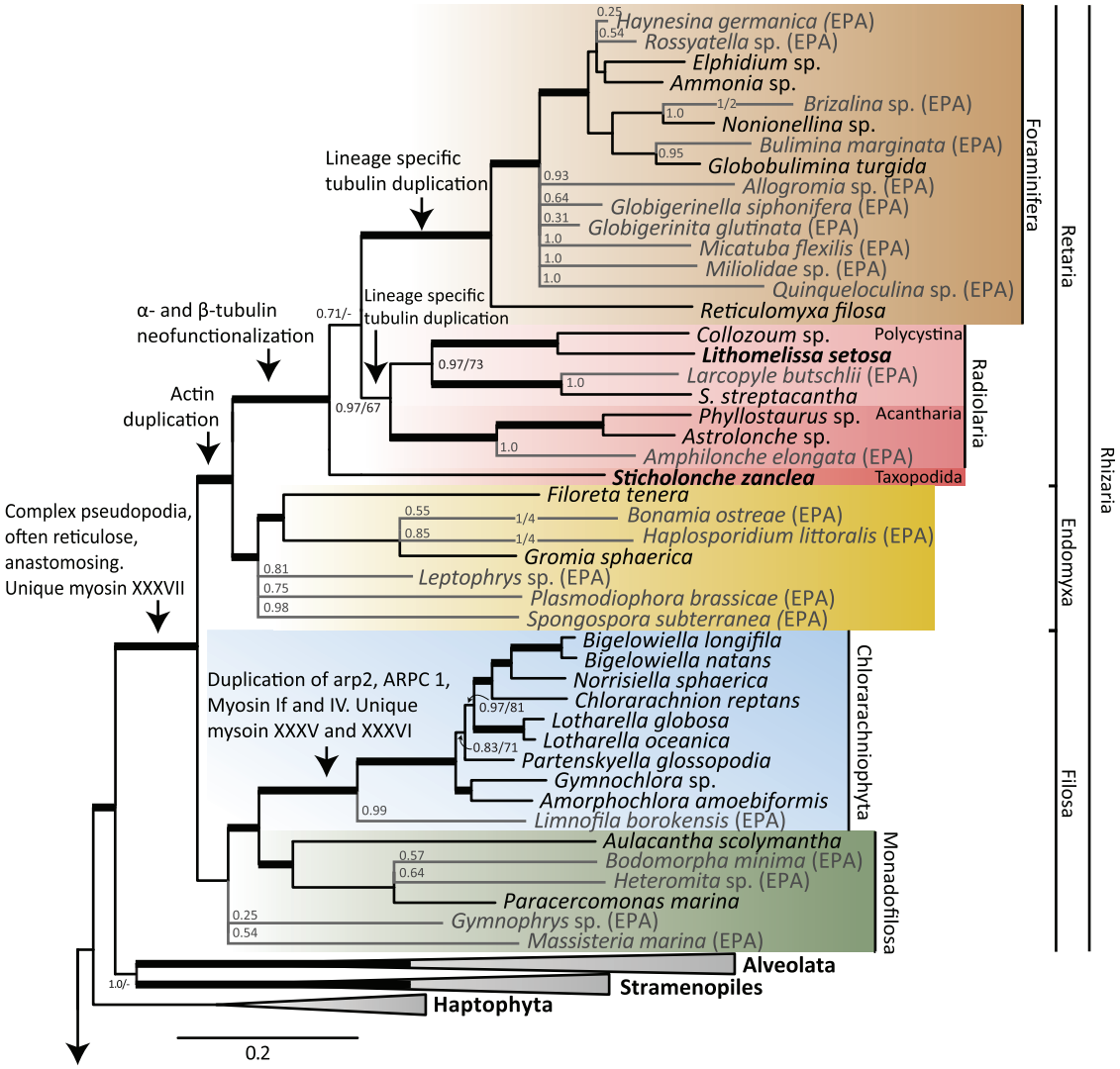

Fig. 6. Summary of rhizarian genetic and morphological evolution mapped on the phylogeny, combining results from Bayesian and ML analyses (i.e., fig.1 and supplementary fig. S3, Supplementary Material online). The basic branching pattern is presented as inferred by ML, since this enabled the use of Evolutionary Placement Algorithm (EPA) to place taxa with large portions of missing data (Berger et al. 2011). Branches in grey are the most likely placement of taxa from EPA with numbers showing the expected likelihood weights for the placement. The placement of *Sticholonche zanclea*, varied between ML and Bayesian inferences; here shown as sister to Retaria as inferred with the most favoured CATGTR Bayesian inference (fig. 1). Rhizaria is shown as sister to Stramenopiles and Alveolates, as in the Bayesian analysis. Taxa in bold are sequenced for this study. Arrows mark important morphological and genetic innovations. Thick branches are highly supported with bootstrap support >85% and posterior probability >0.9.

estimates the evolutionary substitution patterns in amino acids by taking into account across site heterogeneities in the amino acid substitution process (Lartillot and Philippe 2004; Lartillot et al. 2013), making it preferable over the LG model.

