Published in final edited form as:
Mol Biol Evol. 2016 November ; 33(11): 2890–2898. doi:10.1093/molbev/msw168.

A phylogenomic framework to study the diversity and evolution of stramenopiles (=heterokonts)

Romain Derelle[*], **Purificación López-García**, **Hélène Timpano**[1], and **David Moreira**[1,*]

[1]Unité d'Ecologie, Systématique et Evolution, Centre National de la Recherche Scientifique (CNRS), Université Paris-Sud/Paris-Saclay, AgroParisTech, 91400 Orsay, France

Abstract

Stramenopiles or heterokonts constitute one of the most speciose and diverse clades of protists. It includes ecologically important algae (such as diatoms or large multicellular brown seaweeds), as well as heterotrophic (e.g. bicosoecids, MAST groups) and parasitic (e.g. *Blastocystis*, oomycetes) species. Despite their evolutionary and ecological relevance, deep phylogenetic relationships among stramenopile groups, inferred mostly from small-subunit (SSU) rDNA phylogenies, remain unresolved, especially for the heterotrophic taxa. Taking advantage of recently released stramenopile transcriptome and genome sequences, as well as data from the genomic assembly of the MAST-3 species *Incisomonas marina* generated in our laboratory, we have carried out the first extensive phylogenomic analysis of stramenopiles, including representatives of most major lineages. Our analyses, based on a large dataset of 339 widely distributed proteins, strongly support a root of stramenopiles lying between two clades, Bigyra and Gyrista (Pseudofungi plus Ochrophyta). Additionally, our analyses challenge the Phaeista-Khakista dichotomy of photosynthetic stramenopiles (ochrophytes) as two groups previously considered to be part of the Phaeista (Pelagophyceae and Dictyochophyceae), branch with strong support with the Khakista (Bolidophyceae and Diatomeae). We propose a new classification of ochrophytes within the two groups Chrysista and Diatomista to reflect the new phylogenomic results. Our stramenopile phylogeny provides a robust phylogenetic framework to investigate the evolution and diversification of this group of ecologically relevant protists.

Keywords

Bigyra; Gyrista; Diatomista; Chrysista; phylogenomics

Introduction

Stramenopiles (Patterson 1989), also known as Heterokonts (Cavalier-Smith 1986a), constitute one of the major eukaryotic clades, branching with Rhizaria and Alveolata within the ‘supergroup’ SAR (Burki, et al. 2007; Adl, et al. 2012; Burki 2014), also called Harosa (Cavalier-Smith 2010). The stramenopiles encompass an extremely large diversity of organisms that include, among others, free-living flagellates, parasites such as *Blastocystis hominis*, organisms resembling fungi regarding their cytology and ecology, and a myriad of

[*]corresponding authors: romain.derelle@u-psud.fr and david.moreira@u-psud.fr.

Europe PMC Funders Author Manuscripts

photosynthetic lineages that range from single-cell diatoms to giant multicellular brown algae (kelp). In addition, recent environmental surveys have revealed that stramenopiles represent a significant fraction of the poorly known marine picoeukaryotes, most of which are heterotrophic organisms (Massana, et al. 2014; Pernice, et al. 2015). These unicellular organisms, originally named MAST for MArine STramenopiles, form several independent lineages among stramenopiles (Massana, et al. 2004; Massana, et al. 2014), several of which are also present in freshwater systems (Simon et al., 2015). This extraordinary diversity makes of stramenopiles an ideal choice for evolutionary studies aimed at exploring the genomic modifications underlying drastic morphological changes or adaptations to different lifestyles, for instance the acquisition of new genes by lateral gene transfer (Bowler, et al. 2008; Richards, et al. 2011; Tsaousis, et al. 2012; Roy, et al. 2014). However, such large-scale comparative genomic studies require a reliable phylogenetic framework that is still lacking for this eukaryotic group.

Numerous stramenopile groups have been defined since a long time ago based on the presence of distinctive phenotypic characters (e.g. diatoms). For many others, in particular the extremely diverse heterotrophic flagellate species, morphological differences are not always discriminatory so that, currently, classification is largely based on small-subunit (SSU) rDNA phylogenies. These phylogenies consistently point to the monophyly of photosynthetic stramenopiles, a clade named Ochrophyta, and the position of Pseudofungi (i.e. oomycetes and their flagellate relatives) as sister-group to Ochrophyta. These two groups, Ochrophyta and Pseudofungi, form the clade Gyrista (Cavalier-Smith 1998). Ochrophyta are most often divided in the two groups Khakista (diatoms and bolidophytes) and Phaeista (including the remaining photosynthetic groups)(e.g. Brown and Sorhannus 2010; Gomez, et al. 2011; Cavalier-Smith and Scoble 2013; Massana, et al. 2014), although this classification has been challenged by several phylogenetic studies based on different markers (Riisberg, et al. 2009; Yang, et al. 2012; Ševčíková, et al. 2015). The relationships among the rest of stramenopile groups, all of them heterotrophic (e.g., labyrinthulomycetes and thraustochytrids, bicosoecids, the parasite *Blastocystis*, and most MAST lineages), are still unclear as they vary from one phylogenetic analysis to another. In some analyses, those non-Gyrista lineages form two monophyletic groups, Opalozoa and Sagenista, branching in a successive pattern at the base of stramenopiles (Massana, et al. 2004; Gomez, et al. 2011; Massana, et al. 2014). Alternatively, Opalozoa and Sagenista group together in some studies to form a clade called Bigyra, the stramenopiles being in this scenario divided into Bigyra and Gyrista (Cavalier-Smith and Chao 2006; Riisberg, et al. 2009). Those analyses retrieved a very weak statistical support for those deep relationships, which explains the contradictory results observed.

