Molecular Phylogenetics and Evolution 198 (2024) 108120

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Phylogenomic analyses of ochrophytes (stramenopiles) with an emphasis on neglected lineages

Anna Cho *, Gordon Lax, Patrick J. Keeling

Department of Botany, University of British Columbia, Vancouver V6T 1Z4, British Columbia, Canada

ARTICLE INFO

Keywords:
Stramenopile
Gene filtering
Transcriptome
Informative genes
Ochrophytes
Phylogenomic

ABSTRACT

Ochrophyta is a photosynthetic lineage that crowns the phylogenetic tree of stramenopiles, one of the major eukaryotic supergroups. Due to their ecological impact as a major primary producer, ochrophytes are relatively well-studied compared to the rest of the stramenopiles, yet their evolutionary relationships remain poorly understood. This is in part due to a number of missing lineages in large-scale multigene analyses, and an apparently rapid radiation leading to many short internodes between ochrophyte subgroups in the tree. These short internodes are also found across deep-branching lineages of stramenopiles with limited phylogenetic signal, leaving many relationships controversial overall. We have addressed this issue with other deep-branching stramenopiles recently, and now examine whether contentious relationships within the ochrophytes may be resolved with the help of filling in missing lineages in an updated phylogenomic dataset of ochrophytes, along with exploring various gene filtering criteria to identify the most phylogenetically informative genes. We generated ten new transcriptomes from various culture collections and a single-cell isolation from an environmental sample, added these to an existing phylogenomic dataset, and examined the effects of selecting genes with high phylogenetic signal or low phylogenetic noise. For some previously contentious relationships, we find a variety of analyses and gene filtering criteria consistently unite previously unstable groupings with strong statistical support. For example, we recovered a robust grouping of Eustigmatophyceae with Raphidophyceae-Phaeophyceae-Xanthophyceae while Olisthodiscophyceae formed a sister-lineage to Pinguiophyceae. Selecting genes with high phylogenetic signal or data quality recovered more stable topologies. Overall, we find that adding under-represented groups across different lineages is still crucial in resolving phylogenetic relationships, and discrete gene properties affect lineages of stramenopiles differently. This is something which may be explored to further our understanding of the molecular evolution of stramenopiles.

1. Introduction

Ochrophyta is a group of protists that are often used as an example of the vast molecular and morphological diversity of stramenopiles. They include the giant multicellular brown algae, the intricate frustule-covered diatoms, some golden algae that have lost the ability to photosynthesize, and dozens of other distinct subgroups (Cavalier-Smith and Chao, 2006; Graf et al., 2020; Riisberg et al., 2009; Yang et al., 2012). Because of their ecological importance and morphological diversity, there have been many studies reconstructing ochrophyte phylogeny and trying to understand their evolutionary relationships. Yet, despite this attention, phylogenomic analyses of ochrophytes remain incongruent with one another (Azuma et al., 2022; Burki et al., 2016; Cho et al., 2022; Derelle et al., 2016; Di Franco et al., 2022; Noguchi et al., 2016; Thakur et al., 2019), especially between trees reconstructed from nuclear and plastid genes (Barcytė et al., 2021; Di Franco et al., 2022; Dorrell et al., 2021; Ševčíková et al., 2015). Additionally, even with publicly available genomic and transcriptomic data and with many ochrophytes readily available in culture collections (Yang et al., 2012), the diversity of ochrophytes in supermatrices used in phylogenomic analyses has remained mostly under-represented and has been somewhat static over the last few years (Azuma et al., 2022; Burki et al., 2016; Cho et al., 2022; Derelle et al., 2016; Driskell et al., 2004; Noguchi et al., 2016; Thakur et al., 2019) (for an exception, see Terpis et al., 2024).

Current ochrophyte phylogenomic analyses differ in dataset composition and size, processing approaches, and phylogenetic inference methods. Although there is some consensus around the backbones of the ochrophyte phylogeny (Azuma et al., 2022; Cho et al., 2022;

* Corresponding author.
E-mail address: acho@mail.ubc.ca (A. Cho).

https://doi.org/10.1016/j.ympev.2024.108120
Received 4 January 2024; Received in revised form 13 May 2024; Accepted 4 June 2024
Available online 7 June 2024

Derelle et al., 2016; Thakur et al., 2019), numerous recalcitrant relationships characterized by short internodes leave the positioning of some important lineages contentious. These short internodes in stramenopile phylogeny are likely caused by ancient rapid radiation carrying limited phylogenetic signal (Di Franco et al., 2022; Pardo-De La Hoz et al., 2023; Whitfield and Lockhart, 2007). To make matters worse, these short internodes are commonly found across deep, divergent lineages of stramenopiles (i.e., long-branching taxa) where data sites (i.e., nucleotide or amino acid sequences) tend to experience saturation leading to underestimation of actual sequence substitutions (Lartillot et al., 2007; Philippe et al., 2011). Consequently, these branches are prone to long branch attraction (LBA) artefacts (Felsenstein, 1978; Hendy and Penny, 1989; Wägele and Mayer, 2007). Another challenge is phylogenetic incongruence among gene trees (including organellar and nuclear gene trees) caused by non-neutral (adaptive) selection (Dorrell et al., 2019; Edwards, 2009; Stiller et al., 2003), incomplete lineage sorting (ILS), introgression via hybridization, and horizontal gene transfers (Dong et al., 2022; Dorrell et al., 2021; Maddison, 1997; Nichols, 2001).

Several phylogenomic approaches are available to remediate the effects of these issues: incrementally removing fast-evolving sites, genes, and taxa, or increasing taxon sampling and the number of sites (Hedtke et al., 2006; Hillis, 1998; Hillis et al., 2003; Pick et al., 2010; Superson and Battistuzzi, 2022; Zwickl and Hillis, 2002). More recently, applying the CAT-PMSF phylogenetic method (Szantho et al., 2023) was reported to be robust against LBA, while significantly decreasing computing resources. Furthermore, the importance of characterizing phylogenetically informative genes has been highlighted in resolving short internodes in ancient radiation (Salichos and Rokas, 2013; Shen et al., 2016; Smith et al., 2018). Using high variable length bootstrap values as a proxy for phylogenetic signal, ochrophyte plastid genes have been shown to have more phylogenetic signal than nuclear genes with comparable numbers of sites (Di Franco et al., 2022). However, plastid datasets are not suitable for inferring evolutionary history of stramenopiles as a whole, as many stramenopiles lack plastid and plastid-associated genes.

In this study, we aim to resolve relationships within ochrophytes, and by extension stramenopiles as a whole, by first updating the ochrophyte dataset to include a number of neglected, but potentially informative lineages, and by comprehensively assessing nuclear genes to identify those most phylogenetically informative and those with most noise. To update the dataset, we added ten new transcriptomes from ochrophytes, some of which had not been well-represented in previous phylogenomic analyses, along with including all other current publicly available data. The updated dataset now represents 14 out of 17 major ochrophyte classes (Cavalier-Smith and Chao, 2006; Graf et al., 2020; Riisberg et al., 2009; Yang et al., 2012) including members of the Olisthodiscophyceae (Barcytė et al., 2021), Phaeothamniophyceae (Andersen et al., 1998), Schizocladiophyceae (Kawai et al., 2003), and Picophagea (Guillou et al., 1999). We particularly focused on "breaking" long branches leading to known lineages with conflicting placement, such as Eustigmatophyceae, Actinophrydae, and Pinguiophyceae. To identify phylogenetically informative genes and investigate a source of incongruence among various phylogenomic analyses, we explored different gene filtering criteria. We used a previously established method (Mongiardino Koch, 2021; Mongiardino Koch and Thompson, 2021), which calculates phylogenetic signal, noise, and data quality for a given set of marker genes. Overall, we report robust support for previously controversial placements, and some of these relationships were recovered in the majority of trees reconstructed from various subsets of genes. Phylogenetically informative genes could not be unambiguously identified, however we observed that using genes with high phylogenetic signal results in the most stable tree topologies, as opposed to selecting genes with low phylogenetic noise.

2. Materials and methods

2.1. Ochrophyte sample collection and processing

Nine cultures of under-represented ochrophytes were obtained from various culture collections (Table 1). Except for *Actinosphaerium* sp. (which was processed immediately and the culture not maintained), we sub-cultured all cultures every two weeks in 30 mL and kept at 20°C with a 12 h:12 h light:dark cycle. Both *Olisthodiscus luteus* and *O. tomasii* were kept in TL30 media; *Schizocladia ischiensis* was maintained in L1-Si (Guillard, 1975; Guillard and Ryther, 1962); *Phaeothamnion confervicola* in $MiEB_{12}$ (Andersen, 1991); *Pseudostaurastume enorme* in DYV-m (Lehman, 1967); V*acuoliviride crystalliferum* in AF6 with f/2 vitamin solution (Watanae et al., 2000); *Chrysamoeba. radians* in URO + soil (Provasoli and Pintner, 1959); and *Picophagus flagellatus* in 0.22 μm filtered seater water (30 ‰) with an autoclaved rice grain.

We extracted RNA with TRIzol™ LS for all cultures except the two *Olisthodiscus* spp., *P. confervicola*, and *Actinosphaerium* sp. Forty milliliters of each culture was centrifuged at 3000 rpm for 20 min at 4°C to pellet cells at the bottom of the centrifuge tubes. After carefully removing supernatant media, 1 mL of TRIzol™ LS was added to the cells and the mixture was transferred to Lysing Matrix Y bead tubes (MP Biomedicals, USA). The mixture in the bead tubes were subjected to physical lysis using a VWR™ Mini Bead Mill at 5 m/s for 30 sec followed by 30 sec on ice. This step was repeated once more. The solution was then transferred to Phasemaker™ (Invitrogen) tubes to minimize interphase contamination during the aqueous-organic layer separation using chloroform. The precipitated and washed RNA pellets were resuspended in 30 μL PCR-grade water.

For both *Olisthodiscus* cultures, we used a cetyltrimethylammonium bromide (CTAB)-based RNA extraction protocol (Apt et al., 1995; Yao et al., 2009) to prevent co-precipitation of phenolic compounds which can hinder downstream cDNA synthesis. Briefly, 40 mL of each of the culture was centrifuged in 15 mL Falcon™ tubes for 10 min at 4°C, 3000 rpm. After discarding supernatant media, 2 mL of CTAB buffer was added directly to the pelleted cells. While gently agitating the mixture, 25 % v/v of 100 % ethanol and 11 % v/v of potassium acetate (3M, pH 4.8) were slowly added. The remainder of RNA extraction and precipitation were followed as described by Yao et al., 2009. Each of the RNA pellets were resuspended in 200 μL of PCR-grade water, followed by RNA purification using NucleoSpin® RNA XS Kit (Takara Bio USA, Inc.) with 10 μL elution volume.

For *P. confervicola* and *Actinosphaerium* sp., we manually isolated each single cell (or a small filamentous colony of *P. confervicola*) using a glass micropipette under a Leica DLIM inverted microscope, followed by rinsing three times in PCR-grade water. Rinsed cells were then transferred into 0.2 mL PCR tube containing lysis buffer (Picelli et al., 2014) and stored at −80°C until cDNA synthesis. Similarly, we isolated three single cells of *Vicicitus globosus* from marine plankton near-shore tows at Hakai Institute, Quadra Island, BC Canada (50°06′54.6″N, 125°13′10.8″W) on August 7th and September 12th, 2021.

The quality and quantity of the RNA extracts from TRIzol™ LS and CTAB-based methods were assessed using a NanoDrop 1000 Spectrophotometer v3.8.1 (Thermo Fisher Scientific) and Qubit™ RNA High Sensitivity Assay Kits (Thermo Fisher Scientific).

2.2. cDNA synthesis, library preparation and sequencing

We followed the poly-A selection based Smart-Seq2 protocol for cDNA synthesis (Picelli et al., 2014). For RNA extracts, 4 μL was used for each cDNA synthesis while single-cell isolates were subject to 2–3 rounds of freeze–thaw cycles (Onsbring et al., 2020) prior to Smart-Seq2. The quantity of cDNA was measured using Qubit™ dsDNA HS Assay Kits (Thermo Fisher Scientific). To confirm taxonomic identities, we performed small subunit ribosomal DNA (SSU rDNA) polymerase chain reaction (PCR) on each cDNA sample (except *V. globosus*), using

Table 1
List of ochrophyte cultures obtained from various culture collections.

