

1 **Single cell transcriptomics, mega-phylogeny and the genetic basis of**
2 **morphological innovations in Rhizaria**

3

4 Anders K. Krabberød¹, Russell J. S. Orr¹, Jon Bråte¹, Tom Kristensen¹, Kjell R. Bjørklund² & Kamran
5 Shalchian-Tabrizi^{1*}

6

7 ¹Department of Biosciences, Centre for Integrative Microbial Evolution and Centre for Epigenetics,
8 Development and Evolution, University of Oslo, Norway

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

10

11 *Corresponding author:

12 Kamran Shalchian-Tabrizi

13 Kamran@ibv.uio.no

14 Mobile: + 47 41045328

15

16

17 **Keywords:** cytoskeleton, phylogeny, protists, Radiolaria, Rhizaria, SAR, single-cell, transcriptomics

18 **Abstract**

19 The innovation of the eukaryote cytoskeleton enabled phagocytosis, intracellular transport and
20 cytokinesis, and is responsible for diverse eukaryotic morphologies. Still, the relationship between
21 phenotypic innovations in the cytoskeleton and their underlying genotype is poorly understood.
22 To explore the genetic mechanism of morphological evolution of the eukaryotic cytoskeleton we
23 provide the first single cell transcriptomes from uncultivable, free-living unicellular eukaryotes: the
24 radiolarian species *Lithomelissa setosa* and *Sticholonche zanclea*. Analysis of the genetic
25 components of the cytoskeleton and mapping of the evolution of these to a revised phylogeny of
26 Rhizaria reveals lineage-specific gene duplications and neo-functionalization of α and β tubulin in
27 Retaria, actin in Retaria and Endomyxa, and Arp2/3 complex genes in Chlorarachniophyta. We
28 show how genetic innovations have shaped cytoskeletal structures in Rhizaria, and how single cell
29 transcriptomics can be applied for resolving deep phylogenies and studying gene evolution of
30 uncultivable protist species.

31

32 Introduction

33 One of the major eukaryotic innovations is the cytoskeleton, consisting of microtubules, actin
34 filaments, actin-related proteins, intermediate filaments and motor proteins. Together these
35 structures regulate the internal milieu of the cell, aid in movement, cytokinesis, phagocytosis, and
36 predation (Dustin 1984, Grain 1986, Vale 2003, Wickstead & Gull 2011, Katz 2012, Cavalier-Smith
37 et al. 2014). Of essential importance, and the main focus of this work, the cytoskeleton of
38 unicellular eukaryotes determines the morphological patterning of the cell.

39 The evolution of the eukaryotic cytoskeleton is an intriguing story of gene evolution. Homologs to
40 actin and tubulin genes can be found in prokaryotes and Archaea, but the origin of the motor
41 proteins is unclear, lacking distinct homologs in prokaryotes (Vale 2003, Wickstead & Gull 2011).
42 All three major types of motor proteins – kinesin, dynein and myosin – are present in all eukaryote
43 supergroups. Therefore, it is likely that they were already present in the last eukaryotic common
44 ancestor (LECA). Early in the evolution of eukaryotes the cytoskeletal filaments of prokaryotes
45 were given new functions and new motor proteins were invented in addition to a large repertoire
46 of molecules that modify and interact with both the cytoskeleton and the motor proteins
47 (Goldstein 2001, Karcher et al. 2002, Schliwa & Woehlke 2003, Vale 2003, Seabra & Coudrier 2004,
48 Wickstead & Gull 2011).

49 Most of what we know about the eukaryotic cytoskeleton comes from studies of humans, plants
50 and fungi (Jékely 2007, Wickstead & Gull 2011), but less is known about the genetic machinery
51 and the molecular architecture of the cytoskeleton in non-model single celled eukaryotes
52 (protists). Our current knowledge about the evolution of cytoskeletal genes in protists stems from
53 human pathogens, e.g. *Plasmodium*, *Toxoplasma* and *Cryptosporidium* (Wickstead & Gull 2011,
54 Burki & Keeling 2014), but virtually nothing is known about how the evolution of these genes has
55 shaped cytoskeletal morphology in other protists.

56 In this paper we therefore trace the evolution of key cytoskeletal genes in a major group of
57 eukaryotes, Rhizaria, consisting predominantly of understudied single celled protists (Burki &
58 Keeling 2014). Rhizaria is a huge eukaryotic group and harbours species displaying a stunning
59 variety of morphological traits, from naked amoebas to species with delicate and spectacular tests
60 or skeletons. Rhizaria as a group was originally established based on molecular phylogenies that
61 placed the three clades Cercozoa, Radiolaria, and Foraminifera together (Cavalier-Smith 2002,

62 Nikolaev et al. 2004). Although no clearly defined phenotypic synapomorphies for Rhizaria have
63 been described (Pawlowski 2008), there is a common theme to many rhizarians: well-developed
64 pseudopodia which are often reticulose or filose. The different groups of rhizarians use their
65 pseudopodia in different ways: Some form complicated reticulose networks, e.g. many
66 chlorarachniophytes, others use pseudopodia stiffened by microtubules to capture prey, e.g.
67 Radiolaria, to move molecules and organelles, e.g. Foraminifera, or even as oars in Taxopodida
68 (Cachon et al. 1977, Anderson 1978, Sugiyama et al. 2008, Bass et al. 2009). But how this widely
69 different application of pseudopodia has evolved and how the morphological evolution is reflected
70 in changes to cytoskeletal genes is unknown. The cytoskeleton and motor proteins are an integral
71 part of pseudopod development and usage. In the formation of pseudopods in mouse melanoma,
72 actin and myosin interacts in order to make a protrusion in the plasma membrane creating the
73 leading edge of the pseudopod. Nucleators anchor actin to the cell membrane and actin-related
74 proteins (i.e. the Arp2/3-complex) recruits additional actin filaments to form the branching
75 network that supports the pseudopod (Giannone et al. 2007, Mogilner & Keren 2009). The same
76 mechanism drives pseudopod growth and development in the amoeba *Dictyostelium* (Ura et al.
77 2012). More rigid pseudopods are made with bundles of microtubules that stiffen and support the
78 pseudopods (called axopodia in Radiolaria and reticulopodia in Foraminifera ; Anderson 1983, Lee
79 & Anderson 1991). The microtubules are typically hollow tubes or helical filament composed of
80 alternating α - and β -tubulin subunits (Welnhofner & Travis 1998). The evolution of the molecular
81 components of the cytoskeleton and pseudopodia in Rhizaria, and protists in general, remains
82 unclear.

83 To understand the evolution of the cytoskeleton and pseudopodia in Rhizaria a fully resolved
84 phylogenetic tree is vital, but getting a stable phylogeny for the entire group has proven
85 problematic. The main issues have been the relationship between Radiolaria and Foraminifera (i.e.
86 Retaria), the monophyly of Cercozoa, as well as the relationship between Rhizaria and its
87 immediate neighbours in the SAR supergroup, Stramenopiles and Alveolata (Burki et al. 2007,
88 2013, 2016, Parfrey et al. 2010, Krabberød et al. 2011, Sierra et al. 2013, 2015, Katz & Grant 2014,
89 Cavalier-Smith et al. 2015).

90 Reconstruction of multi-gene phylogenies have been hindered by lacking molecular data from key
91 rhizarian groups (Burki & Keeling 2014). The main reason for scarce data from Rhizaria is lacking
92 knowledge of how to hold species in culture. We have previously applied single cell genomics
93 (combined with gene-targeted PCR) to study the diversity of Retaria, but this method is not

94 optimal to obtain large numbers of protein coding genes as it also covers intergenic regions (Bråte
95 2012, Krabberød 2011). Other studies targeting the retarian transcriptome have required pooling
96 many, sometimes several hundred cells (Sierra et al. 2013, Balzano et al. 2015), a method not
97 optimal when morphological markers for species identification are missing or hard to define and
98 cultures cannot be established.

99 Here, we use single cell transcriptomics on two key Rhizaria species (*Sticholonche zanclea* and
100 *Lithomelissa setosa*) to build multi-gene phylogenies and investigate the genetic basis of
101 cytoskeletal differences in Rhizaria. Our aim is to reveal processes at the genetic level that may
102 have caused phenotypic changes to the cytoskeleton, and thereby better understand major
103 morphological transitions in this group of organisms. A key aspect is to understand if gene changes
104 are due to co-option processes, where deeply diverging homologs of cytoskeleton genes have
105 been recruited to new functions, or if novelties of morphologies are caused by innovations of new
106 gene families through gene duplication and neofunctionalization. Are genetic changes specific for
107 each subgroup of Rhizaria or are they common between lineages? And can these changes be used
108 to define homological structures and thereby define morphologically distinct categories of
109 organismal lineages in Rhizaria? To address these questions we apply single cell transcriptomics on
110 two free-living radiolarian species.

111

112 **Results**

113 **Single cell transcriptomics of two uncultured protists**

114 We generated cDNA libraries from two radiolarian specimens: *Lithomelissa setosa* and
115 *Sticholonche zanclea* (Figure 1). The cDNA was sequenced on the Illumina MiSeq platform, 300bp
116 paired end. This resulted in 19,894,654 reads for *S. zanclea* and 11,590,658 for *L. setosa*, which
117 were *de novo* assembled using the Trinity platform (Haas et al. 2013). Assembly resulted in two
118 Single Cell Transcriptomes (SCT) with 4,749 predicted genes for *S. zanclea* and 2,122 predicted
119 genes for *L. setosa* (Table 1). Subsampling and re-assembly of reads showed that the sequencing
120 threshold for both libraries was close to maximum (Figure 1 - figure supplement 1). We assessed
121 the suitability of the data for phylogenomic reconstruction by using the BIR pipeline for single
122 gene alignment and tree construction (Kumar et al. 2015). Using 255 seed alignments covering the
123 eukaryote Tree of Life (Burki et al., 2012) we identified 54 and 16 corresponding orthologous gene

124 sequences from *S. zancea* and *L. setosa* respectively. In addition BIR extracted 3,534 gene
125 sequences from Marine Microbial Eukaryote Transcriptome Sequencing Project, MMETSP (Keeling
126 et al., 2014) and 793 proteins from GenBank with TaxID 543769 (Rhizaria) and added these to their
127 corresponding alignments (See supplementary table S1). After concatenation of all gene
128 alignments we had a super-matrix consisting of 91 taxa and 54,898 amino acids (255 genes).

129

130 **Bayesian GTRCAT trees show consistent phylogeny for SAR and subgroups**

131 In the Bayesian analysis of the full dataset using the CATGTR model (255 genes 54,898 AA, 91 taxa,
132 Figure 2), Stramenopiles and Alveolates formed a clade, with Rhizaria as sister; branches for these
133 groups are fully supported (1.00 posterior probability (pp)). Haptophytes appeared as sister to SAR
134 (0.87 pp). The relationship and support values did not change for SAR with the removal of fast
135 evolving sites (Figure 2 – figure supplement 1; supplementary table S4). The haptophytes, jumped
136 to a position basal to Archaeplastida (0.82 pp) when four bins of fast evolving sites were removed
137 (Supplementary table S4).

138 Within Rhizaria, the three groups Foraminifera, Radiolaria and Taxopodida, all monophyletic,
139 formed a cluster (i.e. Retaria) with maximum support even when fast evolving sites were removed
140 (i.e. always 1.00 pp). Radiolaria and Foraminifera were placed together as a monophyletic group
141 (0.71 pp) with *S. zancea* branching off as sister to them both. This topology remained constant
142 after removing fast evolving sites (Figure 2 – figure supplement 1; supplementary table S4). The
143 posterior probability for the monophyly of Radiolaria together with Foraminifera, i.e. excluding *S.*
144 *zancea*, increased to 0.97 when fast evolving sites were removed (Figure 2 – figure supplement 1).
145 Endomyxa was monophyletic (1.00 pp) and always sister to Retaria with full support (1.00 pp),
146 rendering Cercozoa paraphyletic. Filosa was monophyletic in all analyses (1.00 pp).

147

148 **ML trees converged towards Bayesian topology after removal of fast evolving sites**

149 In contrast, the maximum likelihood (ML) analysis of the full dataset using the LG model (255
150 genes, 54,898 AA, 91 taxa, Figure 3), grouped Alveolata with Rhizaria instead of the Stramenopiles
151 (96% bootstrap support (bs)). Retaria was recovered with high support (88% bs) as in the Bayesian
152 tree, but *S. zancea* was no longer placed ancestrally to radiolarians and Foraminifera. Instead *S.*
153 *zancea* was sister to Radiolarians (88% bs). Importantly, however, *S. zancea* changed to a basal

154 position in Retaria after removal of fast evolving sites, consistent with all the GTRCAT Bayesian
155 trees (Figure 3 - figure supplement 1; Supplementary table S4).

156 Removal of fast evolving sites did not change the monophyly of Foraminifera and Radiolaria
157 (excluding *S. zanclea*) or the sister relation between alveolates and Rhizaria, but the support
158 values were reduced in both instances to 50% bs for Radiolaria together with Foraminifera and
159 67% bs for the alveolates together with Rhizaria (Figure 3 – figure supplement 1; Supplementary
160 table S4). Endomyxa and Retaria group together with full support as in the Bayesian analysis
161 (100% bs), making Cercozoa paraphyletic. As in the Bayesian phylogeny haptophytes appeared as
162 sister to SAR (77% bs) and changed position basal to the plants, glaucophytes and cryptomonads
163 (73% bs) after removal of fast evolving sites. Species with more than 10% missing data in the final
164 concatenated data matrix were placed on the ML phylogeny using the Evolutionary Placement
165 Algorithm (Berger & Stamatakis 2011). Five species were placed in Endomyxa, five in Filosa, two in
166 Radiolaria, and finally ten species in Foraminifera (Figure 14).

167

168 **Influence of fast evolving sites and the choice of model on the phylogeny**

169 The discrepant topologies of the Bayesian (CATGTR) and ML (LG) trees could be due to the
170 different models implemented in these two approaches. We assessed the influence of these two
171 models by running Bayesian inferences using the LG model (the opposite: running ML with a
172 CATGTR model is currently not possible). This was done on a smaller alignment to reduce the
173 computational burden (146 genes, 33,081 AA, 91 taxa, see methods for further explanation). The
174 resulting Bayesian tree showed an important result: *S. zanclea* now grouped with Radiolaria (0.67
175 pp, Figure 3 – figure supplement 1) as in the ML (LG) tree, and not as sister to Foraminifera and
176 Radiolaria, as in all Bayesian trees with the CATGTR model. Other branching patterns in the
177 Rhizaria phylogeny were unaffected.

178 We repeated the ML (LG) analyses after removing fast evolving sites on the full dataset as well as
179 the reduced dataset. While alveolates and Rhizaria formed a clade in the full and small dataset
180 (85% bs, Figure 3- figure supplement 1), removal of four categories of fast evolving site moved
181 alveolates to the stramenopiles in the dataset with 146 genes (74% bs, Figure 3 - figure
182 supplement 1; Supplementary table S4), a result congruent with the Bayesian topology. The
183 support for alveolates together with Rhizaria was also weakened in the dataset with 255 gene
184 when four categories of fast evolving sites were removed, from 96% bs, to 67% bs. When

185 Foraminifera was excluded from the 255 gene dataset with four categories removed,
186 Stramenopiles and Alveolata formed a group with Rhizaria as sister (57% bs. Table S4).

187

188 **Actin radiation in Rhizaria and unique duplications in Retaria and Endomyxa**

189 We identified 6 actin sequences in our SCTs. From MMETSP we identified 18 foraminiferan actin
190 and 18 chlorarachniophyte sequences. Phylogenetic analysis of these and other available actin
191 sequences retrieved from GenBank and Pfam revealed that Retaria (including *S. zanglea*) have two
192 distinct paralogs of actin – actin1 and 2 – where actin2 is fully supported (Figure 4). Actin1 is
193 supported in the Bayesian analysis (0.87 pp) but not in by ML analysis. Actins from Endomyxa form
194 a weakly supported monophyletic group with retarian actin2 (14% bs/0.68 pp). This clade, in turn,
195 groups together with retarian actin1 (59% bs/ 0.97 pp), and is a synapomorphy for Retaria and
196 Endomyxa. There are possibly three paralogs of endomyxean actin, (named a-c in Figure 4) albeit
197 lacking support (a: 32%bs /0.6 pp, b: 16 %bs /- and c: 22 %bs /0.62 pp, Figure 4).

198

199 **Arp2/3 complex gene duplication in Chlorarachniophyta**

200 Of the seven genes in the Arp2/3 complex, which is responsible for branching of actin filaments
201 and recruitment of new actin, we identified *Arp2*, *Arp3*, *ARPC2* and *ARPC5* from *S. zanglea*, but
202 only *Arp2* from *L. setosa* (Figure 5). From MMETSP we identified sequences of all seven genes
203 from both Chlorarachniophyta and Foraminifera. Phylogenetic analysis of these genes revealed
204 that all chlorarachniophytes have two distinct paralogs of both *Arp2* and *ARPC1* (Figure 5A),
205 recovered with maximum support (100% bs/1.0 pp). No other species of Rhizaria has undergone
206 the same gene duplication in the Arp2/3 complex as the chlorarachniophytes (Figure 5B).

207

208 **Neofunctionalization of Arp2/3 in Chlorarachniophytes**

209 Comparative evolutionary analyses of the duplicated Arp2/3 complex genes (*Arp2* and *ARPC1*)
210 were performed by examining the evolutionary rates for each paralog, and then mapping the
211 genes to structural models using Consurf (Ashkenazy et al. 2010, Celniker et al. 2013). The analysis
212 showed that the two different forms of *Arp2* (*Arp2a* and *Arp2b*, Figure 6) and the two different
213 forms of *ARPC1* (*ARPC1a* and *ARPC1b*, Figure 7) follow a pattern where the most conserved sites
214 are localized inside the protein structure. Comparison of the surface between the two *Arp2*

215 proteins (Arp2a and Arp2b, Figure 6) show shared conserved residues in contact surfaces against
216 other proteins in the Arp2/3 complex (colored green in Figure 8). Similarly, the two different
217 paralogs of ARPC1 (ARPC1a and ARPC1b, Figure 8) show shared conserved sites localized inside
218 the complex. In contrast, the surfaces of the two Arp2 and ARPC1 copies show more variable
219 substitution rates, and all paralogs have patches with mutually exclusive conserved residues
220 (figure 8). As the surfaces of Arp2 and ARPC1 are responsible for the recruitment of the daughter
221 filament and are important for anchoring the two actin strands to each other, the divergent
222 substitution patterns imply that the pairs of paralogs have evolved into different directions from
223 the ancestral gene and thereby undergone sub-functionalization.

224

225 **Myosin evolution in Rhizaria**

226 We identified 133 myosin transcripts from MMETSP with rhizarian origin. A phylogenetic
227 reconstruction of the newly identified rhizarian myosins together with already published myosin
228 classes spanning a broad taxonomical distribution of eukaryotes (Richards & Cavalier-Smith 2005,
229 Sebé-Pedrós et al. 2014) revealed the presence of two known classes (I_f and IV) and three
230 previously unknown classes of myosin in Rhizaria (XXXV, XXXVI and XXXVII, following the naming
231 scheme of Sebé-Pedrós et al. (2014)). Myosin XXXVII is unique for Rhizaria, and marks the first
232 known synapomorphy for the group (Figure 9). It is highly supported (100 % bs and 1.0 pp) and has
233 a molecular signal distinct from other described myosins (Richards & Cavalier-Smith 2005, Sebé-
234 Pedrós et al. 2014). In this rhizarian-unique class there has been an additional radiation within the
235 chlorarachniophytes into three separate paralogs, all fully supported (100 % bs and 1.0 pp, Figure
236 9). Rhizarians have also gained a large repertoire of myosin IV, with six paralogs in
237 Chlorarachniophyta and two in Foraminifera. All paralogs were well supported phylogenetically (bs
238 > 90 % and pp >0.9) and differed from each other in functional domains (Figure 9). There was also
239 a class unique to Chlorarachniophyta that resembled myosin IV by having a MYTH4 domain at the
240 C-terminal, but with additional domains at the N-terminal usually not present in myosin IV
241 (Richards & Cavalier-Smith 2005, Sebé-Pedrós et al. 2014). However both paralogs of
242 Chlorarachniophyta in this class were phylogenetically distinct from myosin IV to warrant them to
243 be given a new class (myosin XXXV). We also found myosin I_f in Chlorarachniophyta, which have
244 been suggested to be present in the last common ancestor of all eukaryotes and recently lost in
245 Rhizaria (Sebé-Pedrós et al. 2014). However the presence of two paralogs of myosin I_f in
246 chlorarachniophytes show that it was present in the ancestor to Rhizaria, but that it might have

247 been lost in Endomyxa and Retaria after they separated from the chlorarachniophytes. Finally
248 there was a group unique to Chlorarachniophyta with two paralogs, named myosin XXXVI (Figure
249 9).

250

251 **α - and β -tubulin gene duplications in Retaria**

252 We report 16 new α -tubulin and 19 new β -tubulin sequences from our two retarian
253 transcriptomes: 4 α -tubulin, and 9 β -tubulin from *L. setosa*, 12 α -tubulin and β -tubulin from *S.*
254 *zancelea*. Additionally, we identified 26 α -tubulin and 42 β -tubulin sequences from other rhizarian
255 species in the MMETSP data (i.e. 12 Chlorarachniophyta and 14 Foraminifera α -tubulins; 10
256 Chlorarachniophyta and 32 Foraminifera β -tubulin). All these genes and other homolog sequences
257 identified in GenBank were added to the Pfam seed alignment for α and β -tubulin (Finn et al.
258 2014). The phylogenetic tree of rhizarian α -tubulin revealed two different version of the gene: the
259 canonical version of the α -tubulin gene (α 1-tubulin; α 1) and a novel group (α 2-tubulin; α 2) found
260 only in Retaria (Figure 10). The split separating the two versions received maximal support (100%
261 bs/ 1.0 pp). We identified α 2-tubulin in the SCTs from both *L. setosa* and *S. zancelea*. Together with
262 the available data from other Rhizaria, we confirm that this paralog is unique for Retaria. In
263 Foraminifera there were several paralogs of α 2, with most copies in *Reticulomyxa filosa* (25
264 copies). Foraminifera α 2 was paraphyletic with a clade branching at the base of Retaria (100% bs/
265 1.0 pp), before the Radiolaria and Taxopodida α 2 clade. The bootstrap support was low (70 % bs),
266 while the posterior probability was high (0.97 pp) for the branch that separated this group from
267 the rest of the Foraminifera (Figure 10).

268 Similarly, the β -tubulin trees contained a clearly divergent clade (i.e. β 2-tubulins) with several
269 copies for each Retarian group (Figure 11). All β 2-tubulin copies were grouped together with high
270 support (100% bs/ 1.0 pp) in agreement with earlier studies (Hou et al. 2013). We also found that
271 the β 2 copies were present in Taxopodida as well as in Foraminifera and Radiolaria.

272

273 **Neofunctionalization of tubulin genes in Retaria**

274 Comparative evolutionary analyses of the tubulin paralogs were done to identify patterns of
275 functional change. This was performed by estimation of evolutionary rates and mapping site rates
276 to tubulin structural models with ConSurf (Ashkenazy et al. 2010, Celniker et al. 2013). Highly
277 conserved amino acid residues were assumed to be functionally important and variable residues

278 to be of less importance for function. We therefore compared separately $\alpha 1$ with $\alpha 2$, and $\beta 1$ with
279 $\beta 2$; identifying sites conserved in one paralog and variable in the other. Such sites were believed
280 to have undergone functional shifts and therefore considered important for cytoskeleton
281 evolution. We also examined regions of the α - and β -tubulin structures known to be important for
282 microtubule function and dynamics. Evolutionary changes in these areas are likely to affect the
283 overall function of the microtubules.

284 Tracing evolutionary rates on the molecular structure of α - and β -tubulin (figure 12 and 13)
285 revealed two patterns of functional change between the conventional and new tubulin genes:
286 First, areas that are considered to be functionally important and conserved in α - and β -tubulin in
287 general were conserved in $\alpha 1$ and $\beta 1$ genes while being highly variable in $\alpha 2$ and $\beta 2$, with
288 differences being most prominent in α -tubulin. This pattern was observed for both inter- and
289 intra-dimerization surfaces between monomers (i.e. longitudinal interactions important for
290 protofilament assembly and disassembly), as well as lateral interactions between protofilaments.
291 Both the T7-loop and H8 helix, important for longitudinal interactions, are extremely conserved in
292 the original variant, while highly variable in the novel paralog. And similarly for the lateral contact
293 points, helix H12, which forms a ridge on the outside of the microtubule and therefore affects
294 binding and movement of motor proteins along the filament (Löwe et al. 2001) is much more
295 variable in the novel $\alpha 2$ -tubulin paralog than $\alpha 1$. Even the highly conserved residues in the M-loop
296 of the original tubulin variants are highly variable for $\alpha 2$ and $\beta 2$. These are residues directly
297 interacting between neighboring protofilaments (Löwe et al. 2001). Taken together, all major
298 contact areas both for lateral and longitudinal interactions are less conserved in the novel paralogs
299 compared to the original, especially for $\alpha 2$.