Over the last decade, the phylogenomic approach has proven to be an efficient alternative to rDNA phylogenies when this marker reaches its limits of resolution at deep eukaryotic relationships (e.g. Hampl, et al. 2009; Sierra, et al. 2013; Gentekaki, et al. 2014; Cavalier-Smith, et al. 2015). Its more popular variant consists on combining dozens to hundreds of protein alignments into one single matrix, increasing the quantity of phylogenetic signal to be analyzed. This approach also allows the use of complex evolutionary models with a very precise parameter estimation (e.g. time or site heterogeneous models) (Delsuc, et al. 2005). However no phylogenomic study of stramenopiles has been carried out so far due to the lack

Europe PMC Funders Author Manuscripts

of sequencing data for most non-Gyrista lineages. Indeed, stramenopiles are mostly represented by a few ochrophyte and oomycete species in recent phylogenomic studies of eukaryotes (e.g. Grant and Katz 2014; Cavalier-Smith, et al. 2015; Derelle, et al. 2015; Janouškovec, et al. 2015; Katz and Grant 2015). Therefore, a better taxonomic sampling of non-Gyrista species is of paramount importance to infer a reliable phylogenomic tree of stramenopiles.

In this study we performed the first phylogenomic study specifically focused on stramenopiles. Using a large collection of eukaryotic phylogenetic markers (339 conserved proteins), we took advantage of recent sequencing projects to assemble a large phylogenomic matrix representing most stramenopile lineages. In addition to the few stramenopile species with complete genome sequences available, our dataset was enriched with the transcriptomes of ochrophytes and heterotrophic species produced by the Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling, et al. 2014), the single-cell genome of a MAST-4 species (Roy, et al. 2014), and the genome of *Incisomonas marina*, a species belonging to the group MAST-3 (Cavalier-Smith and Scoble 2013), which we sequenced and assembled in our laboratory. We analyzed this phylogenomic matrix using state-of-the-art phylogenomic methods. This allowed us to resolve two major questions: the position of the root of stramenopiles, and the relationships among ochrophytes.

Results

Phylogenomic dataset and strategy of analyses

We have built a phylogenomic dataset of 339 protein alignments selected from a large in-house collection of eukaryotic phylogenetic markers, resulting in a large concatenated matrix of 88,456 conserved amino acids for 45 species (39 stramenopiles + 6 outgroup species; see Materials and methods). Particular attention was paid to the detection and removal of contaminants, as this type of outlier is predominant in some of the data analyzed here (see Supplementary file 1). Our taxonomic sampling included an outgroup restricted to slow-evolving species belonging to the two closest lineages to stramenopiles, namely alveolates and rhizaria, and an ingroup composed of species representing most stramenopile lineages and for which genomic or large transcriptomic data were available. As a result, our matrix showed a high level of completeness with an average of 84.9% of data per species (see Supplementary file 1). The only species with less than 70% of data was the MAST-4 representative because its genome, obtained by single-cell genomics techniques, was very incomplete (Roy, et al. 2014).

The relatively restricted number of species in our large sequence dataset allowed us to perform calculation-intensive Bayesian analyses under the site-heterogeneous models (CAT-GTR and CAT-Poisson models) assumed to be the best-fitting models as shown by some of the phylogenetic studies from which a large part of the markers used in our work originated (see Material and Methods). The results of these Bayesian analyses are shown in Figure 1. Maximum-likelihood (ML) analyses were performed under the mixture model LG4X (Le, et al. 2012) and are summarized in Figure 2 (ML trees are available in Supplementary file 2). Some ML analyses were repeated under the classical GTR model to test the influence of model selection on our phylogenetic inferences. ML trees obtained under the GTR model

Europe PMC Funders Author Manuscripts

were identical to those obtained under the LG4X model (identical topologies and similar support values; see Supplementary file 2), indicating that the ML results presented below were not affected by the choice of the evolutionary model.

The root of stramenopiles

The consensus tree obtained in Bayesian analysis under the CAT-GTR model, which is expected to be the best fitting model to this dataset, showed the expected monophyly of Gyrista (Figure 1). Opalozoa (i.e. Placidozoa and Bikosia) and Sagenista (i.e. MAST-4 species and Labyrinthulea) were also found to be monophyletic, and these two clades branched together to form the group Bigyra. All branches received maximal support (i.e. posterior probabilities equal to 1). In contrast, the two MCMC chains ran using the CAT-Poisson model did not converge (maxdiff=1): the clade Bigyra was recovered by one chain while the other chain showed a position of Opalozoa as sister-group to all other stramenopiles (see Supplementary file 2). In both chains, all other stramenopile relationships were identical to those obtained under the CAT-GTR model. Finally, we ran a third chain that converged with the one showing the monophyly of Bigyra (Figure 1). These Bayesian analyses therefore suggest a position of the root that divides the stramenopiles in the two clades Bigyra and Gyrista.

The ML tree differed from the Bayesian topology by the paraphyly of Bigyra, with Opalozoa branching as sister group to all other stramenopiles (Supplementary file 2). This topology was however poorly supported as it was observed in only 73% of bootstrap replicates, whereas the other 27% bootstrap replicates supported the dichotomy Bigyra-Gyrista. To check if these conflicting results reflected the presence of noise in our dataset, we performed a removal of fast evolving sites in 2% increments, up to 20% of the complete dataset, and analyzed bootstrap replicates at each shortening step to measure the change in support for each of the two topologies shown in Figure 2A (see Material and Methods). The rationale of this approach was to progressively increase the ratio of phylogenetic signal versus noise in our dataset, expecting that the support for the correct topology will progressively increase along this set of analyses. As shown in the first graphic of Figure 2B, the support for the position of Opalozoa as sister-group of all other stramenopiles decreased in these analyses, with the dichotomy Bigyra-Gyrista becoming the best ML topology at 8% of fast evolving sites removed. However, neither of the alternative topologies became highly supported, with the support for both topologies later converging to a range of 35% to 65%. These results suggested that either our dataset contained a weak phylogenetic signal insufficient to discriminate between the two topologies, or that an artifact was affecting the ML analyses. Looking at the Bayesian and ML trees, it was noticeable that Opalozoa included the most divergent stramenopile species of our dataset, namely the two *Blastocystis* species and *Cafeteria roenbergensis*. We can therefore hypothesize that, due to the presence of these divergent lineages, a Long Branch Attraction (LBA) artifact tended to attract Opalozoa towards the relatively distant outgroup. In order to test this hypothesis, we successively removed these divergent lineages from our dataset and repeated all ML analyses to estimate their impact on the topology and statistical supports. The ML tree obtained from the full dataset without *Blastocystis* showed a shift of topology, with Bigyra becoming monophyletic (Supplementary file 2). As shown in the second graphic of Figure