Species	Class	Culture collection centre (location)	Culture ID	Media
Actinospherium sp.	Actinophrydae	Carolina Biological Supply (USA)	item#131302	Carolina™ Springwater
Chrysamoeba radians	Chrysophyceae	National Institute for Environmental Studies (Japan)	NIES-2890	URO + soil
Olisthodiscus luteus	Olisthodiscophyceae	Norwegian Culture Collection – Scandinavian Culture Collection (Norway)	K-0444	TL30
Olisthodiscus tomasii	Olisthodiscophyceae	National Institute for Environmental Studies (Japan)	NIES-15	TL30
Phaeothamnion confervicola	Phaeothamniophyceae	Roscoff Culture Collection (France)	RCC7139	$MiEB_{12}$
Picophagus flagellatus	Picophagea	Roscoff Culture Collection (France)	RCC22	FSW
Pseudostaurastume enorme	Eustigmatophyceae	Culture Collection of Algae at Göttingen University (Germany)	SAG11.85	DYV-m
Schizocladia ischiensis	Schizocladiophyceae	Roscoff Culture Collection (France)	RCC7138	L1-Si
Vacuoliviride crystalliferum	Eustigmatophyceae	National Institute for Environmental Studies (Japan)	NIES-2860	AF6

18SFU-18SRU primers (Tikhonenkov et al., 2016), followed by purification using QIAquick® PCR Purification Kit (Qiagen), and Sanger dideoxy sequencing (University of British Columbia, UBC BC Canada).

Library preparation was done by the Sequencing and Bioinformatics Consortium (UBC, BC Canada), using the Illumina DNA Flex Library Preparation Kit, and sequenced on a NextSeq platform with 150 bp paired-end library constructs. For some cultures, RNA extraction, cDNA synthesis, library preparation and the subsequent sequencing were repeated to obtain higher completeness of the transcriptome, using the same parameters and methods. The raw transcriptome data is deposited under NCBI accession SRR27254659-SRR27254668, under BioProject PRJNA1050613.

2.3. Transcriptome processing and phylogenomic matrix construction

Along with the ten newly generated transcriptomes, we also processed publicly available transcriptomes of *Saccharina* sp. (ERR2861927), *Sargassum* sp. (DRR042036), *Uroglena* sp. (ERR1368708), *Glossomastix* sp. (ERR3497268), *Synura* sp. (ERR1368706), *Heterococcus* sp. (SRR1099987), *Vischeria* sp. (SRR14572414), *Monodopsis* sp. (SRR14581548), *Eustigmatos polyphem* (SRR397983), *Poteriospumella lacustris* (ERR1368700) as described below. All other pre-processed (i.e., predicted open reading frames, ORFs) genomic level data were obtained from previous publications (Azuma et al., 2022; Cho et al., 2022, 2024; Labarre et al., 2021; Thakur et al., 2019), the EukProt V3 database (Richter et al., 2022), and the Marine Microbial Eukaryote Transcriptome Sequencing Project, MMETSP (Keeling et al., 2014). Many of these transcriptomes represent sub-groups of ochrophytes that were otherwise represented by small numbers of taxa in previous phylogenomic analyses.

First, the quality of all raw sequencing data was evaluated using FastQC v0.11.9 (Andrews, 2010), followed by random sequencing error correction using *k-mer* based Rcorrector v3 (Song and Florea, 2015). The corrected reads were then trimmed and filtered (−phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36) using Trimmomatic v0.39 (Bolger et al., 2014) to remove transposase-inserts, Smart-Seq2 IS-primers and Nextera™ DNA Flex adaptors from library preparation. The resulting forward, reverse, and unpaired transcripts were assembled (or co-assembled if multiple transcriptomes from the same culture were generated) using *de novo* rnaSPAdes v3.15.1 (Bushmanova et al., 2019). The single-cell transcriptome data of *V. globosus* was co-assembled once the species identities were confirmed by extracting SSU rDNA sequences using barrnap v0.9 (Seemann, 2007). To evaluate assembly results (e.g., coverage and taxonomic assignments), we used BlobTools v2.3.3 (Challis et al., 2020; Laetsch and Blaxter, 2017). Taxonomic assignments were determined by searching assembled transcripts against the NCBI nt database using megaBLAST followed by a diamond BLASTX against the Uniprot reference database (Haas et al., 2009), both with e-value cut-offs 1e-25. All bacterial, Viridiplantae, Metazoa, and archaeal reads were removed. Open reading frames (ORFs) were predicted using TransDecoder v5.5.0 (Haas, 2015) and the longest ORFs were annotated with a BLASTP search against UniProt database (e-value 1e-5). To assess the completeness of each transcriptome, BUSCO v5.2.2 (Simão et al., 2015) was used with database 'stramenopiles_odb10'.

2.4. Phylogenomic supermatrices

The predicted ORFs of the newly added transcriptomes were added to an existing supermatrix using PhyloFisher v1.1.2 (Tice et al., 2016). Briefly, to identify homologs from the ORFs of each transcriptome, we searched against 241 genes compiled in PhyloFisher. The identified homolog candidates were then added to their respective gene alignments, followed by sequence processing using PREQUAL (Whelan et al., 2018), MAFFT (Katoh and Standley, 2013), Divvier (Ali et al., 2019) and trimAl (Capella-Gutiérrez et al., 2009) incorporated in PhyloFisher. Each alignment was then used to construct a single gene tree under the L + G4 + X model with 1000 replicates of ultrafast bootstraps (UB), using IQ-TREE v1.6.12 (Nguyen et al., 2015). To ensure correct orthologs were identified for each gene from each transcriptome, we manually screened 241 single-gene trees using ParaSorter v1.0.4. To generate a concatenated supermatrix, we selected 139 taxa (including 14 outgroup taxa) with 231 orthologs (≥39 % taxa completeness) ('231-supermatrix'). An additional supermatrix was generated with orthologs from MAST-1, MAST-7, MAST-8, MAST-9 and MAST-11 (Labarre et al., 2021), consisting of 146 taxa (including 14 outgroup) with 233 orthologs (≥39 % taxa completeness), resulting in 73,440 sites ('233-supermatrix').

2.4.1. Filtering by gene occupancy, fast-evolving and random sites

To investigate the effect of fast-evolving sites, 7,000 fast-evolving amino acid (aa) sites were incrementally removed to exhaustion from the '231-supermatrix', using PhyloFisher, resulting in 10 additional supermatrices ('fsite-supermatrix'). Similarly, 7,000 random sites were incrementally removed, resulting in yet another 10 supermatrices ('randSite-supermatrix'). We also randomly removed genes in 20 % increments to compare with trees recovered from different gene filtering criteria ('randGene-supermatrix'). The average bootstrap (BS) values of phylogenomic trees from each of randSite- and randGene-supermatrices were calculated and used to determine minimum data size (i.e. number of amino acid sites) required to reduce the effect of small data size and distinguish from the effect of different gene-filtering criteria (see below). Based on the condition of recovering a paraphyletic Bigyra and several well-supported clades of ochrophytes (e.g., Chrysista or Diatomista), we decided the cut-off BS values to be > 89 %. Based on this cut-off, we determined that approximately 22,000 amino acid sites is the minimum required.

2.4.2. Conceptual design for phylogenomic gene filtering

To identify phylogenetically informative genes and investigate incongruence among different phylogenomic analyses, we calculated different gene properties based on previously established methods (Mongiardino Koch, 2021; Mongiardino Koch and Thompson, 2021). The calculated properties were then used to rank the genes by noise or

signal (some include data quality, see below) based on correlation significance and contribution to an ordination axis (i.e. PC loadings). Phylogenomic analyses inferred from different sets of selected genes were then used to evaluate whether removing genes with high phylogenetic noise, selecting genes with low noise or high phylogenetic signal would resolve lineages that were previously conflicting, ultimately with the goal of finding the most informative set of genes. Furthermore, we sought to replicate alternative placements of contentious lineages (e.g. placement of Eustigmatophyceae or Pinguiophyceae found in phylogenomic trees inferred from plastid genes), by selecting nuclear genes with high phylogenetic noise.

2.4.3. Filtering by phylogenetic noise, signal, and other data quality

To evaluate the effects of some of the known sources of noise such as average pair-wise patristic distance (av_patristic, a proxy for LBA) (Mongiardino Koch and Thompson, 2021; Struck, 2014), variance of root-to-tip distances (root_tip_var, a proxy for inferring deviation from clock-like evolution) (Mongiardino Koch and Thompson, 2021; Smith et al., 2018), saturation (Kocot et al., 2016; Nosenko et al., 2013), and relative composition frequency variability (RCFV, a proxy for amino acid compositional heterogeneity) (Shen et al., 2016; Whelan et al., 2015; Zhong et al., 2011), and phylogenetic signal such as treeness (length of internal branches) (Lanyon, 1988), average bootstrap supports (average_BS_support), Robinson-Foulds similarity (robinson_sim, distance between a gene and species tree; proxy for incongruencies) (Robinson and Foulds, 1981; Salichos and Rokas, 2013), we applied the measurement method put together by Koch (2021) and Koch and Thompson (2021), which calculates these properties in all the genes used for constructing '231-supermatrix' and visualizes them with principal component analysis (PCA). Other information that is indicative of the dataset quality such as alignment lengths, the proportion of missing data per taxon, completeness/occupancy of genes, total tree length, and tree-based evolutionary rate were also calculated (Mongiardino Koch, 2021; Mongiardino Koch and Thompson, 2021).

We estimated the known possible sources of phylogenetic noise (av_patristic, root-tip-var, saturation, RCFV), signal (treeness, average_BS_support, robinson_sim), and data quality or information (rate, missing data, tree and gene length, proportion of variable sites, and occupancy) using a published R-script (https://github.com/mongiardino/genesortR) (Mongiardino Koch, 2021), with some modifications. Although the '233-supermatrix' has the most up-to-date collections of stramenopile taxa, due to the timing of data analysis, we calculated phylogenetic noise, signal and quality in all genes of the '231-supermatrix'. The resulting measures were plotted onto two principal component axes using the 'factoextra' R-package. Two genes (GDI and NSF1-I) were considered as outliers based on the estimated Mahalanobis distances and were excluded from downstream analyses. To visualize how each of the measured properties are correlated to one another and to calculate correlation coefficients and significance, we generated Pearson correlation graphs using the R-packages 'corrr', 'ggcorrplot', 'GGally', 'ggfortify' and 'FactoMineR'. Based on the correlation analysis and PC loadings of each properties, we subsampled genes using eight criteria: A) high values of treeness and occupancy; B) high values of average_BS_support, robinson_sim, and gene length; C) low values of av_patristic, evolutionary rate, and total tree length; D) filtering out high values of av_patristic, evolutionary rate, and total tree length; E) high values of PC1-associated noise (root_tip_var, av_patristic, and saturation); F) high values of all noise; S) high values of signal (treeness, average_BS_support, robinson_sim); and Q) high values of data quality (occupancy and gene length). Because each criterion is a combination of multiple properties, we extracted shared genes that are found with the properties of a given criterion by searching the top 60 to180 genes of the highest or the lowest values. For example, 43 genes were present in the top 80 highest values of both treeness and occupancy (criterion A80) while 60 genes were present in the top 80 lowest values for each properties in criterion C (criterion C80). We also combined extracted genes from criteria A to C, with the top 60-160 highest values in criteria A and B and, the lowest values in criterion C (i.e., ABC60-160). Finally, we also subsampled genes that are not well represented by any of the two PCA axes (i.e., genes with low cos2 values) (criterion N).

The size of different supermatrices generated from each criterion is summarized in Table 2. For each of the gene sets that were filtered by different criterion or a combination of them, we generated supermatrices as described in 2.4.

2.5. Phylogenomic trees: C60-PMSF, CAT-PMSF, CAT-GTR

For all the supermatrices generated above, we inferred maximum likelihood (ML) trees using IQ-TREE v2.1.2, under the profile mixture model LG + C60 + F + G4 (C60) with posterior mean site frequencies (PMSF) used to generate 100 replicates of non-parametric standard bootstraps (BS) (Quang et al., 2008; Wang et al., 2018). This method involves a two-step process incorporated in IQ-TREE, first by generating initial ML trees under the LG + C60 + F + G4 model with 1000 ultrafast bootstraps (UFB). The estimated guide-topologies of these initial ML trees were then used to estimate PMSF, which were then used to reconstruct the final C60-PMSF trees (Wang et al., 2018). To check whether exchangeabilities were not mis-specified with the F-class, we verified that the F-class values are < 0.11 (Baños et al., 2023), and repeated the tree reconstruction under the LG + C60 + G4 model. All relevant files for each of the supermatrices, phylogenomic trees and calculated properties generated from different filtering criteria are available on Dryad (https://doi.org/10.5061/dryad.f4qrfj73q) (Cho et al., 2024a).

For the '231-supermatrix', we inferred a phylogenomic tree with Bayesian estimation using PhyloBayes-MPI v4.0.3, under the CAT-GTR mixture model with four independent Markov Chain Monte Carlo (MCMC) chains. These chains were run in parallel for 20,000 generations each. After discarding the first 10 % of generations as burn-in, we checked for convergence using bpcomp, and estimated the consensus posterior probability and topology by subsampling every second tree. Finally, we reconstructed an additional phylogenomic tree using the CAT-PMSF pipeline (Szantho et al., 2023) to compare with our C60-PMSF analysis. Both of these two methods assess the effects of potential artefacts derived from compositional heterogeneity across amino acid sites however, CAT-PMSF estimates site-specific amino acid frequency using a non-parametric Bayesian approach while C60-PMSF uses a fixed amino acid frequency vector (Szantho et al., 2023; Wang et al., 2018). CAT-PMSF involves three steps: 1) construct an initial ML tree under a site-homogeneous model, LG + F + G4; 2) correct potential LBA artefacts using Bayesian estimation (PhyloBayes-MPI v4.0.3), under the CAT-LG model with the two Markov chains until convergence (~6,000 generations, 20 % discarded as burn-in, convergence assessed with maxdiff = 0); 3) using site-specific stationary distributions obtained from step 2 to fit the tree to PMSF with IQ-TREE, as described above for C60-PMSF. Each chain was used to generate the final two PMSF trees (CAT-PMSF trees) for step 3.