Taxopodida and Acantharia have been grouped together as Spasmaria based on the existence of contractile myonemes in both groups (Cavalier-Smith 1993), a grouping also supported in combined 18S and 28S rDNA phylogeny (Krabberød et al. 2011). Myonemes give taxopodidans the ability to swim using their pseudopodia like oars, while giving acantharians the ability to regulate their buoyancy by altering their cell volume (Cachon et al. 1977; Febvre 1981). However, if Taxopodida is sister to both Radiolaria and Foraminifera, this implies that contractile myonemes and flexible pseudopodia were an ancestral trait of Retaria that later has been lost or modified in Radiolaria and Foraminifera.

Endomyxa was originally defined as a clade within Cercozoa (Cavalier-Smith 2002). In our trees, however, Endomyxa was consistently excluded from the filose Cercozoa and placed as sister to Retaria. This implies that Rhizaria is split into three lineages: Filosa, Endomyxa, and Retaria. Taxopodida, Foraminifera, and Radiolaria constitute Retaria. This new branching order of rhizarian lineages forms the framework we here use to map changes of the cytoskeleton-related gene families and establish the order of macroevolutionary changes in Rhizaria.

Expansion of Actin, Myosin, and Subfunctionalization of Arp2/3 in Chlorarachniophyta

The chlorarachniophytes can form extensive networks of reticulose actin-based pseudopodia that they rely on for foraging and movement (Margulis et al. 1990). The evolution of these extensive pseudopodial networks seems to have been made possible by gene duplications of proteins controlling actin network dynamics as well as several duplications of the actin gene, and in chlorarachniophytes duplications of the myosin gene. The interaction between actin, the Arp2/3 complex, and myosin is important for pseudopod formation and branching. Branching points between two actin filaments are formed as the Arp2/3 complex recruits actin filaments into networks (Volkmann et al. 2001; Goley and Welch 2006; Mattila and Lappalainen 2008). Here we present evidence for a duplication ancestral to chlorarachniophytes of genes for two of the proteins in the complex: Arp2 and ARPC1. Both proteins are involved in the initial binding of nucleation promoting factors (NPFs), factors that are essential for the formation of protrusions that eventually lead to pseudopodia at the leading edge of motile cells (Boczkowska et al. 2014; Kast et al. 2015). Although the exact nature and conformation of the Arp2/3 complex are still under investigation, it seems clear that actin NPFs bind first to Arp2 and ARPC1, then extend the daughter filament by adding an actin subunit at the barbed end of Arp2 and Arp3 (Boczkowska et al. 2008, 2014). This in turn creates attachment points for daughter actin filaments to bind to the existing mother filament (Rouiller et al. 2008). In chlorarachniophytes, the Arp2 and ARPC1 paralogs have undergone divergent substitution patterns. The differences between the two Arp2 paralogs as well as the two ARPC1 paralogs are mainly found on the surface areas of the Arp2/3 complex where the actin recruiting proteins, NTPs and ultimately the newly formed actin filaments attach. Sites that are conserved and shared between both paralogs (marked green in fig. 3C and fig. 3D) are most likely important for the original function of the complex, while the sites that are conserved in one of the paralogs, but not the other, point to functional differentiation and innovation. In addition, myosin duplications have occurred ancestrally to Rhizaria before several independent events in chlorarachniophytes and Foraminifera.

Over evolutionary time scales these genetic innovations have likely formed the molecular basis of cellular and morphological differentiation in chlorarachniophytes. In turn, this has given them a larger repertoire of Arp2 and ARPC1 and an increased potential to recruit actin filaments to facilitate a reticulate cell and a gliding lifestyle.

Unique Duplication and Neofunctionalization of α- and β-Tubulin in Retaria

Similar to chlorarachniophytes, many species in Retaria and Endomyxa can form highly branched pseudopodial networks (Lee and Anderson 1991; Suzuki and Aita 2011). This is also reflected in the actin gene repertoire: Retaria has two distinct subfamilies of actin genes (fig. 2). Unlike chlorarachniophytes, retarians have additional pseudopods supported by microtubules called axopodia (Anderson 1983; Travis and Bowser 1986; Suzuki and Aita 2011). The axopodia in Radiolaria are often contractile and withdraw upon contact; rapid movement can cause prey to be drawn towards the cytoplasm of the cell where digestion occurs (Sugiyama et al. 2008). Similarly, Foraminifera have stiffened pseudopods called reticulopodia. These microtubule mediated pseudopods can extend and retract at a speed two orders of magnitude faster than in animal cells (Travis and Allen 1981; Bowser 2002). The extraordinary speed at which the microtubules can nucleate in Foraminifera has been linked to a duplication and neofunctionalization of β-tubulin (Habura et al. 2005; Hou et al. 2013). The discovery of the aberrant β2-tubulin in Retaria represented a paradox, since a corresponding α-tubulin paralog of the heterodimer could not be detected (Hou et al. 2013). The question is how an aberrant β-tubulin can function without a correspondingly deviant α-tubulin. Here, we solve this paradox by the detection of α2-tubulin in the single cell transcriptomes of *Sticholonche zanclea* and *Lithomelissa setosa*, enabling identification of homologs from other Retarian species. We also add new β-tubulin data from both *S. zanclea* and *L. setosa*, confirming gene expansion in all major Radiolaria lineages, and the origin of new paralogs in the common ancestor of Retaria. The overall pattern is that the new α2-tubulin paralog presented here evolved in a similar mode to that of the β2-tubulin gene (see also Habura et al. 2005; Hou et al. 2013). Interestingly, none of the α2-tubulin and β2-tubulin paralogs could be identified in available Endomyxa data, suggesting that these gene duplications are synapomorphic for Retaria, with an origin after the division of Retaria and Endomyxa (fig. 6).