Europe PMC Funders Author Manuscripts

2B, support for this topology increased from 71% up to 94% along the removal of fast evolving sites. Finally the additional removal of *C. roenbergensis* strengthened this pattern, with the dichotomy Bigyra-Gyrista receiving up to 99% of support (third graphic of Figure 2B), albeit with only three opalazoan species left in the dataset. Altogether, these results strongly indicate (i) the presence of a LBA artifact in the ML analyses including *C. roenbergensis* and *Blastocystis*, and (ii) that modifications of our dataset to minimize the impact of this artifact led to the stramenopile root obtained in Bayesian analyses, i.e. the dichotomy Bigyra-Gyrista.

Relationships among ochrophyte lineages

Bayesian and ML analyses converged to the same primary dichotomy of ochrophytes, in all cases with maximal statistical support: Pelagophyceae, Dictyochophyceae, Bolidophyceae and diatoms on one side, and all other ochrophytes on the other side (see Figure 1 and Supplementary file 2). This result was in contradiction with the commonly observed, but weakly supported, dichotomy between Khakista (Bolidophyceae plus diatoms) and Phaeista (all other ochrophytes) found in SSU rDNA-based trees (e.g. Brown and Sorhannus 2010; Gomez, et al. 2011; Cavalier-Smith and Scoble 2013; Massana, et al. 2014). By contrast, it agreed with previous multigene phylogenetic analyses (Riisberg, et al. 2009; Yang, et al. 2012; Ševčíková, et al. 2015) although, since our tree was based on a much larger gene sampling than those previously used, we were able to retrieve maximal support for all branches within this group (Figure 1). We thus conclude that the two clades dividing Ochrophyta in our analysis are likely to be correct. We propose to name these two groups Diatomista and Chrysista, respectively (see Discussion below for the rationale for these two new names).

To further asses the robustness of the Diatomista-Chrysista division, we screened the ML bootstrap trees obtained along the removal of fast evolving sites and the removal of divergent Bigyra species to look for occurrences of the Phaeista-Khakista dichotomy, but we did not find any since the Diatomista-Chrysista dichotomy receives maximal support in all of these analyses (see for instance ML trees in Supplementary file 2). We finally repeated ML analyses based on our dataset without divergent lineages (i.e. alveolates, rhizaria, *Blastocystis* and *C. roenbergensis*), given that long branches might have a negative impact on the ingroup relationships. Again, the Diatomista-Chrysista dichotomy was obtained with maximal support (see Supplementary file 2). We can therefore conclude that the topology obtained here is not an artifact due to the presence of a distant outgroup in our phylogenomic dataset.

Discussion

Deep stramenopile relationships in the light of phylogenomics

Despite impressive taxonomic sampling, phylogenetic studies based on SSU rDNA have failed to resolve many uncertainties in the phylogeny of stramenopiles. Here, we report the first phylogenomic study focusing on this group of protists that resolves the inconsistencies encountered by previous phylogenetic studies.

First, our phylogenomic analyses clarify the deep relationships of stramenopiles, pinpointing a root lying between the two clades Bigyra and Gyrista (Figure 1). The monophyly of Gyrista is obtained with maximal support in all of our analyses, in agreement with all phylogenetic studies published so far. In contrast, the monophyly of Bigyra, a grouping that has never been recovered as a strongly supported clade (Cavalier-Smith and Chao 2006; Riisberg, et al. 2009), is obtained here with maximal support in our Bayesian analyses, although with convergence issues (only 2 out of 3 chains converge) under the CAT-Poisson model. The clade Bigyra is also recovered in ML analyses when a LBA artifact caused by very divergent Placidozoa lineages is alleviated by removing these lineages from our dataset. These results illustrate the power of the phylogenomic approach: given the high number of positions analyzed, it allows the use of site-heterogeneous models less sensitive to LBA artifacts (Lartillot, et al. 2007) in a Bayesian framework and the possibility to eliminate such artifacts by progressively removing fast evolving sites or species from the dataset when analyzed in ML under simpler models (e.g. Brinkmann, et al. 2005; Hampl, et al. 2009). Importantly enough, the genus *Cafeteria* was not found monophyletic in our analyses, with *C. roenbergensis* being part of Bikosea whereas '*Cafeteria* sp.' branched with *Incisomonas* and *Blastocystis*. The species '*Cafeteria* sp.' has not been formally described and, as suggested by our analyses, its transfer to a different genus will be required to correct this misidentification. Indeed, 18S rDNA gene phylogeny with a large taxonomic sampling indicates that '*Cafeteria* sp.' is closely related to *Wobblia*, a genus within the Placidozoa (see Supplementary file 3).

The second important result emerging from our study concerns the relationships among Ochrophyte lineages. While all phylogenies based on SSU rDNA have shown the Khakista-Phaeista dichotomy (e.g. Brown and Sorhannus 2010; Gomez, et al. 2011; Cavalier-Smith and Scoble 2013; Massana, et al. 2014), our Bayesian and ML analyses all converge with maximal support to the paraphyly of Phaeista, with Pelagophyceae and Dictyochophyceae branching as sister group of the Khakista. Interestingly, a possible relationship between diatoms, pelagophytes and dictyochophytes had already been proposed based on their reduced flagellar apparatus (Saunders, et al. 1995). These relationships within ochrophytes strongly confirm the results obtained from phylogenetic analyses based on much smaller datasets: SSU and LSU rDNA plus five proteins (Riisberg, et al. 2009); SSU rDNA plus four proteins (Yang, et al. 2012); and 34 plastid proteins (Ševčíková, et al. 2015). Finally, phylogenomic analyses of eukaryotes that include Dictyophyceae or Pelagophyceae species also agree with this topology and contradict with the Khakista-Phaeista dichotomy (e.g. Burki, et al. 2012; Grant and Katz 2014; Katz and Grant 2015). However, in contrast with the publications previously cited, this topology is obtained in all of our analyses with maximal statistical support. This result has implications for our view of the evolution of a very characteristic morphological feature: the transition helix found in the flagella of many species (Cavalier-Smith 1998). It was a major character used to differentiate the Phaeista (with a single transition helix) from the Khakista (without transition helix) (Cavalier-Smith and Chao 2006). Our results support the hypothesis that the ancestral ochrophyte had one transition helix (Cavalier-Smith and Chao 2006) so that its absence in Khakista simply represents a loss.