3. Results and discussion

3.1. The phylogenomic tree of stramenopiles

3.1.1. Updating the ochrophytes tree with under-represented classes

We generated ten new transcriptomes to update the taxon sampling for ochrophytes, including six taxa belonging to four classes that had not been previously represented in phylogenomic analyses (Table 1). The updated phylogenomic supermatrix resulted in 72,932 amino acid (aa) sites ('231-supermatrix'), with 93 Gyrista (70 ochrophyte taxa), 32 Bigyra, and 14 outgroup taxa (Fig. 1). When we included MAST-1, −7, −8, −9, and MAST-11 in the supermatrix ('233-supermatrix'), the resulting dataset consisted of 73,440 aa sites from 96 Gyrista and 36 Bigyra. The addition of MAST-1, −7, −8, −9, and MAST-11 did not

Table 2
Summary of supermatrices generated using different filtering criteria, reporting their total amino acid sites and number of genes (in brackets). 'Top n-value' indicates common genes found in the top n-list for all the properties of a criterion. Each criterion is denoted by A = selecting for genes with high values of treeness and occupancy; B = selecting for genes with high values average_BS_support, robinson_sim, and gene length; C = selecting for genes with low values of av_patristic, rate, and treelength; D = filter out genes with high values of av_patristic, rate, and treelength; ABC = combination of criteria A-C with corresponding 'Top n-values'; E = selecting genes with high values of PC1 axes associated biases (saturation, av_patristic, and root_tip_var); F = selecting genes with high values of all biases (RCFV, saturation, av_patristic, and root_tip_var), S = selecting genes with high signals (average_BS_support, robinson_sim, treeness); Q = selecting genes with high data quality (gene length and occupancy); N = genes that are not explained well by the PC axes (low cos2); C60- & CAT-PMSF, Bayesian = the same 231-supermatrix was used for constructing C60-PMSF tree, CAT-PMSF tree and Bayesian trees.

Top n-value	A	B	C	D	ABC	E	F	S	Q	N	C60- & CAT-PMSF Bayesian
60	4,816 (26)	5,116 (12)	12,932 (49)	57,819 (186)	20,817 (77)	—	—	—	—	11,353 (43)	72,932 (231)
80	9,203 (43)	9,673 (23)	17,636 (60)	52,151 (167)	32,756 (109)	—	—	—	—		
100	15,794 (64)	16,118 (38)	22,884 (79)	46,180 (148)	45,376 (144)	—	—	—	—		
120	22,070 (81)	22,030 (54)	30,955 (102)	40,342 (130)	53,922 (169)	21,576 (70)	14,587 (53)	18,967 (47)	23,265 (54)		
140	29,095 (105)	29,914 (76)	36,544 (120)	33,211 (111)	59,940 (187)	29,228 (92)	20,234 (71)	26,804 (70)	34,067 (84)		
160	36,203 (130)	40,593 (107)	44,134 (139)	25,781 (88)	66,530 (207)	37,817 (121)	28,884 (100)	35,684 (99)	40,540 (109)		
180	—	—	—	—	—	46,973 (149)	38,92 (128)	45,154 (130)	49,794 (141)		

change the topology of the rest of the stramenopiles, except the placement of Nanomonadea and Placididea (Fig. 1). The phylogenomic trees inferred from these two supermatrices are summarized in Fig. 1.

In both trees of '231-supermatrix', C60-PMSF and CAT-PMSF (Figs. 1 and 2), the newly added ochrophyte transcriptomes showed similar topologies as ones reported in previous phylogenetic analyses based on SSU rDNA sequences and conserved plastid genes. With robust node support, we recovered Chrysophyceae + Synurophyceae + Synchromophyceae (CSS) + Picophagea (Pico) as monophyletic in all trees examined, as previously reported in Barcytė et al. (2021) and Guillou et al. (1999) (Fig. 1; Table 3). This relationship was also observed in the only other phylogenomic analysis with a comprehensive ochrophyte dataset (Terpis et al., 2024). Schizocladiophyceae is sister to Phaeophyceae, while Phaeothamniophyceae is a sister-lineage to Phaeophyceae-Xanthophyceae-Schizocladiophyceae (Fig. 1). This placement of Schizocladiophyceae is found in previous studies (Barcytė et al., 2021; Graf et al., 2020; Yang et al., 2012). However, the placement of Phaeothamniophyceae showed more inconsistency within Raphidophyceae-Phaeophyceae-Xanthophyceae (RPX) clades. As we found here, Phaeothamniophyceae falls sister to PX-Schizocladiophyceae in a five-gene maximum-likelihood (ML) tree in Graf et al. (2020), which had extensive taxon sampling across RPX lineages. In other studies, Phaeothamniophyceae was the sister-lineage to PX in a two-gene ML tree (Barcytė et al., 2021) or Xanthophyceae in a 10-gene ML tree (Riisberg et al., 2009; Wetherbee et al., 2019).

Our present dataset is still missing representatives of three ochrophyte classes (Aurearenophyceae, Chrysoparadoxophyceae, and Phaeosacciophyceae). These missing classes have been shown to belong to the PX clade, which forms a monophyletic group in previous multigene phylogenetic analyses, along with Raphidophyceae (Yang et al., 2012; Wetherbee et al., 2019; Graf et al., 2020). A recent phylogenomic study that included the latter two ochrophyte classes showed Phaeothamniophyceae as the sister group of Phaeosacciophyceae while Chrysoparadoxophyceae to Xanthophyceae, both with strong BS supports (Terpis et al., 2024). The absence of these classes therefore, account for the low BS values for PX in our phylogenomic analyses (53 % BS in '231-supermatrix' C60-PMSF; 95 % in CAT-PMSF) (Fig. 1).

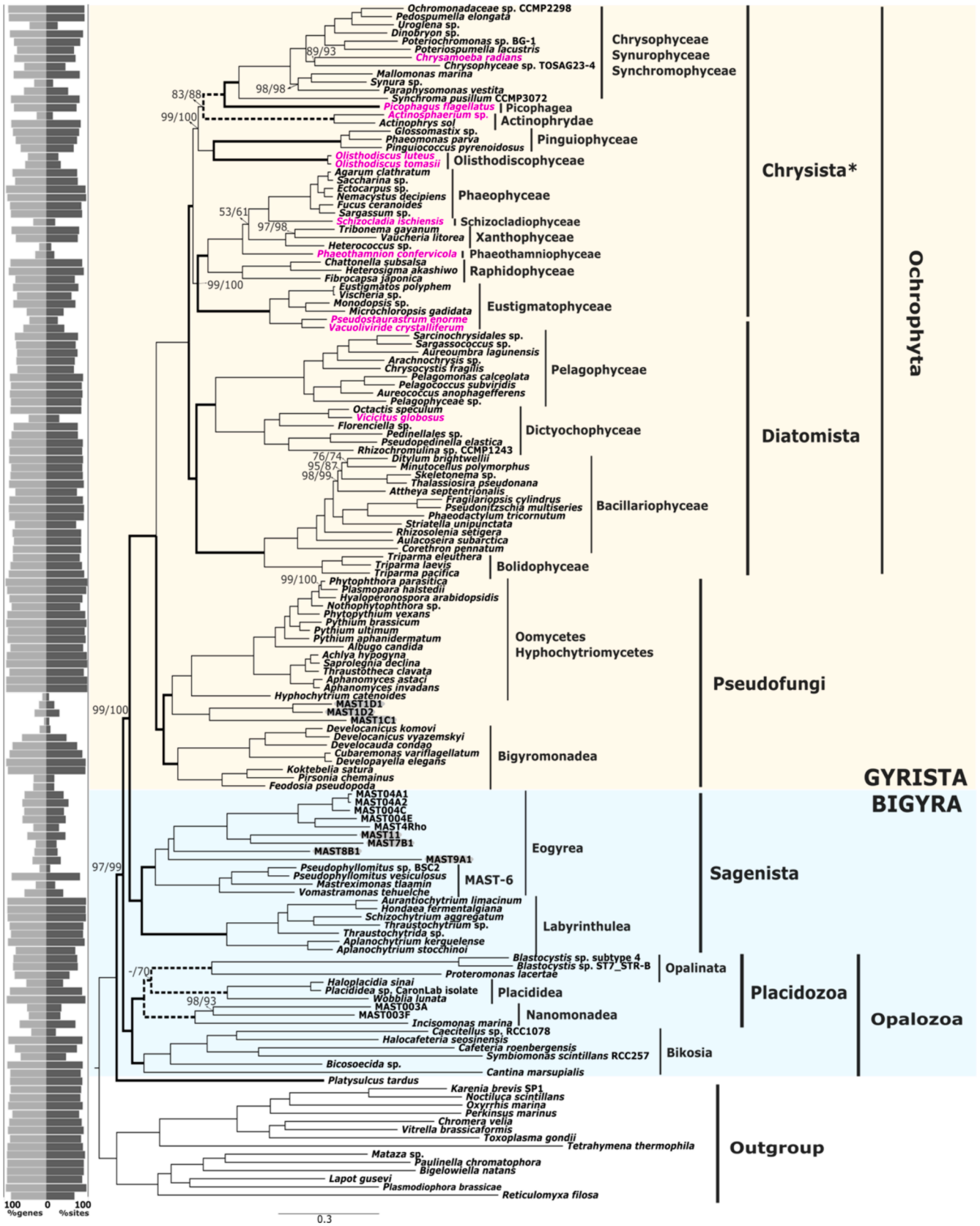
The two Actinophrydae taxa are sister to CSS + Pico, although with a modest BS support of 83 % (Fig. 1). This relationship was also recovered in Cho et al. (2024), but only when genes with a minimum 39 % completeness were selected. This instability was likely due to erosion of phylogenetic signal in Actinophrydae in our dataset. Our newly generated transcriptome of *Vicicitus globosus* was nested within the Dictyochophyceae (D) with 100 % BS support. *Vicicitus globosus* is known to produce a fast-acting cytotoxin (Chang, 2015) and its transcriptome was included in our analyses due to its availability at the time.

3.1.2. *Robust support for contentious lineages while breaking long branches*

Eustigmatophyceae (Eustig) is composed of the sub-groups Eustigmataceae, Monodopsidaceae, Neomonodaceae, and Goniochloridales (Amaral et al., 2020), but had been frequently represented only by a single taxon from Monodopsidaceae (i.e., *Microchloropsis gaditana*) (for an exception, see Terpis et al., 2024). Pinguiophyceae has been represented by one or two taxa, and is sometimes omitted entirely (Derelle et al., 2016; Thakur et al., 2019). To "break" these long branches, we added newly generated and publicly available transcriptomes belonging to different Eustigmatophyceae sub-groups and Pinguiophyceae.

We recovered a robust monophyly of RPX and Eustigmatophyceae (RPX + Eustig) in a majority of our trees (Figs. 1 and 2; Table 3), a previously contentious topology (Di Franco et al., 2022). This relationship was also observed in the recent phylogenomic analysis that included more Eustigmatophyceae subgroups (Terpis et al., 2024). Eustigmatophyceae is the sister lineage to CSS in a phylogenomic tree inferred from plastid genes (Di Franco et al., 2022; Ševčíková et al., 2015), while it is sister to RPX in a nuclear phylogeny (Azuma et al., 2022; Burki et al., 2016; Cho et al., 2022, 2024; Derelle et al., 2016; Di Franco et al., 2022; Noguchi et al., 2016; Terpis et al., 2024; Thakur et al., 2019). However, the latter studies only included a single taxon from Eustigmatophyceae, which likely contributes to weak bootstrap supports. Two chains of our Bayesian analysis did recover the Eustigmatophyceae grouping close to CSS, along with Olisthodiscophyceae and Actinophrydae (Fig. S1 Chain 1 and 2), however with lower average posterior probabilities (PP = 1 and 0.71), while the two other chains with the Eustig + RPX grouping both had PP = 1. We observed close groupings of Eustigmatophyceae with CSS in only two trees generated from different supermatrices For example, clades comprising [(CSS + Pico) + Olis] + Eustig and (CSS + Pico)+(Eustig + Actino) were observed in trees inferred from C60 and F140 supermatrices, respectively (Table 3).