Modeling of evolutionary rates on the tubulin structure shows global changes of the molecule along two different paths: first, several conserved and functionally important regions in α1 and β1 have become more variable, and probably therefore less functionally important in α2 and β2. This pattern is particularly obvious at the interface between the α and β heterodimers (which is the basic unit of protofilaments), and in the lateral surfaces between protofilaments that build up the microtubule. Second, many variable residues localized outside of the classical contact surfaces in the conventional α1 and β1 have become conserved in α2 and β2 and have probably gained new functional roles.

Retaria is unique among eukaryotes in having such divergent tubulin genes. It is not clear how retarians combine the four tubulin variants α1, α2, β1, and β2 into heterodimers, but the presence of these variants certainly enables modularity. The different affinities between the α and β tubulins will likely affect assembly and disassembly of microtubules, and may be used to adjust flexibility, strength, and conformation of the axopodia or reticulopodia (Löwe et al. 2001). In addition, we observe that many of the sites that have undergone evolutionary change are located on the surface of the heterodimer. They may represent binding sites for microtubule associated proteins (MAPs) as well as motor proteins, further expanding the range and flexibility of cytoskeletal structures (Brouhard and Rice 2014). Taken together, all major contact areas for both lateral and longitudinal interactions are less

conserved in the novel paralogs compared to the original, and both the α2 and β2 tubulins have undergone dramatic evolutionary changes and are likely to be functionally distinct from their α1 and β1 counterparts.

Single Cell Transcriptomics for Macroevolutionary Studies of Unculturable Protists

One of the main challenges of applying single cell transcriptomics to protists is the optimization of cell lysis. This is of special importance in the study of species with rigid skeletons and tough cell walls. Here we present modified lysis procedures for single cell transcriptomics (Picelli et al. 2014). Radiolaria species have a tough cellular wall that protects the endoplasm, and successful lysis of these indicates that the method can be applied to less hardy unicellular species. The number of predicted genes from our single cell transcriptomes is comparable to that generated from colonies or pooling of hundreds of cells from other radiolarian species (Burki et al. 2010; Balzano et al. 2015), as well as other experiments where similar methods have been applied on cells in culture (Kolisko et al. 2014; Macaulay et al. 2016; Liu et al. 2017). Subsampling of sequence reads showed sufficient sequencing depth, suggesting that an incomplete transcriptome was likely due to stochastic loss of mRNA. Thus, transcriptomes of sufficient quality for phylogenomic and molecular evolutionary analyses can be generated from single cells isolated from natural samples. This protocol can undoubtedly be applied to other uncultured protists, adding resolution to the relationships between eukaryotes, in addition to revealing the evolution of morphologically related genes.

Conclusion

Data generated from these transcriptomes demonstrate that genetic innovations through multiple gene duplication and neofunctionalization processes, rather than co-option of deep gene homologs, have taken place in cytoskeletal genes of Chlorarachniophyta and Retaria. Differential expansion of genes in chlorarachniophytes and Retaria shows that underlying genetic changes in cytoskeletal evolution have taken different routes in morphologically distinct groups; the overall pattern of the data reveals extensive gene duplications of actin-related proteins in chlorarachniophytes and of α- and β-tubulins in Retaria, with group-specific expansions of myosin and actin in both groups (fig. 6). The hypothesized connection between the evolutionary changes in cytoskeletal genes and the cellular morphology of the cells suggests that genetic innovations occurred in the ancestors of the respective groups, forming the basis for morphological, and species diversification. While the actin-related proteins and the myosin motor proteins that use them have driven changes in chlorarachniophytes, tubulin has directed central aspects of Retaria evolution. Subsequent to the initial innovation, additional expansions of functional genes crucial to cytoskeletal formation have impacted on the morphological diversification of Chlorarachniophyta and Retaria. Our analyses elucidate relationships between genotype and phenotype of these organisms, linking gene evolution to evolution of cell morphology. Better understanding of macroevolution in these organisms will require functional studies regarding what types of actin branching the new Arps can form in chlorarachniophytes and how Retaria combine the two sets of α and β tubulin proteins in their protofilaments. Such studies should be complemented with more data from other gene families known to be involved in cytoskeleton development, regulation, and transportation, such as MAPs, GTPases, dynein, and kinesin (Hammer and Wu 2002; Kollmar et al. 2012; Rojas et al. 2012; Brouhard and Rice 2014) .