Insights into the taxonomy of stramenopiles

Adl and colleagues published in 2012 an updated version of their classification of eukaryotes (Adl, et al. 2012) in which stramenopiles were represented by a large polytomy, illustrating the poor resolution observed at the base of this group in SSU rDNA phylogenies. By contrast, Cavalier-Smith developed a classification of stramenopiles (Cavalier-Smith 1998; Cavalier-Smith and Chao 2006; Cavalier-Smith and Scoble 2013), in which all stramenopile lineages were classified into hierarchical groups. We recovered in our phylogenomic analyses the monophyly of most of those groups (e.g. Bigyra, Opalozoa, Placidozoa, Sagenista, Gyrista), the only exception being the group Phaeista (Ochrophyta) found paraphyletic in our analyses. Therefore, in the present study we have followed the nomenclature used by Cavalier-Smith and Scoble 2013 for those deep-branching groups, with two deviations: (i) we did not follow the Linnaean ranking system above genera (e.g. phylum, class, order) and (ii) we did not use the name Phaeista.

The relationships among ochrophytes revealed by our study call for a change of the existing nomenclature. Riisberg and colleagues have proposed to keep the two names Phaeista and Khakista, and to modify their meaning by transferring Pelagophyceae and Dictyochophyceae from Phaeista to Khakista (Riisberg, et al. 2009). We find this option rather confusing, as identical names would have different definitions depending on the publication cited. For this reason, we propose to keep using the name Khakista in its original sense (i.e. for the Diatomeae plus Bolidophyceae group ; characterized by the absence of flagellar transition helix as explained above), to abandon the name Phaeista, and to define two new formal taxa using branch-based phylogenetic definitions. These two taxa are defined by the phylogeny of ochrophytes obtained in this study as follows:

> **Diatomista**: The most inclusive clade containing *Thalassiosira pseudonana*, Hasle and Heimdal (1970) (Diatomeae); *Pelagomonas calceolata*, Andersen and Saunders (1993) (Pelagophyceae); and *Dictyocha speculum*, Ehrenberg (1839) (Dictyophyceae); but not *Ectocarpus siliculosus*, Lyngbye (1819) (Phaeophyceae); *Synchroma pusillum*, Schmidt *et al.* (2012) (Synchromophyceae); and *Fibrocapsa japonica*, Toriumi and Takano (1973) (Raphidophyceae).
>
> **Chrysista** (Cavalier-Smith 1986a): The most inclusive clade containing *Ectocarpus siliculosus*, Lyngbye (1819) (Phaeophyceae); *Synchroma pusillum*, Schmidt *et al.* (2012) (Synchromophyceae); and *Fibrocapsa japonica*, Toriumi and Takano (1973) (Raphidophyceae); but not *Thalassiosira pseudonana*, Hasle and Heimdal (1970) (Diatomeae); *Pelagomonas calceolata*, Andersen and Saunders (1993) (Pelagophyceae); and *Dictyocha speculum*, Ehrenberg (1839) (Dictyophyceae).

The name Chrysista was proposed by Cavalier-Smith for the group formed by all the Phaeista except the subgroups containing the Pelagophyceae and the Dictyochophyceae (Cavalier-Smith 1986a). This corresponds to the monophyletic group that we retrieve in our phylogenomic analysis, so we consider natural to keep this name. In the same way as Chrysista is reminiscent of the Chrysophyceae, one major algal group within the Chrysista, Diatomista is reminiscent of the diatoms, the best-known and first group to be described within the Diatomista. As noted above, we propose to keep the name Khakista in its current version to refer to the subgroup within the Diatomista containing the Bolidophyceae and the

Diatomeae. Given the use of branch-based definitions, the corollary of this classification is that all ochrophyte lineages not included in our analyses (e.g. Eustigmatophyceae, Pinguiophyceae) would belong to one of the two groups Chrysista or Diatomista unless they branch as sister group to all other ochrophytes.

Perspectives

Due to the limited amount of available genomic data for stramenopiles, the taxonomic sampling used in this study only represents a subset of the real diversity of this clade, especially among bigyran lineages. Because our phylogenomic analysis is endowed with robust statistical support, we predict that this topology will be recovered by further phylogenetic analyses. Nonetheless, this will be tested in the future by sequencing additional genomes of heterotrophic stramenopiles. Most of these species are currently not in culture, or are cultivated with bacteria (e.g. *I. marina* sequenced in this study). In this respect, single-cell transcriptomics and genomics techniques represent a promising avenue (Kolisko, et al. 2014).

Nevertheless, the phylogenetic backbone of stramenopiles presented in this study already raises interesting evolutionary questions that could be further addressed by comparative genomic analyses. For instance, the new phylogeny of stramenopiles has important implications concerning plastid evolution in this group. In addition to recent losses of photosynthesis within heterotrophic lineages of Ochrophyta, the classical rDNA-based phylogeny entailed a large number of plastid losses in basal-branching heterotrophic lineages in the case that the last common ancestor of stramenopiles was photosynthetic (as proposed by Cavalier-Smith (Cavalier-Smith 1986b, 1999)). However, according to our phylogenomic tree, only two losses are necessary (one in the ancestor of Bigyra and one in the ancestor of Oomycota/Pseudofungi). This scenario would be almost as parsimonious as a single late plastid acquisition by the ancestor of Ochrophyta in the alternative scenario of a non-photosynthetic ancestral stramenopile (Sanchez-Puerta and Delwiche 2008). A detailed inspection of the genomes of bigyran stramenopiles, especially free-living species, looking for possible genes of plastid origin might help to settle this issue.