300 Second, and in contrast to the pattern above, areas of the tubulin molecules considered to be
301 functionally less important typically evolve faster than the contact surfaces. Several residues
302 outside of the conventional longitudinal and lateral binding sites are highly conserved in both $\alpha 2$
303 and $\beta 2$ while highly variable in the original $\alpha 1$ and $\beta 1$ genes (Supplementary alignment). Many of
304 these residues are exposed on the surface of the monomers and could represent new sites for
305 other tubulin interactions or surfaces for motor protein attachment and movement.

306 Altogether, both the $\alpha 2$ - and $\beta 2$ -tubulins have undergone dramatic evolutionary changes and are
307 likely functionally distinct from their $\alpha 1$ and $\beta 1$ counterparts (Figure 12 and 13).

308

309

310 **Discussion**

311 The last common ancestor of Rhizaria was most likely a naked, heterotrophic flagellate, who relied
312 extensively on its pseudopodia to explore the environment and to catch prey (Cavalier-Smith
313 2009). Its pseudopods were supported by actin and at least one group of myosins unique to
314 Rhizaria (Figure 4 and 9, summarized in Figure 14). The Rhizarian cytoskeletons have since
315 undergone evolutionary changes and their diversification follows a pattern where the major
316 groups have their own favoured filament: the chlorarachniophytes have relied on actin to support
317 their reticulose pseudopodia while the axopodia and reticulopodia in Retaria have been stiffened
318 by microtubules composed of tubulin. Although some structural differences between lineages are
319 known, little is established about the genetic basis of these phenotypes. Here we investigate the
320 genetic evolution responsible for diversification of the cytoskeleton and present a hypothesis that
321 relate these evolutionary changes to the varying morphology of Rhizaria groups. In order to fully
322 understand the evolution of the cytoskeleton in Rhizaria we started by constructing a robust
323 phylogenetic tree on which to map the evolutionary events.

324

325 **Placing Rhizaria in the Tree of Life**

326 Rhizaria has an evolutionary origin at the intersection between two very morphologically diverse
327 and abundant lineages in the Tree of Life: the alveolates and stramenopiles. Together the three
328 lineages form the supergroup SAR (Burki et al. 2007). Although there is little doubt that these
329 three lineages are closely related to each other, the exact relationship between them remains
330 debated. Three main hypotheses exists: either Stramenopiles and Rhizaria are monophyletic and
331 sister to Alveolates (Burki et al. 2007, 2013, Katz & Grant 2014), alveolates and stramenopiles
332 constitute a monophyletic group with Rhizaria as sister (Burki et al. 2008, 2010, 2012, 2016,
333 Parfrey et al. 2010, Cavalier-Smith et al. 2015), or finally Rhizaria and Alveolates are monophyletic
334 with Stramenopiles as sister group (Sierra et al. 2013, 2015, He et al. 2016).

335 Our Bayesian and ML inferences resulted in two different phylogenies. The Bayesian tree inferred
336 with the CATGTR model grouped alveolates and stramenopiles, whereas the ML tree with the LG
337 model clustered alveolates with Rhizaria. This inconsistency was evaluated by removing fast
338 evolving sites, which have been suggested to contain misleading phylogenetic information
339 (Philippe et al. 2005, Townsend 2007, Cummins & McInerney 2011, Townsend et al. 2012).

340 Removing such sites, caused no changes in the Bayesian phylogeny of the SAR groups, but the ML
341 analyses converged towards the Bayesian tree by grouping alveolates and stramenopiles (Figure
342 4A and Suppl. table S4). The CATGTR model is a better representation of amino acid substitution
343 patterns than the LG model, because it takes into account substitution pattern heterogeneity
344 (Lartillot & Philippe 2004). In addition, the Bayesian inferences were less affected by site selection
345 and were always reconstructing essentially identical phylogenies with the CATGTR model,
346 altogether strongly supporting the grouping of alveolates and stramenopiles with exclusion of
347 Rhizaria.

348 One of the remaining challenges about the SAR phylogeny is to identify the closest sister group of
349 SAR in the global phylogeny of eukaryotes. Recently it has been proposed that haptophytes
350 together with Centrohelida make up the sister clade to SAR (Burki et al. 2016). This is congruent to
351 our Bayesian and ML trees where the haptophytes branch at the base of SAR, at least prior to the
352 removal of fast evolving sites. However, removal of fast evolving sites typically groups
353 haptophytes together with cryptophytes at the base of the Archaeplastida, as seen in other recent
354 multi-gene phylogenies (Supplementary table S4; Parfrey et al., 2010; Brown et al., 2012; Katz &
355 Grant, 2014; Cavalier-Smith et al., 2015). This shifting position between SAR and the
356 Archaeplastida may reflect that haptophytes diverged at the base of both groups and therefore is
357 a key group for understanding the origin and evolution of this huge diversity of eukaryotes.

358

359 **Resolving Rhizarian Relationships**

360 Within Rhizaria, it has been suspected for some time that Foraminifera and Radiolaria are closely
361 related, and they have therefore been grouped together as Retaria (Cavalier-Smith 2002, Moreira
362 et al. 2007, Krabberød et al. 2011, Ishitani et al. 2011, Sierra et al. 2013). In phylogenies based on
363 ribosomal DNA, Foraminifera groups within radiolarians, although this placement has been
364 contested based on the aberrant nature of both the small (18S) and large (28S) subunit of
365 ribosomal genes in Foraminifera (Pawlowski & Burki 2009, Krabberød et al. 2011). Recent
366 phylogenomic analyses place Foraminifera either within Radiolaria implying Radiolaria to be a
367 paraphyletic group (Burki et al. 2013, Sierra et al. 2013, 2015) or as sister to Radiolaria (Cavalier-
368 Smith et al. 2015, Burki et al. 2016). However, these analyses lack two crucial pieces in the puzzle;
369 representatives from Nassellaria, one of the major polycystine radiolarian orders, and *S. zanglea*,
370 the only species of Taxopodida. Including both in combined 18S and 28S rDNA phylogenies,
371 divided the Radiolaria in two main groups, Polycystina and Spasmaria, where the latter contained

372 Taxopodida, but the position of Foraminifera was unresolved (Krabberød et al. 2011). Here, we
373 have generated transcriptome data and protein sequences from both the missing Radiolaria
374 groups in our multi-gene analyses, *L. setosa* (Nassellaria) and *S. zanclea* (Taxopodida). In addition,
375 we have reduced the impact of missing data in earlier phylogenomic analyses (Sierra et al. 2013,
376 2015, Cavalier-Smith et al. 2015, Burki et al. 2016) by adding genes to Foraminifera and a
377 substantially larger sampling of other Rhizaria species.

378 Using these data, our analyses always cluster Radiolaria, Foraminifera and Taxopodida into
379 Retaria. We find that Radiolaria (excluding Taxopodida) is monophyletic (congruent with Cavalier-
380 Smith et al. 2015). Endomyxa and Retaria form a monophyletic group, revealing Cercozoa as
381 paraphyletic. But in our multi-gene alignments, as in those of Sierra et al. (2013), data from two
382 important endomyxean clades (i.e. Haplosporida and Vampyrellida) are absent. However, we
383 included representatives from the two clades on the ML tree with the Evolutionary Placement
384 Algorithm (Berger & Stamatakis 2011) and they fall inside the endomyxean clade, strengthening
385 the monophyly of Retaria and Endomyxa (Figure 14).

386

387 **Taxopodida and Endomyxa revealed as sister lineages to Foraminifera and Radiolaria**

388 Taxopodida have previously been placed within Radiolaria (Nikolaev et al. 2004, Krabberød et al.
389 2011), but has two different positions in our trees dependent on the analysis. The Bayesian
390 CATGTR trees show Taxopodida as the sister to Radiolaria and Foraminifera, while ML LG place the
391 species as sister to Radiolaria. We assessed the basis for this discrepancy by running Phylobayes
392 with the substitution model used in the ML analyses (i.e. the LG model). The resulting Bayesian LG
393 tree placed Taxopodida as sister to Radiolaria – congruent with the ML tree – clearly
394 demonstrating the impact of the model on the phylogeny. It should also be noted that removing
395 fast evolving sites in the ML LG analysis changed the tree correspondingly by placing Taxopodida
396 at the base of Retaria (Figure 4B). While all the Bayesian inferences were highly congruent, the ML
397 topologies were less stable and converged towards the Bayesian tree with removal of fast evolving
398 sites. The stability of the Bayesian results may be due to the use of the CATGTR model which more
399 realistically estimates the evolutionary substitution patterns in amino acids by taking into account
400 across site heterogeneities in the amino acid substitution process (Lartillot & Philippe 2004,
401 Lartillot et al. 2013) and therefore preferable over the LG model.

402

403 All evidence taken into account, Taxopodida most likely diverged early in the radiation of Retaria
404 and before the separation of Radiolaria and Foraminifera. This has consequences for the
405 interpretation of the cytoskeleton and morphological evolution of Retaria. Taxopodida and
406 Acantharia were grouped together as Spasmaria based on the existence of contractile myonemes
407 in both groups (Cavalier-Smith 1993), a grouping also supported in 18S and 28S rDNA phylogenies
408 (Krabberød et al. 2011). Myonemes give taxopodidans the ability to swim using their pseudopodia
409 like oars while giving acantharians the ability to regulate their buoyancy by altering their cell
410 volume (Cachon et al. 1977, Febvre 1981). However, if Taxopodida is sister to both Radiolaria and
411 Foraminifera it implies that contractile myonemes and flexible pseudopodia, were an ancestral
412 trait of Retaria, and have later been lost or modified in Radiolaria and Foraminifera.

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

421

422 **Expansion of actin, myosin and subfunctionalization of Arp2/3 in Chlorarachniophyta**

423 The chlorarachniophytes can form extensive networks of reticulose actin-based pseudopodia that
424 they rely on for foraging and movement (Margulis 1990). The evolution of these extensive
425 pseudopodial networks seems to have been made possible by gene duplications of proteins
426 controlling actin network dynamics as well as several duplications of the actin gene, and of myosin
427 specific to chlorarachniophytes. The interaction between actin, the Arp2/3 complex, and myosin is
428 important for pseudopod formation and branching. Branching points between two actin filaments
429 are formed as the Arp2/3 complex recruits actin filaments into networks (Volkman et al. 2001,
430 Goley & Welch 2006, Pollard 2007, Mattila & Lappalainen 2008, Xu et al. 2011). Here we present
431 evidence for a duplication ancestral to chlorarachniophytes for two of the proteins in the complex:
432 Arp2 and ARPC1. Both proteins are involved in the initial binding of nucleation promoting factors
433 (NPFs) that are essential for the formation of protrusions that eventually leads to pseudopodia at
434 the leading edge of motile cells (Boczkowska et al. 2008, 2014, Xu et al. 2011, Ura et al. 2012, Kast

435 et al. 2015). Although the exact nature and conformation of the Arp2/3 complex are still under
436 investigation, it seems clear that actin NPFs bind first to Arp2 and ARPC1, then extend the
437 daughter filament by adding an actin subunit at the barbed end of Arp2 and Arp3 (Boczkowska et
438 al. 2008, 2014). This in turn creates attachment points for daughter actin filaments to bind to the
439 existing mother filament (Rouiller et al. 2008). In chlorarachniophytes the Arp2 and ARPC1
440 paralogs have undergone divergent substitution patterns. The differences between the two Arp2
441 paralogs as well as the two ARPC1 paralogs are mainly found on the surface areas of the Arp2/3
442 complex where the actin recruiting proteins, NTPs and ultimately the newly formed actin filaments
443 attach. Sites that are conserved and shared between both paralogs (marked green in Figure 8) are
444 most likely important for the original function of the complex, while the sites that are conserved in
445 one of the paralogs but not the other points to functional differentiation and innovation. In
446 addition myosin duplications have occurred ancestrally to Rhizaria before several independent
447 events in chlorarachniophytes and Foraminifera.

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

452

453 **Unique duplication and neofunctionalization of α - and β -tubulin in Retaria**

454 Similar to Chlorarachniophytes many species in Retaria and Endomyxa can form highly branched
455 pseudopodial networks (Anderson 1976a, b, Lee & Anderson 1991, Suzuki & Aita 2011). This is also
456 reflected in the expansion of actin genes: Retaria has two distinct subfamilies of actin genes, one
457 grouping with actin homologs from Endomyxa. In Endomyxa the actin diversity is extensive with
458 three possible paralogs (Figure 4). Unlike Chlorarachniophyta however, Retaria have additional
459 pseudopods supported by microtubules called axopodia (Anderson 1983, Travis & Bowser 1986,
460 Lee & Anderson 1991, Suzuki & Aita 2011). The axopodia in Radiolaria are often contractile and
461 withdraw upon contact; rapid movement can cause prey to be drawn towards the cytoplasm of
462 the cell where digestion occurs (Sugiyama et al. 2008). Similarly Foraminifera have stiffened
463 pseudopods called reticulopodia. These microtubule mediated pseudopods can extend and retract
464 at a speed two orders of magnitude faster than in animal cells (Travis & Allen 1981, Bowser 2002).
465 The extraordinary speed at which the microtubules can nucleate in Foraminifera has been linked
466 to a duplication and neo-functionalization of β -tubulin (Habura et al. 2005, Hou et al. 2013). The

467 discovery of the aberrant β 2-tubulin was a paradox, because the corresponding α -tubulin paralog
468 of the heterodimer was absent in all Retaria (Hou et al. 2013). The question is how an aberrant β -
469 tubulin can function without a correspondingly deviant α -tubulin. Here we solve this paradox by
470 presenting α 2-tubulin in the single cell transcriptomes of *Sticholonche zanclea* and *Lithomelissa*
471 *setosa*, which enabled identification of homologs from other Retarian species. We also add new β -
472 tubulin data from both *S. zanclea* and *L. setosa*, confirming gene expansion in all major Radiolaria
473 lineages, and the origin of new paralogs in the common ancestor to Retaria. Interestingly, none of
474 the α 2-tubulin and β 2-tubulin paralogs could be identified in available Endomyxa data, suggesting
475 that these gene duplications are synapomorphic for Retaria, with an origin after Retaria and
476 Endomyxa diverged (Figure 14).

477 Both the α 2- and β 2-tubulin genes form distinct phylogenetic clades and diverged from the
478 conventional α 1- and β 1-tubulin genes through changes in functionally important residues.
479 Subsequent to the initial duplication, these paralogs have undergone repeated duplications to
480 form subgroups of each gene family. These duplications and increased evolutionary rate have
481 developed both α 2- and β 2-tubulin as more divergent genes than the conventional α 1- and β 1-
482 tubulins.

483 Modelling of evolutionary rates on the tubulin structure shows global changes of the molecule
484 along two different paths: Firstly, a large number of conserved and functionally important residues
485 in α 1 and β 1 have become more variable, and probably therefore less functionally important in α 2
486 and β 2. This pattern is particularly clear at the interface between the α and β heterodimers (which
487 is the basic unit of protofilaments), and in the lateral surfaces between protofilaments that create
488 microtubule (i.e. the M-loop, the T3-loop and 8H helix etc.). Secondly, many variable sites localized
489 outside of the classical contact surfaces in the conventional α 1 and β 1, have become conserved in
490 α 2 and β 2 and have probably gained new functional roles. In addition, tracing the evolutionary
491 rates of all tubulin paralogs show higher evolutionary rate at the interface between the α and β
492 heterodimers (which are the basic units of protofilaments), and in the lateral surfaces between
493 protofilaments that create microtubules (i.e. the M-loop, the T3-loop and 8H helix etc.). The
494 overall pattern is that the new α 2-tubulin paralog presented here evolved with a similar mode to
495 that of the β 2-tubulin gene (Habura et al. 2005, Hou et al. 2013).

496 Retaria is unique among eukaryotes in having such divergent tubulin genes. It is not clear how
497 Retaria combines the four tubulin variants α 1, α 2, β 1 and β 2 into heterodimers, but it certainly
498 enables modularity. We hypothesize that Retaria can assemble four types (type 1-4) of

499 heterodimers; i.e. *type1*: $\alpha1+\beta1$, *type2*: $\alpha1+\beta2$, *type3*: $\alpha2+ \beta1$, *type4*: $\alpha2+ \beta2$. These four
500 heterodimers can function as modules and can be combined to develop protofilaments with
501 different properties. The different affinities between the α and β tubulins will certainly affect
502 assembly and disassembly of microtubules, and may be used to adjust flexibility, strength and
503 conformation of the axopodia or reticulopodia (Löwe et al. 2001). Retaria is known to develop
504 elaborate pseudopodia stiffened by bundles of microtubules. The axopodial microtubules in
505 Taxopodida and Acantharia attach laterally to other microtubules and form multiple hexagonal
506 rings (Cachon et al. 1973). The microtubules of polycystine radiolarians axopodia on the other
507 hand form a branching pattern (Cachon & Cachon 1971, Grain 1986). In some Foraminifera (e.g.
508 *Astrammina*) the microtubules coil tightly around one around another increasing the tensile
509 strength of the pseudopodia used to capture prey (Lee & Anderson 1991). Having several types of
510 α and β heterodimers allows a large repertoire of architectural structures to be drawn upon when
511 forming microtubules. In addition, we observe that many of the sites that have undergone
512 evolutionary change are located on the surface of the heterodimer. This can be linked to binding
513 sites for microtubule associated proteins (MAPs) as well as motor proteins further expanding the
514 range and flexibility of cytoskeletal structures (Brouhard & Rice 2014).

515

516 **Single cell transcriptomics for macroevolutionary studies of unculturable protists**

517 Here we have applied single cell transcriptomics as a new approach for phylogenomic
518 reconstruction of SAR, and the genetic basis of cytoskeleton and morphological evolution in
519 Rhizaria. Presently single cell transcriptomics has been applied to animal model organisms, such as
520 cell differentiation studies in mouse (Liang et al. 2014, Liu et al. 2014). To date, the only
521 application on protists has been on single cells grown from culture, confirming that the method
522 gives comparable results to that of sequencing many thousand cells (Kolisko et al. 2014). Although
523 single cell transcriptomic studies from cultured cells confirm the approach, they do not address
524 how efficient the method works on free-living protists, with highly divergent cell types, from
525 natural samples.

526

527 One of the main challenges of applying single cell transcriptomics to protists is the optimization of
528 cell lysis. This is emphasized when studying species with rigid skeletons and tough cell walls. Here
529 we modified lysis procedures for single cell transcriptomics (Picelli et al. 2014). Radiolaria species

530 have a tough cellular wall that protects the endoplasm; successful lysis demonstrates how the
531 method can be applied to less hardy unicellular species. The number of predicted genes from our
532 single cell transcriptomes are comparable to that generated from colonies, or pooling of hundreds
533 of cells from other Radiolarian species (Burki et al. 2010, Balzano et al. 2015). Subsampling of
534 sequence reads showed sufficient sequencing depth, suggesting that an incomplete transcriptome
535 was likely due to stochastic loss of mRNA. Despite these challenges we have shown that
536 transcriptomes of sufficient quality for phylogenomic and molecular evolutionary analyses can be
537 generated from single cells isolated from natural samples. This protocol can undoubtedly be
538 applied to other uncultivable protists, adding resolution to the relationships between eukaryotes,
539 in addition to revealing the evolution of morphologically related genes.

540

541 **Morphological diversification by lineage specific innovation of cytoskeleton genes in Rhizaria**

542 Data generated from these transcriptomes demonstrate that genetic innovation through multiple
543 gene duplication and neo-functionalization processes, rather than co-option of deep gene
544 homologs, have taken place in cytoskeletal genes of Chlorarachniophyta and Retaria. Differential
545 expansion of genes in chlorarachniophytes and Retaria show that underlying genetic changes to
546 cytoskeletal evolution have taken different routes in morphologically distinct groups; the overall
547 pattern of the data reveals extensive gene duplications of actin-related proteins in
548 chlorarachniophytes and of α - and β -tubulins in Retaria, with group specific expansions of myosin
549 in both groups (Figure 14). The hypothesized connection between the evolutionary changes to
550 cytoskeletal genes and the cellular morphology of the cells suggest that genetic innovations
551 occurred in the ancestor of the respective groups, subsequently forming the basis for
552 morphological and species diversification. While the actin-related proteins, and the myosin motor
553 proteins that use them have driven changes in chlorarachniophytes; tubulin has directed central
554 components of Retaria evolution. Subsequent to the initial innovation, additional expansions of
555 functional genes crucial to cytoskeletal formation have impacted on the morphological
556 diversification of Chlorarachniophyta and Retaria. Our analyses elucidate relationships between
557 genotype and phenotype of these organisms, linking gene evolution to evolution of cell
558 morphology. Better understanding of macroevolution in these organisms will require functional
559 studies of what types of actin branching the new Arps can form in chlorarachniophytes and how
560 Retaria combine the two sets of α and β tubulin proteins in their protofilaments. Such studies
561 should be complemented with more data from other gene families known to be involved in

562 cytoskeleton development, regulation, and transportation, such as MAPs, GTPases, dynein and
563 kinesin (Hammer & Wu 2002, Kollmar et al. 2012, Rojas et al. 2012, Brouhard & Rice 2014), .
564 Using the transcriptome data, we addressed uncertainties in the phylogeny of SAR by
565 phylogenomic analyses of supermatrices and by identification of synapomorphic gene
566 duplications. The phylogenomic analyses strongly support Rhizaria as sister to Stramenopiles and
567 Alveolates, and reveal Endomyxa and Taxopodida as two sister lineages to Foraminifera and
568 Radiolaria, with the latter two being divided into two distinct clades. Foraminifera is not placed
569 within Radiolaria as earlier reported (Krabberød et al. 2011, Sierra et al. 2013, 2015). In addition,
570 we identified independent synapomorphic gene duplications characters on several taxonomic
571 levels in Rhizaria, including a myosin family unique to Rhizaria. The actin-2 paralog of
572 Foraminifera, Radiolaria and Taxopodida is shared with Endomyxa, arguing for the grouping of
573 Endomyxa with Retaria, instead of being placed within Cercozoa. However, Retaria is divided from
574 Endomyxa by the Retaria-specific $\alpha 2$ and $\beta 2$ tubulin synapomorphies. The duplication of genes in
575 the Arp2/3 complex is a synapomorphy for Chlorarachniophyta. Altogether the phylogenomic
576 trees and synapomorphic gene-duplications shown here, form a new framework for future
577 revisions of the classification of SAR and Rhizaria. In total, we demonstrate single cell
578 transcriptomics as a promising approach for inclusion of a larger diversity of uncultured protists in
579 macroevolutionary studies and phylogenomic inferences.

580

581 **Methods:**

582 **Sampling and transcriptome amplification**

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

600

601 **Sequencing and assembly**

602 Library preparation and sequencing of the cDNA with Illumina MiSeq were performed at the
603 Natural History Museum in London. The sample was prepared using the Illumina TruSeq Nano
604 DNA LT Library Preparation Kit (FC-121-4001). The standard Illumina protocol was followed with
605 fragmentation on a Covaris M220 Focused-ultrasonicator. The finished library was quality checked
606 using an Agilent Tapestation to check the size of the library fragments, and a qPCR in a Corbett
607 RotorGene instrument to quantify the library. This was repeated for two MiSeq 600 cycle runs,
608 2*300 cycle paired end sequencing. The MiSeq platform was chosen over HiSeq since the longer
609 reads would provide an easier assembly when dealing with a possible metatranscriptomic library.
610 The raw reads (19,894,654 for *S. zanclea* and 11,590,658 for *L. setosa*) were quality filtered and

611 pairwise assembled with PEAR (Zhang et al. 2013) using default parameters. The reads were
612 further cleaned with Trimmomatic (Bolger et al. 2014) and then *de novo* assembled into contigs
613 with Trinity (Haas et al. 2013) using default settings. TransDecoder in the Trinity package was used
614 to predict genes from the assembled cDNA (Haas et al. 2013).

615

616 To check if all transcripts in the library had been sequenced, the raw reads were randomly split up
617 in 10 different datasets representing 10%, 20%, up to 90% of the original raw reads. The sub-
618 sampled datasets were assembled and new gene predictions were independently performed using
619 PEAR, Trimmomatic and Trinity as for the full dataset. Accumulation curves obtained by plotting
620 the predicted gene number against increasing partition size show that the slope of the curves
621 decrease with increasing partition size and more or less flattens when it reaches 100% of the total
622 dataset for both libraries (Figure 1 – figure supplement 1). We therefore assume that acceptable
623 sequencing depth for each library has been achieved, and that a further sequencing effort would
624 not have increased the number of predicted genes significantly.