Materials and Methods

Sampling and Transcriptome Amplification

Plankton samples were collected from the inner part of the Oslo fjord (May 2014) using a net haul with a mesh size of 60 µm. The seawater samples were stored overnight in an incubator holding the same temperature as the fjord to let living cells recover and self-clean. Radiolarian cells were manually extracted from the plankton samples by capillary isolation with Pasteur pipettes and an inverted microscope. Cells were individually photographed and then thoroughly washed in sterile PBS to remove possible surface contamination (supplementary fig. S1, Supplementary Material online). Immediately after isolation, cells were placed in Nucleospin RNA XS lysis buffer (Macherey-Nagel) and processed further. Total RNA was isolated from the free-living radiolarian cells using Nucleospin RNA XS (Macherey-Nagel) following standard protocol, with on-column DNase treatment and eluting with 5 µl elution buffer. Hybridization of oligo(dT) primer, reverse transcription, template switching, and PCR amplification of cDNA were performed by modification of a protocol outlined in (Picelli et al. 2014) called Smart-seq2; we used 7 µl of mRNA mix (5 µl isolated RNA, 1 µl oligo(dT) primer, and 1 µl 10 mM dNTPs) which was added to 9 µl of reverse transcriptase (RT) mix). All 16 µl (mRNA + RT mix) was used for PCR amplification employing 20 cycles. The quality of the resulting cDNA was assessed using a Bioanalyzer (Agilent) with a high-sensitivity DNA chip, in addition to visualization on a 1% TAE gel. cDNA concentrations were measured using a Qubit fluorometer (Life Technologies) and the dsDNA HS assay kit.

Sequencing and Assembly

Library preparation and sequencing of the cDNA with Illumina MiSeq were performed at the Natural History Museum in London. The sample was prepared using the Illumina TruSeq Nano DNA LT Library Preparation Kit (FC-121-4001). The standard Illumina protocol was followed with fragmentation on a Covaris M220 Focused-ultrasonicator. The finished library was quality checked using an Agilent Tapestation to check the size of the library fragments, and a qPCR in a Corbett RotorGene instrument to quantify the library. This was repeated for two MiSeq 600 cycle runs, 2*300 cycle paired end sequencing. The MiSeq platform was chosen over HiSeq since the longer reads would provide an easier assembly when dealing with a possible

metatranscriptomic library. The raw reads (19,894,654 for *S. zanclea* and 11,590,658 for *L. setosa*) were quality filtered and pairwise assembled with PEAR (Zhang et al. 2014) using default parameters. The reads were further cleaned with Trimmomatic (Bolger et al. 2014) and then *de novo* assembled into contigs with Trinity (Haas et al. 2013) using default settings. TransDecoder in the Trinity package was used to predict genes from the assembled cDNA (Haas et al. 2013).

To check if all transcripts in the library had been sequenced, the raw reads were randomly split up in 10 different datasets representing 10%, 20%, up to 90% of the original raw reads. The sub-sampled datasets were assembled and new gene predictions were independently performed using PEAR, Trimmomatic and Trinity as for the full dataset. Accumulation curves obtained by plotting the predicted gene number against increasing partition size show that the slope of the curves decreases with increasing partition size and more or less flattens when it reaches 100% of the total dataset for both libraries (supplementary fig. S2, Supplementary Material online). We therefore assume that acceptable sequencing depth for each library has been achieved, and that a further sequencing effort would not have increased the number of predicted genes significantly.

Alignment Construction, Paralog Identification, and Phylogenetic Inference

The BIR Pipeline

We used the BIR pipeline (www.bioportal.no; Kumar et al. 2015) to extract genes and prepare single gene alignments to be used in multi-gene phylogenetic analyses. As seed alignments for the BIR pipeline we used 258 genes previously published in multi-gene phylogenies (Burki et al. 2012). As a query database we used the generated transcripts from our single cell transcriptomes (6898 in total), all proteins in GenBank with Rhizaria as TaxID (44278 sequences at the time of retrieval, October 2014), all 16 transcriptomes assigned to Rhizaria from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling et al. 2014. See table 2), as well as all rhizarian sequences from Sierra et al. (2013). In addition, seven reference genomes are included in the BIR pipeline (*Arabidopsis thaliana, Bigelowiella natans, Dictyostelium discoideum, Guillardia theta, Homo sapiens, Monosiga brevicollis, Naegleria gruberi, Paramecium tetraurelia, Saccharomyces cerevisiae*, and *Thalassiosira pseudonana* (Kumar et al. 2015). In short, the BIR pipeline will screen the query sequences against the database consisting of one or more seed alignments, using BLAST, and assign the sequences that match the criteria set by the user to the corresponding alignment [for details, see Kumar et al. (2015)].

Single Gene Analyses

Maximum Likelihood (ML) trees for all single genes were constructed with RAxML v 8.0.2, with the program calculating the best fitting model for each gene (the option -m PROTGAMMAAUTO), and with the automatic bootstrapping criteria MRE (option -I autoMRE) (Pattengale et al. 2010; Stamatakis 2014). The Tree Certainty index (Salichos et al. 2014) was calculated for each tree separately, and all trees were run through a custom-made R script to decide whether the following clades were monophyletic or not: Opisthokonta, Fungi, Alveolata, Stramenopiles, Haptophyta, Rhizaria, Viridiplantae, Excavata, Fungi, and Rhodophyta. This allowed us to screen for genes containing artefacts and dubious sequences, such as sequences that had been assigned to the wrong species and sequences that originated from contamination and possible paralogs (Struck 2013). Three genes (β-tubulin, actin, and rac1) were found to have paralogs and deemed not suitable for multi-gene phylogenies. We therefore proceeded with 255 genes for the multi-gene analysis.