Materials and methods

Culture of *Incisomonas marina*, genome sequencing and assembly

Incisomonas marina strain CCAP 997/1 was cultured in artificial seawater with boiled barley grains at 20°C in the dark. In these conditions, *I. marina* grows feeding on the bacteria that degrade the cereal grains. To remove those bacteria before DNA extraction, we used fluorescence activated cell sorting (FACS) using a BD FacsAria III to sort cells based on their size, in order to retain the biggest ones (corresponding to *I. marina*). DNA was extracted from the sorted *I. marina* cells using the ARCTURUS PicoPure DNA Extraction Kit (Applied Biosystems).

DNA was used to construct a paired-end library (insert size of ca. 300 bp) that was sequenced using a paired-end strategy (2 x 125 bp) in one run of Illumina HiSeq 2500 with chemistry v4 (Eurofins Genomics, Ebersberg, Germany). Reads were filtered using

Trimmomatic (Bolger, et al. 2014) and assembled using Velvet (Zerbino and Birney 2008), yielding a draft assembly of 136.1 Mb (23,926 contigs; N50 = 20.5 kb). The reads are available at the NCBI SRA database (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRR2962707. We inferred protein sequences of interest from the genome of *I. marina* using a simple custom pipeline: for each alignment, a protein profile was created and used to infer gene prediction using Augustus-PPX (Keller, et al. 2011). The protein sequences of *I. marina* used in this study are available in Supplementary file 4.

Selection and curation of phylogenetic markers

This phylogenetic project took advantage of an in-house collection of 421 eukaryotic phylogenetic markers, manually inspected and curated over the last five years. These protein alignments mostly originate from published phylogenomic datasets (Philippe, et al. 2009; Capella-Gutierrez, et al. 2012; Torruella, et al. 2012; Derelle, et al. 2015), and contain sequences from most eukaryotic groups and their prokaryotic homologs when available.

A large range of publicly available stramenopile protein sequences was gathered (see Supplementary file 1) and blasted against our collection of phylogenetic markers (Blastp; threshold of 1e-6; 5 best blast hits were retained for each species). For each single-gene alignment, we performed several rounds of phylogenetic analyses to detect outliers (i.e. contaminants, lateral gene transfers and paralogs) as follows: multiple sequence alignments were produced using Muscle (Edgar 2004), trimmed by trimAl (Capella-Gutierrez, et al. 2009) to remove positions with more than 30% of gaps and finally analyzed using RAxML version 8 (Stamatakis 2014) under the LG + Γ4 model. Single-gene trees were checked manually to combine partial orthologous sequences and remove outliers from the alignments. We provide a list of the main contamination sources for each species in Supplementary file 1. In cases where several sequences of a given species were present in the alignment, the slowest evolving one was selected (according to the branch lengths in RAxML trees) leading to the presence of one orthologous sequence per species in each alignment. A final and automated round of outlier detection was carried out using Phylo-MCOA ((de Vienne, et al. 2012); default parameters) using RAxML trees (same model).

Construction of the phylogenomic matrix

We first performed a selection of phylogenetic markers widely represented in stramenopiles and suitable to study their relationships. From the initial set of 421 alignments, we selected those that contained at least 24 stramenopile species and that did not show any deep duplication event within stramenopiles. These criteria led to a final dataset of 339 protein alignments. In order to obtain a combined phylogenetic matrix, these 339 alignments were first aligned with T-coffee (Notredame 2010) by masking all characters that had a consistency index lower than 9 (which corresponds to the highest value), and trimmed using Gblocks (Castresana 2000) under the following parameters: maximum proportion of gaps equal to 20%, minimum size of a block equal to 5, and maximum number of contiguous non conserved positions equal to 3. Trimmed alignments were finally concatenated into the phylogenetic matrix using a custom-made script. The phylogenomic matrix has been deposited in the TreeBase database (http://treebase.org; accession no. 18548).

Phylogenetic analyses

Bayesian inferences were performed with the CAT-GTR + Γ4 and CAT-Poisson + Γ4 models, using the "-dc" option, by which constants sites are removed, implemented in the program PhyloBayes-MPI (Rodrigue and Lartillot 2014). For the plain posterior estimation, two independent runs were performed with a total length of 2,500 and 12,000 cycles under the CAT-GTR and CAT-Poisson models, respectively. Convergence between the two chains was ascertained by calculating the difference in frequency for all their bipartitions using a threshold maxdiff <0.1. The first 1,000 and 4,000 points were discarded as burn-in in the CAT-GTR and CAT-Poisson analyses, respectively, and the posterior consensus was computed by selecting 1 tree every 10 over both chains. ML analyses were performed using RAxML as follows: searches for the best ML tree were conducted under the LG4X + Γ4 model from three random starting trees, and 200 ML bootstraps were analyzed with the rapid BS algorithm under the same model. ML analyses were repeated under the GTR + Γ4 model using the same approach.

Removal of fast evolving sites

Fast evolving sites were removed using a tree-independent approach: positions of the concatenated matrix were ranked according to their conservation value as calculated by Trimal, and removed in 2% increment up to 20%. At each shortening step, 100 ML bootstraps were analyzed with the rapid BS algorithm implemented in RAxML under the LG4X + Γ4 model. Node supports, as defined by Derelle and Lang (Derelle and Lang 2012), for the two alternative roots of stramenopiles were calculated from bootstrap trees using the ETE package (Huerta-Cepas, et al. 2010).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Thomas Cavalier-Smith for providing the culture of *Incimonas marina* and Marie Ragon for carrying out initial flow cytometry experiments at the UNICELL single-cell genomics platform (Orsay, France) and Fyodor Kondrashov for giving us access to the CRG bioinformatics cluster. This work was supported by the European Research Council under the European Union's Seventh Framework Program (ERC Grant Agreement 322669 'ProtistWorld') and the French Agence Nationale de la Recherche (project ANR-15-CE32-0003 'ANCESSTRAM'). We acknowledge the Conseil régional Ile-de-France (SESAME project) for supporting the creation of the UNICELL facility.