625

626 **Alignment construction, paralog identification, and phylogenetic inference**

627 **The BIR pipeline:** We used the BIR pipeline (www.bioportal.no; Kumar *et al.*, 2015) to extract
628 genes and prepare single gene alignments to be used in multi-gene phylogenetic analyses. As seed
629 alignments for the BIR pipeline we used 258 genes previously published in multi-gene phylogenies
630 (Burki et al. 2012). As a query database we used the generated transcripts from our single cell
631 transcriptomes (6898 in total), all proteins in GenBank with Rhizaria as TaxID (44278 sequences at
632 the time of retrieval, October 2014), all 16 transcriptomes assigned to Rhizaria from the Marine
633 Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling et al., 2014. See table 1)
634 , as well as all Rhizarian sequences from Sierra et al., (2013). In addition seven reference genomes
635 are included in the BIR pipeline (*Arabidopsis thaliana*, *Bigeloviella natans*, *Dictyostelium*
636 *discoideum*, *Guillardia theta*, *Homo sapiens*, *Monosiga brevicollis*, *Naegleria gruberi*, *Paramecium*
637 *tetraurelia*, *Saccharomyces cerevisiae* and *Thalassiosira pseudonana* (Kumar et al. 2015). In short
638 the BIR pipeline will screen the query sequences against the database consisting of one or more
639 seed alignments, using BLAST, and assign the sequences that match the criteria set by the user to
640 the corresponding alignment (for details see Kumar et al., (2015)).

641

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

654

655 **Supermatrix construction:** After screening we were left with 255 genes that were concatenated
656 using ScaFos (Roure et al. 2007). We also merged close species into composite sequences when
657 they covered different parts of the supermatrix (see table S2). The final matrix had a length of
658 54,898 amino acids with 124 taxa.

659

660 **Removal of jumping and long branched taxa:** *Mikrocytos mackini* was not included in the analysis
661 due to an extremely long branch (Burki et al. 2013) and RogueNaRok, using default parameters
662 (Aberer et al. 2013) was used to identify jumping taxa, which also were excluded from further
663 analysis (Supplementary table S2)

664

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

670

671 **Missing data:** To assess the impact of missing data we excluded taxa with low coverage in
672 increments from the two concatenated dataset. First we set the lowest allowed percentage of

673 missing data for a taxon to be 10% of the total characters (i.e. if a taxon had more than 90% data
674 missing it was excluded), the next cut-off at 20%, and finally at 30%. The number of characters in
675 the matrix was held constant. The Tree Certainty index (Salichos et al. 2014) was calculated for
676 each increment see supplementary table S2 and S3 for details. The relative Tree Certainty index
677 increased markedly when the threshold was set at 90%, but did not increase significantly after
678 that, in fact there seem to be a decrease in the relative TC value as the number of taxa drops
679 (Supplementary table S1).

680 **Influence of taxa with low coverage, or uncertain position:** we also remove taxa and clades from
681 Rhizaria that had a consistently low bootstrap value (< 75%) or low posterior probability (< 75 pp),
682 but that had not been flagged by RogueNaRok to see if they affected the topology of the
683 phylogenetic inference. *Spongosphaera streptacantha* and *Sticholonche zanclea* were removed
684 one by one and together from both the full and the reduced dataset. Foraminifera were also
685 removed in analyses (see table S3).

686 **Removal of fast evolving sites:** TIGER (Cummins & McInerney 2011) was used with default
687 settings to produce categories of fast evolving sites, 10 categories in total for each dataset.
688 Categories of fast evolving sites were removed in increments, starting with the category with the
689 fastest evolving sites, subsequently removing the category with the second fastest evolving sites
690 etc. Up to 4 categories were removed from all datasets before phylogenetic analyses.

691 **Phylogenetic analyses:** Phylogenetic trees were inferred for all concatenated datasets, with
692 RAxML choosing the best fitting model, and with the automatic bootstrapping criteria as
693 previously described. The preferred model was always LG+ Γ (see supplementary table S3). Due to
694 the heavy demand on computational resources from Bayesian inference only six of the alignments
695 were included for analysis with the CATGTR model in Phylobayes MPI version 1.5a (Lartillot et al.
696 2013), as well as 1 dataset with the LG model. For these we ran 2 chains in parallel for at least
697 15.000 iteration only stopping when the maxdiff was >0.3 (see supplementary table S3).

698 **Evolutionary Placement Algorithm:** In order to place rhizarian species that had been excluded
699 when the cut-off threshold for missing data had been raised on the phylogenetic tree we used the
700 Evolutionary Placement Algorithm (EPA) included in RAxML 8.0.26 (Stamatakis et al. 2010, Berger
701 et al. 2011, Stamatakis 2014). As reference tree we used the 255-gene maximum likelihood tree
702 with a 10% missing data cut off.

703 **Genes related to cytoskeleton formation and motor proteins**

704 The assembled transcriptomes from the single cells were annotated with InterProscan 5 (Jones et
705 al. 2014) as implemented in Geneious 8 (Kearse et al. 2012). The annotations were screened for
706 genes commonly involved in the formation and development of the cytoskeleton, as well as the
707 most common motor proteins using the cytoskeleton. In particular we looked for α - and β -tubulin,
708 myosin, actin, the actin regulating Arp2/3-complex consisting of seven actin-related proteins
709 (arp2, arp3, ARPC1, ARPC2, ARPC3, ARPC4 and ARPC5). Reference alignments and sequences were
710 downloaded from PFAM (<http://pfam.xfam.org/>), as well as relevant other recently published
711 alignments (Hou et al. 2013, Seb e-Pedr os et al. 2014, Cavalier-Smith et al. 2015) and used in BIR
712 as seed alignments with the same query database as before. In addition representatives for all the
713 genes were blasted against 6 additional non-rhizarian transcriptomes from MMETSP
714 (MMETSP0039 *Eutreptiella gymnastica*, MMETSP0046 *Guillardia theta*, MMETSP0308 *Gloeochaete*
715 *wittrockiana*, MMETSP0380 *Alexandrium tamarense*, MMETSP0902 *Thalassiosira Antarctica*, and
716 MMETSP1150 *Emiliana huxleyi*), as well as against the non-redundant protein database in
717 GenBank. For each gene ML trees were constructed with RAxML as before, and manually curated
718 for any confounding artefacts. Redundant and short sequences were manually removed in
719 Geneious 8 (Kearse et al. 2012) before another round of ML analysis with RAxML and a Bayesian
720 analysis with the CATGTR model implemented in Phylobayes MPI version 1.5a (Lartillot et al.
721 2013). Comparative evolutionary analyses of tubulin and the duplicated genes in the Arp2/3
722 complex were performed by examining the evolutionary rates of the paralogs separately and then
723 mapping the genes to structural models using Consurf (Ashkenazy et al. 2010, Celniker et al. 2013).
724 InterPro annotations of functional domains of myosin was performed with InterProscan 5 (Jones
725 et al. 2014, Mitchell et al. 2015).

726

727 All sequences from *S. zanclea* and *L. setosa* used in this study have been deposited in GenBank
728 with accession numbers (xxxx-xxxx). All data, alignments, and trees can be downloaded from
729 www.bioportal.no.

730 **Acknowledgment**

731 We would like to thank Bente Edvardsen, UiO, for providing the plankton haul from which the
732 sampled cells were isolated, the Sequencing centre at Natural History Museum in London for

733 performing the Illumina library preparations and sequencing. We would also like to thank Simon
734 Picelli for answering questions related to the Smart-seq2 method, Fabien Burki for providing single
735 gene alignments and The Gordon and Betty Moore Foundation for making all those wonderful
736 protists transcriptomes available for the public. All analyses were run either on the Abel
737 supercomputer at The High Performance Computing cluster at University Of Oslo, or on Lifeportal
738 (www.lifeportal.uio.no). For more information on BIR see www.bioportal.no. This project was
739 funded by University of Oslo. This work was supported by grants from University of Oslo and from
740 Research Council of Norway to Shalchian-Tabrizi (NFR216475).

741

742 **References**

- 743 Aberer AJ, Krompass D, Stamatakis A (2013) Pruning rogue taxa improves phylogenetic accuracy: An efficient
744 algorithm and webservice. *Syst Biol* 62:162–166
- 745 Anderson OR (1976a) Fine structure of a collodarian radiolarian (*Sphaerozoum punctatum* Müller 1858) and
746 cytoplasmic changes during reproduction. *Mar Micropaleontol* 1:287–297
- 747 Anderson OR (1976b) Ultrastructure of a colonial radiolarian *collozoum inerme* and a cytochemical determination of
748 the role of its zooxanthellae. *Tissue Cell* 8:195–208
- 749 Anderson OR (1978) Light and electron microscopic observations of feeding behavior, nutrition, and reproduction in
750 laboratory cultures of *Thalassicolla nucleata*. *Tissue Cell* 10:401–12
- 751 Anderson OR (1983) *Radiolaria*, 2nd edn. Springer-Verlag, New York, USA
- 752 Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: Calculating evolutionary conservation in
753 sequence and structure of proteins and nucleic acids. *Nucleic Acids Res* 38
- 754 Balzano S, Corre E, Decelle J, Sierra R, Wincker P, Silva C, Poulain J, Pawlowski J, Not F (2015) Transcriptome analyses
755 to investigate symbiotic relationships between marine protists. *Front Microbiol* 6:98: 1–14
- 756 Bass D, Chao EE-Y, Nikolaev S, Yabuki A, Ishida K-I, Berney C, Pakzad U, Wylezich C, Cavalier-Smith T (2009) Phylogeny
757 of novel naked Filose and Reticulose Cercozoa: Granofilosea cl. n. and Proteomyxidea revised. *Protist* 160:75–
758 109
- 759 Berger S a., Krompass D, Stamatakis A (2011) Performance, accuracy, and Web server for evolutionary placement of
760 short sequence reads under maximum likelihood. *Syst Biol* 60:291–302
- 761 Berger SA, Stamatakis A (2011) Aligning short reads to reference alignments and trees. *Bioinformatics* 27:2068–75
- 762 Boczkowska M, Rebowski G, Kast DJ, Dominguez R (2014) Structural analysis of the transitional state of Arp2/3
763 complex activation by two actin-bound WCAs. *Nat Commun* 5:3308
- 764 Boczkowska M, Rebowski G, Petoukhov M V., Hayes DB, Svergun DI, Dominguez R (2008) X-ray scattering study of
765 activated Arp2/3 complex with bound actin-WCA. *Structure* 16:695–704
- 766 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*
767 30:2114–2120
- 768 Bowser SS (2002) Reticulopodia: Structural and Behavioral Basis for the Suprageneric Placement of Granuloreticulosan
769 Protists. *J Foraminifer Res* 32:440–447
- 770 Brouhard GJ, Rice LM (2014) The contribution of $\alpha\beta$ -tubulin curvature to microtubule dynamics. *J Cell Biol* 207:323–
771 334
- 772 Brown MW, Kolisko M, Silberman JD, Roger AJ (2012) Aggregative multicellularity evolved independently in the
773 eukaryotic supergroup Rhizaria. *Curr Biol* 22:1123–1127
- 774 Burki F, Corradi N, Sierra R, Pawlowski J, Meyer GR, Abbott CL, Keeling PJ (2013) Phylogenomics of the intracellular

- 775 parasite *Mikrocytos mackini* reveals evidence for a mitosome in Rhizaria. *Curr Biol* 23:1–7
- 776 Burki F, Kaplan M, Tikhonenkov D V, Zlatogursky V, Minh BQ, Radaykina L V, Smirnov A, Mylnikov P, Keeling PJ, Keeling
777 PJ (2016) Untangling the early diversification of eukaryotes : a phylogenomic study of the evolutionary origins of
778 Centrohelida , Haptophyta and Cryptista. :1–10
- 779 Burki F, Keeling PJ (2014) Rhizaria. *Curr Biol* 24:R103–R107
- 780 Burki F, Kudryavtsev A, Matz M V, Aglyamova G V, Bulman S, Fiers M, Keeling PJ, Pawlowski J (2010) Evolution of
781 Rhizaria: new insights from phylogenomic analysis of uncultivated protists. *BMC Evol Biol* 10:377
- 782 Burki F, Okamoto N, Pombert J-F, Keeling PJ (2012) The evolutionary history of haptophytes and cryptophytes:
783 phylogenomic evidence for separate origins. *Proc R Soc B Biol Sci* 279:2246–2254
- 784 Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J (2007) Phylogenomics
785 reshuffles the eukaryotic supergroups. *PLoS One* 2:e790
- 786 Burki F, Shalchian-Tabrizi K, Pawlowski J (2008) Phylogenomics reveals a new “megagroup” including most
787 photosynthetic eukaryotes. *Biol Lett* 4:366–9
- 788 Cachon J, Cachon M (1971) Le system axopodial des Radiolaires Nassellaires. *Arch für Protistenkd* 113:80–97
- 789 Cachon J, Cachon M, Febvre-Chevalier C, Febvre J (1973) Determinisme de l’edification des systemes microtubulaires
790 stereoplasmiques d’Actinopodes. *Arch für Protistenkd* 115:137–153
- 791 Cachon J, Cachon M, Tilney LG, Tilney MS (1977) Movement generated by interactions between the dense material at
792 the ends of microtubules and non-actin-containing microfilaments in *Sticholonche zancolea*. *J Cell Biol* 72:314–
793 338
- 794 Cavalier-Smith T (1993) Kingdom Protozoa and its 18 phyla. *Microbiol Rev* 57:953–994
- 795 Cavalier-Smith T (2002) The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst*
796 *Evol Microbiol* 52:297–354
- 797 Cavalier-Smith T (2009) Megaphylogeny, cell body plans, adaptive zones: causes and timing of eukaryote basal
798 radiations. *J Eukaryot Microbiol* 56:26–33
- 799 Cavalier-Smith T, Chao EE, Lewis R (2015) Multiple origins of Heliozoa from flagellate ancestors: New cryptist
800 subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista.
801 *Mol Phylogenet Evol* 93:331–362
- 802 Cavalier-Smith T, Guy L, Saw JH, Ettema TJG, Eme L, Sharpe SC, Brown MW, Irimia M, Roy SW (2014) The neomuran
803 revolution and phagotrophic origin of eukaryotes and cilia in the light of intracellular coevolution and a revised
804 Tree of Life. *Cold Spring Harb Perspect Biol* 6:a016006–a016006
- 805 Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, Mayrose I, Pupko T, Ben-Tal N (2013) ConSurf: Using
806 evolutionary data to raise testable hypotheses about protein function. *Isr J Chem* 53:199–206
- 807 Cummins C a, Mclnerney JO (2011) A method for inferring the rate of evolution of homologous characters that can
808 potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Syst Biol*
809 60:833–844

- 810 Dustin P (1984) *Microtubules*, 2nd edn. Springer-Verlag, Berlin
- 811 Febvre J (1981) The myoneme of the Acantharia (Protozoa): A new model of cellular motility. *Biosystems* 14:327–336
- 812 Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J,
813 Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. *Nucleic Acids Res* 42:D222–30
- 814 Giannone G, Dubin-Thaler BJ, Rossier O, Cai Y, Chaga O, Jiang G, Beaver W, Döbereiner H-G, Freund Y, Borisy G, Sheetz
815 MP (2007) Lamellipodial actin mechanically links myosin activity with adhesion-site formation. *Cell* 128:561–75
- 816 Goldstein LS (2001) Molecular motors: from one motor many tails to one motor many tales. *Trends Cell Biol* 11:477–
817 82
- 818 Goley ED, Welch MD (2006) The ARP2/3 complex: an actin nucleator comes of age. *Nat Rev Mol Cell Biol* 7:713–726
- 819 Grain J (1986) The cytoskeleton in protists: nature, structure, and functions. *Int Rev Cytol* 104:153–249
- 820 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes
821 MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD,
822 Friedman N, Regev A (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity
823 platform for reference generation and analysis. *Nat Protoc* 8:1494–1512
- 824 Habura A, Wegener L, Travis JL, Bowser SS (2005) Structural and functional implications of an unusual foraminiferal
825 beta-tubulin. *Mol Biol Evol* 22:2000–2009
- 826 Hammer J a., Wu XS (2002) Rabs grab motors: Defining the connections between Rab GTPases and motor proteins.
827 *Curr Opin Cell Biol* 14:69–75
- 828 He D, Sierra R, Pawlowski J, Baldauf SL (2016) Reducing long-branch effects in multi-protein data uncovers a close
829 relationship between Alveolata and Rhizaria. *Mol Phylogenet Evol*
- 830 Hou Y, Sierra R, Bassen D, Banavali NK, Habura A, Pawlowski J, Bowser SS (2013) Molecular evidence for β -tubulin
831 neofunctionalization in Retaria (Foraminifera and Radiolarians). *Mol Biol Evol* 30:2487–2493
- 832 Ishida KI, Yabuki A, Ota S (2011) Research note: *Amorphochlora amoebiformis* gen. et comb. nov.
833 (*Chlorarachniophyceae*). *Phycol Res* 59:52–53
- 834 Ishitani Y, Ishikawa S, Inagaki Y, Tsuchiya M, Takishita K (2011) Multigene phylogenetic analyses including diverse
835 radiolarian species support the “Retaria” hypothesis - the sister relationship of Radiolaria and Foraminifera. *Mar*
836 *Micropaleontol* 81:32–42
- 837 Jékely G (ed) (2007) *Eukaryotic Membranes and Cytoskeleton*. Springer New York, New York, NY
- 838 Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn
839 AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: Genome-scale
840 protein function classification. *Bioinformatics* 30:1236–1240
- 841 Karcher RL, Deacon SW, Gelfand VI (2002) Motor-cargo interactions: The key to transport specificity. *Trends Cell Biol*
842 12:21–27
- 843 Kast DJ, Zajac AL, Holzbaaur ELF, Ostap EM, Dominguez Correspondence R, Dominguez R (2015) WHAMM Directs the
844 Arp2/3 Complex to the ER for Autophagosome Biogenesis through an Actin Comet Tail Mechanism. *Curr Biol*

- 845 25:1791–1797
- 846 Katz L a (2012) Origin and Diversification of Eukaryotes. *Annu Rev Microbiol*:411–427
- 847 Katz LA, Grant JR (2014) Taxon-rich phylogenomic analyses resolve the eukaryotic Tree of Life and reveal the power of
848 subsampling by sites. *Syst Biol* 64:406–415
- 849 Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer
850 T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: An integrated and extendable desktop software
851 platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649
- 852 Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral Zettler L, Armbrust EV, Archibald JM, Bharti AK, Bell CJ,
853 Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron D a., Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps
854 P, Dyhrman ST, Edvardsen B, Gates RD, Gobler CJ, Greenwood SJ, Guida SM, Jacobi JL, Jakobsen KS, James ER,
855 Jenkins B, John U, Johnson MD, Juhl AR, Kamp A, Katz L a., Kiene R, Kudryavtsev A, Leander BS, Lin S, Lovejoy C,
856 Lynn D, Marchetti A, McManus G, Nedelcu AM, Menden-Deuer S, Miceli C, Mock T, Montresor M, Moran MA,
857 Murray S, Nadathur G, Nagai S, Ngam PB, Palenik B, Pawlowski J, Petroni G, Piganeau G, Posewitz MC, Rengefors
858 K, Romano G, Rumpho ME, Ryneerson T, Schilling KB, Schroeder DC, Simpson AGB, Slamovits CH, Smith DR,
859 Smith GJ, Smith SR, Sosik HM, Stief P, Theriot E, Twary SN, Umale PE, Vaultot D, Wawrik B, Wheeler GL, Wilson
860 WH, Xu Y, Zingone A, Worden AZ (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project
861 (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome
862 sequencing (RG Roberts, Ed.). *PLoS Biol* 12:e1001889
- 863 Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ (2014) Single-cell transcriptomics for microbial eukaryotes. *Curr Biol*
864 24:R1081–R1082
- 865 Kollmar M, Lbik D, Enge S (2012) Evolution of the eukaryotic ARP2/3 activators of the WASP family: WASP, WAVE,
866 WASH, and WHAMM, and the proposed new family members WAWH and WAML. *BMC Res Notes* 5:88
- 867 Krabberød AK, Bråte J, Dolven JK, Ose RF, Klaveness D, Kristensen T, Bjørklund KR, Shalchian-Tabrizi K (2011)
868 Radiolaria Divided into Polycystina and Spasmaria in Combined 18S and 28S rDNA Phylogeny. *PLoS One*
869 6:e23526
- 870 Kumar S, Krabberød AK, Neumann RS, Michalickova K, Zhang X, Zhao S, Shalchian-Tabrizi K (2015) BIR Pipeline for
871 Preparation of Phylogenomic Data. *Evol Bioinforma* 11:79–83
- 872 Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement
873 process. *Mol Biol Evol* 21:1095–1109
- 874 Lartillot N, Rodrigue N, Stubbs D, Richer J (2013) Phylobayes mpi: Phylogenetic reconstruction with infinite mixtures of
875 profiles in a parallel environment. *Syst Biol* 62:611–615
- 876 Lee JJ, Anderson OR (Eds) (1991) *Biology of Foraminifera*. Academic Press, London, UK, London
- 877 Liang J, Cai W, Sun Z (2014) Single-Cell sequencing technologies: current and future. *J Genet Genomics* 41:513–528
- 878 Liu N, Liu L, Pan X (2014) Single-cell analysis of the transcriptome and its application in the characterization of stem
879 cells and early embryos. *Cell Mol Life Sci* 71:2707–2715
- 880 Löwe J, Li H, Downing KH, Nogales E (2001) Refined structure of alpha beta-tubulin at 3.5 Å resolution. *J Mol Biol*

- 881 313:1045–1057
- 882 Margulis L (1990) Handbook of protozoa : the structure, cultivation, habitats, and life histories of the eukaryotic
883 microorganisms and their descendants exclusive of animals, plants and fungi : a guide to the algae, ciliates,
884 foraminifera, sporozoa, water molds, slime m (L Margulis, Ed.). Jones and Bartlett Publishers, Boston
- 885 Mattila PK, Lappalainen P (2008) Filopodia: molecular architecture and cellular functions. *Nat Rev Mol Cell Biol* 9:446–
886 454
- 887 Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S,
888 Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJA, Redaschi
889 N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu
890 CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD (2015) The InterPro protein families database: The
891 classification resource after 15 years. *Nucleic Acids Res* 43:D213–D221
- 892 Mogilner A, Keren K (2009) The Shape of Motile Cells. *Curr Biol* 19:R762–R771
- 893 Moreira D, Heyden S von der, Bass D, López-García P, Chao E, Cavalier-Smith T (2007) Global eukaryote phylogeny:
894 Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata.
895 *Mol Phylogenet Evol* 44:255–266
- 896 Nikolaev SI, Berney C, Fahrni JF, Bolivar I, Polet S, Mylnikov AP, Aleshin V V, Petrov NB, Pawlowski J (2004) The twilight
897 of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. *Proc Natl Acad Sci U S A*
898 101:8066–8071
- 899 Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morrison HG, Sogin ML, Patterson DJ, Katz L a (2010) Broadly
900 sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst Biol* 59:518–533
- 901 Pattengale ND, Swenson KM, Moret BME (2010) Uncovering hidden phylogenetic consensus. In: *Lecture Notes in*
902 *Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*
903 *Bioinformatics)*.p 128–139
- 904 Pawlowski J (2008) The twilight of Sarcodina: a molecular perspective on the polyphyletic origin of amoeboid protists.
905 *Protistology* 5:281–302
- 906 Pawlowski J, Burki F (2009) Untangling the phylogeny of amoeboid protists. *J Eukaryot Microbiol* 56:16–25
- 907 Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005) Heterotachy and long-branch attraction in
908 phylogenetics. *BMC Evol Biol* 5:50
- 909 Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R (2014) Full-length RNA-seq from single cells
910 using Smart-seq2. *Nat Protoc* 9:171–181
- 911 Pollard TD (2007) Regulation of actin filament assembly by Arp2/3 complex and formins. *Annu Rev Biophys Biomol*
912 *Struct* 36:451–477
- 913 Richards T a, Cavalier-Smith T (2005) Myosin domain evolution and the primary divergence of eukaryotes. *Nature*
914 436:1113–8
- 915 Rojas AM, Fuentes G, Rausell A, Valencia A (2012) The Ras protein superfamily: Evolutionary tree and role of
916 conserved amino acids. *J Cell Biol* 196:189–201