Although we replicated a similar placement of Eustigmatophyceae that would be observed in trees inferred from plastid genes, we speculate that these groupings are due to small data size (C60) and/or LBA



(caption on next page)

Fig. 1. Combined Maximum-likelihood (ML) multi-gene trees of stramenopiles with 10 new transcriptomes from under-represented ochrophyte lineages (pink): '231-supermatrix' C60-PMSF and '233-supermatrix' C60-PMSF. The trees were constructed from a 231 gene-alignment of 125 stramenopiles and 14 outgroup taxa (72,932 aa sites), and a 233 gene-alignment of 132 stramenopiles and 14 outgroup (73,440 aa sites), under model LG + C60 + F + G4 + PMSF with 100 non-parametric bootstrap replicates each (BS). Only nodes with ≤ 99 % support, and support values that were different between the two analyses ('231-supermatrix' and '233-supermatrix') are labelled. All other nodes indicate BS = 100. Dashed line in the BS value indicates the topology was not recovered for the corresponding supermatrix ('231-supermatrix'/'233-supermatrix'). The bold black branches indicate the topologies of major classes or sub-groups that were found in a majority of phylogenomic trees that were constructed using various gene filtering criteria. The dotted lines of the tree branches indicate that the relationships were not recovered in the majority of the phylogenomic trees constructed from difference supermatrices (see Fig. 2 and Table 3). The taxa names with the gray highlights are the additional taxa used to concatenate '233-supermatrix', and not included in the gene-filtering analysis. The asterisk (*) denotes Chrysista Cavalier-Smith, 1986, its description did not include Eustigmatophyceae, Actinophrydae, Pinguiophyceae, and Olisthodiscophyceae. The percent genes (light grey) and sites (dark grey) occupied for each taxon are shown on the mirrored bar plot. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

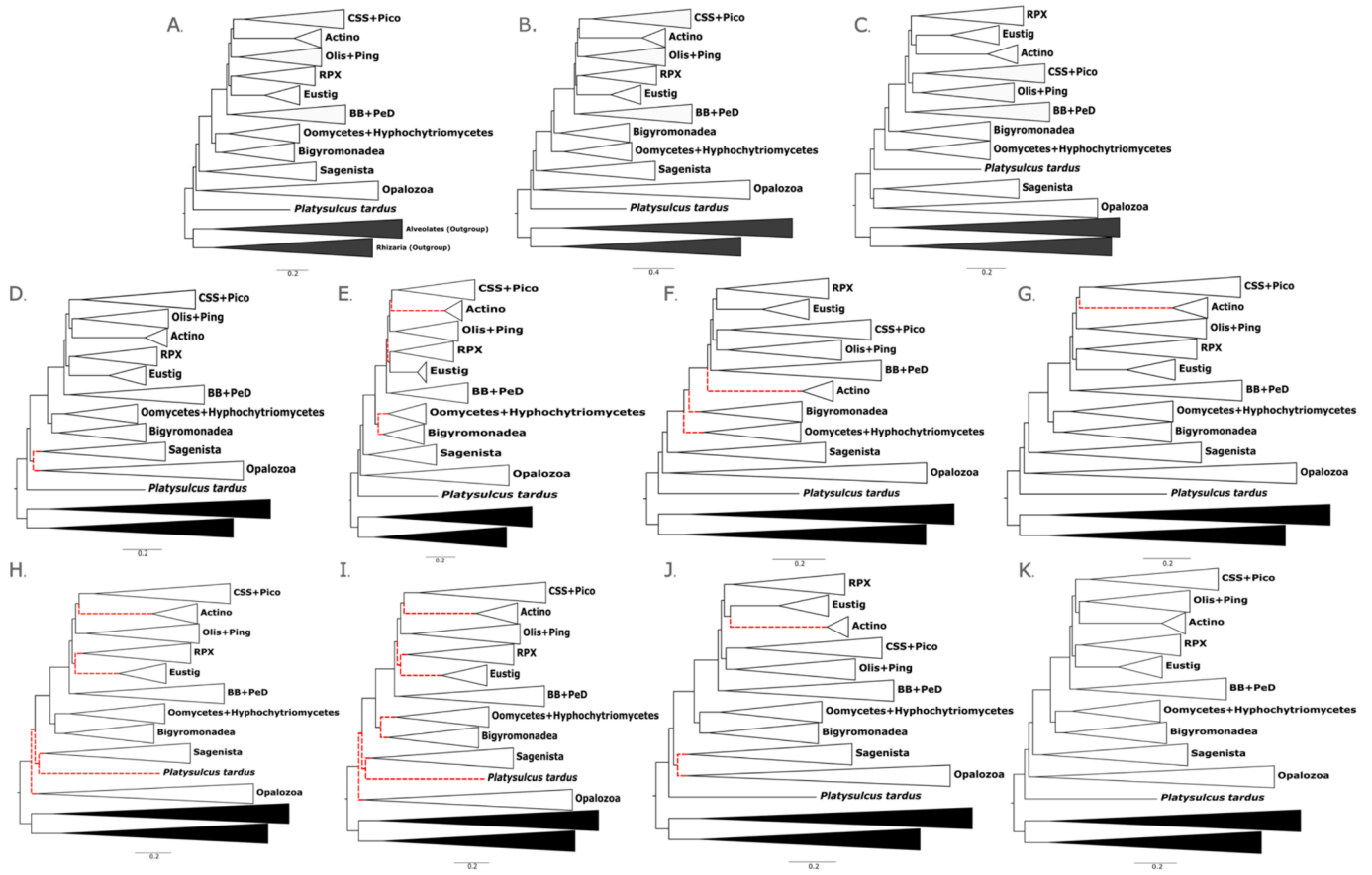


Fig. 2. Schematic representation of stramenopile topologies recovered from phylogenomic analyses reconstructed with various gene-filtering criteria and inference methods. A = unfiltered '231-supermatrix' C60-PMSF, '233-supermatrix' C60-PMSF; B = CAT-PMSF; C = criterion N; D = criteria A and B120-160; E = C120-160; F = D120-160; G = ABC120-160; H = E120-180; I = F140-180; J = S140-180; K = Q120-180. The sub-group topologies within the collapsed groups were ignored (e.g., placements of taxa within Opalozoa, RPX, and BB + PeD). For unstable topologies within the same criterion, the branches are marked with dotted red lines, otherwise, all other branches were consistently recovered in the phylogenomic trees generated within each criterion. Black groupings indicate outgroups. CSS = Chrysophyceae-Synurophyceae-Synchromophyceae; Pico = Picophagea; Olis = Olisthodiscophyceae; Ping = Pinguiophyceae; BB = Bolidophyceae-Bacillariophyceae; PeD = Pelagophyceae-Dictyochophyceae; RPX = Raphidophyceae-Phaeophyceae-Xanthophyceae; Actino = Actinophrydae; Eustig = Eustigmatophyceae. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

artefacts (Eustig + Actino in F140), rather than replicating effects of phylogenetic signals or bias of plastid genes. Instead, we suspect that the incongruence observed in nuclear versus plastid trees is the result of molecular convergence arising from non-neutral (adaptive) selection force. Molecular convergence arising from neutral or random mutations (e.g., homoplasy) can be remediated with current phylogenetic site-heterogeneous mixture models (Lartillot and Philippe, 2004; Wang et al., 2018, 2008). However, adaptive force on plastid genes across eukaryotes can result in strong phylogenetic signal in these genes (Edwards, 2009; Stiller et al., 2003). For example, plastids of chrysophytes are under directional selection in their genome reduction (Dorrell et al., 2019; Kim et al., 2020) while balancing selection maintains the same suites of plastid genes observed in both apicomplexans and chrysophytes (Dorrell et al., 2019). Therefore, one should be conservative when inferring phylogeny of ochrophytes using plastid genes. Further investigation on the effects on non-neutral forces on plastid and nuclear genes may help understanding the incongruence between the two datasets (Castoe et al., 2009; Stiller et al., 2003). Additionally, it may be worthwhile examining the gene properties of plastid genes and compare them with those of nuclear genes.

We observed a clade comprising Olisthodiscophyceae + Pinguiophyceae (Olis + Ping) in almost all trees examined, including the ones with fast-evolving sites, random sites, and random genes removed. (Figs. 1 and 2; Table 3; Fig. S2). This clade was the sister group of CSS,

Table 3

List of stramenopile groupings and their standard bootstrap support from the highest to the lowest prevalence observed in trees constructed from supermatrices obtained with different criteria (A-F, ABC, N, S, and Q), along with '231-supermatrix' C60-PMSF and CAT-PMSF. The numbers in brackets indicate the number of occurrences out of all 16 trees considered in the table. For each criterion, we selected shared genes within top 60 to 180 highest or lowest values found in all corresponding properties. Controversial groupings are bolded and underlined. Each criterion is denoted by A = selecting for genes with high values of treeness and occupancy; B = selecting for genes with high values average_BS_support, robinson_sim, and gene length; C = selecting for genes with low values of av_patristic, rate, and treelength; D = filter out gens with high values of av_patristic, rate, and treelength; ABC = combination of A-C criteria with corresponding top cut-off values; N = genes that are not explained well by the PC axes (low cos2); E = selecting genes with high values of PC1 associated biases (saturation, av_patristic, and root_tip_var); F = selecting genes with high values of all biases (RCFV, saturation, av_patristic, and root_tip_var), S = selecting genes with high signals (average_BS_support, robinson_sim, treeness); Q = selecting genes with high data quality (gene length and occupancy). CSS = Chrysophyceae-Synurophyceae-Synchromophyceae; Pico = Picophagea; Olis = Olisthodiscophyceae; Ping = Pinguiophyceae; BB = Bolidophyceae-Bacillariophyceae; PeD = Pelagophyceae-Dictyochophyceae; Bigyro = Bigyromonadea; Oomy = Oomycetes-Hyphochytriomycetes; Platy = Platysulcidae; RPX = Raphidophyceae-Phaeophyceae-Xanthophyceae; Actino = Actinophryidae; Ochro = Ochrophyta; Eustig = Eustigmatophyceae. For Diatomista + Chrysita*, the relationship only considered general grouping of (CSS + RPX)+(BB + PeD), regardless of the placements of Eustig, Actino, Olis, and Ping.

Groupings	Criterion A						Criterion B						Criterion C						Criterion D						N
	60	80	100	120	140	160	60	80	100	120	140	160	60	80	100	120	140	160	60	80	100	120	140	160	
CSS + Pico (46)	100	100	100	100	100	100	99	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Olis + Ping (44)	89	100	99	99	100	100	62	73	83	94	99	95	—	88	99	100	100	100	100	100	100	100	100	99	79
BB + PeD (43)	95	100	100	100	100	100	—	—	73	88	92	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Diatomista + Chrysista (40)*	—	100	100	100	100	100	—	—	73	100	100	100	—	89	72	—	100	100	100	100	100	100	80	74	100
RPX + Eustig (35)	85	94	92	99	100	99	38	67	65	86	94	92	—	—	—	64	94	99	99	100	100	93	84	69	67
Bigyro + Oomy (32)	—	—	94	100	91	99	—	—	67	92	96	96	74	—	—	—	86	93	99	100	100	83	—	—	—
Platy + rest (26)	—	100	95	100	100	100	—	—	—	100	100	100	—	100	100	100	100	100	100	100	100	100	100	100	—
[CSS + Pico] + [Olis + Ping] (16)	—	—	—	—	—	—	—	77	59	—	—	—	—	77	70	73	—	—	—	—	—	—	76	71	72
Bigyro + Ochro (15)	72	74	—	—	—	—	49	63	—	—	—	—	—	79	94	98	—	—	—	—	—	—	96	93	78
[CSS + Pico] + Actino (14)	—	91	63	—	—	—	—	—	—	—	—	—	—	—	—	—	81	86	67	84	87	93	—	—	—
[[CSS + Pico] + Actino]+ [Olis + Ping]]+ [RPX + Eustig] (14)	—	76	82	—	—	—	—	—	—	—	—	—	—	—	—	—	77	83	100	99	93	83	—	—	—
[Ping + Olis] + Actino (14)	—	—	—	71	63	75	—	—	—	73	58	72	—	—	—	—	—	—	—	—	—	—	—	—	—
[CSS + Pico] + [[Ping + Olis]+ Actino] (14)	—	—	—	95	98	94	—	—	—	93	98	95	—	—	—	—	—	—	—	—	—	—	—	—	—
Sagenista + Opalozoa (12)	67	96	100	100	92	—	—	—	—	95	92	—	—	—	—	—	—	—	—	—	—	—	—	—	90
Platy + Sagenista (8)	—	—	—	—	—	—	—	69	78	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Actino + Ochro (7)	—	—	—	—	—	—	—	—	—	—	—	—	100	100	100	100	—	—	—	—	—	—	100	100	—
Eustig + Actino (7)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	64
[[CSS + Pico] + [Olis + Ping]]+ RPX (4)	—	—	—	—	—	—	—	—	—	—	—	—	—	50	63	—	—	—	—	—	—	—	—	—	—
RPX + [Eustig + Actino] (3)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	67
[[CSS + Pico] + [Olis + Ping]]+ Actino (3)	—	—	—	—	—	—	—	47	71	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[Gyrista + Sagenista]+ Platy (3)	—	—	—	—	—	—	72	—	—	—	—	—	86	—	—	—	—	—	—	—	—	—	—	—	—
[BB + PeD] + Eustig (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	52	82	—	—	—	—	—	—	—	—	—	—
BB + Ochro (2)	—	—	—	—	—	—	84	100	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[BB + PeD] + [CSS + Pico] (2)	70	—	—	—	—	—	52	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[PeD + BB]+ [RPX + Eustig] (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	56	—	—	—	—	—	—	—	—	—
[[[PeD + [CSS + Pico]]+ [[RPX + Eustig]+ [Ping + Olis]]] + Actino (1)	—	—	—	—	—	—	30	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[[CSS + Pico] + Olis] +Eustig (1)	—	—	—	—	—	—	—	—	—	—	—	—	42	—	—	—	—	—	—	—	—	—	—	—	—