References

Adl SM, Simpson AG, Lane CE, Lukes J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, et al. The revised classification of eukaryotes. J Eukaryot Microbiol. 2012; 59:429–493. [PubMed: 23020233]

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30:2114–2120. [PubMed: 24695404]

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, et al. The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. 2008; 456:239–244. [PubMed: 18923393]

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol. 2005; 54:743–757. [PubMed: 16243762]

Brown JW, Sorhannus U. A molecular genetic timescale for the diversification of autotrophic stramenopiles (Ochrophyta): substantive underestimation of putative fossil ages. PLoS One. 2010; 5

Burki F. The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb Perspect Biol. 2014; 6:a016147. [PubMed: 24789819]

Burki F, Okamoto N, Pombert JF, Keeling PJ. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc Biol Sci. 2012; 279:2246–2254. [PubMed: 22298847]

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, Jakobsen KS, Pawlowski J. Phylogenomics reshuffles the eukaryotic supergroups. PLoS One. 2007; 2:e790. [PubMed: 17726520]

Capella-Gutierrez S, Marcet-Houben M, Gabaldon T. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. BMC Biol. 2012; 10:47. [PubMed: 22651672]

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25:1972–1973. [PubMed: 19505945]

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17:540–552. [PubMed: 10742046]

Cavalier-Smith, T. The kingdom Chromista: origin and systematics. Progress in Phycological Research. Round, FE., Chapman, DJ., editors. Elsevier Biomedical Press; 1986a. p. 309-347.

Cavalier-Smith T. The kingdoms of organisms. Nature. 1986b; 324:416–417. [PubMed: 2431320]

Cavalier-Smith T. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. Biol Lett. 2010; 6:342–345. [PubMed: 20031978]

Cavalier-Smith T. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol. 1999; 46:347–366. [PubMed: 18092388]

Cavalier-Smith T. A revised six-kingdom system of life. Biol Rev Camb Philos Soc. 1998; 73:203–266. [PubMed: 9809012]

Cavalier-Smith T, Chao EE. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J Mol Evol. 2006; 62:388–420. [PubMed: 16557340]

Cavalier-Smith T, Chao EE, Lewis R. Multiple origins of Heliozoa from flagellate ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. Mol Phylogenet Evol. 2015; 93:331–362. [PubMed: 26234272]

Cavalier-Smith T, Fiore-Donno AM, Chao E, Kudryavtsev A, Berney C, Snell EA, Lewis R. Multigene phylogeny resolves deep branching of Amoebozoa. Mol Phylogenet Evol. 2015; 83:293–304. [PubMed: 25150787]

Cavalier-Smith T, Scoble JM. Phylogeny of Heterokonta: Incisomonas marina, a uniciliate gliding opalozoan related to Solenicola (Nanomonadea), and evidence that Actinophryida evolved from raphidophytes. Eur J Protistol. 2013; 49:328–353. [PubMed: 23219323]

de Vienne DM, Ollier S, Aguileta G. Phylo-MCOA: a fast and efficient method to detect outlier genes and species in phylogenomics using multiple co-inertia analysis. Mol Biol Evol. 2012; 29:1587–1598. [PubMed: 22319162]

Delsuc F, Brinkmann H, Philippe H. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 2005; 6:361–375. [PubMed: 15861208]

Derelle R, Lang BF. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol Biol Evol. 2012; 29:1277–1289. [PubMed: 22135192]

Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček C, Lang BF, Eliáš M. Bacterial proteins pinpoint a single eukaryotic root. Proc Natl Acad Sci U S A. 2015; 112:E693–699. [PubMed: 25646484]

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

Gentekaki E, Kolisko M, Boscaro V, Bright KJ, Dini F, Di Giuseppe G, Gong Y, Miceli C, Modeo L, Molestina RE, et al. Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon Protocruzia and unravels the deep phylogenetic affinities of the ciliate lineages. Mol Phylogenet Evol. 2014; 78:36–42. [PubMed: 24814356]

Gomez F, Moreira D, Benzerara K, Lopez-Garcia P. Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. Environ Microbiol. 2011; 13:193–202. [PubMed: 20722698]

Grant JR, Katz LA. Building a phylogenomic pipeline for the eukaryotic tree of life - addressing deep phylogenies with genome-scale data. PLoS Curr. 2014; 6

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG, Roger AJ. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A. 2009; 106:3859–3864. [PubMed: 19237557]

Huerta-Cepas J, Dopazo J, Gabaldon T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010; 11:24. [PubMed: 20070885]

Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Kolisko M, Mylnikov AP, Keeling PJ. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. Proc Natl Acad Sci U S A. 2015; 112:10200–10207. [PubMed: 25717057]

Katz LA, Grant JR. Taxon-rich phylogenomic analyses resolve the eukaryotic tree of life and reveal the power of subsampling by sites. Syst Biol. 2015; 64:406–415. [PubMed: 25540455]

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014; 12:e1001889. [PubMed: 24959919]

Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011; 27:757–763. [PubMed: 21216780]

Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. Single-cell transcriptomics for microbial eukaryotes. Curr Biol. 2014; 24:R1081–1082. [PubMed: 25458215]

Lartillot N, Brinkmann H, Philippe H. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 2007; 7(Suppl 1):S4.

Le SQ, Dang CC, Gascuel O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. Mol Biol Evol. 2012; 29:2921–2936. [PubMed: 22491036]

Massana R, Castresana J, Balague V, Guillou L, Romari K, Groisillier A, Valentin K, Pedros-Alio C. Phylogenetic and ecological analysis of novel marine stramenopiles. Appl Environ Microbiol. 2004; 70:3528–3534. [PubMed: 15184153]

Massana R, del Campo J, Sieracki ME, Audic S, Logares R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. ISME J. 2014; 8:854–866. [PubMed: 24196325]

Notredame C. Computing multiple sequence/structure alignments with the T-coffee package. Curr Protoc Bioinformatics. 2010 Chapter 3:Unit 3 8 1-25.