- 917 Rouiller I, Xu XP, Amann KJ, Egile C, Nickell S, Nicastro D, Li R, Pollard TD, Volkman N, Hanein D (2008) The structural
918 basis of actin filament branching by the Arp2/3 complex. *J Cell Biol* 180:887–895
- 919 Roure B, Rodriguez-Ezpeleta N, Philippe H (2007) SCaFoS: a tool for selection, concatenation and fusion of sequences
920 for phylogenomics. *BMC Evol Biol* 7 Suppl 1:S2
- 921 Salichos L, Stamatakis A, Rokas A (2014) Novel information theory-based measures for quantifying incongruence
922 among phylogenetic trees. *Mol Biol Evol* 31:1261–1271
- 923 Schliwa M, Woehlke G (2003) Molecular motors. *Nature* 422:759–765
- 924 Seabra MC, Coudrier E (2004) Rab GTPases and myosin motors in organelle motility. *Traffic* 5:393–9
- 925 Sebé-Pedrós A, Grau-Bové X, Richards TA, Ruiz-Trillo I (2014) Evolution and classification of myosins, a paneukaryotic
926 whole-genome approach. *Genome Biol Evol* 6:290–305
- 927 Sierra R, Cañas-Duarte SJ, Burki F, Schwelm A, Fogelqvist J, Dixelius C, González-García LN, Gile GH, Slamovits CH,
928 Klopp C, Restrepo S, Arzul I, Pawlowski J (2015) Evolutionary origins of rhizarian parasites. *Mol Biol Evol*
929 33:msv340–
- 930 Sierra R, Matz M V., Aglyamova G, Pillet L, Decelle J, Not F, Vargas C de, Pawlowski J (2013) Deep relationships of
931 Rhizaria revealed by phylogenomics: A farewell to Haeckel’s Radiolaria. *Mol Phylogenet Evol* 67:53–59
- 932 Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies.
933 *Bioinformatics* 30:1312–1313
- 934 Stamatakis A, Komornik Z, Berger SA (2010) Evolutionary placement of short sequence reads on multi-core
935 architectures. In: 2010 ACS/IEEE International Conference on Computer Systems and Applications, AICCSA
936 2010.p 1–44
- 937 Sugiyama K, Hori RS, Kusunoki Y, Matsuoka A (2008) Pseudopodial features and feeding behavior of living
938 nassellarians *Eucyrtidium hexagonatum* Haeckel, *Pterocorys zancleus* (Müller) and *Dictyocodon prometheus*
939 Haeckel. *Paleontol Res* 12:209–222
- 940 Suzuki NO, Aita YO (2011) Radiolaria: achievements and unresolved issues: taxonomy and cytology. *Plankt Benthos*
941 *Res* 6:69–91
- 942 Townsend JP (2007) Profiling phylogenetic informativeness. *Syst Biol* 56:222–231
- 943 Townsend JP, Su Z, Tekle YI, Ownsend JEPT, Huo ZSU (2012) Phylogenetic signal and noise: predicting the power of a
944 data set to resolve phylogeny. *Syst Biol* 61:835–49
- 945 Travis JL, Allen RD (1981) Studies on the Motility of the Foraminifera .1. Ultrastructure of the Reticulopodial Network
946 of *Allogromia-Laticollaris* (Arnold). *J Cell Biol* 90:211–221
- 947 Travis JL, Bowser SS (1986) A new model of reticulopodial motility and shape: evidence for a microtubule-based motor
948 and an actin skeleton. *Cell Motil Cytoskeleton* 6:2–14
- 949 Ura S, Pollitt AY, Veltman DM, Morrice N a., MacHesky LM, Insall RH (2012) Pseudopod growth and evolution during
950 cell movement is controlled through SCAR/WAVE dephosphorylation. *Curr Biol* 22:553–561
- 951 Vale RD (2003) The molecular motor toolbox for intracellular transport. *Cell* 112:467–480

- 952 Volkmann N, Amann KJ, Stoilova-McPhie S, Egile C, Winter DC, Hazelwood L, Heuser JE, Li R, Pollard TD, Hanein D
953 (2001) Structure of Arp2/3 complex in its activated state and in actin filament branch junctions. *Science*
954 293:2456–2459
- 955 Welnhof E a, Travis JL (1998) Evidence for a direct conversion between two tubulin polymers--microtubules and
956 helical filaments--in the foraminiferan, *Allogromia laticollaris*. *Cell Motil Cytoskeleton* 41:107–116
- 957 Wickstead B, Gull K (2011) The evolution of the cytoskeleton. *J Cell Biol* 194:513–525
- 958 Xu X-P, Rouiller I, Slaughter BD, Egile C, Kim E, Unruh JR, Fan X, Pollard TD, Li R, Hanein D, Volkmann N (2011) Three-
959 dimensional reconstructions of Arp2/3 complex with bound nucleation promoting factors. *EMBO J* 31:236–247
- 960 Zhang J, Kobert K, Flouri T, Stamatakis A (2013) PEAR: A fast and accurate Illumina Paired-End reAd mergeR.
961 *Bioinformatics* 30:1–7
- 962
- 963

964 **Table 1:** Single cell transcriptome statistics

Species name	Raw Reads	Contigs ¹	GC-content (%)	Predicted Genes ²
<i>Sticholonche zanclea</i>	19,894,654	19,509	53.5	4,749
<i>Lithomelissa setosa</i>	11,590,658	12,212	48.8	2,122

965 ¹Number of contigs assembled by Trinity (Haas et al. 2013). ²The number of genes predicted by TransDecoder in the
 966 Trinity platform.

967

968 **Table 2:** Rhizarian transcriptomes from MMETSP (Keeling et al. 2014) used in this study.

Sample Id	Phylum	Species	Strain	Transcripts ¹
MMETSP0040	Chlorarachniophyta	<i>Lotharella oceanica</i>	CCMP622	17,354
MMETSP0041	Chlorarachniophyta	<i>Lotharella globosa</i>	LEX01	25,644
MMETSP0042	Chlorarachniophyta	<i>Amorphochlora amoebiformis</i> *	CCMP2058	23,387
MMETSP0045	Chlorarachniophyta	<i>Bigelowiella natans</i>	CCMP 2755	22,651
MMETSP0109	Chlorarachniophyta	<i>Chlorarachnion reptans</i>	CCCM449	26,481
MMETSP0110	Chlorarachniophyta	<i>Gymnochlora</i> sp.	CCMP2014	15,507
MMETSP0111	Chlorarachniophyta	<i>Lotharella globosa</i>	CCCM811	19,670
MMETSP0112	Chlorarachniophyta	<i>Lotharella globosa</i>	CCCM811	11,910
MMETSP0113	Chlorarachniophyta	<i>Norrisiella sphaerica</i>	BC52	14,550
MMETSP0186	Cercozoa	<i>Minchinia chitonis</i>	Missing	461
MMETSP1052	Chlorarachniophyta	<i>Bigelowiella natans</i>	CCMP623	24,186
MMETSP1318	Chlorarachniophyta	<i>Partenskyella glossopodia</i>	RCC365	15,025
MMETSP1358	Chlorarachniophyta	<i>Bigelowiella natans</i>	CCMP1242	18,273
MMETSP1359	Chlorarachniophyta	<i>Bigelowiella longifila</i>	CCMP242	15,959
MMETSP1384	Foraminifera	<i>Ammonia</i> sp.	Missing	31,225
MMETSP1385	Foraminifera	<i>Elphidium margaritaceum</i>	Missing	25,184

969 ¹The number of amino acid sequences for the sample. **Amorphochlora amoebiformis* is called *Lotharella*
 970 *amoebiformis* in MMETSP, but it was moved to the genus *Amorphochlora* by Ishida et al., (2011)

971

972 **Figure Legends**

973 **Figure 1:** The two specimens sequenced. A) *Lithomelissa setosa*, B) *Sticholonche zanclea*. Scale bar
974 50 μm .

975 **Figure 1 – figure supplement 1:** Gene accumulation curve. The number of predicted genes is
976 plotted against subsamples of the original dataset. Each subsample is independently assembled
977 before gene prediction.

978

979 **Figure 2:** Bayesian phylogeny with the CATGTR model, 255 genes, 54,898 AA, and 91 taxa, maxdiff
980 0.2666. Species sequenced for this paper in bold. Thick branches represent maximal support
981 (posterior probability = 1). Number after '@' is concatenated sequence length. Important clades in
982 Rhizaria are coloured for easier identification: brown = Foraminifera, dark red = Taxopodida, red =
983 Radiolaria, yellow = Endomyxa, blue= Chlorarachniophyta (Filosa), and green = Mondaofilosa
984 (Filosa). The scale bar equals the mean number of substitutions per site. The changing support
985 values for selected branches depending on the number of genes, and the number of fast evolving
986 sites removed are shown in Figure 2 – figure supplement 1.

987 **Figure 2 – figure supplement 1:** Influence of number of genes, and fast evolving sites on the
988 Bayesian analysis using the CATGTR model, with the exception of the dataset 146 (LG) where the
989 LG model was used. See text for discussion. **Number of genes** is the number of genes used in the
990 concatenated data set. **Number of bins removed** equals the number of bins of fast evolving sites
991 removed by TIGER (Cummins & McInerney 2011). The numbers in the matrix represent posterior
992 probability for the branch marked with an asterisk. (A) The internal relationship in SAR, the first
993 tree represents the monophyly of Stramenopiles and Alveolates, with Rhizaria as sister. The
994 second tree represents the monophyly of Alveolates and Rhizaria, with Stramenopiles as sister. (B)
995 The placement of Taxopodida: The first tree is the support value for Taxopodida as sister to
996 Radiolarians, with Foraminifera as outgroup. In the second tree Taxopodida is basal in Retaria with
997 the values showing support for the monophyly of Radiolaria and Foraminifera, excluding
998 Taxopodida.

999

1000 **Figure 3:** Maximum likelihood with the LG model, 255 genes, 54,898 AA, and 91 taxa. Species
1001 sequenced for this paper in bold. Thick branches represent maximal support (bootstrap = 100 %).
1002 Number after '@' is concatenated sequence length. As in figure 2 important clades in Rhizaria are

1003 coloured for easier identification: brown = Foraminifera, dark red = Taxopodida, red=Radiolaria,
1004 yellow = Endomyxa, blue= Chlorarachniophyta (Filosa), and green = Mondaofilosa (Filosa). The
1005 scale bar equals the mean number of substitutions per site. The changing support values for
1006 selected branches depending on the number of genes, and the number of fast evolving sites
1007 removed are shown in Figure 3 – figure supplement 1.

1008 **Figure 3 – figure supplement 1:** Influence of number of genes, and fast evolving sites on the ML
1009 analysis. **Number of genes** is the number of genes used in the concatenated data set. **Number of**
1010 **bins removed** equals the number of bins of fast evolving sites removed by TIGER (Cummins &
1011 McInerney 2011). The numbers in the matrix represent ML bootstrap values for the branch
1012 marked with an asterisk. (A) The internal relationship in SAR, the first tree represents the
1013 monophyly of Stramenopiles and Alveolates, with Rhizaria as sister. The second tree represents
1014 the monophyly of Alveolates and Rhizaria, with Stramenopiles as sister. (B) The placement of
1015 Taxopodida: The first tree is the support value for Taxopodida as sister to Radiolarians, with
1016 Foraminifera as outgroup. In the second tree Taxopodida is basal in Retaria with the values
1017 showing support for the monophyly of Radiolaria and Foraminifera, excluding Taxopodida.

1018

1019 **Figure 4:** Actin phylogeny (229 taxa, 374 AA). Thick branches represents bootstrap > 75% and
1020 posterior probability > 0.9. Some branches are collapsed to save space. Support values for selected
1021 nodes discussed in the text added for clarity. The scale bar equals the mean number of
1022 substitutions per site. The colouring scheme is the same as in figure 2. (Brown = Foraminifera, red=
1023 Radiolaria/Taxopodida, yellow = Endomyxa and blue= Filosa)

1024

1025 **Figure 5:** Phylogenies of the seven genes in the Arp2/3 complex. *Arp2* (39 taxa, 373 AA), *Arp3* (33
1026 taxa, 403 AA), *ARPC1* (34 taxa, 328 AA), *ARPC2* (24 taxa, 303 AA), *ARPC3* (30 taxa, 181 AA), *ARPC4*
1027 (23 taxa, 169 AA) and *ARPC5* (29 taxa, 151 AA). Colouring of groups as in figure 2 (Brown =
1028 Foraminifera, red= Radiolaria, yellow = Endomyxa and blue= Filosa). Thick branches represents
1029 bootstrap > 75% and posterior probability > 0.9 and the scale bar equals the mean number of
1030 substitutions per site. (A) The two genes with a recent duplication in Chlorarachniophyta (*Arp2*
1031 and *ARPC1*). (B) The five genes without duplication in Chlorarachniophyta (*ARP3*, *ARPC2*, *ARPC3*,
1032 *ARPC4*, *ARPC5*).

1033

1034 **Figure 6:** Molecular models of the Arp2a and Arp2b paralogs in Chlorarachniophyta with
1035 evolutionary rates from Consurf superimposed on the Arp2/3 complex (PDB accession 4JD2; Arp2
1036 red, Arp3 orange, ARPC1 green, ARPC2 cyan, ARPC3 pink, ARPC4 blue, ARPC5 yellow). Residues are
1037 coloured according to the evolutionary rates calculated by Consurf. Turquoise residues are highly
1038 variable and maroon means conserved residues.

1039

1040 **Figure 7:** Molecular models of the ARPC1a and ARPC1b paralogs in Chlorarachniophyta with
1041 evolutionary rates from Consurf superimposed on the Arp2/3 complex (PDB accession 4JD2; Arp2
1042 red, Arp3 orange, ARPC1 green, ARPC2 cyan, ARPC3 pink, ARPC4 blue, ARPC5 yellow). Residues are
1043 coloured according to the evolutionary rates calculated by Consurf. Turquoise residues are highly
1044 variable and maroon means conserved residues.

1045

1046 Figure 8: Comparison of conserved residues between the two paralogs of Arp2 and the two
1047 paralogs of ARPC1 in Chlorarachniophytes superimposed on PDB accession 4JD2. A) Conserved
1048 residues from the two different paralogs of Arp2 (Arp2a and Arp2b). Red represent residues
1049 conserved in Arp2a only, blue are conserved in Arp2b only, while green residues are conserved in
1050 both paralogs. B) Conserved site from the two different paralogs of ARPC1 (ARPC1 and ARPC1b).
1051 Red sites represent residues conserved in ARPC1 only, blue are conserved in ARPC1 only, while
1052 green residues are conserved in both paralogs.

1053

1054 **Figure 9:** Myosin maximum likelihood phylogeny, (830 taxa, 754 AA). Groups coloured according
1055 to taxonomic affinity as in figure 2 (blue= Chlorarachniophyta, brown= Foraminifera). Branches
1056 collapsed according to myosin class affiliation and following the nomenclature of Sebé-Pedrós et
1057 al., (2014). The tree is midpoint-rooted and thick branches represents bootstrap > 75%, and
1058 Bayesian support > 0.8 pp. The scale bar equals the mean number of substitutions per site. The
1059 domain architectures for each class with representatives from Rhizaria are shown. IPR annotation
1060 of functional domains is listed Figure 9 – figure supplement 1. A complete ML tree without
1061 collapsed branches can be found in Figure 9 – figure supplement 2.

1062 **Figure 9 – figure supplement 1:** InterPro domains of myosins annotated with InterProscan (Jones
1063 et al. 2014, Mitchell et al. 2015).

1064 **Figure 9 – figure supplement 2:** Maximum likelihood tree of myosin showing all branches. Names
1065 of taxa and myosin classes are from of Sebé-Pedrós et al. (2014), except newly discovered
1066 sequences from MMETSP and new myosin classes. The scale bar equals the mean number of
1067 substitutions per site

1068

1069 **Figure 10:** Phylogeny of rhizarian α -tubulin (75 taxa, 453 AA). Thick branches represents bootstrap
1070 > 75% and posterior probability > 0.9. Some branches are collapsed to save space. Support values
1071 for selected nodes discussed in the text are added for clarity. The colouring scheme is the same as
1072 in figure 2. (Brown = Foraminifera, red= Radiolaria, yellow = Endomyxa, and blue= Filosa). The
1073 scale bar equals the mean number of substitutions per site.

1074

1075 **Figure 11:** Phylogeny of rhizarian β -tubulin (104 taxa, 456 AA). Thick branches represents
1076 bootstrap > 75% and posterior probability > 0.9. Some branches are collapsed to save space.
1077 Support values for selected nodes discussed in the text are added for clarity. The colouring scheme
1078 is the same as in figure 2. (Brown = Foraminifera, red= Radiolaria, yellow = Endomyxa, and blue=
1079 Filosa). The scale bar equals the mean number of substitutions per site.

1080

1081 **Figure 12** Molecular models of paralogs of α -tubulin (α 1- and α 2-tub) in Retaria using PDB
1082 accession 3du7 as template. Residues are coloured according to the evolutionary rates calculated
1083 by ConSurf. Turquoise residues are highly variable and maroon means conserved residues.

1084 **Figure 13:** Molecular models of paralogs of β -tubulin (β 1- and β 2-tub) in Retaria using PDB
1085 accession 3du7 as template. Residues are coloured according to the evolutionary rates calculated
1086 by ConSurf. Turquoise residues are highly variable and maroon means conserved residues.

1087

1088 **Figure 14.** Phylogenetic tree of Rhizaria based on the full dataset, 255 genes, summarizing the
1089 major evolutionary events. Taxa with large portions of missing data are placed on the maximum
1090 likelihood reference tree with the Evolutionary Placement Algorithm (EPA; Berger et al., 2011).
1091 Taxa in bold are sequenced for this study. Arrows mark important evolutionary events,
1092 morphological changes and gene duplications. Thick branches are highly supported with
1093 bootstrap support >90% and posterior probability > 0.9. Branches in grey are the most likely
1094 placement of taxa from EPA with numbers showing the expected likelihood weights for the

1095 placement. Branches that differ between maximum likelihood and Bayesian trees are marked
1096 with dashed lines. For *Sticholonche zancelea* the blue line represent the Bayesian CATGR placement
1097 with posterior probability, red dashed line represents the maximum likelihood LG placement with
1098 bootstrap support. For a further discussion of the placement of *Sticholonche zancelea* and the
1099 relationship between Alveolates, Stramenopiles and Rhizaria see the text. The scale bar equals the
1100 mean number of substitutions per site.

1101

Figure 1

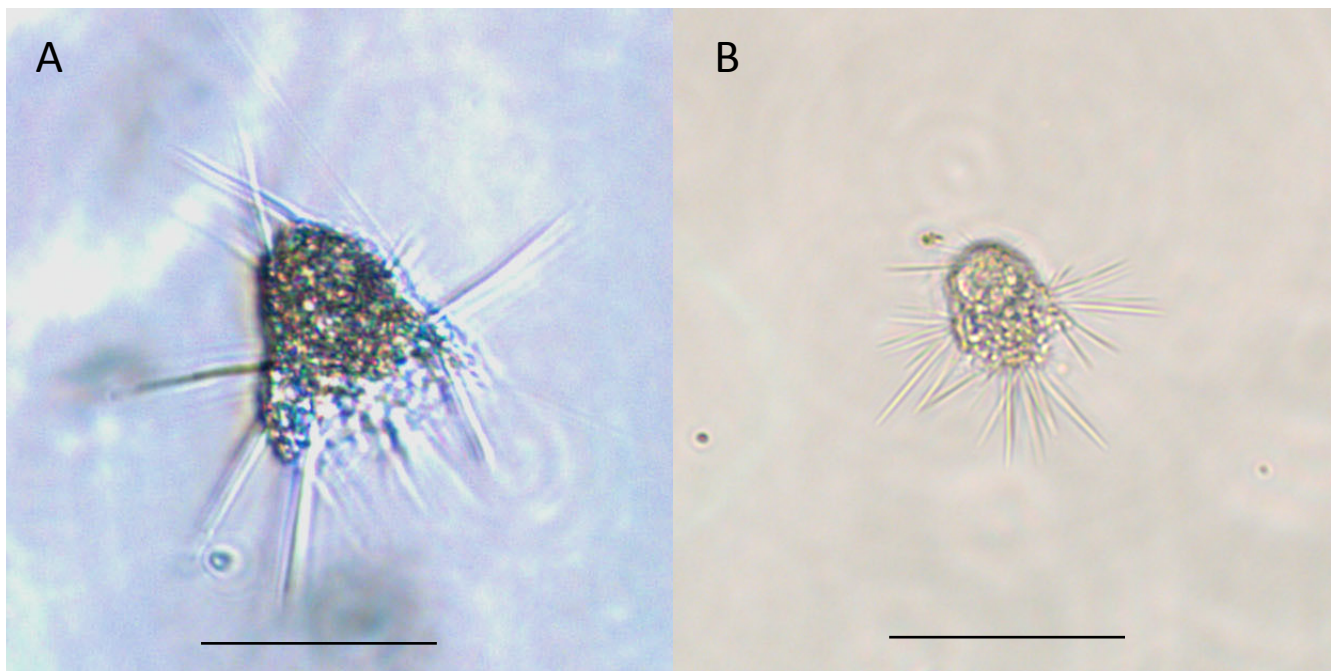


Figure 1 - figure supplement 1

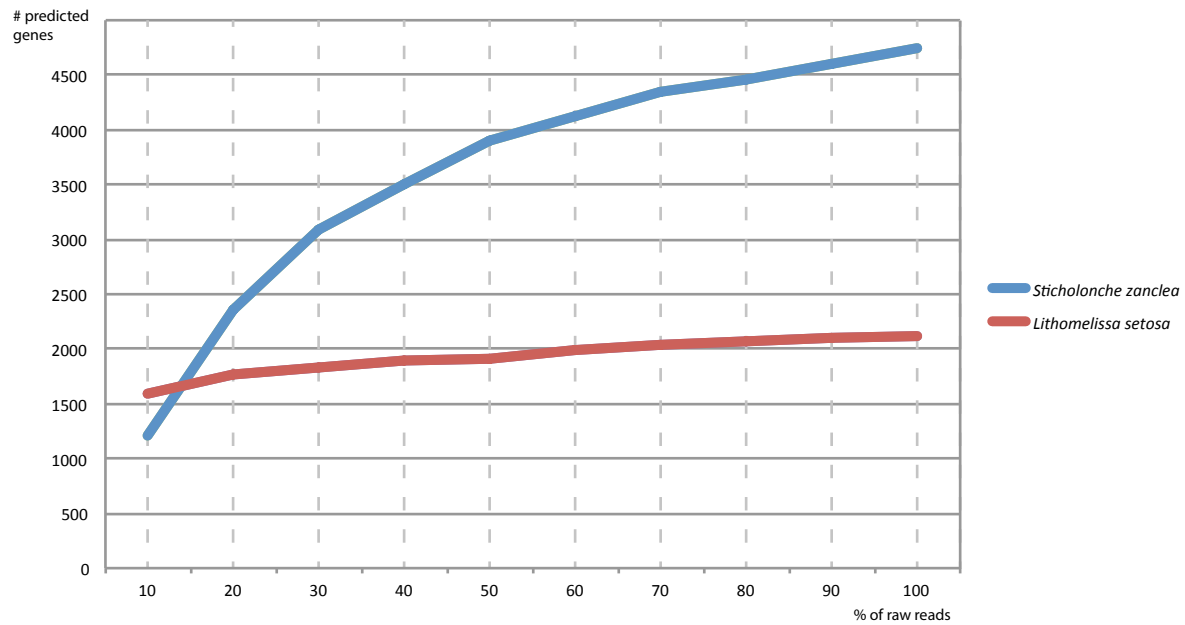
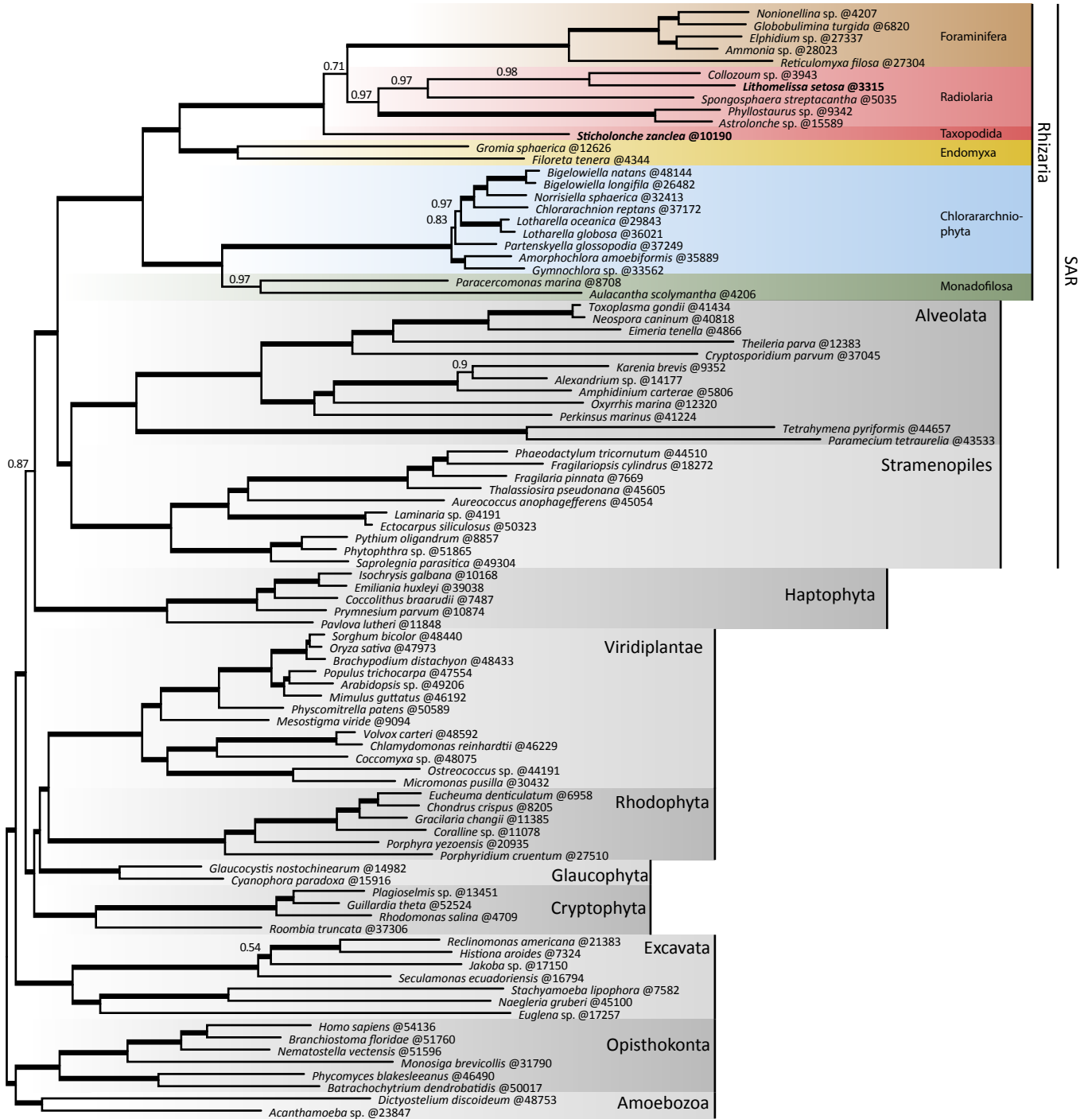