(*continued on next page*)

Table 3 (*continued*)

Groupings	Criterion A						Criterion B						Criterion C						Criterion D						N
	60	80	100	120	140	160	60	80	100	120	140	160	60	80	100	120	140	160	60	80	100	120	140	160	
[CSS + Pico] + PeD (1)	—	—	—	—	—	—	52	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[RPX + Eustig] + PeD (1)	—	—	—	—	—	—	—	46	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[RPX + Eustig] + [Olis + Ping] (1)	—	—	—	—	—	—	22	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[[BB + PeD] + [CSS + Pico]]+ Actino (1)	45	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Platy + Gyrista (1)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	99
[CSS + Pico] + [Eustig + Actino] (1)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Groupings	**Criterion E**				**Criterion F**		**Criterion S**						**Criterion Q**				**Criterion ABC**						**231-supermatrix ML-PMSF**	**CAT-PMSF**
	120	**140**	**160**	**180**	**120**	**140**	**160**	**180**	**120**	**140**	**160**	**180**	**120**	**140**	**160**	**180**	**60**	**80**	**100**	**120**	**140**	**160**		
CSS + Pico (46)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Olis + Ping (44)	75	96	100	100	68	98	100	100	70	98	99	100	91	93	98	100	96	100	100	100	100	100	100	100
BB + PeD (43)	99	100	100	100	99	98	97	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Diatomista + Chrysista (40)*	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	—	100	100	100	100	100	100	100
RPX + Eustig (35)	—	—	97	96	—	—	96	99	67	89	—	—	98	97	100	99	87	100	100	100	100	100	99	98
Bigyro + Oomy (32)	99	100	100	100	—	—	98	100	89	97	93	98	92	100	93	100	—	—	—	93	96	98	100	—
Platy + rest (26)	—	—	—	100	—	—	—	100	100	100	100	100	100	100	100	100	—	100	100	100	100	100	100	100
[CSS + Pico] + [Olis + Ping] (16)	47	60	—	—	49	—	—	—	—	—	94	91	—	—	—	—	75	66	—	—	—	—	—	—
Bigyro + Ochro (15)	—	—	—	—	82	87	—	—	—	—	—	—	—	—	—	—	96	97	99	—	—	—	—	88/85
[CSS + Pico] + Actino (14)	—	—	88	91	—	—	96	80	—	—	—	—	—	—	—	—	—	—	—	—	69	66	83	78/81
[[CSS + Pico] + Actino]+ [Olis + Ping]]+ [RPX + Eustig] (14)	—	—	99	100	—	—	99	99	—	—	—	—	—	—	—	—	—	—	—	—	100	100	100	100
[Ping + Olis] + Actino (14)	—	—	—	—	—	—	—	—	71	69	—	—	80	50	70	77	—	—	100	66	—	—	—	—
[CSS + Pico] + [[Ping + Olis]+ Actino] (14)	—	—	—	—	—	—	—	—	88	84	—	—	96	89	93	100	—	—	98	98	—	—	—	—
Sagenista + Opalozoa (12)	—	—	—	90	—	—	—	91	95	95	95	—	—	—	—	—	—	—	—	—	—	—	—	—
Platy + Sagenista (8)	100	95	99	—	65	97	100	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Actino + Ochro (7)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	100	—	—	—	—	—	—	—
Eustig + Actino (7)	53	72	—	—	68	66	—	—	—	—	90	81	—	—	—	—	—	—	—	—	—	—	—	—
[[CSS + Pico] + [Olis + Ping]]+ RPX (4)	46	60	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
RPX + [Eustig + Actino] (3)	—	—	—	—	—	—	—	—	—	—	84	74	—	—	—	—	—	—	—	—	—	—	—	—
[[CSS + Pico] + [Olis + Ping]]+ Actino (3)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	81	—	—	—	—	—	—
[Gyrista + Sagenista]+ Platy (3)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	89	—	—	—	—	—	—	—
[BB + PeD] + Eustig (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
BB + Ochro (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[BB + PeD] + [CSS + Pico] (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[PeD + BB] + [RPX + Eustig] (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	49	—	—	—	—	—	—	—
[[[PeD + [CSS + Pico]]+ [[RPX + Eustig]+ [Ping + Olis]]] + Actino (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[[CSS + Pico] + Olis] +Eustig (1)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

(*continued on next page*)

Table 3 (*continued*)

Groupings	Criterion A					Criterion B					Criterion C					Criterion D					N
	60	80	100	120	140	160	60	80	100	120	140	160	60	80	100	120	140	160	60	80	100	120	140	160	
[CSS + Pico] + PeD (1)	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	
[RPX + Eustig] + PeD (1)	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	
[RPX + Eustig] + [Olis + Ping] (1)	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	
[[BB + PeD] + [CSS + Pico]]+ Actino (1)	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	
Platy + Gyrista (1)	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	
[CSS + Pico] + [Eustig + Actino] (1)	|	|	|	|	|	61	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	

often with strong branch support (Fig. 1; Table 3), and was also observed in a previous phylogenomic study (Terpis et al., 2024). The close relationship between Pinguiophyceae and CSS has been demonstrated in several studies, including those with plastid genes, however these either only used a single taxon representing Pinguiophyceae or recovered lower bootstrap supports for this relationship (Burki et al., 2016; Cho et al., 2022; Di Franco et al., 2022; Noguchi et al., 2016). As with Eustig + RPX, half of the Bayesian chains (Fig. S1: Chain 1 and 2) had different placements of Pinguiophyceae (branching sister to Diatomista, consisting of Pelagophyceae, Dictyochophyceae, Bolidophyceae, and Bacillariophyceae).

The newly added ochrophyte data broke many long branches leading to Eustigmatophyceae, CSS, Pinguiophyceae, and Actinophrydae. Pseudofungi (Oomycotes, Hyphochytriomycetes, and Bigyromonadea) is a clade branching sister to the rest of the Ochrophyta with 100 % BS support. The same topology was observed in the tree recovered from the '233-supermatrix' analysis, but most branches had higher BS supports (Fig. 1). We observed a clade comprising Bigyromonadea and Ochrophyta in the CAT-PMSF tree (Fig. 2; Table 3) with up to 88 % BS.

3.1.3. Examining phylogenomic relationships with the Bayesian analysis

Overall, the Bayesian analysis was inconclusive even with 20,000 generations, as none of the chains converged (maxdiff = 1). However, the topology of chain 1 and 2 were identical except for the outgroup (Fig. S1), while the topology of chain 3 and 4 had the same topology in Gyrista (Fig. S1). The topology of the ochrophytes were almost the same (except for the placement of *Attheya septentrionalis*; Bacillariophyceae) between chains 1–4 and the C60-PMSF tree inferred from '231-supermatrix' (Fig. 1). In all chains of the Bayesian analysis, *A. septentrionalis* is sister to a clade of pennate diatoms while it is sister to a clade of centric diatoms in C60-PMSF (Fig. 1). This conflicting placement of *A. septentrionalis* can also be found in previous studies (Dorrell et al., 2021; Parks et al., 2018; Theriot et al., 2015, 2010) where different set sizes of genes were sampled; small subunit ribosomal genes and plastid genes (Theriot et al., 2015, 2010), high occupancy orthologs (58,294 sites) found in diatoms (Parks et al., 2018) or ochrophytes (26,399 sites) (Dorrell et al., 2021).

For Bigyra, we found paraphyly similar to that observed by Cho et al., (2024) in addition to the unstable groupings within Placidozoa (Fig. 1; Fig. S1). In all consensus trees from the Bayesian analysis and the '233-supermatrix', Nanomonadea (MAST-3) is sister to the rest of the Placidozoa (data not shown), as was also observed in Cho et al. (2024). This is likely because of a LBA artefact due to lack of taxon sampling in Opalinata and MAST-12 (Cho et al., 2024; Kolodziej and Stoeck, 2007; Okamura and Kondo, 2015).

3.2. No filtering criteria to select "good" or "bad" genes for phylogenomic analyses

Due to the presence of many phylogenetically contentious lineages in stramenopiles, particularly in Ochrophyta, we initially aimed to resolve phylogenomic relationships by selecting genes with high phylogenetic signal and/or low noise, while also increasing taxon sampling. A principal component analysis (PCA) of 13 gene properties that are proxies for sources of known phylogenetic noise, signal, and data quality, revealed far a more complex relationship (all values of the 13 properties are summarized in Table S1). As a result, it was challenging to devise suitable filtering criteria that could discern "good" or "bad" gene properties (Fig. 3A; Fig. S4). In contrast to the results from the work of Mongiardino Koch (2021), who established this method by testing on more recently diverged (121.8 to 479.1 million years old) organisms (Mongiardino Koch, 2021), our stramenopile dataset did not have a clear separation between phylogenetic signal and noise affecting genes along the two PC axes. Moreover, the two PC axes only explained 51.8 % of the total variance while some gene properties have high loadings on an additional PC axis (Fig. S3). This made the delineation of "good" or

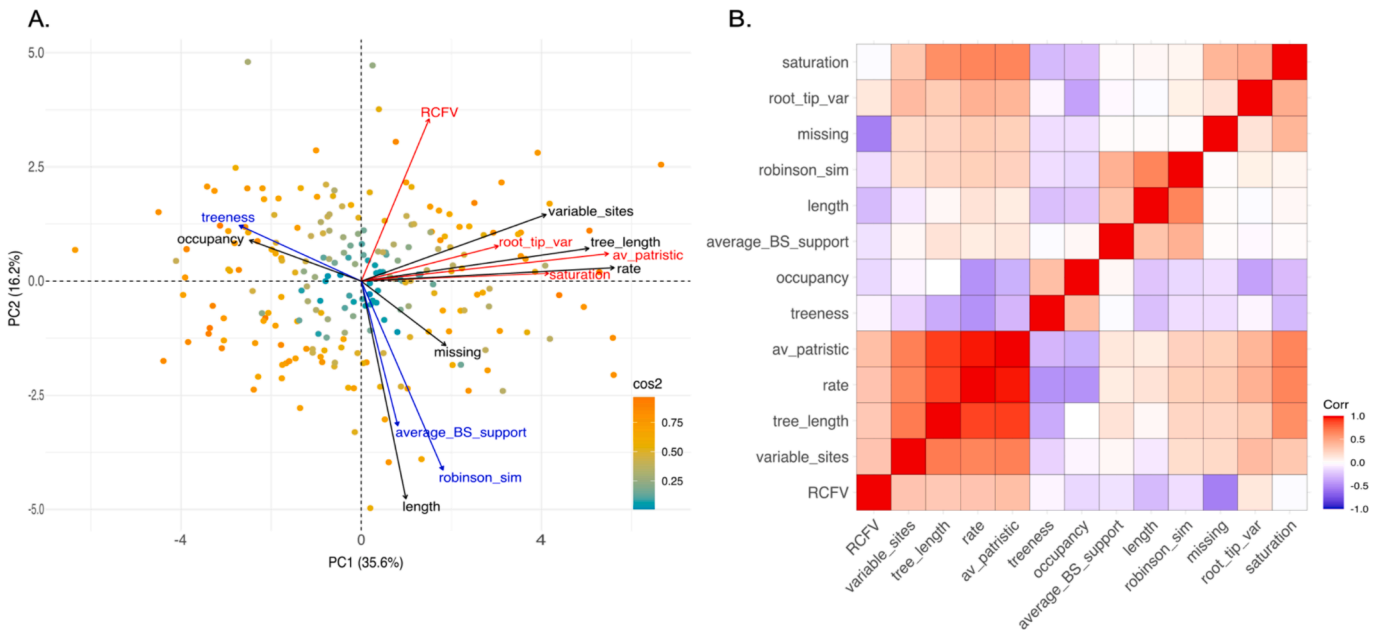


Fig. 3. Thirteen gene properties summarized in a principal component analysis (PCA) plot and a correlation matrix. (A) PCA plot of 229 genes. Each coloured dot indicates a gene, plotted onto two principal component (PC1 and PC2) axes. High cos2 values are orange and low cos2 values are blue. Higher cos2 values indicate the genes are represented well by the two PC axes. The 13 properties are shown as variables each coloured by noise (red), signal (blue), and data quality (black). (B) Correlation matrix with hierarchical clustering of 13 gene properties. Positive correlations are indicated by red and negative correlations are indicated by blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

"bad" genes further challenging.