Patterson, DJ. Stramenopiles: Chromophytes from a protistan perspective. The Chromophyte Algae: Problems and Perspectives. Green, JC.Leadbeater, BSC., Diver, WL., editors. Clarendon Press; 1989. p. 357-379.

Pernice MC, Giner CR, Logares R, Perera-Bel J, Acinas SG, Duarte CM, Gasol JM, Massana R. Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. ISME J. 2015

Philippe H, Derelle R, Lopez P, Pick K, Borchiellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Queinnec E, et al. Phylogenomics revives traditional views on deep animal relationships. Curr Biol. 2009; 19:706–712. [PubMed: 19345102]

Richards TA, Soanes DM, Jones MD, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci U S A. 2011; 108:15258–15263. [PubMed: 21878562]

Riisberg I, Orr RJ, Kluge R, Shalchian-Tabrizi K, Bowers HA, Patil V, Edvardsen B, Jakobsen KS. Seven gene phylogeny of heterokonts. Protist. 2009; 160:191–204. [PubMed: 19213601]

Rodrigue N, Lartillot N. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. Bioinformatics. 2014; 30:1020–1021. [PubMed: 24351710]

Roy RS, Price DC, Schliep A, Cai G, Korobeynikov A, Yoon HS, Yang EC, Bhattacharya D. Single cell genome analysis of an uncultured heterotrophic stramenopile. Sci Rep. 2014; 4:4780. [PubMed: 24759094]

Sanchez-Puerta MV, Delwiche CF. A hypothesis for plastid evolution in chromoalveolates. Journal of Phycology. 2008; 44:1097–1107. [PubMed: 27041706]

Saunders GW, Potter D, Paskind MP, Andersen RA. Cladistic analyses of combined traditional and molecular data sets reveal an algal lineage. Proc Natl Acad Sci U S A. 1995; 92:244–248. [PubMed: 7816825]

Ševčíková T, Horák A, Klimeš V, Zbránková V, Demir-Hilton E, Sudek S, Jenkins J, Schmutz J, Přibyl P, Fousek J, et al. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci Rep. 2015; 5:10134. [PubMed: 26017773]

Simon M, Jardillier L, Deschamps P, Moreira D, Restoux G, Bertolino P, López-García P. Complex communities of small protists and unexpected occurrence of typical marine lineages in shallow freshwater systems. Environ Microbiol. 2015; 17:3610–3627. [PubMed: 25115943]

Sierra R, Matz MV, Aglyamova G, Pillet L, Decelle J, Not F, de Vargas C, Pawlowski J. Deep relationships of Rhizaria revealed by phylogenomics: a farewell to Haeckel's Radiolaria. Mol Phylogenet Evol. 2013; 67:53–59. [PubMed: 23280368]

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. [PubMed: 24451623]

Torruella G, Derelle R, Paps J, Lang BF, Roger AJ, Shalchian-Tabrizi K, Ruiz-Trillo I. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. Mol Biol Evol. 2012; 29:531–544. [PubMed: 21771718]

Tsaousis AD, Ollagnier de Choudens S, Gentekaki E, Long S, Gaston D, Stechmann A, Vinella D, Py B, Fontecave M, Barras F, et al. Evolution of Fe/S cluster biogenesis in the anaerobic parasite Blastocystis. Proc Natl Acad Sci U S A. 2012; 109:10426–10431. [PubMed: 22699510]

Yang EC, Boo GH, Kim HJ, Cho SM, Boo SM, Andersen RA, Yoon HS. Supermatrix data highlight the phylogenetic relationships of photosynthetic stramenopiles. Protist. 2012; 163:217–231. [PubMed: 22001261]

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]