Figure 2



0.2

Figure 2 - figure supplement 1



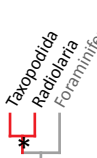
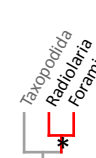
Bayesian		A) SAR - relationship		B) Placement of Taxopodida	
Number of genes	Number of bins removed				
255	0	1	-	-	0.71
	4	1	-	-	0.97
146	0	1	-	-	0.81
146 (LG)	0	1	-	0.67	-

Figure 3
Maximum likelihood

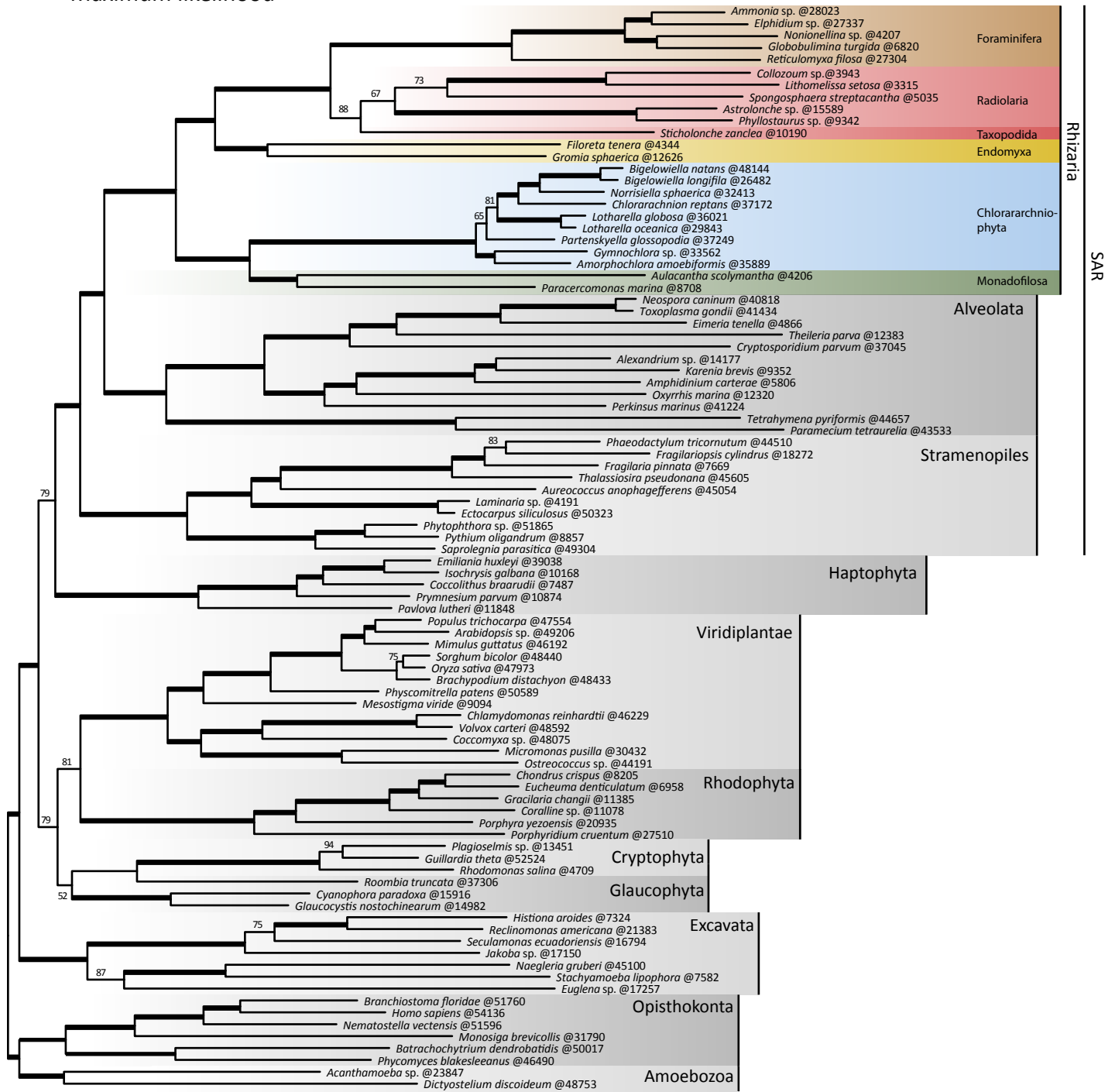


Figure 3 - figure supplement 1

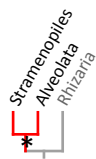

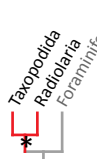

Maximum Likelihood		A) SAR - relationship		B) Placement of Taxopodida	
Number of genes	Number of bins removed				
		255	0	-	96
4	-		67	-	50
146	0	-	85	87	-
	4	74	-	-	75

Figure 4

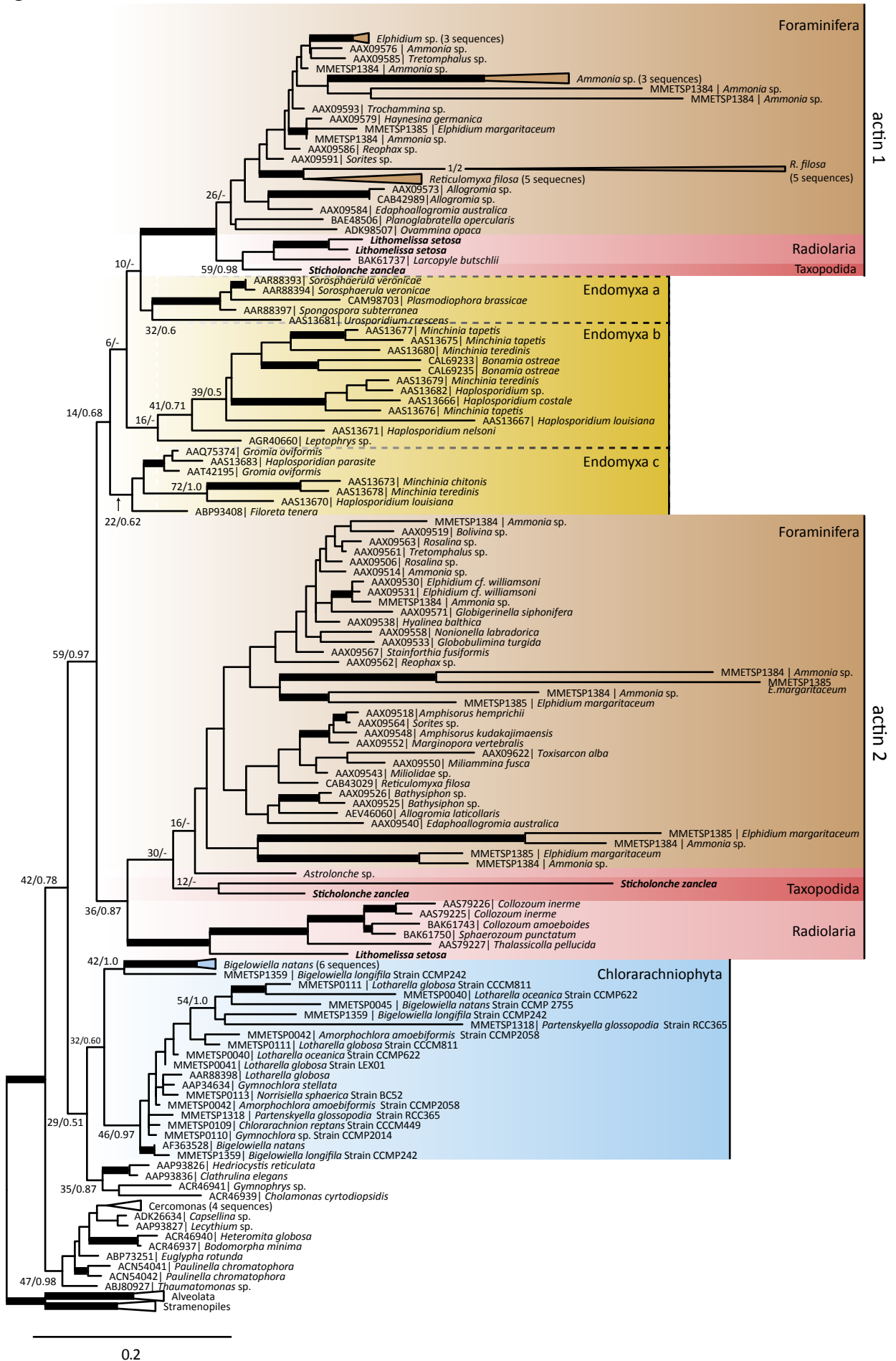


Figure 5

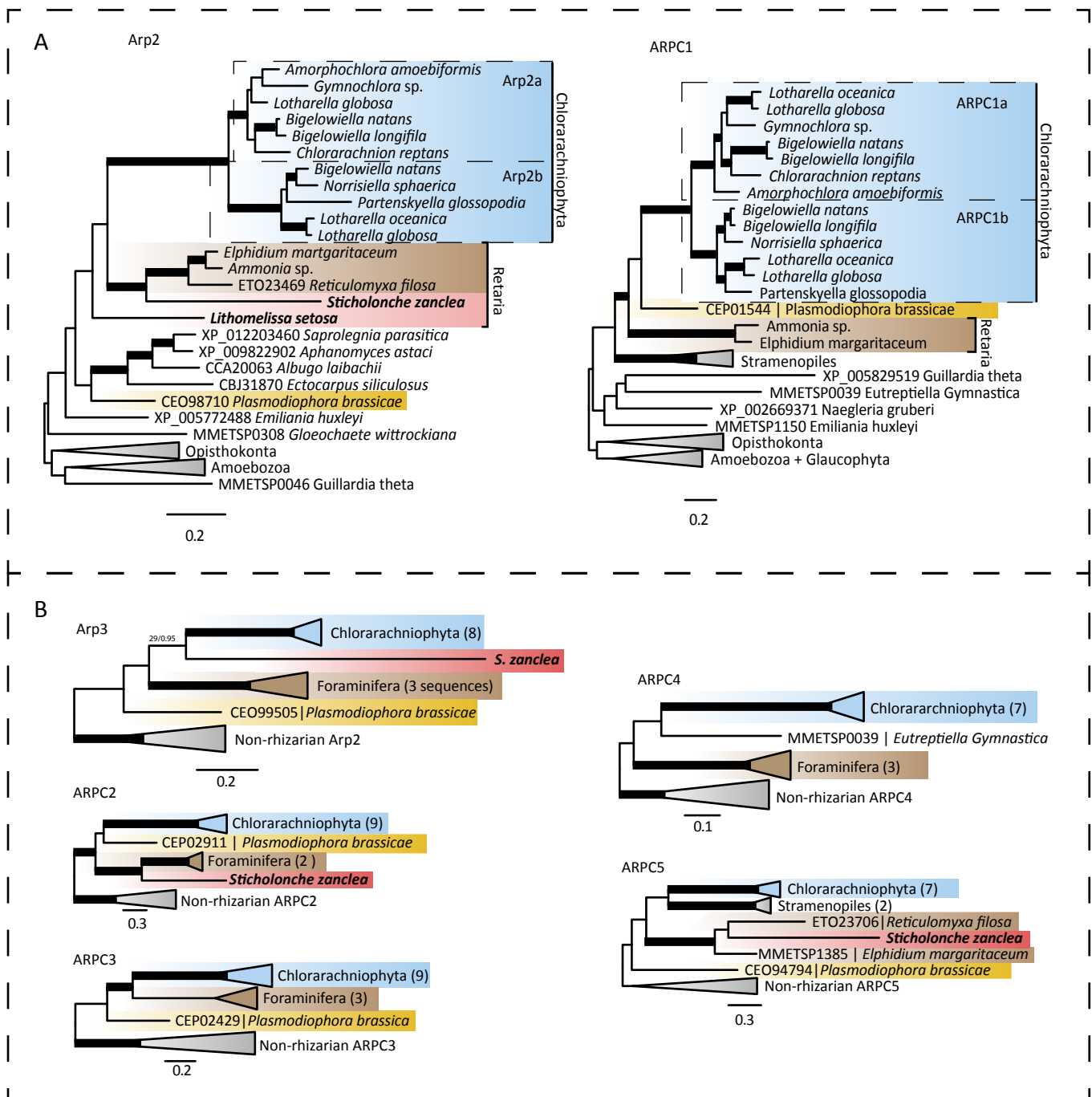


Figure 6

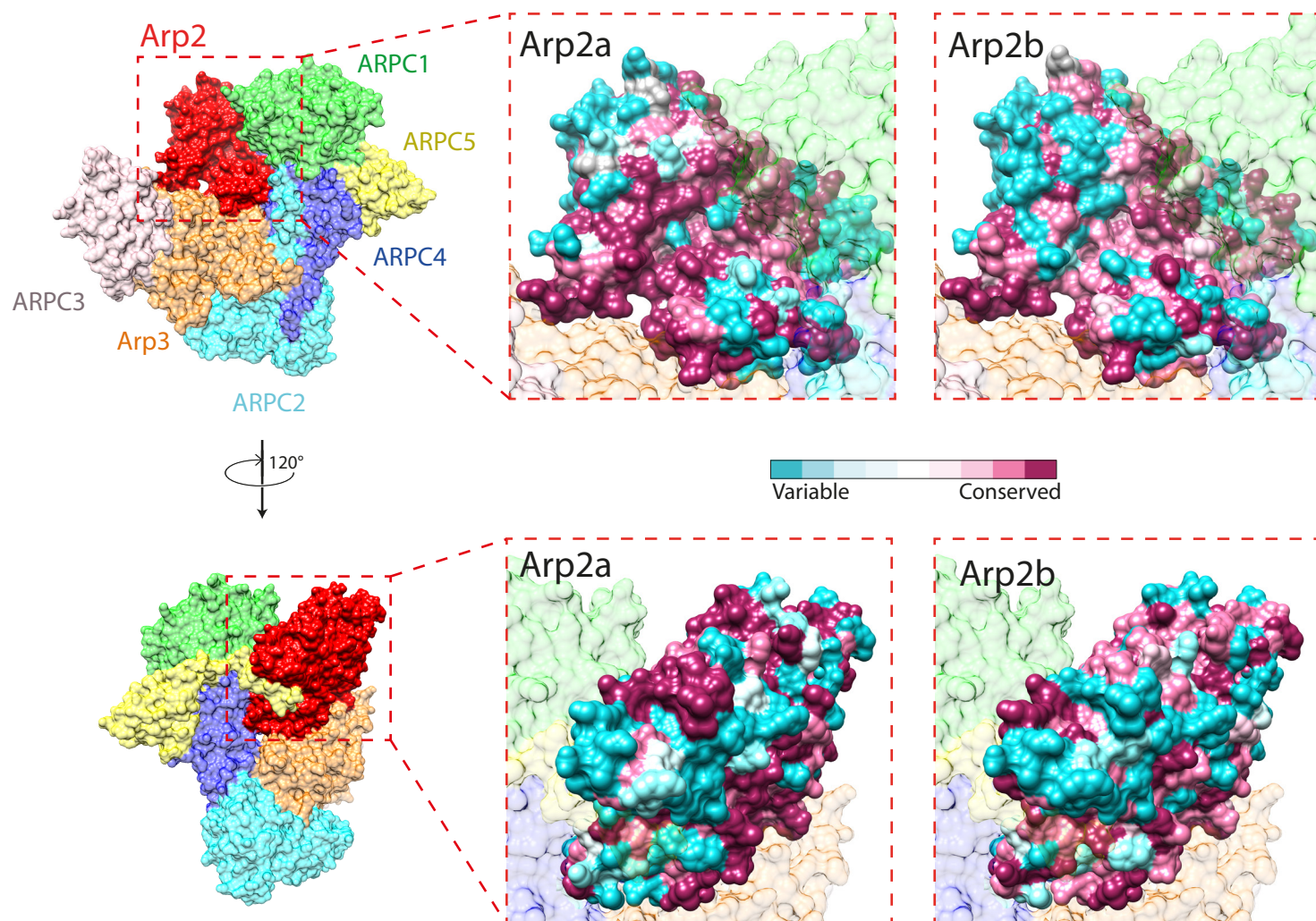


Figure 7

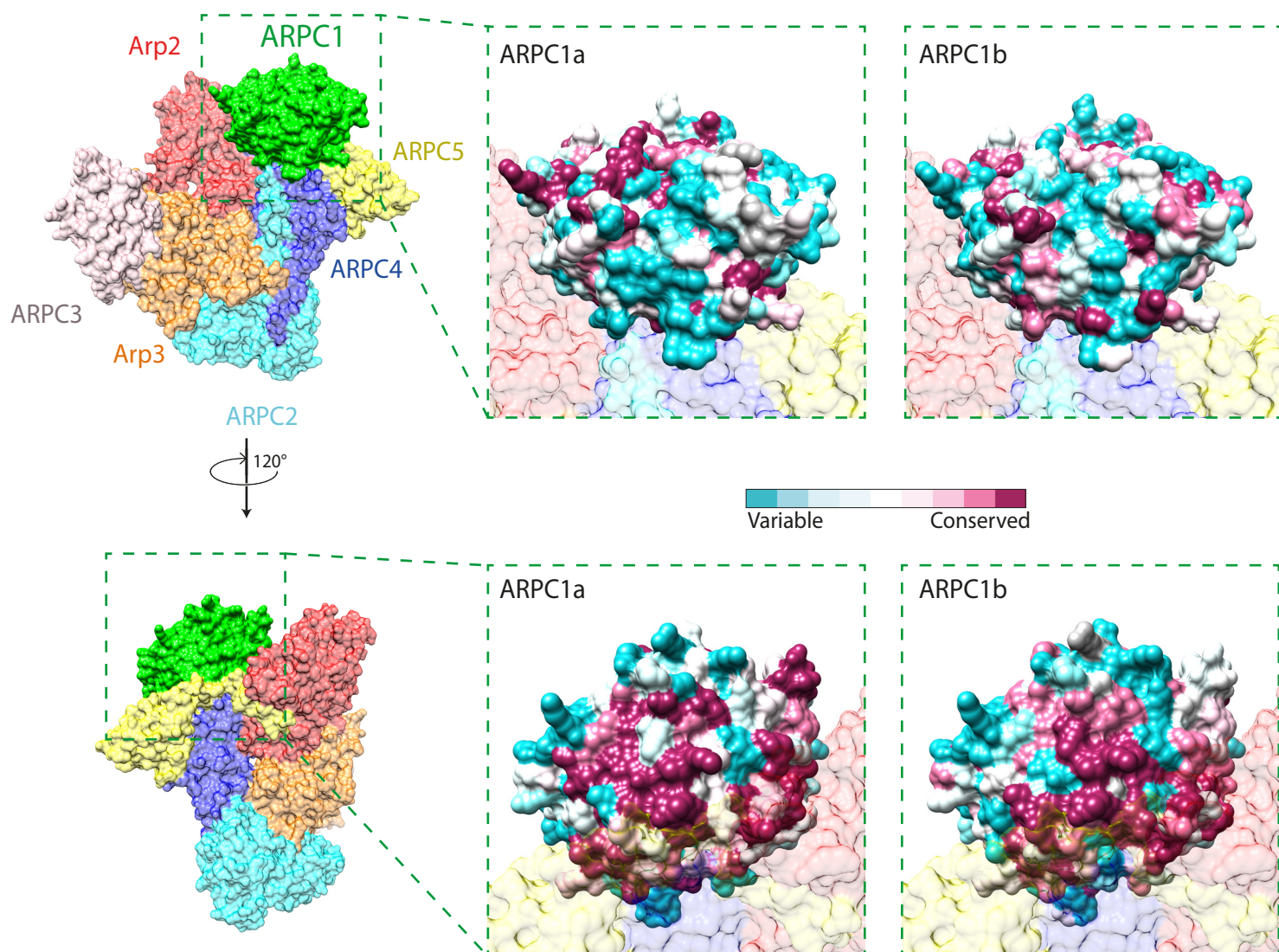


Figure 8

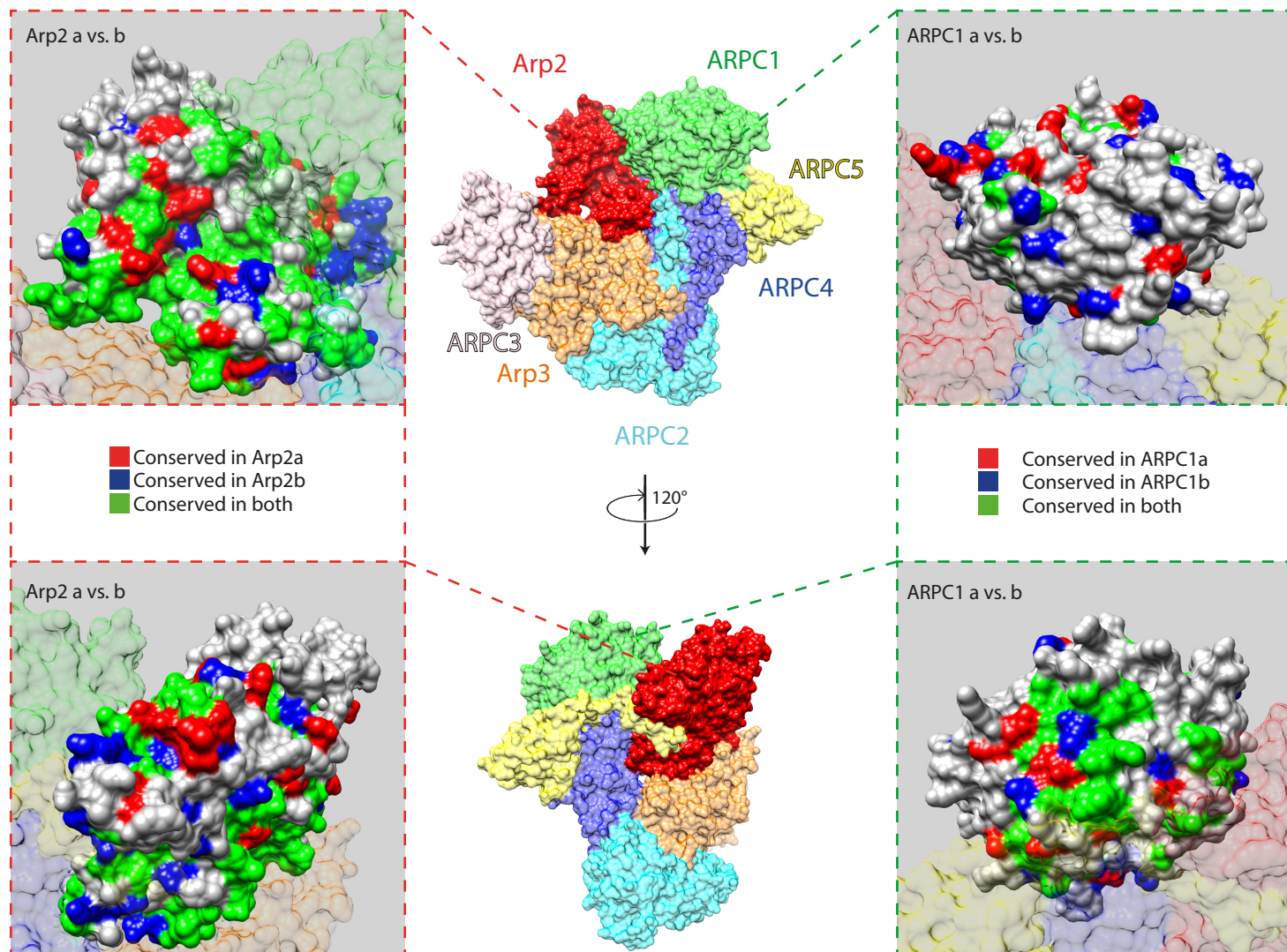


Figure 9 - Figure supplement 1

Functional InterPro domains	
Myosin head	IPR001609
WW	IPR001202
IQ	IPR000048
Myosin TH1	IPR010926
EF	IPR002048
SH3	IPR001452
CH	IPR001715
LRR_RI	IPR032675
MyTH4	IPR000857
DUF	IPR025640
MN	IPR004009
DH	IPR000219
PH	IPR011993
FYVE	IPR013083
MORN	IPR003409
RCC1	IPR000408