Supermatrix Construction

After screening we were left with 255 genes that were concatenated using ScaFos (Roure et al. 2007). We also merged close species into composite sequences when they covered different parts of the supermatrix (see supplementary table

Table 2. Rhizarian Transcriptomes from MMETSP (Keeling et al. 2014) Used in This Study.

Sample ID	Phylum	Species	Strain	Transcripts[a]
MMETSP0040	Chlorarachniophyta	*Lotharella oceanica*	CCMP622	17,354
MMETSP0041	Chlorarachniophyta	*Lotharella globosa*	LEX01	25,644
MMETSP0042	Chlorarachniophyta	*Amorphochlora amoebiformis*[b]	CCMP2058	23,387
MMETSP0045	Chlorarachniophyta	*Bigelowiella natans*	CCMP 2755	22,651
MMETSP0109	Chlorarachniophyta	*Chlorarachnion reptans*	CCCM449	26,481
MMETSP0110	Chlorarachniophyta	*Gymnochlora* sp.	CCMP2014	15,507
MMETSP0111	Chlorarachniophyta	*Lotharella globosa*	CCCM811	19,670
MMETSP0112	Chlorarachniophyta	*Lotharella globosa*	CCCM811	11,910
MMETSP0113	Chlorarachniophyta	*Norrisiella sphaerica*	BC52	14,550
MMETSP0186	Cercozoa	*Minchinia chitonis*	Missing	461
MMETSP1052	Chlorarachniophyta	*Bigelowiella natans*	CCMP623	24,186
MMETSP1318	Chlorarachniophyta	*Partenskyella glossopodia*	RCC365	15,025
MMETSP1358	Chlorarachniophyta	*Bigelowiella natans*	CCMP1242	18,273
MMETSP1359	Chlorarachniophyta	*Bigelowiella longifila*	CCMP242	15,959
MMETSP1384	Foraminifera	*Ammonia* sp.	Missing	31,225
MMETSP1385	Foraminifera	*Elphidium margaritaceum*	Missing	25,184

[a]The number of amino acid sequences for the sample.
[b]*Amorphochlora amoebiformis* is called *Lotharella amoebiformis* in MMETSP, but it was moved to the genus *Amorphochlora* by Ishida et al. (2011).

S3, Supplementary Material online). The final matrix had a length of 54,898 amino acids with 124 taxa.

Removal of Jumping and Long Branched Taxa

Mikrocytos mackini was not included in the analysis due to an extremely long branch (Burki et al. 2013). RogueNaRok, using default parameters (Aberer et al. 2013) was used to identify additional jumping taxa, which also were excluded from further analysis (supplementary table S3, Supplementary Material online)

Reduced Dataset

We also constructed a concatenated dataset consisting of 146 representative genes for easier and faster analysis. The selection of genes was made to meet several criteria: we excluded genes that had less than 45 taxa (50% of the inferred taxa), a low relative Tree Certainty index (Salichos et al. 2014), or that failed to group at least two of the major clades mentioned above.

Missing Data

To assess the impact of missing data we incrementally excluded taxa with low coverage from the two concatenated datasets. First we set the highest allowed percentage of missing data for a taxon to be 90% of the total characters (i.e., if a taxon had more than 90% data missing it was excluded), the next cut-off at 80%, and finally at 70%. The number of characters in the matrix was held constant. The Tree Certainty index (Salichos et al. 2014) was calculated for each increment (see supplementary table S4, Supplementary Material online). The relative Tree Certainty index increased markedly when the threshold was set at 90%, but did not increase significantly after that, in fact there seems to be a decrease in the relative value of the Tree Certainty index as the number of taxa drops (supplementary table S4, Supplementary Material online).

Influence of Taxa with Low Coverage, or Uncertain Position

We also removed taxa and clades from Rhizaria that had a consistently low bootstrap value ($<$75%) or low posterior probability ($<$0.75 pp), but that had not been flagged by RogueNaRok, to see if they affected the topology of the phylogenetic inference. *Spongosphaera streptacantha* and *Sticholonche zanclea* were removed one by one and together from both the full and the reduced datasets. Foraminifera were also removed in some analyses (see supplementary table S4, Supplementary Material online).

Removal of Fast Evolving Sites

TIGER (Cummins and McInerney 2011) was used with default settings to produce categories of fast evolving sites, 10 categories in total for each dataset. Categories of fast evolving sites were removed in increments, starting with the category with the fastest evolving sites, subsequently removing the category with the second fastest evolving sites etc. Up to four categories were removed from all datasets before phylogenetic analyses.

Phylogenetic Analyses of Concatenated Dataset

Phylogenetic trees were inferred for all concatenated datasets, with RAxML choosing the best fitting model, and with the automatic bootstrapping criteria as previously described. The preferred model was always LG + Γ (see supplementary table S4, Supplementary Material online). Due to the heavy demand on computational resources from Bayesian inference, only five of the alignments were included for analysis with the CATGTR model in Phylobayes MPI version 1.5a (Lartillot et al. 2013), as well as 1 dataset with the LG model. For these, we ran two chains in parallel for at least 15,000 iterations, only stopping when the maxdiff was less than 0.3 (see supplementary table S4, Supplementary Material online).