We observed that the majority of noise (e.g., saturation, av_patristic, and root_tip_var – coloured in red in Fig. 3A) had higher vector loadings with principal component 1 (PC1), however the two groups of phylogenetic signal (criteria A and B) were explained with different PC axes (Fig. 3; Fig. S3). The rest of the noise, RCFV (coloured in red in Fig. 3A), a proxy for aa composition bias, was explained mostly by PC2 (i.e., higher vector loading with PC2) along with some properties that are potential indicators of phylogenetic signal (e.g., average_BS_support, robinson_sim – coloured in blue in Fig. 3A), although in an opposing direction (i.e., negative correlation). The two properties—treeness and occupancy—were explained by PC1 but negatively correlated with the noise and some of the proxies for data quality and information (Fig. 3; Fig. S3). Consequently, we included various filtering criteria (criteria A-D) by PC loadings and their correlations (Fig. S3-4) among different properties regardless of the nature (e.g., noise, signal, or data quality) of the gene properties, to maximize the genes sampled. Additionally, not all the gene properties of the same nature showed strong positive correlations (Fig. 3; Fig. S4). We also observed that the higher data quality does not necessarily correlate with indicators of phylogenetic signal. For example, average_BS_support and occupancy are negatively correlated while robinson_sim and rate are positively correlated (Fig. 3; Fig. S4). The presence of many recalcitrant nodes, an older evolutionary history with the estimated origin of 719–414 million years ago (Ma) for ochrophytes (Brown and Sorhannus, 2010; Choi et al., 2024) and 1077–1025 Ma for the rest of the stramenopiles (Yoon et al., 2004), and early rapid radiations are likely some cause for the difference between our stramenopile dataset and the dataset analysed by the initial research that established this method (Mongiardino Koch, 2021).

3.2.1. *Evolutionary rate provides phylogenetic signal but correlates with noise*

Among all the gene properties calculated, 'evolutionary rate' had the highest vector loading (0.448) along PC1, followed closely by 'av_patristic' and 'tree_length' (0.446 and 0.415, respectively) (Fig. S3). Strictly speaking, 'evolutionary rate' and 'tree length' are a measure of information. However, due to strong positive correlations among the 'evolutionary rate' and 'tree length' with noise (e.g., 'saturation', 'av_patristic', and 'root_tip_var'), and neutral or negative correlation with most of phylogenetic signal, we treated them as noise in our analyses (Fig. 3B; Fig. S4). Similarly, we treated 'gene alignment' as an indicator of phylogenetic signal based on its strong positive correlation with 'average_BS_support' and 'robinson_sim'. Along PC2, 'alignment length' had the highest vector loading (0.571) followed by 'robinson_sim' (0.513) (Fig. S3).

Rapid evolutionary rate has been previously reported to cause saturation as the number of possible mutation states for each nucleotide or amino acid character is limited (Felsenstein, 1978; Philippe et al., 2005; Superson and Battistuzzi, 2022). As a result, without significantly limiting the number of sites, removal of fast-evolving sites and genes has been used to minimize noise (Bapteste et al., 2007; Edwards, 2016; Philippe et al., 2005; Superson and Battistuzzi, 2022). However, despite their correlation with other noise in this study (Fig. 3B; Fig. S4), rate and tree length (both used to estimate rate) should not be solely regarded as sources of noise. In a simplified simulation of evolutionary processes, Revell et al. (2008) showed that under weak stabilizing selection, high mutation rate can provide a more informative signal, while observing no correlation with rate and phylogenetic signal under a constant genetic drift. The authors proposed that phylogenetic signal is affected by the non-neutral selection force, rather than just the rate, as it can be significantly decreased by divergent selection (leading to speciation) or increased with an initially high rate that slowed over time (i.e., rate variation), or high rate of niche occupancy (Revell et al., 2008). This means that filtering by criteria A (selecting for genes with high values of treeness and occupancy), B (selecting for genes with high values of average_BS_support, robinson_sim, and gene length), and C (selecting for genes with low values root_tip_var, av_patristic, rate, and saturation) might have resulted in significant losses of these phylogenetic signal.

3.3. *Phylogenomic analyses using different filtering criteria*

Based on the 13 gene properties calculated, we generated a total of 46 supermatrices and subsequent phylogenomic trees to examine the effects of gene properties on phylogenomic analyses (Table 3; Supp 1).

To minimize the effect of small data size (i.e., number of amino acid sites) on our phylogenomic analyses, we compared the average BS support of all trees reconstructed from random site or gene removal datasets to the C60-PMSF tree reconstructed from the 231-supermatrix (Fig. 1). Based on the change of backbone topologies and their average BS supports, supermatrices with an average BS less than 89 % were deemed too small to sufficiently distinguish from the effects of different gene-filtering criteria and small data size. Therefore, we only considered the topologies of supermatrices with size larger than ~ 22,000 sites (e.g., criteria A120-160; B120-160; C100-160; all D and ABC) (Fig. 2; Table 2-3).

For criteria A (selecting high values of treeness and occupancy) and B (selecting for high values of average_BS_support, robinson_sim, and gene length), the ochrophyte topology was similar in general to the '231-supermatrix' under C60-PMSF (Fig. 2A and D; Table 3).

To investigate the effects of signal, noise, and data quality alone, we included additional filtering criteria (criteria S, E, F, and Q) to compare the trees with those reconstructed from supermatrices A-D and ABC (Fig. 2). When we compared the topologies of trees reconstructed from criteria A, B, and S, most of the topologies (including the instability of Sagenista and Opalozoa) were the same, except the placement of Actinophrydae (Fig. 2A and J). These criteria all selected for high signal while criteria A and B distinguished the signal associated PC axis in addition to other highly correlated gene properties (i.e. data quality and information). Interestingly, trees reconstructed from high data quality (criterion Q) had the most stable topologies (Fig. 2K), all of which were identical to the'231-supermatrix' C60-PMSF, except the placement of Actinophrydae. For trees reconstructed from supermatrices C120-160 (select genes with low noise and associated properties), there were more unstable topologies (including Pseudofungi and Actinophrydae) compared to the ones reconstructed with criteria A, B, S, and Q (Fig. 2D, E, J, K). Similarly, the trees reconstructed from supermatrix D120-160 (Fig. 2F) showed unstable topology of Pseudofungi and the placement of Actinophrydae (Fig. 2F).

When we examined the trees reconstructed from supermatrices E and F (selecting genes with high noise), the placement of *Platysulcus tardus* became unstable, no longer branching sister to the rest of the ochrophytes (Fig. 2H and I; Table 3). Other "deep-branching" lineages such as Opalozoa and Sagenista were also affected, although the same instability was observed in trees reconstructed from different criteria (e.g., N, A and B120-160). It is likely that these "deep-branching" lineages maybe more sensitive to data size and phylogenetic noise, probably due to phylogenetic signal present in a smaller set of genes compared to the later-diverged lineages. This was also observed when random sites and genes were removed – many lineages belonging to Gyrista remained consistent with more sites or genes removed, compared to Opalozoa and Sagenista. For some instances, Eustigmatophyceae was sister to Actinophrydae, in which the clade branched sister to Chrysista or CSS (Fig. 2H and I; Table 3), a latter topology observed in plastid multi-gene trees (Barcytė et al., 2021; Di Franco et al., 2022; Ševčíková et al., 2019).

The majority of the Actinophrydae (Actino) placement was observed to be sister to CSS + Picophagea (CSS + Pico) or Olis + Ping, each relationship with the same frequency (14 occurrences) (Fig. 2; Table 3). The latter relationship was present in supermatrices A and B120-160, S120-140, and Q120-180, selecting for genes with high signal, data quality and other properties that were correlated. The clade of Actinophrydae with CSS + Pico was observed in trees reconstructed from supermatrices E and F160-180, A80-100, C140-160, D60-120, even though some criteria select for genes with high noise (criteria E and F) while others select for low noise (criterion C), or remove ones with high noise (criterion D). It is likely that as the data size increases for each criterion, there are more overlapping genes sampled (Fig. S5). However, Actino + [CSS + Pico] was also recovered in '231-supermatrix' C60-PMSF (Fig. 1). We suspect that this particular topology is influenced by a small number genes (Shen et al., 2017) and various filtering criteria that removed any of these genes may have recovered alternative placements of Actinophrydae. The placement of Actinophrydae to the rest of the ochrophytes were observed in seven out of 46 trees, mostly from supermatrices C and D with lower data size (C60-120 and D140-160) (Table 3) and this is the topology observed in Azuma et al. (2022). The placement of Acitnophrydae being sister to the rest of the ochrophytes is likely due to selecting for slow evolving genes thereby eroding phylogenetic signal and its effect likely more pronounced in smaller data size. Micrographs of *Actinosphaerium* sp. is available in Fig. S6.

To lessen the loss of rate-derived phylogenetic signal that might be present in genes affected by high rate or tree length, we combined the filtered genes of each criterion's top-ranking values (i.e., criterion ABC60-160). Excluding the placements of Actinophrydae, the rest of the topologies had the same relationships as the ones found in '231-supermatrix' C60-PMSF (Fig. 2A and G).

The placement of Bigyromonadea being sister to ochrophytes was observed with criteria that had relatively small data sizes (e.g. N, A and B60-80, C80-120, D140-160, ABC60-100), and with ones that select for genes with high noise (F120-140) (Table 2-3). Thus, the grouping of Bigyromonadea + Ochrophytes may have been the result of lack of phylogenetic signal arising from small data size or the effect of compositional bias (Fig. 2I). When we incrementally removed fast-evolving sites, we observed the monophyly of Pseudofungi (oomycetes, hyphochytriomycetes, and Bigyromonadea) in trees with up to 67 % aa sites removed (Fig. S2A). Even when we randomly removed amino acid sites, bigyromonads formed a monophyly with oomycetes, and Platysulcidae remained sister to rest of the stramenopiles in most cases (Fig. S2B). When we randomly removed genes in 20 % increments, monophyly of Bigyromonadea + Oomycetes were observed most of the times, even when up to 60 % of genes (139 genes) were removed (Fig. S2C). Values of saturation and missing data for each of the supermatrices using to compare topologies (Fig. 2) are summarized in Table S2.

3.3.1. *Different types of compositional heterogeneity may recover different topologies*

Compositional heterogeneity in phylogenomic inferences has been known to cause LBA, mainly due to lack of models that account for this (Jimenez et al., 2018; Koshi and Goldstein, 1995; Szantho et al., 2023). We used relative composition frequency variability (RCFV) as a proxy for compositional heterogeneity among branch terminals, to evaluate disproportionate amino acid composition across different taxa. However, compositional variation also occurs across sites and through time as a result of selection pressures, constraints on protein folding sites or preferential traits due to environmental factors (Boussau et al., 2008; Jimenez et al., 2018; Koshi and Goldstein, 1995; Szantho et al., 2023). To account for across-site compositional heterogeneity, we followed the C60-PMSF (Quang et al., 2008; Wang et al., 2018) and CAT-PMSF pipelines (Szantho et al., 2023). The resulting trees largely showed the same topology, except for the placement of Bigyromonadea (Fig. 2B). When we compared the trees inferred from supermatrices E and F, the monophyly of Bigyromondea and Oomycetes were no longer stable in F (selecting for genes with all the high noise, including RCFV) (Fig. 2H and I). It is beyond the scope of this work to account how the two different inference methods (CAT-PMSF vs C60-PMSF) may have influenced compositional biases across sites and taxa. However, based on trees inferred from other selecting criteria (Fig. 2), we speculate that paraphylectic relationship of Bigyomonadea and Oomycetes is an artefact of across-taxa amino acid compositional bias (i.e., RCFV).

4. Conclusion

To resolve the placement of several contentious lineages of stramenopiles, we increased taxon-sampling for the group and conducted phylogenomic analyses using various phylogenetic inference methods. We recovered robust relationships of previously phylogenomically

under-studied or contentious lineages such as Eustigmatophyceae, Olisthodiscophyceae, and Pinguiophyceae. Additionally, based on 13 proxies for phylogenetic noise, signal, and qualities for each gene, we constructed numerous supermatrices based on different criteria selecting for genes with high signal or low noise. We found the tree topologies were more stable when we select for genes with high signal and data quality. Selecting the most conserved and slowest evolving genes resulted in the most variable and incongruent tree topologies across the trees examined. Furthermore, when considering the effect of compositional heterogeneity on phylogenomic inferences, we should be conservative in our interpretation as different types of compositional variations exist along with different methods to remediate it. Future efforts should include devising systematic evaluation criteria that select for genes with high signal and quality while removing genes highly affected by noise. Additionally, finding the minimum set of genes that encompasses all these criteria may lessen computational resources and time, a challenge inherent to phylogenomic analyses.

CRediT authorship contribution statement

Anna Cho: Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Gordon Lax:** Writing – review & editing, Validation, Investigation. **Patrick J. Keeling:** Writing – review & editing, Validation, Supervision, Resources, Project administration.