Figure 10

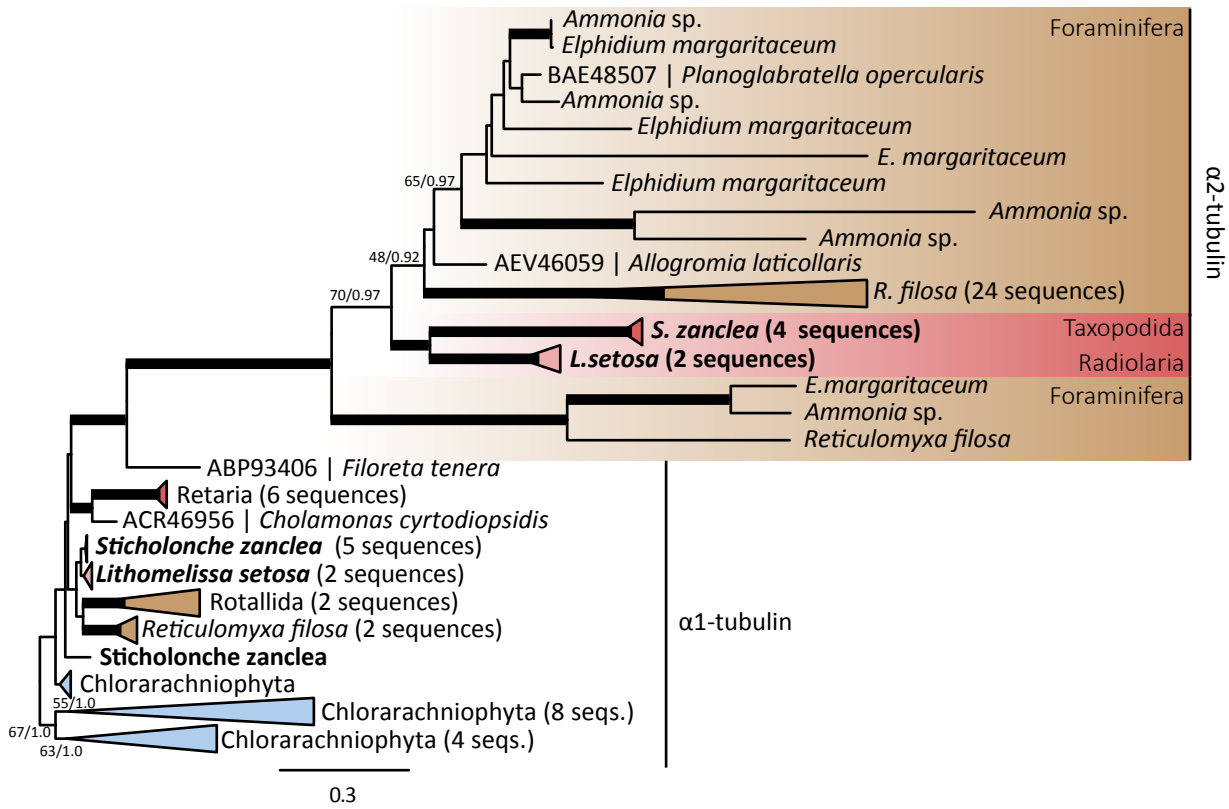


Figure 11

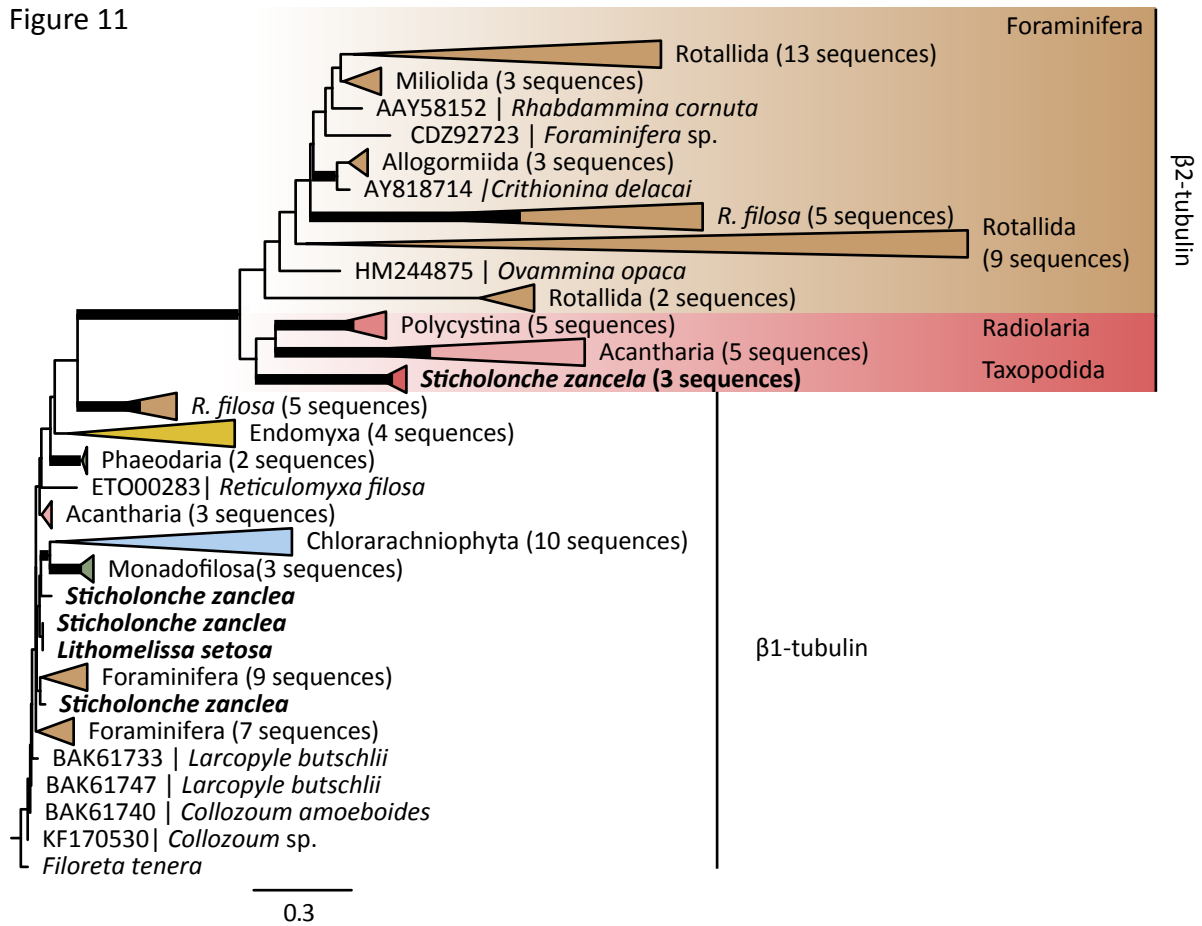


Figure 12

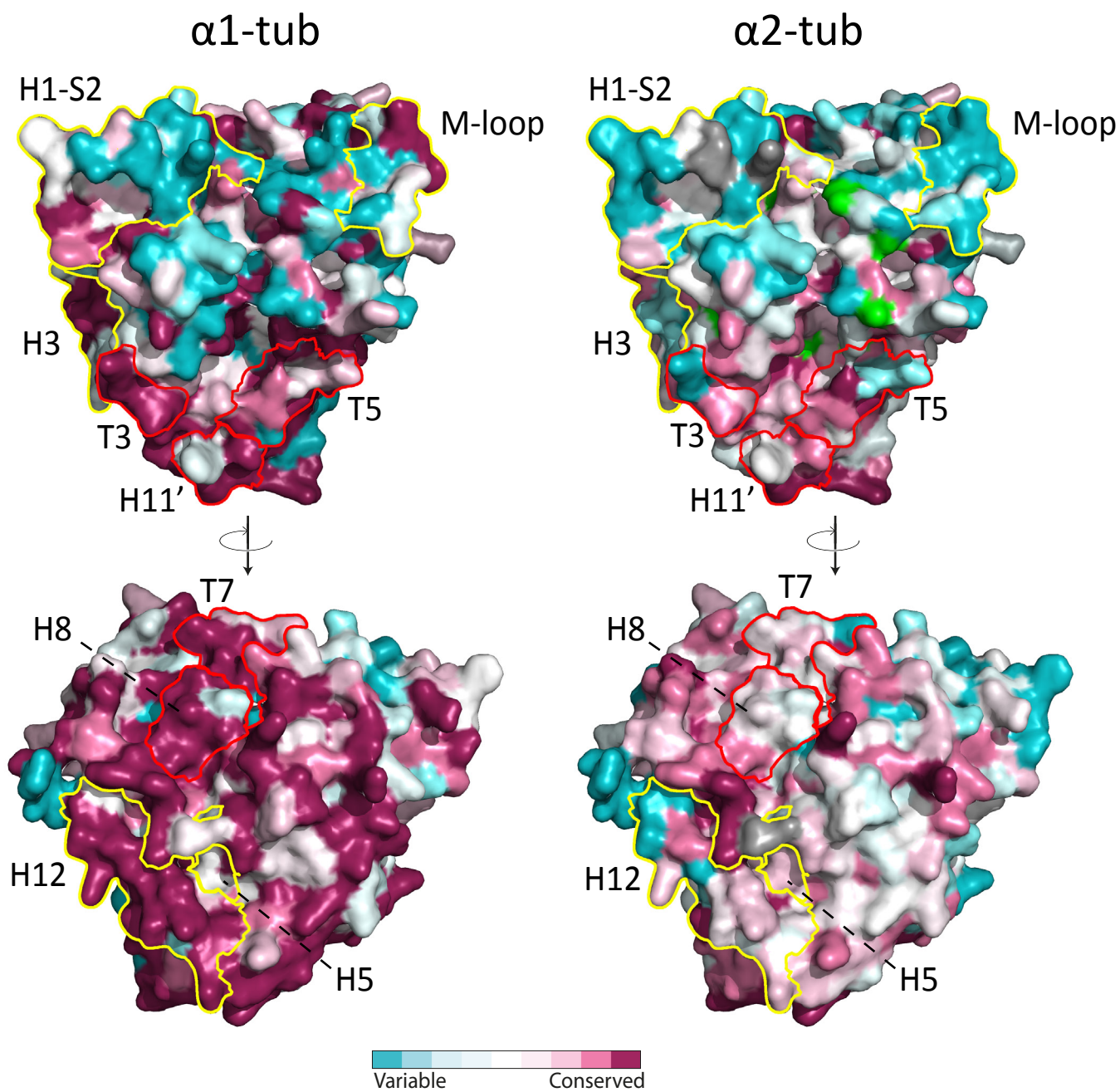


Figure 13

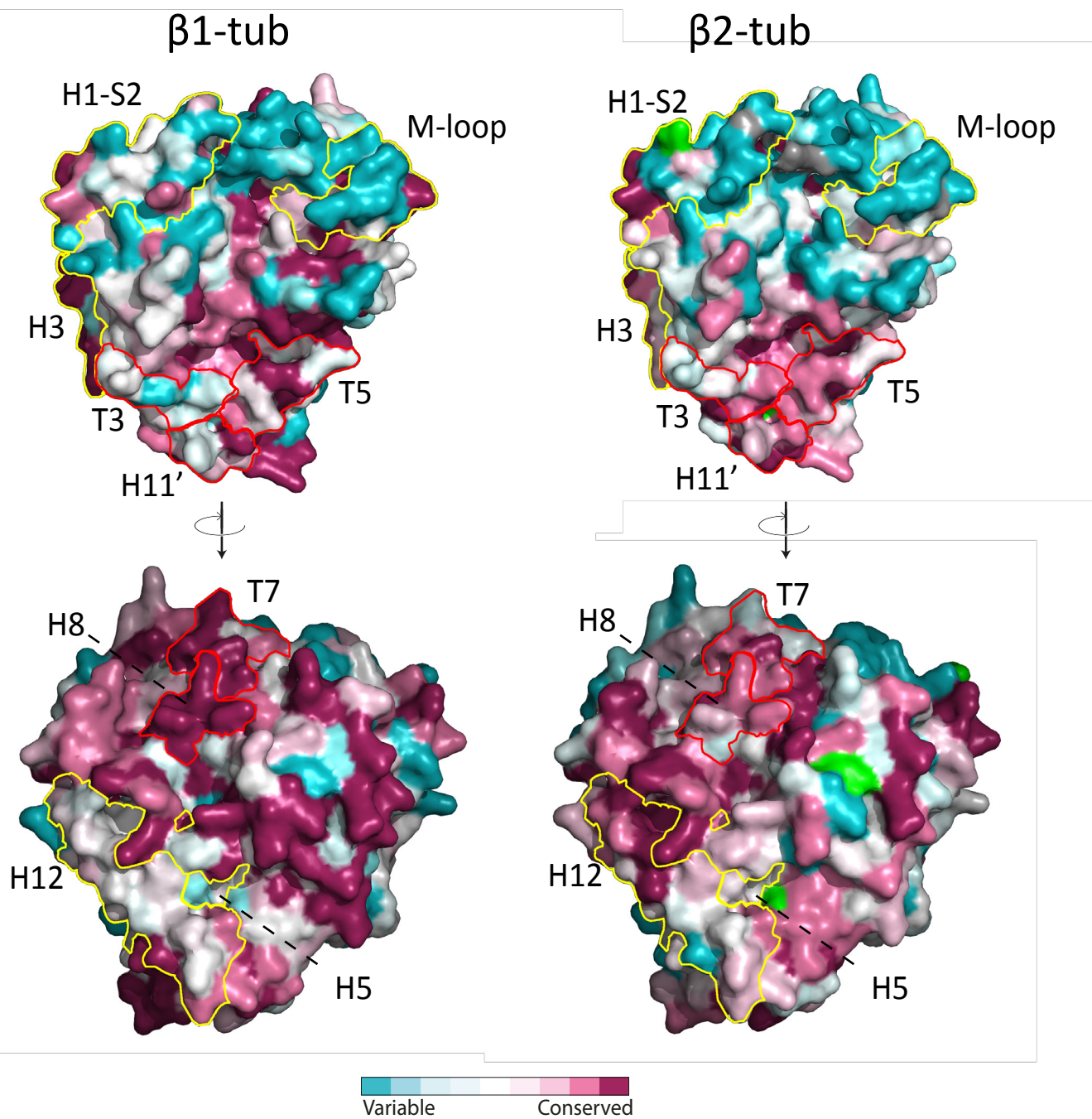
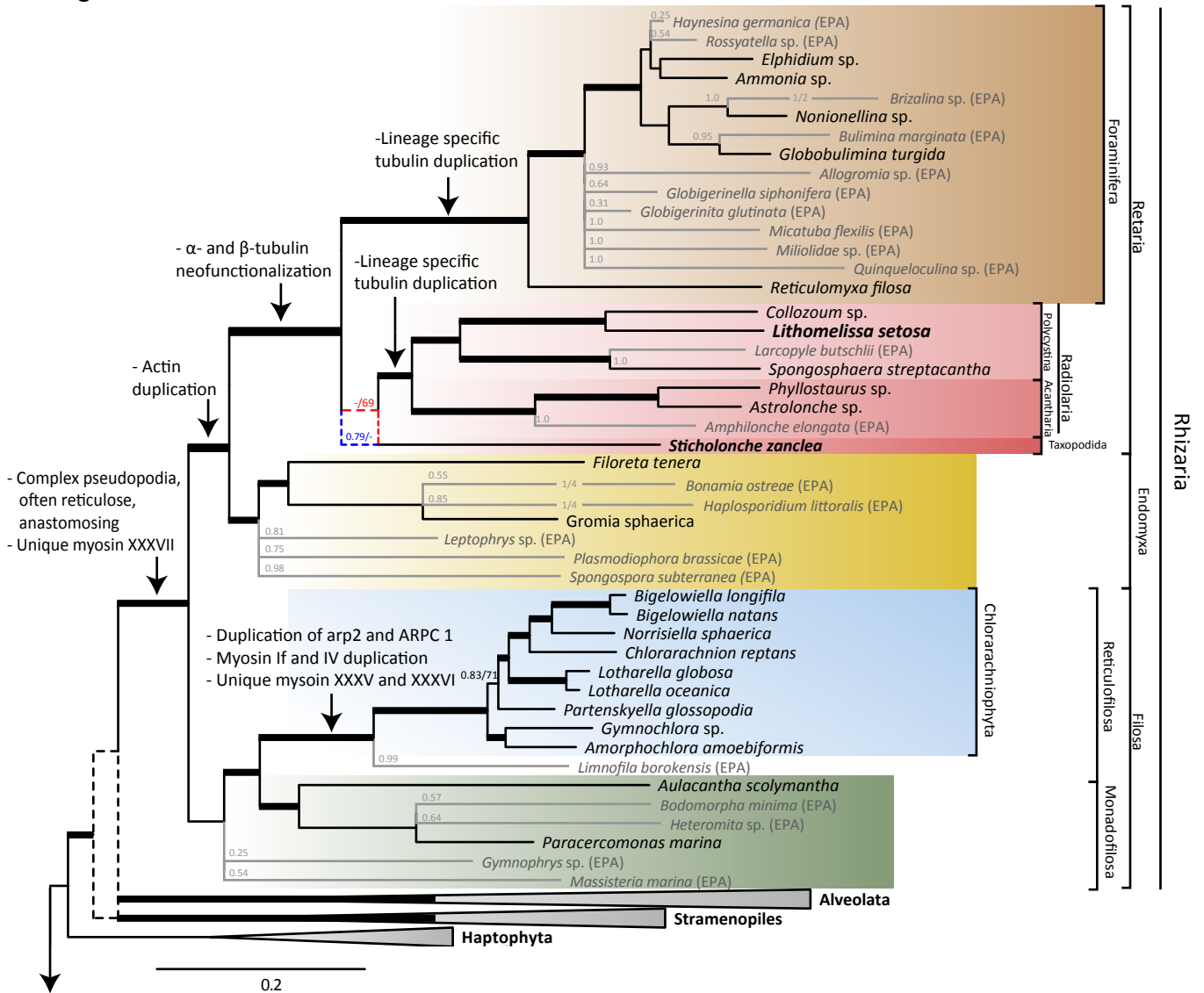


Figure 14



For the rest of the tree, see Fig. 2 (maximum likelihood) and Fig. 3 (Bayesian)

S1 Single gene statistics

Supplementary table S1: Statistics for single gene alignments. **Gene name:** Name of the gene alignment, adopted from Burki et. al (2012). **Number of taxa:** The number of taxa in the gene alignment after BIR and selection in ScaFos. **Min seq. length:** The length of the shortest sequence in the alignment in AA. **Max seq. length:** Length of the longest sequence in the alignment in AA. **Pairwise Identity:** Identity calculated for all pairs of sequences in the alignment. **Model:** The model chosen as the best fitting the sequence data in the alignment by RAxML (Stamatakis 2014). **Likelihood of model:** The likelihood of the model chosen by RAxML. **Relative TC and Relative TCA:** Relative Tree Certainty for the phylogenetic tree of the gene calculate from the bootstrap trees (Salichos et al. 2014). **S. zanzlea:** indicates if the newly sequenced *Sticholonche zanzlea* is present in the gene alignment **L. setosa:** indicates if the newly sequenced *Lithomelissa setosa* is present in the gene alignment. **Selected for reduced dataset:** about half of the genes were selected for a smaller dataset. Genes marked with ACCEPT was selected to the 147-genes dataset.

Gene name	Number of taxa	Min. Seq. length	Max seq. length	Pairwise Identity	Model	Likelihood of model	Relative TC	Relative TCA	S. zanzlea	L. setosa	Selected for red.dataset
ABHD13	36	47	137	57,00%	LG	-5760,505024	0,318832	0,357927	0	0	
AGX	43	131	270	53,30%	LG	-13578,0947	0,319509	0,337786	0	0	
ALG11	33	86	209	58,00%	LG	-8078,540125	0,230892	0,301337	0	0	
ap1m1	58	62	256	70,60%	LG	-8282,246972	0,450315	0,463256	0	yes	ACCEPTED
AP1S2	51	75	109	65,40%	LG	-3948,823321	0,344491	0,386934	0	0	ACCEPTED
ap2m1	48	40	202	60,00%	LG	-7015,548224	0,46415	0,477977	0	0	ACCEPTED
ap3m1	37	63	189	52,80%	LG	-6785,573219	0,51729	0,521267	0	0	
AP3S1	46	47	116	64,70%	LG	-3638,753749	0,290083	0,315354	0	0	
AP4S1	43	79	114	67,80%	LG	-3717,694654	0,353887	0,356413	0	0	
ARL6	21	83	126	59,40%	LG	-2850,923763	0,48422	0,492966	0	0	
ARP2	42	27	157	67,20%	LG	-4811,863091	0,369159	0,38038	0	0	
arpc3	30	43	107	50,70%	LG	-3988,11484	0,416454	0,431791	0	0	
arpc4	45	35	143	60,80%	LG	-5306,773776	0,289444	0,306523	0	0	ACCEPTED
atad1	34	106	152	61,70%	LG	-4634,564786	0,354555	0,385238	0	0	
atp6	77	47	127	75,70%	MTZOA	-4423,911594	0,2374	0,278178	0	0	ACCEPTED
ATP6V0A1	61	78	455	52,50%	LG	-25700,3205	0,381807	0,406443	0	0	ACCEPTED
atp6v1a	76	37	447	74,10%	LG	-14189,67727	0,388532	0,397548	0	0	ACCEPTED
atp6v1b	78	69	433	76,30%	LG	-14137,02466	0,461456	0,473029	yes	yes	ACCEPTED
atp6v1d	64	65	167	55,20%	LG	-10167,97542	0,339726	0,349941	0	0	ACCEPTED
atp6v1e	55	83	138	41,30%	LG	-11233,94332	0,386724	0,422531	0	0	ACCEPTED
bat1	73	88	258	71,80%	LG	-10630,98133	0,404242	0,438108	yes	0	ACCEPTED
C16orf80	48	91	165	88,40%	LG	-2248,848997	0,109672	0,189948	yes	0	ACCEPTED
C22orf28	38	91	466	76,40%	LG	-9894,363438	0,281505	0,336485	yes	0	ACCEPTED
calm	78	49	129	90,00%	LG	-2110,01212	0,148728	0,192604	yes	0	ACCEPTED
calr	60	57	170	67,00%	WAG	-7815,156224	0,231478	0,268941	0	0	ACCEPTED
capzb	45	52	139	59,00%	LG	-5504,958066	0,320148	0,343714	0	0	
CCDC113	18	82	237	51,10%	LG	-5508,47554	0,539594	0,536236	0	0	
CCDC37	30	90	303	46,10%	LG	-11786,2832	0,47835	0,503777	0	0	
CCDC40	19	115	467	42,60%	LG	-12842,83032	0,491146	0,525936	0	0	
CCDC65	25	82	321	42,20%	LG	-10626,90619	0,420871	0,438333	0	0	
CDK5	42	65	280	63,40%	LG	-9418,462225	0,417232	0,422685	yes	0	ACCEPTED
clgn	34	48	145	62,00%	LG	-4899,706222	0,114411	0,173907	0	0	
COP-beta	48	61	243	60,00%	LG	-9593,223689	0,294955	0,308558	0	0	ACCEPTED
COPE	44	123	223	41,50%	LG	-12970,30913	0,517069	0,526119	0	0	ACCEPTED
COPG2	55	70	509	56,20%	LG	-23616,91598	0,521554	0,536455	yes	0	ACCEPTED
COP52	44	89	309	58,50%	LG	-11230,39468	0,527241	0,541656	0	0	
COP56	28	123	187	50,70%	LG	-6346,544893	0,537547	0,560168	0	0	
COQ4_mito	36	109	151	49,00%	LG	-7280,812689	0,401101	0,422436	0	0	
CORO1C	31	97	222	47,20%	LG	-8167,847504	0,374021	0,403971	0	0	
crfg	51	76	292	69,20%	LG	-10103,6544	0,406951	0,411905	0	0	ACCEPTED
CS	64	77	322	63,40%	LG	-16434,00411	0,371514	0,398614	yes	0	ACCEPTED
CTU1	21	88	268	73,10%	LG	-4782,043678	0,323508	0,330796	0	0	
D2HGDH_mito	44	90	309	58,60%	LG	-14265,88235	0,283087	0,324249	0	0	ACCEPTED
DCAF13	52	68	300	55,20%	LG	-15029,30276	0,40634	0,422661	0	0	ACCEPTED
DIMT1L	55	63	183	68,60%	LG	-6863,39766	0,412767	0,445715	0	0	ACCEPTED
DNAI2	28	43	230	54,30%	LG	-6502,660093	0,52915	0,534576	0	0	
DNAL1	25	130	153	54,40%	LG	-5228,202384	0,446404	0,448189	0	0	
dpagt1	45	76	170	65,60%	LG	-6228,299749	0,359216	0,402348	0	0	
DPP3	15	131	528	41,70%	LG	-12511,48145	0,401157	0,423351	0	0	
DRG2	52	63	245	64,70%	LG	-9836,526877	0,385974	0,421983	0	0	ACCEPTED
eef2	81	94	562	67,30%	LG	-24983,51597	0,4763	0,494258	yes	yes	ACCEPTED
EFG_mito	41	51	281	75,00%	LG	-6495,903547	0,223644	0,236228	0	0	
EFTUD1	31	36	270	60,60%	LG	-7839,893152	0,363824	0,388091	0	0	
eftud2	47	103	322	70,00%	LG	-9505,018307	0,412589	0,428361	0	0	ACCEPTED
eif1a	66	47	108	69,70%	LG	-4471,341828	0,297396	0,3331	0	0	ACCEPTED
eif2b	44	72	94	57,20%	LG	-3973,460453	0,328507	0,348898	0	0	
eif2g	58	70	366	76,40%	LG	-9647,829098	0,36947	0,389087	0	0	ACCEPTED
EIF3I	52	52	221	51,60%	LG	-11650,70359	0,425295	0,457061	0	0	ACCEPTED
EIF4A3	63	60	203	80,80%	LG	-4512,305984	0,193142	0,209894	yes	yes	ACCEPTED