Evolutionary Placement Algorithm

In order to place rhizarian species that had been excluded when the cut-off threshold for missing data had been raised on the phylogenetic tree, we used the Evolutionary Placement Algorithm (EPA) included in RAxML 8.0.26 (supplementary table S3, Supplementary Material online; Stamatakis et al. 2010; Berger et al. 2011; Stamatakis 2014). As reference tree, we used the 255 gene maximum likelihood tree with a 10% missing data cut off. EPA was used to obtain a broad representation of species from Rhizaria in the phylogenetic inferences. EPA assigns sequence fragments (short reads) to edges of a given phylogenetic tree under the maximum-likelihood (ML) model, allowing taxa with limited sequence data to be placed in a phylogenetic context.

Genes Related to Cytoskeleton Formation and Motor Proteins

The assembled transcriptomes from the single cells were annotated with InterProscan 5 (Jones et al. 2014) as implemented in Geneious 8 (Kearse et al. 2012). The annotations were screened for genes commonly involved in the formation and development of the cytoskeleton, as well as the most common cytoskeletal motor proteins. In particular, we looked for α- and β-tubulin, myosin, actin, and the actin regulating Arp2/3-complex that consists of seven actin-related proteins (arp2, arp3, ARPC1, ARPC2, ARPC3, ARPC4, and ARPC5). Reference alignments and sequences downloaded from PFAM (http://pfam.xfam.org/), as well as relevant other recently published alignments (Hou et al. 2013; Sebé-Pedrós et al. 2014; Cavalier-Smith et al. 2015), were used in BIR as seed alignments with the same query database as before. In addition, representatives for all the genes were compared to six additional non-rhizarian transcriptomes from MMETSP (MMETSP0039 *Eutreptiella gymnastica*, MMETSP0046 *Guillardia theta*, MMETSP0308 *Gloeochaete wittrockiana*, MMETSP0380 *Alexandrium tamarense*, MMETSP0902 *Thalassiosira Antarctica*, and MMETSP1150 *Emiliania huxleyi*), as well as against the nonredundant protein database in GenBank, using Blast. For each gene, ML trees were constructed with RAxML as before and manually curated for any confounding artefacts. Redundant and short sequences were manually removed in Geneious 8 (Kearse et al. 2012) before another round of ML analysis using RAxML and a Bayesian

analysis with the CATGTR model implemented in Phylobayes MPI version 1.5a (Lartillot et al. 2013). Comparative evolutionary analyses of tubulin and the duplicated genes encoding members of the Arp2/3 complex were performed by examining the evolutionary rates of the paralogs separately and then mapping the genes to structural models using Consurf (Ashkenazy et al. 2010; Celniker et al. 2013). InterPro annotations of functional domains of myosin were performed with InterProscan 5 (Jones et al. 2014; Mitchell et al. 2015). All sequences from *S. zanclea* and *L. setosa* used in this study have been deposited in European Nucleotide Archive (http://www.ebi.ac.uk/ena) with accession numbers (LT673657–LT673754). Alignments and trees can be downloaded from www.bioportal.no.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank Bente Edvardsen, UiO, for providing the plankton haul from which the sampled cells were isolated, and the Sequencing centre at Natural History Museum in London for performing the Illumina library preparations and sequencing. We would also like to thank Simon Picelli for answering questions related to the Smart-seq2 method, Fabien Burki for providing single gene alignments, and The Gordon and Betty Moore Foundation for making all those wonderful protists transcriptomes available for the public and the reviewers whose comments greatly improved the quality of the manuscript. All analyses were run either on the Abel supercomputer at The High Performance Computing (HPC) cluster at University Of Oslo, or on Lifeportal (www.lifeportal.uio.no). For more information on BIR see www.bioportal.no. This project was funded by University of Oslo. This work was supported by grants from University of Oslo and from Research Council of Norway to Shalchian-Tabrizi (NFR216475).

References

Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst Biol*. 62:162–166.

Anderson OR. 1978. Light and electron microscopic observations of feeding behavior, nutrition, and reproduction in laboratory cultures of *Thalassicolla nucleata*. *Tissue Cell* 10:401–412.

Anderson OR. 1983. Radiolaria. 2nd ed. New York: Springer-Verlag.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*. 38:W529–W533.

Balzano S, Corre E, Decelle J, Sierra R, Wincker P, Silva C, Poulain J, Pawlowski J, Not F. 2015. Transcriptome analyses to investigate symbiotic relationships between marine protists. *Front Microbiol*. 6:1–14. 98:

Bass D, Chao EE-Y, Nikolaev S, Yabuki A, Ishida K-I, Berney C, Pakzad U, Wylezich C, Cavalier-Smith T. 2009. Phylogeny of novel naked Filose and Reticulose Cercozoa: Granofilosea cl. n. and Proteomyxidea revised. *Protist* 160:75–109.

Berger SA, Krompass D, Stamatakis A. 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol*. 60:291–302.

Berger SA, Stamatakis A. 2011. Aligning short reads to reference alignments and trees. *Bioinformatics* 27:2068–2075.

Boczkowska M, Rebowski G, Kast DJ, Dominguez R. 2014. Structural analysis of the transitional state of Arp2/3 complex activation by two actin-bound WCAs. *Nat Commun*. 5:3308.