Acknowledgement

We thank Donald Wong for advice on RNA extraction, Juan Saldarriaga for encouraging the initiation of the study, Elizabeth Cooney for improving the manuscript better; Corey Holt and Jan Finke for help with R; Andrew Roger and Hector Baños for advice on phylogenomic analyses. We also thank the Statistical Opportunity for Students (SOS) program in UBC for their insights on gene filtering criteria. We also like to express our gratitude toward Sean Graham, Dolph Schluter, Heroen Verbruggen, and the reviewers for thorough and insightful discussions and inquisition. This work was funded by grants from NSERC Grant 2014-03994 to PJK. AC was also supported by NSERC-CGSD and UBC 4YF.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2024.108120.

References

Ali, R.H., Bogusz, M., Whelan, S., 2019. Identifying clusters of high confidence homologies in multiple sequence alignments. Mol. Biol. Evol. 36, 2340–2351. https://doi.org/10.1093/molbev/msz142.

Amaral, R., Fawley, K.P., Němcová, Y., Ševčíková, T., Lukešová, A., Fawley, M.W., Santos, L.M.A., Eliáš, M., 2020. Toward modern classification of eustigmatophytes, including the description of neomonodaceae Fam. Nov. and Three New Genera [1]. J. Phycol. 56, 630–648. https://doi.org/10.1111/jpy.12980.

Andersen, R.A., 1991. Algal culturing techniques. Elsevier Academic Press, Oxford, UK.

Andersen, R.A., Potter, D., Bidigare, R.R., Latasa, M., Rowan, K., O'Kelly, C.J., 1998. Characterization and phylogenetic position of the enigmatic golden alga Phaeothamnion confervicola: ultrastructure, pigment composition and partial SSU rDNA sequence. J. Phycol. 34, 286–298. https://doi.org/10.1046/j.1529-8817.1998.340286.x.

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Apt, K.E., Clendennen, S.K., Powers, D.A., Grossman, A.R., 1995. The gene family encoding the ucoxanthin chlorophyll proteins from the brown alga Macrocystis pyrifera. Mol Gen Genet 246, 455–464. https://doi.org/10.1007/BF00290449.

Azuma, T., Pánek, T., Tice, A.K., Kayama, M., Kobayashi, M., Miyashita, H., Suzaki, T., Yabuki, A., Brown, M.W., Kamikawa, R., 2022. An Enigmatic Stramenopile Sheds Light on Early Evolution in Ochrophyta Plastid Organellogenesis. Mol. Biol. Evol. 39, msac065. https://doi.org/10.1093/molbev/msac065.

Baños, H., Susko, E., Roger, A.J., 2023. Is Over-parameterization a Problem for Profile Mixture Models? Systematic Biology syad063. https://doi.org/10.1093/sysbio/syad063.

Bapteste, E., Susko, E., Leigh, J., Ruiz-Trillo, I., Bucknam, J., Doolittle, W.F., 2007. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. Mol. Biol. Evol. 25, 83–91. https://doi.org/10.1093/molbev/msm229.

Barcytė, D., Eikrem, W., Engesmo, A., Seoane, S., Wohlmann, J., Horák, A., Yurchenko, T., Eliáš, M., 2021. *Olisthodiscus* represents a new class of Ochrophyta. J. Phycol. 57, 1094–1118. https://doi.org/10.1111/jpy.13155.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N., Gouy, M., 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature 456, 942–945. https://doi.org/10.1038/nature07393.

Brown, J.W., Sorhannus, U., 2010. A molecular genetic timescale for the diversification of autotrophic stramenopiles (ochrophyta): substantive underestimation of putative fossil ages. PLoS One 5, e12759.

Burki, F., Kaplan, M., Tikhonenkov, D.V., Zlatogursky, V., Minh, B.Q., Radaykina, L.V., Smirnov, A., Mylnikov, A.P., Keeling, P.J., 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc. r. Soc. b. 283, 20152802. https://doi.org/10.1098/rspb.2015.2802.

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 8, giz100. https://doi.org/10.1093/gigascience/giz100.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

Castoe, T.A., De Koning, A.P.J., Kim, H.-M., Gu, W., Noonan, B.P., Naylor, G., Jiang, Z.J., Parkinson, C.L., Pollock, D.D., 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. USA 106, 8986–8991. https://doi.org/10.1073/pnas.0900233106.

Cavalier-Smith, T., Chao, E.-E.-Y., 2006. Phylogeny and megasystematics of phagotrophic heterokonts (Kingdom Chromista). J Mol Evol 62, 388–420. https://doi.org/10.1007/s00239-004-0353-8.

Challis, R., Richards, E., Rajan, J., Cochrane, G., Blaxter, M., 2020. BlobToolKit – interactive quality assessment of genome assemblies. Genes Genomes Genetics 10, 1361–1374.

Chang, F., 2015. Cytotoxic effects of vicicitus globosus (class dictyochophyceae) and chattonella marina (class raphidophyceae) on rotifers and other microalgae. JMSE 3, 401–411. https://doi.org/10.3390/jmse3020401.

Cho, A., Lax, G., Keeling, P., 2024a. Phylogenomic analyses of ochrophytes (stramenopiles) with an emphasis on neglected lineages . https://doi.org/10.5061/dryad.f4qrfj73q.

Cho, A., Tikhonenkov, D.V., Hehenberger, E., Karnkowska, A., Mylnikov, A.P., Keeling, P.J., 2022. Monophyly of diverse Bigyromonadea and their impact on phylogenomic relationships within stramenopiles. Mol. Phylogenet. Evol. 171, 107468 https://doi.org/10.1016/j.ympev.2022.107468.

Cho, A., Tikhonenkov, D.V., Lax, G., Prokina, K.I., Keeling, P.J., 2024. Phylogenomic position of genetically diverse phagotrophic stramenopile flagellates in the sediment-associated MAST-6 lineage and a potentially halotolerant placididean. Mol. Phylogenet. Evol. 190, 107964 https://doi.org/10.1016/j.ympev.2023.107964.

Choi, S.-W., Graf, L., Choi, J.W., Jo, J., Boo, G.H., Kawai, H., Choi, C.G., Xiao, S., Knoll, A.H., Andersen, R.A., Yoon, H.S., 2024. Ordovician origin and subsequent diversification of the brown algae. Curr. Biol. 34, 740–754.e4. https://doi.org/10.1016/j.cub.2023.12.069.

Derelle, R., López-García, P., Timpano, H., Moreira, D., 2016. A Phylogenomic framework to study the diversity and evolution of stramenopiles (=Heterokonts). Mol Biol Evol 33, 2890–2898. https://doi.org/10.1093/molbev/msw168.

Di Franco, A., Baurain, D., Glöckner, G., Melkonian, M., Philippe, H., 2022. Lower statistical support with larger datasets: insights from the Ochrophyta radiation. Mol. Biol. Evol. 39, msab300. https://doi.org/10.1093/molbev/msab300.

Dong, S., Wang, Y., Xia, N., Liu, Y., Liu, M., Lian, L., Li, N., Li, L., Lang, X., Gong, Y., Chen, L., Wu, E., Zhang, S., 2022. Plastid and nuclear phylogenomic incongruences and biogeographic implications of *Magnolia* s.l. (Magnoliaceae). J. Sytemat. Evolut. 60, 1–15. https://doi.org/10.1111/jse.12727.

Dorrell, R.G., Azuma, T., Nomura, M., Audren De Kerdrel, G., Paoli, L., Yang, S., Bowler, C., Ishii, K., Miyashita, H., Gile, G.H., Kamikawa, R., 2019. Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. Proc. Natl. Acad. Sci. USA 116, 6914–6923. https://doi.org/10.1073/pnas.1819976116.

Dorrell, R.G., Villain, A., Perez-Lamarque, B., Audren De Kerdrel, G., McCallum, G., Watson, A.K., Ait-Mohamed, O., Alberti, A., Corre, E., Frischkorn, K.R., Pierella Karlusich, J.J., Pelletier, E., Morlon, H., Bowler, C., Blanc, G., 2021. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. Proc. Natl. Acad. Sci. USA 118. https://doi.org/10.1073/pnas.2009974118 e2009974118.

Driskell, A.C., Ané, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306, 1172–1174. https://doi.org/10.1126/science.1102036.

Edwards, S.V., 2009. Natural selection and phylogenetic analysis. Proc. Natl. Acad. Sci. USA 106, 8799–8800. https://doi.org/10.1073/pnas.0904103106.

Edwards, S.V., 2016. Phylogenomic subsampling: a brief review. Zool Scr 45, 63–74. https://doi.org/10.1111/zsc.12210.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systemic Biology 27, 401–410. https://doi.org/10.1093/sysbio/27.4.401.

Graf, L., Yang, E.C., Han, K.Y., Küpper, F.C., Benes, K.M., Oyadomari, J.K., Herbert, R.J. H., Verbruggen, H., Wetherbee, R., Andersen, R.A., Yoon, H.S., 2020. Multigene phylogeny, morphological observation and re-examination of the literature lead to the description of the phaeosacciophyceae classis nova and four new species of the heterokontophyta SI clade. Protist 171, 125781. https://doi.org/10.1016/j.protis.2020.125781.

Guillard, R.R.L., Ryther, J.H., 1962. Studies of marine planktonic diatoms: I. Cyclotella nana Hustedt, and Detonula confervacea (Cleve) Gran. Can. J. Microbiol. 8, 229–230. https://doi.org/10.1139/m62-029.

Guillard, R.R.L., 1975. Culture of phytoplankton for feeding marine invertebrates., in: Culture of Marine Invertebrate Animals. Springer, Boston, MA, pp. 29–60.

Guillou, L., Chrétiennot-Dinet, M.-J., Boulben, S., Moon-van Der Staay, S.Y., Vaulot, D., 1999. Symbiomonas scintillans gen. et sp. nov. and Picophagus flagellatus gen. et sp. nov. (Heterokonta): Two New Heterotrophic Flagellates of Picoplanktonic Size. Protist 150, 383–398. https://doi.org/10.1016/S1434-4610(99)70040-4.

Haas, B.J., Kamoun, S., Zody, M.C., Jiang, R.H.Y., Handsaker, R.E., Cano, L.M., Grabherr, M., Kodira, C.D., Raffaele, S., Torto-Alalibo, T., Bozkurt, T.O., Ah-Fong, A. M.V., Alvarado, L., Anderson, V.L., Armstrong, M.R., Avrova, A., Baxter, L., Beynon, J., Boevink, P.C., Bollmann, S.R., Bos, J.I.B., Bulone, V., Cai, G., Cakir, C., Carrington, J.C., Chawner, M., Conti, L., Costanzo, S., Ewan, R., Fahlgren, N., Fischbach, M.A., Fugelstad, J., Gilroy, E.M., Gnerre, S., Green, P.J., Grenville-Briggs, L.J., Griffith, J., Grünwald, N.J., Horn, K., Horner, N.R., Hu, C.-H., Huitema, E., Jeong, D.-H., Jones, A.M.E., Jones, J.D.G., Jones, R.W., Karlsson, E.K., Kunjeti, S.G., Lamour, K., Liu, Z., Ma, L., MacLean, D., Chibucos, M.C., McDonald, H., McWalters, J., Meijer, H.J.G., Morgan, W., Morris, P.F., Munro, C.A., O'Neill, K., Ospina-Giraldo, M., Pinzón, A., Pritchard, L., Ramsahoye, B., Ren, Q., Restrepo, S., Roy, S., Sadanandom, A., Savidor, A., Schornack, S., Schwartz, D.C., Schumann, U.D., Schwessinger, B., Seyer, L., Sharpe, T., Silvar, C., Song, J., Studholme, D.J., Sykes, S., Thines, M., Van De Vondervoort, P.J.I., Phuntumart, V., Wawra, S., Weide, R., Win, J., Young, C., Zhou, S., Fry, W., Meyers, B.C., Van West, P., Ristaino, J., Govers, F., Birch, P.R.J., Whisson, S.C., Judelson, H.S., Nusbaum, C., 2009. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature 461, 393–398. https://doi.org/10.1038/nature08358.

Hedtke, S.M., Townsend, T.M., Hillis, D.M., 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Syst. Biol. 55, 522–529. https://doi.org/10.1080/10635150600697358.

Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38, 297. https://doi.org/10.2307/2992396.

Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47, 3–8. https://doi.org/10.1080/106351598260987.

Hillis, D.M., Pollock, D.D., McGuire, J.A., Zwickl, D.J., 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52, 124–126. https://doi.org/10.1080/10635150390132911.

Jimenez, M.J., Arenas, M., Bastolla, U., 2018. Substitution rates predicted by stability-constrained models of protein evolution are not consistent with empirical data. Mol. Biol. Evol. 35, 743–755. https://doi.org/10.1093/molbev/msx327.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software Version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. https://doi.org/10.1093/molbev/mst010.

Kawai, H., Maeba, S., Sasaki, H., Okuda, K., Henry, E.C., 2003. Schizocladia ischiensis: A new filamentous marine chromophyte belonging to a new class, schizocladiophyceae. Protist 154, 211–228.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Rynearson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaulot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z., 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol 12, e1001889.