S1 Single gene statistics

EIF4E	33	72	107	54,00%	LG	-4350,586754	0,41534	0,420327	0	0	
EIF5A	59	43	100	59,60%	LG	-5896,500255	0,236079	0,266349	0	0	
EIF5B	52	66	379	65,00%	LG	-14629,03823	0,369459	0,387896	0	0	ACCEPTED
EIF6	73	39	189	68,90%	LG	-9203,168463	0,316631	0,323257	0	0	ACCEPTED
EMG1	50	48	137	58,70%	LG	-6251,800505	0,32807	0,364891	0	0	ACCEPTED
ERLIN1	26	58	270	58,10%	LG	-6995,632051	0,446725	0,475001	0	0	
ETF1	62	82	326	74,40%	LG	-10397,59059	0,282949	0,3303	0	0	ACCEPTED
FA2H	22	100	138	56,10%	LG	-3792,536181	0,362447	0,376038	0	0	
FAM18B	39	55	106	48,40%	MTZOA	-5587,027958	0,031813	0,111896	0	0	
FAM96B	40	47	111	66,70%	LG	-3987,056771	0,156244	0,1794	0	0	
FBL	64	50	195	74,00%	LG	-7444,494885	0,389779	0,422724	yes	0	ACCEPTED
FTSJ1	49	88	165	66,80%	LG	-6088,454206	0,366011	0,379057	0	0	ACCEPTED
GAS8	32	81	372	44,90%	LG	-12643,31147	0,491982	0,508781	0	0	
GLI2	72	46	212	61,80%	LG	-11089,82621	0,333202	0,358492	0	0	ACCEPTED
GNB2L1	79	48	192	75,20%	LG	-8732,103713	0,265904	0,306306	yes	yes	ACCEPTED
GPD1L	37	101	221	61,30%	LG	-7689,745256	0,37787	0,406056	0	0	
GPN1	56	45	144	71,20%	LG	-5363,738803	0,212174	0,27348	0	0	ACCEPTED
GPN2	41	32	162	55,70%	LG	-6613,807943	0,30105	0,342609	0	0	
GPN3	46	50	156	61,30%	LG	-6745,689652	0,364161	0,370243	0	0	ACCEPTED
GRWD1	45	85	193	55,90%	LG	-8675,085077	0,328738	0,367185	0	0	ACCEPTED
GSS	34	114	218	54,40%	LG	-9180,431006	0,316714	0,349284	0	0	
H2A	71	81	104	78,90%	LG	-3316,468046	0,179693	0,235261	0	0	
H2B	51	79	98	75,60%	LG	-2796,842548	0,28254	0,293791	0	0	ACCEPTED
H3-2	72	58	129	91,60%	LG	-2055,800843	0,098365	0,155827	yes	0	
H4	55	65	99	92,10%	LG	-1122,507982	0,101839	0,133267	0	0	ACCEPTED
HDCC2	35	74	119	59,60%	LG	-4442,215557	0,164491	0,184876	0	0	
HSP90	83	66	415	75,30%	LG	-17790,87246	0,346099	0,362569	yes	0	ACCEPTED
HYOU1	33	80	223	48,00%	LG	-8724,657423	0,486438	0,488844	0	0	
IFT46	33	88	174	58,20%	LG	-5956,901367	0,264837	0,299679	yes	0	
IFT57	27	93	260	39,40%	LG	-10551,41885	0,404256	0,426612	0	0	
IFT88	34	92	458	51,10%	LG	-16827,41096	0,501749	0,515524	0	0	
IMP4	54	79	202	62,90%	LG	-9421,343366	0,325505	0,371947	0	0	
INO1	61	41	344	70,00%	LG	-14295,91978	0,217782	0,283593	0	yes	ACCEPTED
IPO4	30	84	376	38,80%	LG	-15957,99189	0,591417	0,598396	0	0	
KARS	62	62	264	70,80%	LG	-10600,97116	0,214743	0,2715	0	0	ACCEPTED
KDEL2	49	77	156	59,10%	LG	-7062,524635	0,434685	0,441327	yes	0	ACCEPTED
LRRC48	23	104	293	42,10%	LG	-9179,427104	0,430746	0,43918	0	0	
MAT1A	64	71	260	75,30%	LG	-8605,46176	0,412943	0,422438	0	0	ACCEPTED
MCM2	50	90	263	73,80%	LG	-8012,514112	0,315807	0,340618	yes	yes	ACCEPTED
MCM3	51	93	271	72,50%	LG	-9047,602802	0,30263	0,325999	yes	0	ACCEPTED
MCM4	44	81	296	70,10%	LG	-9420,901756	0,393779	0,405547	yes	0	ACCEPTED
MCM5	45	46	246	74,20%	LG	-7002,009287	0,501695	0,516008	0	yes	ACCEPTED
MCM6	52	77	299	70,30%	LG	-9645,961336	0,429845	0,469787	yes	0	ACCEPTED
MCM7	51	62	263	72,40%	LG	-8287,185832	0,356912	0,366708	0	0	
MCM8	29	76	249	65,10%	LG	-6245,43182	0,354433	0,373885	0	0	
MCM9	29	143	255	64,60%	LG	-6777,807545	0,312351	0,329352	0	0	
MEMO1	38	44	168	60,10%	LG	-6043,619772	0,262226	0,299762	0	0	
METAP2	65	46	258	68,80%	LG	-11415,57471	0,254494	0,270575	0	0	ACCEPTED
METTL1	54	77	160	55,60%	LG	-8932,773544	0,396241	0,40591	0	0	ACCEPTED
MMAA_mito	26	109	250	53,90%	LG	-7212,012798	0,346028	0,362445	0	0	
MTHFR	45	65	334	63,30%	LG	-11874,32317	0,323083	0,343072	0	0	ACCEPTED
NAA15	47	66	431	47,70%	LG	-23715,45616	0,418084	0,435806	0	0	ACCEPTED
NAE1	31	136	236	52,20%	LG	-9548,540186	0,342409	0,373591	0	0	
NAPB	44	93	199	42,50%	LG	-12243,68002	0,332845	0,365541	0	0	
NDUFV1	62	75	402	73,80%	LG	-12540,76853	0,376573	0,403972	0	0	ACCEPTED
NDUFV2_mito	51	54	155	58,60%	LG	-7609,888533	0,398947	0,408587	0	0	ACCEPTED
NFS1_mito	61	46	232	75,50%	LG	-7853,794077	0,193981	0,243005	0	0	ACCEPTED
NHP2	37	41	68	62,10%	LG	-2330,871773	0,167164	0,230068	0	0	
NHP2L1	52	99	115	72,30%	LG	-3873,133239	0,202258	0,248085	0	0	ACCEPTED
NLN_mito	30	78	381	48,60%	LG	-14240,63183	0,421332	0,44309	0	0	
NOP56	60	96	281	67,10%	LG	-11882,5291	0,338887	0,362741	yes	0	ACCEPTED
NOP58	55	45	213	74,40%	LG	-7000,149845	0,231158	0,263766	0	0	ACCEPTED
NSA2	64	68	240	72,40%	LG	-9499,737422	0,303197	0,325586	0	0	ACCEPTED
NSF	53	65	270	69,00%	LG	-9891,769117	0,374889	0,39972	0	0	
OPAH	38	50	564	64,90%	LG	-14938,31009	0,394035	0,411375	0	0	
ORF1	16	57	134	54,50%	LG	-2767,405751	0,422535	0,458527	0	0	
OSGEP	46	59	259	73,60%	LG	-7878,100778	0,251901	0,259963	0	0	
PABPC4	52	80	293	62,30%	LG	-13166,19715	0,439501	0,458006	0	0	
PACRG	38	70	160	67,00%	LG	-3910,462495	0,451179	0,462882	0	0	
PIK3C3	29	81	304	57,80%	LG	-8753,142845	0,378127	0,390866	0	0	
PLS3	46	72	299	56,40%	LG	-12508,93565	0,331283	0,355786	0	0	
PMM2	56	80	196	63,40%	LG	-10196,48796	0,28168	0,318027	0	0	ACCEPTED

S1 Single gene statistics

pno1	52	82	147	64,60%	LG	-5798,528592	0,321841	0,346093	0	0	ACCEPTED
polr2a	58	38	772	67,80%	LG	-28598,87753	0,501619	0,51791	0	0	ACCEPTED
polr3a	44	35	722	64,70%	LG	-25416,11427	0,448526	0,453603	0	0	ACCEPTED
PPP2R3	51	98	216	51,10%	LG	-10956,80171	0,33061	0,357455	yes	0	ACCEPTED
PPP2R5C	53	100	254	65,40%	LG	-9093,0928	0,452387	0,470386	yes	0	ACCEPTED
prmt8	56	70	193	68,40%	LG	-7952,374899	0,115864	0,189454	yes	0	ACCEPTED
PROSC	40	68	146	61,30%	LG	-6082,345294	0,167065	0,217243	0	0	ACCEPTED
psma1	59	60	137	67,50%	LG	-5245,198916	0,280318	0,306409	0	0	ACCEPTED
psma2	58	63	121	76,10%	LG	-3542,825171	0,250677	0,29696	0	0	ACCEPTED
psma3	59	70	135	60,10%	LG	-6965,175678	0,308209	0,344862	0	0	
psma4	52	52	128	72,50%	LG	-4080,020923	0,355296	0,383875	0	0	ACCEPTED
psma6	54	37	126	64,70%	LG	-5607,745264	0,29337	0,322105	0	0	ACCEPTED
psma7	74	93	171	69,20%	LG	-7389,740101	0,387995	0,40465	yes	0	ACCEPTED
psmb1	49	43	107	62,90%	LG	-4522,89283	0,189984	0,264529	0	0	
psmb2	48	57	136	51,30%	LG	-7681,533281	0,299603	0,336806	0	0	ACCEPTED
psmb3	63	38	155	60,60%	LG	-8500,309613	0,206209	0,269064	0	0	ACCEPTED
psmb4	60	31	127	56,10%	LG	-7943,98869	0,345529	0,364564	0	0	ACCEPTED
psmb5	67	39	165	67,70%	LG	-7447,399713	0,320037	0,332597	yes	0	ACCEPTED
psmb6	61	69	156	60,90%	LG	-8273,200252	0,245688	0,285231	yes	0	ACCEPTED
psmb7	59	79	167	71,60%	LG	-6434,285244	0,257649	0,308525	0	0	ACCEPTED
psmc1	62	17	231	85,30%	LG	-4624,959992	0,221607	0,227897	yes	0	ACCEPTED
psmc2	62	69	192	86,10%	LG	-3452,946091	0,257755	0,276869	yes	0	ACCEPTED
psmc3	65	68	205	82,00%	LG	-5119,32424	0,233063	0,257247	yes	yes	ACCEPTED
psmc4	65	62	217	85,70%	LG	-4527,047494	0,228069	0,257664	yes	0	ACCEPTED
psmc5	62	77	227	87,50%	LG	-4087,161508	0,22135	0,245021	yes	0	ACCEPTED
psmc6	57	79	205	85,00%	LG	-3653,729711	0,336373	0,34444	0	0	ACCEPTED
psmd1	60	47	399	59,60%	LG	-19493,56751	0,327354	0,349363	0	0	ACCEPTED
PSMD12	60	58	254	50,40%	LG	-14784,39862	0,386045	0,41888	0	0	ACCEPTED
psmd14	63	63	257	74,40%	LG	-9323,920868	0,316358	0,343751	0	0	ACCEPTED
PYGB	37	95	544	59,90%	LG	-17011,34898	0,452082	0,462753	0	0	
rad51	40	44	256	67,80%	LG	-7401,132592	0,400212	0,420737	0	0	
ran	76	92	165	80,20%	LG	-4827,137813	0,29601	0,325396	yes	0	ACCEPTED
RBX1	42	53	81	84,50%	JTTDCMUT	-1321,38173	0,160756	0,207393	0	0	
RPF1	44	78	174	57,00%	LG	-7822,702904	0,227766	0,307594	0	0	
rpl10	83	61	159	74,40%	LG	-7018,621297	0,335667	0,35485	0	0	ACCEPTED
rpl10a	67	45	161	64,90%	LG	-9013,950253	0,328544	0,354329	0	0	ACCEPTED
rpl11	79	69	152	70,00%	LG	-8075,80399	0,257236	0,298261	yes	0	ACCEPTED
rpl12	67	79	123	66,00%	LG	-6772,058452	0,263227	0,305638	yes	0	ACCEPTED
rpl13	59	33	103	63,60%	LG	-5917,390741	0,142489	0,191396	0	0	
rpl13a	76	49	147	60,00%	LG	-10377,96112	0,181501	0,248287	0	0	ACCEPTED
rpl14	55	58	106	46,50%	LG	-8034,119988	0,091558	0,14002	0	0	
rpl15	84	68	175	69,90%	LG	-9673,89018	0,32926	0,369159	yes	yes	ACCEPTED
rpl17	63	72	113	64,70%	LG	-5723,051921	0,246751	0,280843	0	0	ACCEPTED
rpl18	55	41	112	70,30%	LG	-4767,156378	0,114703	0,152984	0	0	ACCEPTED
rpl18a	66	51	91	60,90%	LG	-5269,811383	0,303035	0,316073	0	0	
rpl19	76	70	141	69,10%	LG	-7501,80192	0,286695	0,336505	0	0	ACCEPTED
rpl21	57	62	97	62,10%	LG	-5129,179991	0,267927	0,299892	0	0	ACCEPTED
rpl24	50	55	74	50,20%	LG	-4410,209795	0,319932	0,32644	0	0	
rpl26	67	44	99	67,70%	LG	-4801,451332	0,214933	0,25063	0	0	ACCEPTED
rpl3	88	55	311	70,50%	LG	-18214,71205	0,36935	0,395013	yes	yes	ACCEPTED
rpl30	59	50	79	68,80%	LG	-3682,374608	0,072009	0,128008	0	0	
rpl31	55	49	74	59,60%	LG	-4354,221301	0,128679	0,147874	0	0	
rpl32	51	57	97	65,50%	LG	-4399,083884	0,184069	0,218971	0	0	
rpl35	53	33	89	58,40%	RTREV	-5252,732522	0,160468	0,225778	0	0	
rpl35a	68	50	90	56,90%	LG	-6333,63494	0,227962	0,272497	0	0	ACCEPTED
rpl36a	61	43	81	72,10%	LG	-3333,656145	0,194011	0,224686	0	0	
rpl37a	56	24	85	67,50%	LG	-3716,721948	0,204181	0,22087	0	0	
rpl4	76	43	198	68,30%	LG	-11227,56917	0,182645	0,252107	yes	yes	ACCEPTED
rpl5	73	43	144	68,10%	LG	-8208,175351	0,212312	0,213534	0	0	
rpl6	53	22	94	67,60%	LG	-4388,032474	-0,043405	0,02437	0	0	
rpl7	82	69	149	62,50%	LG	-10755,8653	0,287812	0,332822	0	0	ACCEPTED
rpl7a	81	36	132	65,90%	LG	-8859,248599	0,274114	0,318738	0	0	ACCEPTED
rpl8	77	124	240	67,90%	LG	-13793,04187	0,311511	0,352076	yes	0	ACCEPTED
rpl9	64	53	115	64,50%	LG	-6691,555067	0,196335	0,260506	0	0	
rplp0	78	40	180	51,10%	LG	-15763,68346	0,381238	0,389908	0	0	ACCEPTED
rps10	40	62	75	63,20%	LG	-2814,409652	0,244271	0,28835	0	0	
rps11	73	50	107	71,80%	LG	-5361,813968	0,219257	0,26323	yes	0	ACCEPTED
rps12	54	57	94	60,40%	LG	-4519,596208	0,314607	0,341976	0	0	
rps13	75	44	128	73,70%	LG	-5701,230706	0,212594	0,241174	0	0	ACCEPTED
rps14	64	48	127	83,50%	LG	-3390,476476	0,180371	0,231673	0	yes	
rps15	61	31	98	71,60%	LG	-4180,540985	0,20494	0,23828	0	0	
rps15a	73	62	129	73,30%	LG	-5860,161627	0,197495	0,227841	yes	0	ACCEPTED

S1 Single gene statistics

rps16	61	65	124	72,70%	LG	-4965,414355	0,136972	0,165743	yes	0	ACCEPTED
rps17	53	38	83	78,40%	LG	-2181,131702	0,246075	0,271472	0	0	
rps18	72	56	117	70,20%	LG	-5474,465755	0,26954	0,318506	yes	0	ACCEPTED
rps2	78	47	190	71,90%	LG	-9334,854998	0,252889	0,298213	yes	yes	ACCEPTED
rps20	64	48	91	74,60%	LG	-3381,434534	0,169691	0,206879	0	0	
rps23	75	62	126	79,50%	LG	-4486,572429	0,318774	0,334894	0	0	ACCEPTED
rps26	61	40	85	61,30%	LG	-4199,599607	0,299643	0,313042	0	0	
rps27	52	35	65	75,00%	LG	-2194,238872	0,156599	0,213324	0	0	
rps3	78	59	178	74,30%	LG	-7468,33651	0,208557	0,247935	0	0	ACCEPTED
rps3a	75	25	204	62,00%	LG	-12542,94964	0,31865	0,355165	0	0	ACCEPTED
rps4y1	81	58	218	63,00%	LG	-15664,68245	0,354728	0,379183	yes	0	ACCEPTED
rps5	70	68	162	75,40%	LG	-5794,456167	0,319957	0,334876	0	0	ACCEPTED
rps6	68	104	169	64,80%	LG	-10147,66672	0,325746	0,348633	yes	yes	ACCEPTED
rps8	68	37	147	70,60%	LG	-7105,316255	0,203985	0,24948	0	0	ACCEPTED
rps9	46	45	162	68,20%	LG	-5623,640554	0,303148	0,332849	0	0	ACCEPTED
rpsa	56	54	169	65,90%	LG	-7072,3259	0,35915	0,38947	0	0	
RRM1	58	36	564	71,10%	LG	-20008,85311	0,423029	0,439685	yes	0	ACCEPTED
sars	56	96	237	67,40%	LG	-9924,277015	0,346782	0,361108	0	0	ACCEPTED
SCO1_mito	42	66	123	54,40%	LG	-5480,379693	0,257199	0,301144	0	0	ACCEPTED
sec61	78	73	399	70,80%	LG	-16785,0376	0,383642	0,39609	yes	0	ACCEPTED
SND1	34	255	443	44,50%	LG	-19473,31849	0,519936	0,539994	0	0	
SORCS3	21	72	207	47,50%	LG	-6614,66172	0,436617	0,448523	0	0	
SPTLC1	28	153	173	53,90%	LG	-5674,688333	0,454531	0,490133	0	0	
srp54	67	58	304	64,30%	LG	-12873,87608	0,274982	0,314687	0	0	
srpr	57	57	157	72,10%	LG	-4987,352766	0,309341	0,331631	0	0	ACCEPTED
STXBP1	30	138	200	45,10%	LG	-8270,480822	0,437103	0,459147	0	0	
suclg1	70	54	238	75,50%	LG	-9519,676276	0,292659	0,335853	yes	0	ACCEPTED
tbp	43	39	142	69,60%	LG	-4720,390277	0,172201	0,213067	yes	0	ACCEPTED
tcp1-alpha	50	70	267	73,50%	LG	-7326,908482	0,324951	0,349667	0	0	
tcp1-beta	57	34	285	71,00%	LG	-8833,546155	0,208771	0,24571	0	0	ACCEPTED
tcp1-delta	63	43	280	70,80%	LG	-9007,103802	0,330709	0,353747	yes	0	ACCEPTED
tcp1-epsilon	68	58	243	72,50%	LG	-7502,59552	0,260491	0,312204	yes	0	ACCEPTED
tcp1-eta	59	54	273	70,70%	LG	-9944,671421	0,315724	0,348581	0	0	ACCEPTED
tcp1-gamma	43	66	226	71,60%	LG	-5856,5992	0,312327	0,328966	0	0	
tcp1-theta	46	30	172	63,20%	LG	-5905,817557	0,343444	0,361182	0	0	
tcp1-zeta	51	42	212	66,50%	LG	-7354,641576	0,297075	0,343551	0	0	
TM9SF1	49	22	218	57,50%	MTZOA	-7848,370323	0,324587	0,350012	0	0	ACCEPTED
tubg	58	95	340	76,70%	LG	-9313,404275	0,345022	0,375534	0	0	
UBA3	45	45	142	69,00%	LG	-4447,790728	0,234178	0,24643	0	0	
ubc	78	50	67	94,00%	JTTDCMUT	-651,559846	0,068042	0,105242	0	0	
UBE2J2	44	59	109	63,60%	LG	-4583,840059	0,267364	0,277833	0	0	ACCEPTED
VAR5	48	66	526	65,90%	LG	-19329,61625	0,459151	0,46621	0	0	
VBP1	43	79	129	47,90%	LG	-7049,32631	0,319827	0,370876	0	0	
vpc	73	54	456	79,70%	LG	-10499,67654	0,294834	0,323803	yes	yes	ACCEPTED
VPS18	27	101	441	46,70%	LG	-15118,53653	0,521031	0,539672	0	0	
VPS26B	60	65	187	66,90%	LG	-6771,809835	0,398249	0,420922	0	0	ACCEPTED
vps4	63	60	194	72,60%	LG	-6591,279974	0,338121	0,354526	0	0	ACCEPTED
wars	59	38	244	66,10%	LG	-11033,81374	0,248931	0,291756	yes	0	ACCEPTED
WBSCR22	50	66	201	59,90%	LG	-10051,88619	0,404664	0,426312	0	0	ACCEPTED
xpb	49	109	285	67,30%	LG	-10121,99666	0,424055	0,438759	0	0	ACCEPTED
XRP2	22	91	129	46,50%	LG	-4369,734607	0,278992	0,306433	0	0	
YKT6	49	102	117	53,00%	LG	-5621,869644	0,39287	0,405895	0	0	ACCEPTED