Boczkowska M, Rebowski G, Petoukhov MV, Hayes DB, Svergun DI, Dominguez R. 2008. X-ray scattering study of activated Arp2/3 complex with bound actin-WCA. *Structure* 16:695–704.

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

Bowser SS. 2002. Reticulopodia: structural and behavioral basis for the suprageneric placement of granuloreticulosan protists. *J Foraminifer Res*. 32:440–447.

Brouhard GJ, Rice LM. 2014. The contribution of $\alpha\beta$-tubulin curvature to microtubule dynamics. *J Cell Biol*. 207:323–334.

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*. 23:1–7.

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 Biol Sci*. 283:1–10.

Burki F, Keeling PJ. 2014. Rhizaria. *Curr Biol*. 24:R103–R107.

Burki F, Kudryavtsev A, Matz MV, Aglyamova GV, Bulman S, Fiers M, Keeling PJ, Pawlowski J. 2010. Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. *BMC Evol Biol*. 10:377.

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.

Cachon J, Cachon M, Tilney LG, Tilney MS. 1977. Movement generated by interactions between the dense material at the ends of microtubules and non-actin-containing microfilaments in *Sticholonche zanclea*. *J Cell Biol*. 72:314–338.

Cavalier-Smith T. 1993. Kingdom Protozoa and its 18 phyla. *Microbiol Rev*. 57:953–994.

Cavalier-Smith T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol*. 52:297–354.

Cavalier-Smith T. 2009. Megaphylogeny, cell body plans, adaptive zones: causes and timing of eukaryote basal radiations. *J Eukaryot Microbiol*. 56:26–33.

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, Guy L, Saw JH, Ettema TJG, Eme L, Sharpe SC, Brown MW, Irimia M, Roy SW. 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:a016006.

Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, Mayrose I, Pupko T, Ben-Tal N. 2013. ConSurf: using evolutionary data to raise testable hypotheses about protein function. *Isr J Chem*. 53:199–206.

Cummins CA, McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Syst Biol*. 60:833–844.

Febvre J. 1981. The myoneme of the Acantharia (Protozoa): a new model of cellular motility. *Biosystems* 14:327–336.

Giannone G, Dubin-Thaler BJ, Rossier O, Cai Y, Chaga O, Jiang G, Beaver W, Döbereiner H-G, Freund Y, Borisy G, et al. 2007. Lamellipodial actin mechanically links myosin activity with adhesion-site formation. *Cell* 128:561–575.