Europe PMC Funders Author Manuscripts

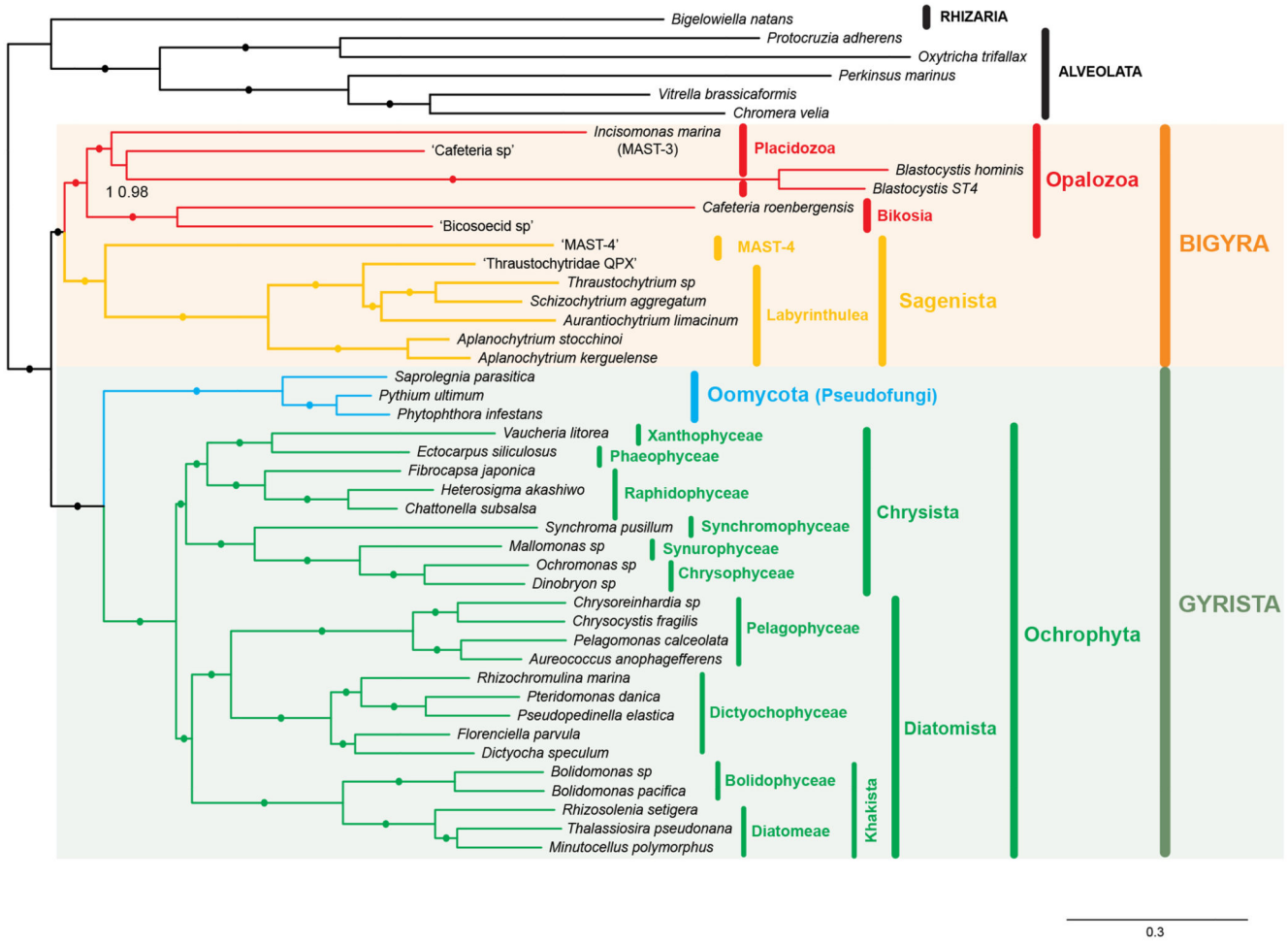

Figure 1. Bayesian phylogenomic tree of stramenopiles.

Bayesian consensus tree obtained from the full data set under the CAT-GTR + Γ4 model. The tree is arbitrary rooted on Rhizaria and Alveolata. Branch supports correspond to Bayesian posterior probabilities obtained under the CAT-GTR + Γ4 (left) and CAT-Poisson + Γ4 (right; inferred from the two converging chains) models. Branches with posterior probabilities equal to 1 in both Bayesian analyses are marked with a bullet.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

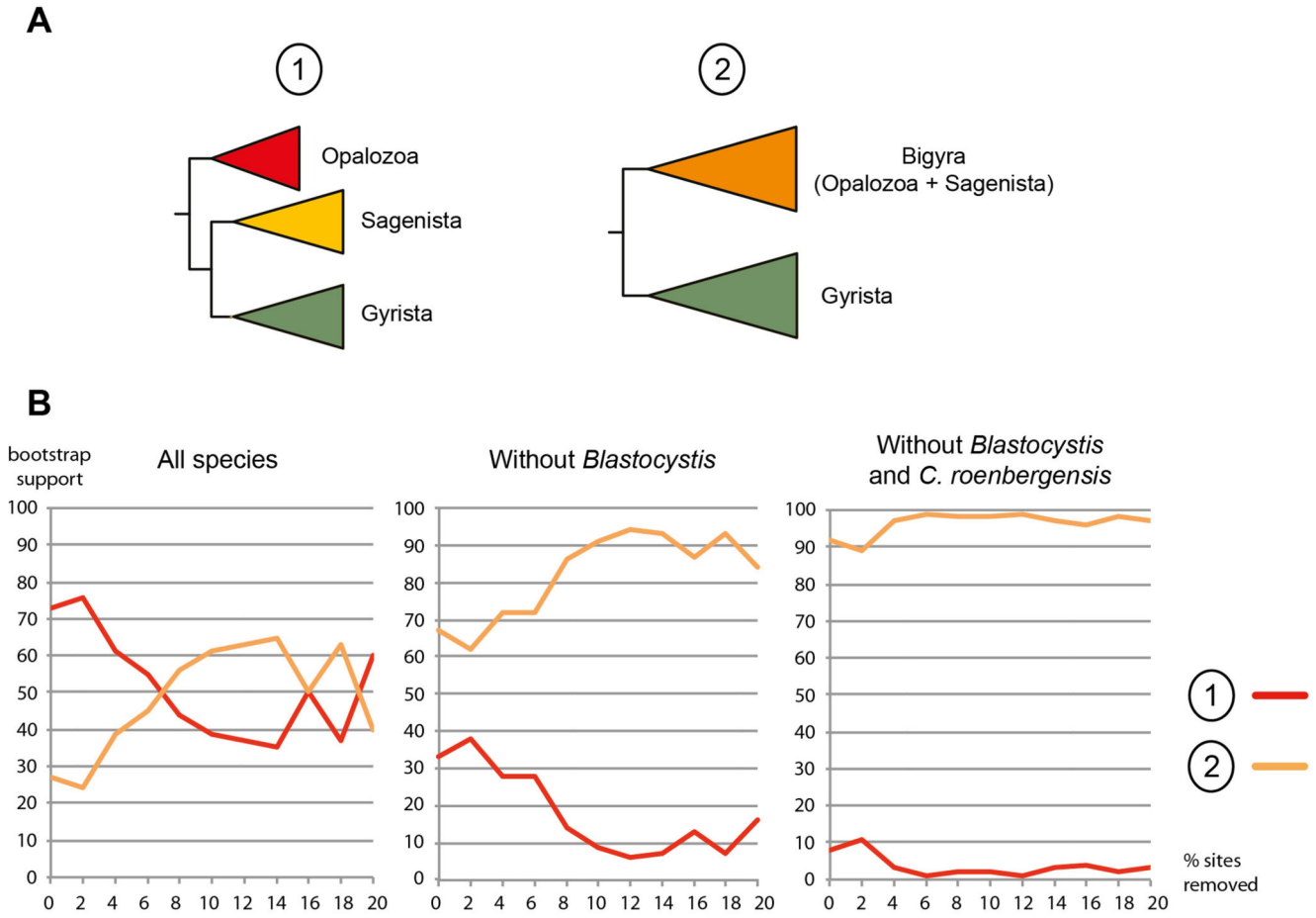

Figure 2. The root of stramenopiles in maximum likelihood (ML) analyses.

A, Schematic representation of the two alternative positions of the root of stramenopiles observed in ML analyses. B, Evolution of bootstrap supports for the two topologies shown in Figure 2A along the removal of fast evolving sites and divergent species.