S2 Taxon information

Supplementary table S2: Statistics for the taxa in the multigene datasets. **Taxon name:** the name of the taxon as used in the multigene trees and alignment. **Composite:** Some closely related species have been merged into one sequence in the alignment. **Missing, 255G:** The percentage of missing data for the given taxon in the 255 gene alignment. **Missing, 146G:** The percentage of missing data for the given taxon in the 146 gene alignment. **Notes:** Taxa marked EPA have been placed with the Evolutionary Placement Algorithm (Berger, *et al.* 2011), taxa marked with RogueNaRok was identified as jumping taxa and removed before concatenation (Aberer *et al.*, 2013)

Taxon name	Composite	Missing, 255G	Missing, 146G	Notes
Acanthamoeba sp.	Acanthamoeba sp, Acanthamoeba astronyxis, Acanthamoeba castellanii	57	46	-
Alexandrium sp.	Alexandrium catenella, Alexandrium fundyense, Alexandrium tamarense	74	69	-
Ammonia sp.	Ammonia sp., Ammonia beccarii	49	32	-
Amorphochlora amoebiformis	-	35	14	-
Amphidinium carterae	-	89	84	-
Arabidopsis sp.	Arabidopsis lyrata, Arabidopsis thaliana	10	2	-
Astrolonche sp	-	72	60	-
Aulacantha scolymantha	-	92	88	-
Aureococcus anophagefferens	-	18	6	-
Batrachochytrium dendrobatidis	-	9	4	-
Bigelowiella longifila	-	52	37	-
Bigelowiella natans	-	12	2	-
Brachypodium distachyon	-	12	3	-
Branchiostoma floridae	-	6	3	-
Calliarthron tuberculosum	-	90	90	-
Cercomonas longicauda	-	84	80	-
Chlamydomonas reinhardtii	-	16	6	-
Chlorarachnion reptans	-	32	9	-
Chondrus crispus	-	85	81	-
Coccolithus braarudii	-	86	82	-
Coccomyxa sp	-	12	1	-
Collozoum sp	Collozoum sp., Collozoum inerme	93	90	-
Coralline	-	80	74	-
Corallomyxa sp	Corallomyxa sp., Corallomyxa tenera	92	88	-
Cryptosporidium parvum	-	33	15	-
Cyanophora paradoxa	-	71	65	-
Dictyostelium discoideum	-	11	4	-
Ectocarpus siliculosus	-	8	4	-
Eimeria tenella	-	91	88	-
Elphidium sp	Elphidium sp., Elphidium margaritaceum	50	32	-
Emiliana huxleyi	-	29	17	-
Eucheuma denticulatum	-	87	85	-
Euglena sp	Euglena gracilis, Euglena mutabilis	69	63	-
Fragilaria pinnata	-	86	79	-
Fragilariopsis cylindrus	-	67	51	-
Glaucocystis nostochinearum	-	73	66	-
Globobulimina turgida	-	88	81	-
Gracilaria changii	-	79	74	-
Gromia sphaerica	-	77	69	-
Guillardia theta	-	4	1	-
Gymnochlora sp	-	39	16	-
Histiona aroides	-	87	84	-

S2 Taxon information

Homo sapiens	-	1	1	-
Isochrysis galbana	-	81	74	-
Jakoba sp	Jakoba bahamensis, Jakoba libera	69	65	-
Karenia brevis	-	83	77	-
Laminaria sp	-	92	89	-
Lithomelissa setosa	-	94	90	-
Lotharella globosa	-	34	17	-
Lotharella oceanica	-	46	27	-
Mesostigma viride	-	83	81	-
Micromonas pusilla	-	45	46	-
Mimulus guttatus	-	16	6	-
Monosiga brevicollis	-	42	27	-
Naegleria gruberi	-	18	14	-
Nematostella vectensis	-	6	5	-
Neospora caninum	-	26	13	-
Nonionellina sp	Nonionellina sp, Nonionella labradorica	92	89	-
Norrsiella sphaerica	-	41	23	-
Oryza sativa	-	13	3	-
Ostreococcus sp	Ostreococcus lucimarinus, Ostreococcus tauri	20	8	-
Oxyrrhis marina	-	78	73	-
Paramecium tetraurelia	-	21	17	-
Partenskyella glossopodia	-	32	12	-
Pavlova lutheri	-	78	74	-
Perkinsus marinus	-	25	7	-
Phaeodactylum tricornutum	-	19	6	-
Phycomyces blakesleeanus	-	15	7	-
Phyllostaurus sp	Phyllostaurus sp., Phyllostaurus siculus	83	76	-
Physcomitrella patens	-	8	3	-
Phytophthora sp	Phytophthora ramorum, Phytophthora sojae	6	3	-
Plagioselmis sp	Plagioselmis sp., Plagioselmis nannoplanctica	75	73	-
Populus trichocarpa	-	13	6	-
Porphyra yezoensis	-	62	50	-
Porphyridium cruentum	-	50	36	-
Prymnesium parvum	-	80	74	-
Pythium oligandrum	-	84	76	-
Reclinomonas americana	-	61	57	-
Reticulomyxa filosa	-	50	36	-
Rhodomonas salina	-	91	89	-
Roombia truncata	-	32	32	-
Saprolegnia parasitica	-	10	6	-
Seculamonas ecuadoriensis	-	69	62	-
Sorghum bicolor	-	12	2	-
Spongosphaera streptacantha	-	91	87	-
Stachyamoeba lipophora	-	86	83	-
Sticholonche zancela	-	81	70	-
Tetrahymena pyriformis	-	19	14	-
Thalassiosira pseudonana	-	17	5	-
Theileria parva	-	77	67	-
Toxoplasma gondii	-	25	10	-
Volvox carteri	-	11	3	-
Allogromia sp.	-	99	99	EPA
Amphilonche elongata	-	95	92	EPA
Bodomorpha minima	-	99	99	EPA
Bonamia ostreae	-	99	99	EPA

S2 Taxon information

Brizalina sp	-	98	97	EPA
Bulimina marginata	-	95	93	EPA
Globigerinella siphonifera	-	99	99	EPA
Globigerinita glutinata	-	99	99	EPA
Gymnophrys sp	-	99	99	EPA
Haplosporidium littoralis	-	99	99	EPA
Haynesina germanica	-	99	99	EPA
Larcopyle butschlii	-	99	99	EPA
Leptophrys sp	-	99	99	EPA
Limnofila borokensis	-	97	96	EPA
Massisteria marina	-	99	99	EPA
Micatuba flexilis	-	99	99	EPA
Miliolidae sp	-	99	99	EPA
Plasmodiophora brassicae	-	94	92	EPA
Quinqueloculina sp	-	96	95	EPA
Rosseyatella sp	-	99	99	EPA
Spongospora subterranea	-	96	94	EPA
Blastocystis hominis	-	-	-	RogueNaRok
Cyanidioschyzon merolae	-	-	-	RogueNaRok
Galdieria sulphuraria	-	-	-	RogueNaRok
Schizochytrium sp.	-	-	-	RogueNaRok

S3 Concatenated alignments

Supplementary table S3: Statistics for the concatenated alignments. **Genes:** the number of genes in the alignment. **Characters:** The number of amino acids in the alignment. **Min.mis:** the lowest allowed percentage for missing data for a taxon in the alignment. **Cat.rem.:** the number of categories of fast evolving sites removed. **Taxa:** the number of taxa in the alignment. **Excluded taxa:** The name of excluded taxa, if any. **Method:** The method used for phylogenetic inference pb-Bayesian inference with Phylobayes mpi version 1.5a (Lartillot *et al.*, 2013), RAxML= maximum likelihood inference with RAxML v 8.0.26 (Stamatakis 2014). **Model:** The model used for the inference. **Max.diff:** (only for Bayesian), maximum difference in posterior probability support between two chains that ran independently. **TCA_rel:** (only for maximum likelihood), The relative Tree certainty index for all nodes (Salichos *et al.*, 2014), calculated on the bootstrap files for the inference.

Table 3A: 255 Genes, Bayesian

alignment name	Genes	Characters	min.mis.	Cat.rem	taxa	Excluded taxa	Method	Model	Max diff
255G_min10	255	54898	10	0	91	0	pb	CATGTR	0,266411
255G_min10_rmcat4	255	54898	10	4	91	0	pb	CATGTR	0,30002

Table 3B: 147 Genes, Bayesian

alignment name	Genes	Characters	min.mis.	#cat.rem	taxa	Excluded taxa	Method	Model	Max diff
146G_min10.CATGTR	146	33081	10	0	91	0	pb	CATGTR	0,286544
146G_min10.LG	146	33081	10	0	91	0	pb	LG	0,355603
146G_min10_exSpongo.CATGTR	146	32952	10	0	90	<i>S. streptacantha</i>	pb	CATGTR	0,171197
146G_min10_exSticho.CATGTR	146	32952	10	0	90	<i>S. zancea</i>	pb	CATGTR	0,29748

Table 3C: 255 Genes, Maximum Likelihood.

alignment name	Genes	Characters	min.mis.	#cat.rem	taxa	Excluded taxa	Method	Model	TCA_rel
255G	255	54898	0	0	124	0	RAxML	LG	0,565162
255G_min10	255	54898	10	0	91	0	RAxML	LG	0,891783
255G_min20	255	54898	20	0	69	0	RAxML	LG	0,879462
255G_min30	255	54898	30	0	55	0	RAxML	LG	0,881501
255G_min10_rmcat1	255	54825	10	1	91	0	RAxML	LG	0,886363
255G_min10_rmcat2	255	54247	10	2	91	0	RAxML	LG	0,899755
255G_min10_rmcat3	255	52502	10	3	91	0	RAxML	LG	0,888871
255G_min10_rmcat4	255	48410	10	4	91	0	RAxML	LG	0,842282
255G_min10_exSpongo	255	54898	10	0	90	<i>S. streptacantha</i>	RAxML	LG	0,893954
255G_min10_exSpongo_rmcat1	255	54814	10	1	90	<i>S. streptacantha</i>	RAxML	LG	0,908864
255G_min10_exSpongo_rmcat2	255	54131	10	2	90	<i>S. streptacantha</i>	RAxML	LG	0,910279
255G_min10_exSpongo_rmcat3	255	52157	10	3	90	<i>S. streptacantha</i>	RAxML	LG	0,897523
255G_min10_exSpongo_rmcat4	255	47572	10	4	90	<i>S. streptacantha</i>	RAxML	LG	0,852508
255G_min10_exSticho	255	54898	10	0	90	<i>S. zancea</i>	RAxML	LG	0,900949
255G_min10_exSticho_rmcat1	255	54831	10	1	90	<i>S. zancea</i>	RAxML	LG	0,904303
255G_min10_exSticho_rmcat2	255	54309	10	2	90	<i>S. zancea</i>	RAxML	LG	0,906108
255G_min10_exSticho_rmcat3	255	52624	10	3	90	<i>S. zancea</i>	RAxML	LG	0,883426
255G_min10_exSticho_rmcat4	255	48585	10	4	90	<i>S. zancea</i>	RAxML	LG	0,817331
255G_min10_exSpongo_exSticho	255	54898	10	0	89	<i>S. strep. and S. zancea</i>	RAxML	LG	0,901578
255G_min10_exSpongo_exSticho_rmcat1	255	54823	10	1	89	<i>S. strep. and S. zancea</i>	RAxML	LG	0,898729
255G_min10_exSpongo_exSticho_rmcat2	255	54211	10	2	89	<i>S. strep. and S. zancea</i>	RAxML	LG	0,902772
255G_min10_exSpongo_exSticho_rmcat3	255	52308	10	3	89	<i>S. strep. and S. zancea</i>	RAxML	LG	0,895606
255G_min10_exSpongo_exSticho_rmcat4	255	47774	10	4	89	<i>S. strep. and S. zancea</i>	RAxML	LG	0,871055
255G_min10_exForams	255	54898	10	0	86	Foraminifera	RAxML	LG	0,896866
255G_min10_exForams_rmcat1	255	54842	10	1	86	Foraminifera	RAxML	LG	0,903465
255G_min10_exForams_rmcat2	255	54188	10	2	86	Foraminifera	RAxML	LG	0,902428
255G_min10_exForams_rmcat3	255	52222	10	3	86	Foraminifera	RAxML	LG	0,887357
255G_min10_exForams_rmcat4	255	47464	10	4	86	Foraminifera	RAxML	LG	0,848266

Table 3D: 147 Genes, Maximum Likelihood.

alignment name	Genes	Characters	min.mis.	#cat.rem	taxa	Excluded taxa	Method	Model	TCA_rel
146G	146	33081	0	0	124	0	RAxML	LG	0,544504
146G_rmcat1	146	25236	0	1	124	0	RAxML	LG	0,436772
146G_rmcat2	146	13424	0	2	124	0	RAxML	LG	0,204149
146G_rmcat3	146	8470	0	3	124	0	RAxML	LG	0,074327
146G_min10	146	32952	10	0	91	0	RAxML	LG	0,854739
146G_min20	146	33081	20	0	74	0	RAxML	LG	0,860809
146G_min30	146	33081	30	0	62	0	RAxML	LG	0,843727
146G_min10_rmcat1	146	33007	10	1	91	0	RAxML	LG	0,856469
146G_min10_rmcat2	146	32344	10	2	91	0	RAxML	LG	0,87058
146G_min10_rmcat3	146	30512	10	3	91	0	RAxML	LG	0,829692
146G_min10_rmcat4	146	26726	10	4	91	0	RAxML	LG	0,77852
146G_min10_exSpongo	146	32952	10	0	90	<i>S. streptacantha</i>	RAxML	LG	0,865018
146G_min10_exSpongo_rmcat1	146	32869	10	1	90	<i>S. streptacantha</i>	RAxML	LG	0,870338
146G_min10_exSpongo_rmcat2	146	32109	10	2	90	<i>S. streptacantha</i>	RAxML	LG	0,873687
146G_min10_exSpongo_rmcat3	146	30202	10	3	90	<i>S. streptacantha</i>	RAxML	LG	0,82786
146G_min10_exSpongo_rmcat4	146	26177	10	4	90	<i>S. streptacantha</i>	RAxML	LG	0,787182
146G_min10_exSticho	146	33081	10	0	90	<i>S. zancea</i>	RAxML	LG	0,873601

S3 Concatenated alignments

146G_min10_exSticho_rmcat1	146	33021	10	1	90	<i>S. zanclea</i>	RAxML	LG	0,890642
146G_min10_exSticho_rmcat2	146	32483	10	2	90	<i>S. zanclea</i>	RAxML	LG	0,888718
146G_min10_exSticho_rmcat3	146	30767	10	3	90	<i>S. zanclea</i>	RAxML	LG	0,863364
146G_min10_exSticho_rmcat4	146	27153	10	4	90	<i>S. zanclea</i>	RAxML	LG	0,791703
146G_exSpongo_exSticho_min10	146	32952	10	0	89	<i>S. strep. and S. zanclea</i>	RAxML	LG	0,878715
146G_exSpongo_exSticho_min10_rmcat1	146	32887	10	1	89	<i>S. strep. and S. zanclea</i>	RAxML	LG	0,876233
146G_exSpongo_exSticho_min10_rmcat2	146	32258	10	2	89	<i>S. strep. and S. zanclea</i>	RAxML	LG	0,879483
146G_exSpongo_exSticho_min10_rmcat3	146	30450	10	3	89	<i>S. strep. and S. zanclea</i>	RAxML	LG	0,85276
146G_exSpongo_exSticho_min10_rmcat4	146	26609	10	4	89	<i>S. strep. and S. zanclea</i>	RAxML	LG	0,795294
146G_min10_exForam	146	33081	10	0	86	Foraminifera	RAxML	LG	0,874225
146G_min10_exForam_rmcat1	146	33026	10	1	86	Foraminifera	RAxML	LG	0,869218
146G_min10_exForam_rmcat2	146	32328	10	2	86	Foraminifera	RAxML	LG	0,878678
146G_min10_exForam_rmcat3	146	30417	10	3	86	Foraminifera	RAxML	LG	0,847643
146G_min10_exForam_rmcat4	146	26443	10	4	86	Foraminifera	RAxML	LG	0,764009

Supplementary table S4: The posterior probability (bayesian) and bootstrap support (maximum likelihood) for selected nodes. The support values has been heatmapped with a three color scheme, Green color= high support, yellow=medium support, red= low support. NM= not monophyletic, i.e. the clade or node does not exist for that particular inference. NA=Not applicable, i.e. the clade or node does not exist because some taxa has been removed. For the specifics for each alignment refer to table S3. Tax=Taxopodida; Rad=Radiolaria; Foram=Foraminifer; SAR=Stramenopile, Alveolates, Rhizaria; Red=Rhodophyta; Green=Chlorophytes; Glauco=Glaucophyta; Crypto=Cryptophyta

Table 4A: 255 Genes, Bayesian

alignment name	Topologies/Clades											
	Rhizaria	(Tax, Rad) *Foram	Tax* (Rad,Foram)	Endomyxa+R etaria	(SA)*R	S*(AR)	(SAR) Hapt (Red + Green)	Green+ Glauco	Red+Green+ Glauco	Red+Green+ Glauco+Crypto	Hapt+Crypto	
255G_min10.CATGTR	1	NM	0,71	1	1	NM	0,87	1	NM	1	1	NM
255G_min10_rmc4.CATGTR	1	NM	0,97	1	1	NM	NM	1	NM	0,96	0,93	NM

Table 4B: 147 Genes, Bayesian

alignment name	Topologies/Clades											
	Rhizaria	(Tax, Rad) *Foram	Tax* (Rad,Foram)	Endomyxa+R etaria	(SA)*R	S*(AR)	(SAR) Hapt (Red + Green)	Green+ Glauco	Red+Green+ Glauco	Red+Green+ Glauco+Crypto	Hapt+Crypto	
146G_min10.CATGTR	1	NM	0,81	1	1	NM	1	0,84	NM	1	1	NM
146G_min10.LG	1	0,67	NM	1	1	NM	1	0,53	NM	1	1	NM
146G_min10_exSpongo.CATGTR	1	1	NM	1	1	NM	1	0,93	NM	0,93	1	NM
146G_min10_exSticho.CATGTR	1	NA	NA	1	1	NM	1	0,82	NM	0,95	1	NM

Table 4C: 255 Genes, Maximum Likelihood.

alignment name	Topologies/Clades											
	Rhizaria	(Tax, Rad) *Foram	Tax* (Rad,Foram)	Endomyxa+R etaria	(SA)*R	S*(AR)	(SAR) Hapt (Red + Green)	Green+ Glauco	Red+Green+ Glauco	Red+Green+ Glauco+Crypto	Hapt+Crypto	
255G	NM	63	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
255G_min10	100	88	NM	98	NM	96	79	81	NM	NM	79	NM
255G_min20	100	51	NM	94	NM	96	90	79	NM	NM	90	NM
255G_min30	100	NA	NA	100	NM	97	77	92	100	92	77	NM
255G_min10_rmc1	100	65	NM	98	NM	94	88	94	NM	NM	90	NM
255G_min10_rmc2	100	69	NM	100	NM	94	77	77	NM	NM	79	NM
255G_min10_rmc3	100	NM	25	100	NM	94	85	85	NM	NM	92	NM
255G_min10_rmc4	100	NM	50	90	NM	67	NM	79	NM	NM	50	NM
255G_min10_exSpongo	100	71	NM	100	NM	94	81	81	NM	NM	81	NM
255G_min10_exSpongo_rmc1	100	67	NM	98	NM	98	90	88	NM	NM	90	NM
255G_min10_exSpongo_rmc2	100	67	NM	98	NM	96	90	96	NM	NM	92	NM
255G_min10_exSpongo_rmc3	100	48	NM	98	NM	96	85	90	NM	NM	88	NM
255G_min10_exSpongo_rmc4	100	49	NM	95	48	NM	NM	67	NM	46	NM	33
255G_min10_exSticho	100	NA	NA	98	NM	94	81	88	NM	NM	83	NM
255G_min10_exSticho_rmc1	100	NA	NA	100	NM	98	90	85	NM	NM	83	NM
255G_min10_exSticho_rmc2	100	NA	NA	100	NM	100	88	87	NM	NM	88	NM
255G_min10_exSticho_rmc3	100	NA	NA	98	NM	92	83	87	NM	NM	90	NM
255G_min10_exSticho_rmc4	100	NA	NA	94	NM	71	NM	83	NM	NM	44	NM
255G_min10_exSpongo_exSticho	100	NA	NA	98	NM	100	83	81	NM	NM	83	NM
255G_min10_exSpongo_exSticho_rmc1	100	NA	NA	98	NM	100	85	83	NM	NM	87	NM
255G_min10_exSpongo_exSticho_rmc2	100	NA	NA	98	NM	100	87	90	NM	NM	90	NM
255G_min10_exSpongo_exSticho_rmc3	100	NA	NA	97	NM	83	77	85	NM	NM	90	NM
255G_min10_exSpongo_exSticho_rmc4	100	NA	NA	100	NM	60	NM	60	NM	44	NM	37
255G_min10_exForams	100	NA	NA	100	NM	92	83	83	NM	NM	85	NM
255G_min10_exForams_rmc1	100	NA	NA	100	NM	85	87	92	NM	NM	88	NM
255G_min10_exForams_rmc2	100	NA	NA	100	NM	92	96	83	NM	NM	96	NM
255G_min10_exForams_rmc3	100	NA	NA	100	NM	85	94	92	NM	NM	96	NM
255G_min10_exForams_rmc4	100	NA	NA	100	57	NM	NM	64	NM	59	NM	42

Table 4D: 147 Genes, Maximum Likelihood.

alignment name	Topologies/Clades											
	Rhizaria	(Tax, Rad) *Foram	Tax* (Rad,Foram)	Endomyxa+R etaria	(SA)*R	S*(AR)	(SAR) Hapt (Red + Green)	Green+ Glauco	Red+Green+ Glauco	Red+Green+ Glauco+Crypto	Hapt+Crypto	
146G	NM	62	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
146G_rmc1	NM	53	NM	NM	60	NM	NM	NM	NM	NM	NM	NM
146G_rmc2	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
146G_rmc3	NM	4	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM
146G_min10	100	87	NM	100	NM	85	69	44	NM	46	67	NM
146G_min20	100	64	NM	100	NM	85	72	NM	55	42	73	NM
146G_min30	100	NA	NA	NA	NM	95	66	NM	58	47	69	NM
146G_min10_rmc1	100	85	NM	100	NM	90	79	48	NM	65	77	NM
146G_min10_rmc2	100	58	NM	100	NM	92	90	NM	46	71	85	NM
146G_min10_rmc3	100	27	NM	96	NM	71	85	NM	60	69	88	NM
146G_min10_rmc4	100	NM	75	97	74	NM	NM	48	NM	52	NM	29
146G_min10_exSpongo	100	86	NM	NM	NM	86	69	39	NM	53	70	NM
146G_min10_exSpongo_rmc1	100	79	NM	100	NM	87	65	42	NM	58	69	NM
146G_min10_exSpongo_rmc2	100	83	NM	100	NM	96	88	50	NM	65	90	NM
146G_min10_exSpongo_rmc3	100	63	NM	98	NM	78	78	56	NM	NM	84	NM
146G_min10_exSpongo_rmc4	100	51	NM	97	70	NM	NM	NM	NM	74	NM	38
146G_min10_exSticho	100	NA	NA	100	NM	90	67	46	NM	52	69	NM
146G_min10_exSticho_rmc1	100	NA	NA	100	NM	92	83	42	NM	65	71	NM
146G_min10_exSticho_rmc2	100	NA	NA	100	NM	94	85	65	NM	62	83	NM
146G_min10_exSticho_rmc3	100	NA	NA	98	NM	75	88	NM	46	48	90	NM
146G_min10_exSticho_rmc4	100	NA	NA	96	67	NM	NM	52	NM	58	NM	35
146G_exSpongo_exSticho_min10	100	NA	NA	100	NM	85	71	40	NM	67	73	NM
146G_exSpongo_exSticho_min10_rmc1	100	NA	NA	100	NM	96	69	42	NM	56	71	NM
146G_exSpongo_exSticho_min10_rmc2	100	NA	NA	100	NM	94	79	52	NM	65	81	NM
146G_exSpongo_exSticho_min10_rmc3	100	NA	NA	100	NM	79	79	NM	52	63	85	NM
146G_exSpongo_exSticho_min10_rmc4	100	NA	NA	98	66	NM	NM	NM	NM	63	NM	38
146G_min10_exForam	100	NA	NA	100	NM	85	75	44	NM	67	71	NM
146G_min10_exForam_rmc1	100	NA	NA	100	NM	94	73	38	NM	52	75	NM
146G_min10_exForam_rmc2	100	NA	NA	100	NM	87	87	NM	62	73	79	NM
146G_min10_exForam_rmc3	100	NA	NA	100	NM	82	83	61	NM	NM	91	NM
146G_min10_exForam_rmc4	100	NA	NA	89	68	NM	52	43	NM	63	56	NM