Goley ED, Welch MD. 2006. The ARP2/3 complex: an actin nucleator comes of age. *Nat Rev Mol Cell Biol*. 7:713–726.
Grain J. 1986. The cytoskeleton in protists: nature, structure, and functions. *Int Rev Cytol*. 104:153–249.
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*. 8:1494–1512.
Habura A, Wegener L, Travis JL, Bowser SS. 2005. Structural and functional implications of an unusual foraminiferal beta-tubulin. *Mol Biol Evol*. 22:2000–2009.
Hammer JA, Wu XS. 2002. Rabs grab motors: defining the connections between Rab GTPases and motor proteins. *Curr Opin Cell Biol*. 14:69–75.
He D, Sierra R, Pawlowski J, Baldauf SL. 2016. Reducing long-branch effects in multi-protein data uncovers a close relationship between Alveolata and Rhizaria. *Mol Phylogenet Evol*. 101:1–7.
Hou Y, Sierra R, Bassen D, Banavali NK, Habura A, Pawlowski J, Bowser SS. 2013. Molecular evidence for β-tubulin neofunctionalization in Retaria (Foraminifera and Radiolarians). *Mol Biol Evol*. 30:2487–2493.
Ishida KI, Yabuki A, Ota S. 2011. Research note: *Amorphochlora amoebiformis* gen. et comb. nov. (Chlorarachniophyceae). *Phycol Res*. 59:52–53.
Jékely G. 2007. Eukaryotic membranes and cytoskeleton. New York, NY: Springer.
Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240.
Kast DJ, Zajac AL, Holzbaur ELF, Ostap EM, Dominguez Correspondence R, Dominguez R. 2015. WHAMM directs the Arp2/3 complex to the ER for autophagosome biogenesis through an actin comet tail mechanism. *Curr Biol*. 25:1791–1797.
Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649.
Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral Zettler L, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*. 12:e1001889.
Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. 2014. Single-cell transcriptomics for microbial eukaryotes. *Curr Biol*. 24:R1081–R1082.
Kollmar M, Lbik D, Enge S. 2012. Evolution of the eukaryotic ARP2/3 activators of the WASP family: WASP, WAVE, WASH, and WHAMM, and the proposed new family members WAWH and WAML. *BMC Res Notes* 5:88.
Krabberød AK, Bråte J, Dolven JK, Ose RF, Klaveness D, Kristensen T, Bjørklund KR, Shalchian-Tabrizi K. 2011. Radiolaria divided into polycystina and spasmaria in combined 18S and 28S rDNA phylogeny. *PLoS One* 6:e23526.
Kumar S, Krabberød AK, Neumann RS, Michalickova K, Zhao S, Zhang X, Shalchian-Tabrizi K. 2015. BIR pipeline for preparation of phylogenomic data. *Evol Bioinform Online* 11:79–83.
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.
Lee JJ, Anderson OR. 1991. Biology of Foraminifera. London: Academic Press.
Liu Z, Hu SK, Campbell V, Tatters AO, Heidelberg KB, Caron DA. 2017. Single-cell transcriptomics of small microbial eukaryotes: limitations and potential. *ISME J*. doi: 10.1038/ismej.2016.190. [Epub ahead of print].
Löwe J, Li H, Downing KH, Nogales E. 2001. Refined structure of alpha beta-tubulin at 3.5 A resolution. *J Mol Biol*. 313:1045–1057.
Macaulay IC, Svensson V, Labalette C, Ferreira L, Hamey F, Voet T, Teichmann SA, Cvejic A. 2016. Single-cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep*. 14:966–977.
Margulis L, Corliss JO, Melkonian M, Hapman DJ. 1990. Handbook of protoctista: the structure, cultivation, habitats, and life histories of the eukaryotic microorganisms and their descendants exclusive of animals, plants and fungi: a guide to the algae, ciliates, foraminifera, sporozoa, water molds, slime m. Boston: Jones and Bartlett Publishers.
Mattila PK, Lappalainen P. 2008. Filopodia: molecular architecture and cellular functions. *Nat Rev Mol Cell Biol*. 9:446–454.
Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, et al. 2015. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res*. 43:D213–D221.
Mogilner A, Keren K. 2009. The shape of motile cells. *Curr Biol*. 19:R762–R771.
Nikolaev SI, Berney C, Fahrni JF, Bolivar I, Polet S, Mylnikov AP, Aleshin VV, 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.
Pattengale ND, Swenson KM, Moret BME. 2010. Uncovering hidden phylogenetic consensus. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 6053 LNBI. p. 128–139.
Pawlowski J. 2008. The twilight of Sarcodina: a molecular perspective on the polyphyletic origin of amoeboid protists. *Protistology* 5:281–302.
Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc*. 9:171–181.
Richards TA, Cavalier-Smith T. 2005. Myosin domain evolution and the primary divergence of eukaryotes. *Nature* 436:1113–1118.
Rojas AM, Fuentes G, Rausell A, Valencia A. 2012. The Ras protein superfamily: evolutionary tree and role of conserved amino acids. *J Cell Biol*. 196:189–201.
Rouiller I, Xu XP, Amann KJ, Egile C, Nickell S, Nicastro D, Li R, Pollard TD, Volkmann N, Hanein D. 2008. The structural basis of actin filament branching by the Arp2/3 complex. *J Cell Biol*. 180:887–895.
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.
Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol*. 31:1261–1271.
Sebé-Pedrós A, Grau-Bové X, Richards TA, Ruiz-Trillo I. 2014. Evolution and classification of myosins, a paneukaryotic whole-genome approach. *Genome Biol Evol*. 6:290–305.
Sierra R, Cañas-Duarte SJ, Burki F, Schwelm A, Fogelqvist J, Dixelius C, González-García LN, Gile GH, Slamovits CH, Klopp C, et al. 2016. Evolutionary origins of rhizarian parasites. *Mol Biol Evol*. 33:980–983.
Sierra R, Matz MV, 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.
Stamatakis A, Komornik Z, Berger SA. 2010. Evolutionary placement of short sequence reads on multi-core architectures. In: 2010 ACS/IEEE International Conference on Computer Systems and Applications, AICCSA 2010. p. 1–44.
Struck TH. 2013. The impact of paralogy on phylogenomic studies—a case study on annelid relationships. *PLoS One* 8:e62892.
Sugiyama K, Hori RS, Kusunoki Y, Matsuoka A. 2008. Pseudopodial features and feeding behavior of living nassellarians *Eucyrtidium hexagonatum* Haeckel, *Pterocorys zancleus* (Müler) and *Dictyocodon prometheus* Haeckel. *Paleontol Res*. 12:209–222.
Suzuki NO, Aita YO. 2011. Radiolaria: achievements and unresolved issues: taxonomy and cytology. *Plankt Benthos Res*. 6:69–91.

Travis JL, Allen RD. 1981. Studies on the motility of the foraminifera. I. Ultrastructure of the reticulopodial network of *Allogromia laticollaris* (Arnold). *J Cell Biol*. 90:211–221.

Travis JL, Bowser SS. 1986. A new model of reticulopodial motility and shape: evidence for a microtubule-based motor and an actin skeleton. *Cell Motil Cytoskeleton* 6:2–14.

Ura S, Pollitt AY, Veltman DM, Morrice NA, MacHesky LM, Insall RH. 2012. Pseudopod growth and evolution during cell movement is controlled through SCAR/WAVE dephosphorylation. *Curr Biol*. 22:553–561.

Vale RD. 2003. The molecular motor toolbox for intracellular transport. *Cell* 112:467–480.

Volkmann N, Amann KJ, Stoilova-McPhie S, Egile C, Winter DC, Hazelwood L, Heuser JE, Li R, Pollard TD, Hanein D. 2001. Structure of Arp2/3 complex in its activated state and in actin filament branch junctions. *Science* 293:2456–2459.

Wickstead B, Gull K. 2011. The evolution of the cytoskeleton. *J Cell Biol*. 194:513–525.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30:614–620.