Kim, J.I., Jeong, M., Archibald, J.M., Shin, W., 2020. Comparative plastid genomics of non-photosynthetic chrysophytes: genome reduction and compaction. Front. Plant Sci. 11, 572703 https://doi.org/10.3389/fpls.2020.572703.

Kocot, K.M., Struck, T.H., Merkel, J., Waits, D.S., Todt, C., Brannock, P.M., Weese, D.A., Cannon, J.T., Moroz, L.L., Lieb, B., Halanych, K.M., 2016. Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. Syst Biol syw079. https://doi.org/10.1093/sysbio/syw079.

Kolodziej, K., Stoeck, T., 2007. Cellular identification of a novel uncultured marine stramenopile (MAST-12 Clade) small-subunit rRNA gene sequence from a norwegian estuary by use of fluorescence in situ hybridization-scanning electron microscopy. Appl. Environ. Microbiol. 73, 2718–2726. https://doi.org/10.1128/AEM.02158-06.

Koshi, J.M., Goldstein, R.A., 1995. Context-dependent optimal substitution matrices. Protein Eng. 8, 641–645. https://doi.org/10.1093/protein/8.7.641. PMID: 8577693.

Labarre, A., López-Escardó, D., Latorre, F., Leonard, G., Bucchini, F., Obiol, A., Cruaud, C., Sieracki, M.E., Jaillon, O., Wincker, P., Vandepoele, K., Logares, R., Massana, R., 2021. Comparative genomics reveals new functional insights in uncultured MAST species. ISME J 15, 1767–1781. https://doi.org/10.1038/s41396-020-00885-8.

Laetsch, D.R., Blaxter, M.L., 2017. BlobTools: Interrogation of genome assemblies. F1000Res 6, 1287. https://doi.org/10.12688/f1000research.12232.1.

Lanyon, S.M., 1988. The stochastic mode of molecular evolution: what consequences for systematic investigations? Auk 105, 565–573. https://doi.org/10.1093/auk/105.3.565.

Lartillot, N., Brinkmann, H., Philippe, H., 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7, S4. https://doi.org/10.1186/1471-2148-7-S1-S4.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109. https://doi.org/10.1093/molbev/msh112.

Lehman, J.T., 1967. Ecological and nutritional studies on Dinobryon Ehrenb.: Seasonal periodicity and the phosphate toxicity problem. Limnol. Oceanogr. 21, 646–658. https://doi.org/10.4319/lo.1976.21.5.0646.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–636. https://doi.org/10.1093/sysbio/46.3.523.

Mongiardino Koch, N., 2021. Phylogenomic subsampling and the search for phylogenetically reliable Loci. Mol. Biol. Evol. 38, 4025–4038. https://doi.org/10.1093/molbev/msab151.

Mongiardino Koch, N., Thompson, J.R., 2021. A Total-evidence dated phylogeny of echinoidea combining phylogenomic and paleontological data. Syst. Biol. 70, 421–439. https://doi.org/10.1093/sysbio/syaa069.

Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274. https://doi.org/10.1093/molbev/msu300.

Nichols, R., 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 16, 358–364. https://doi.org/10.1016/S0169-5347(01)02203-0.

Noguchi, F., Tanifuji, G., Brown, M.W., Fujikura, K., Takishita, K., 2016. Complex evolution of two types of cardiolipin synthase in the eukaryotic lineage stramenopiles. Mol. Phylogenet. Evol. 101, 133–141. https://doi.org/10.1016/j.ympev.2016.05.011.

Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Müller, W.E.G., Nickel, M., Schierwater, B., Vacelet, J., Wiens, M., Wörheide, G., 2013. Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. 67, 223–233. https://doi.org/10.1016/j.ympev.2013.01.010.

Okamura, T., Kondo, R., 2015. Suigetsumonas clinomigrationis gen. et sp. nov., a Novel Facultative Anaerobic Nanoflagellate Isolated from the Meromictic Lake Suigetsu. Japan. Protist 166, 409–421. https://doi.org/10.1016/j.protis.2015.06.003.

Onsbring, H., Tice, A.K., Barton, B.T., Brown, M.W., Ettema, T.J.G., 2020. An efficient single-cell transcriptomics workflow for microbial eukaryotes benchmarked on Giardia intestinalis cells. BMC Genomics 21, 448. https://doi.org/10.1186/s12864-020-06858-7.

Pardo-De La Hoz, C.J., Magain, N., Piatkowski, B., Cornet, L., Dal Forno, M., Carbone, I., Miadlikowska, J., Lutzoni, F., 2023. Ancient rapid radiation explains most conflicts among gene trees and well-supported phylogenomic trees of nostocalean cyanobacteria. Syst. Biol. 72, 694–712. https://doi.org/10.1093/sysbio/syad008.

Parks, M.B., Nakov, T., Ruck, E.C., Wickett, N.J., Alverson, A.J., 2018. Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). Am. J. Bot. 105, 330–347. https://doi.org/10.1002/ajb2.1056.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., Delsuc, F., 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5, 50. https://doi.org/10.1186/1471-2148-5-50.

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9, e1000602.

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R., 2014. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc 9, 171–181. https://doi.org/10.1038/nprot.2014.006.

Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alie, A., Morgenstern, B., Manuel, M., Worheide, G., 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. 27, 1983–1987. https://doi.org/10.1093/molbev/msq089.

Provasoli, L., Pintner, I.J., 1959. Artificial media for fresh-water algae: problems and suggestions. in: Ecology of Algae, 2. University of Pittsburgh, Pittsburgh, pp. 84–96.

Quang, L.S., Gascuel, O., Lartillot, N., 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24, 2317–2323. https://doi.org/10.1093/bioinformatics/btn445.

Revell, L.J., Harmon, L.J., Collar, D.C., 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57, 591–601. https://doi.org/10.1080/10635150802302427.

Richter, D.J., Berney, C., Strassert, J.F.H., Poh, Y.-P., Herman, E.K., Muñoz-Gómez, S.A., Wideman, J.G., Burki, F., De Vargas, C., 2022. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. Peer Commun. J. 2, e56. https://doi.org/10.24072/pcjournal.173.

Riisberg, I., Orr, R.J.S., Kluge, R., Shalchian-Tabrizi, K., Bowers, H.A., Patil, V., Edvardsen, B., Jakobsen, K.S., 2009. Seven Gene Phylogeny of Heterokonts. Protist 160, 191–204. https://doi.org/10.1016/j.protis.2008.11.004.

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147. https://doi.org/10.1016/0025-5564(81)90043-2.

Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331. https://doi.org/10.1038/nature12130.

Seemann, T., 2007. Barrnap 0.9: BAsic Rapid Ribosomal RNA Predictor.

Ševčíková, T., Horák, A., Klimeš, V., Zbránková, V., Demir-Hilton, E., Sudek, S., Jenkins, J., Schmutz, J., Přibyl, P., Fousek, J., Vlček, Č., Lang, B.F., Oborník, M.,

Worden, A.Z., Eliáš, M., 2015. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci Rep 5, 10134. https://doi.org/10.1038/srep10134.
Ševčíková, T., Yurchenko, T., Fawley, K.P., Amaral, R., Strnad, H., Santos, L.M.A., Fawley, M.W., Eliáš, M., 2019. Plastid genomes and proteins illuminate the evolution of eustigmatophyte algae and their bacterial endosymbionts. Genome Biol. Evol. 11, 362–379. https://doi.org/10.1093/gbe/evz004.
Shen, X.-X., Salichos, L., Rokas, A., 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol Evol 8, 2565–2580. https://doi.org/10.1093/gbe/evw179.
Shen, X.-X., Hittinger, C.T., Rokas, A., 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol 1, 0126. https://doi.org/10.1038/s41559-017-0126.
Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
Smith, S.A., Brown, J.W., Walker, J.F., 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. PLoS One 13, e0197433.
Song, L., Florea, L., 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaSci 4, 48. https://doi.org/10.1186/s13742-015-0089-y.
Stiller, J.W., Reel, D.C., Johnson, J.C., 2003. A single origin of plastids revisited: convergent evolution in organellar genome content. J. Phycol. 39, 95–105. https://doi.org/10.1046/j.1529-8817.2003.02070.x.
Struck, T.H., 2014. TreSpEx—detection of misleading signal in phylogenetic reconstructions based on tree information. Evol Bioinform Online 10, EBO.S14239. https://doi.org/10.4137/EBO.S14239.
Superson, A.A., Battistuzzi, F.U., 2022. Exclusion of fast evolving genes or fast evolving sites produces different archaean phylogenies. Mol. Phylogenet. Evol. 170, 107438 https://doi.org/10.1016/j.ympev.2022.107438.
Szantho, L.L., Lartillot, N., Szollosi, G.L., Schrempf, D., 2023. Compositionally constrained sites drive long branch attraction. Syst. Biol. 72, 767–780. https://doi.org/10.1093/sysbio/syad013.
Terpis, K.X., Salomaki, E.D., Barcytė, D., Pánek, T., Verbruggen, H., Kolisko, M., Bailey, J.C., Eliáš, M., Lane, C.E., 2024. Multiple plastid losses within photosynthetic stramenopiles revealed by comprehensive phylogenomics. bioRxiv. https://doi.org/10.1101/2024.02.03.578753.
Thakur, R., Shiratori, T., Ishida, K., 2019. Taxon-rich multigene phylogenetic analyses resolve the phylogenetic relationship among deep-branching stramenopiles. Protist 170, 125682. https://doi.org/10.1016/j.protis.2019.125682.
Theriot, E.C., Ashworth, M., Ruck, E., Nakov, T., Jansen, R.K., 2010. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. Plecevo 143, 278–296. https://doi.org/10.5091/plecevo.2010.418.
Theriot, E.C., Ashworth, M.P., Nakov, T., Ruck, E., Jansen, R.K., 2015. Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. Mol. Phylogenet. Evol. 89, 28–36. https://doi.org/10.1016/j.ympev.2015.03.012.
Tice, A.K., Shadwick, L.L., Fiore-Donno, A.M., Geisen, S., Kang, S., Schuler, G.A., Spiegel, F.W., Wilkinson, K.A., Bonkowski, M., Dumack, K., Lahr, D.J.G., Voelcker, E., Clauß, S., Zhang, J., Brown, M.W., 2016. Expansion of the molecular and morphological diversity of Acanthamoebidae (Centramoebida, Amoebozoa) and identification of a novel life cycle type within the group. Biol Direct 11, 69. https://doi.org/10.1186/s13062-016-0171-0.
Tikhonenkov, D.V., Janouškovec, J., Keeling, P.J., Mylnikov, A.P., 2016. The morphology, ultrastructure and SSU rRNA gene sequence of a new freshwater flagellate, *Neobodo borokensis* n. sp. (Kinetoplastea, Excavata). J. Eukaryot. Microbiol. 63, 220–232. https://doi.org/10.1111/jeu.12271.
Wägele, J., Mayer, C., 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. BMC Evol. Biol. 7, 147. https://doi.org/10.1186/1471-2148-7-147.
Wang, H.-C., Susko, E., Spencer, M., Roger, A.J., 2008. Topological estimation biases with covarion evolution. J. Mol. Evol. 66, 50–60. https://doi.org/10.1007/s00239-007-9062-4.
Wang, H.-C., Minh, B.Q., Susko, E., Roger, A.J., 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67, 216–235. https://doi.org/10.1093/sysbio/syx068.
Watanae, M.M., Kawachi, M., Hiroki, M., Kasai, F., 2000. NIES Collection List of Strains, 6th Ed. ed. NIES, Japan.
Wetherbee, R., Jackson, C.J., Repetti, S.I., Clementson, L.A., Costa, J.F., Meene, A., Crawford, S., Verbruggen, H., 2019. The golden paradox – a new heterokont lineage with chloroplasts surrounded by two membranes. J. Phycol. 55, 257–278. https://doi.org/10.1111/jpy.12822.
Whelan, S., Irisarri, I., Burki, F., 2018. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. Bioinformatics 34, 3929–3930. https://doi.org/10.1093/bioinformatics/bty448.
Whelan, N.V., Kocot, K.M., Moroz, L.L., Halanych, K.M., 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA 112, 5773–5778. https://doi.org/10.1073/pnas.1503453112.
Whitfield, J.B., Lockhart, P.J., 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22, 258–265. https://doi.org/10.1016/j.tree.2007.01.012.
Yang, E.C., Boo, G.H., Kim, H.J., Cho, S.M., Boo, S.M., Andersen, R.A., Yoon, H.S., 2012. Supermatrix data highlight the phylogenetic relationships of photosynthetic stramenopiles. Protist 163, 217–231. https://doi.org/10.1016/j.protis.2011.08.001.
Yao, J., Fu, W., Wang, X., Duan, D., 2009. Improved RNA isolation from Laminaria japonica Aresch (Laminariaceae, Phaeophyta). J. Appl. Phycol. 21, 233–238. https://doi.org/10.1007/s10811-008-9354-0.
Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., Bhattacharya, D., 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol. Biol. Evol. 21, 809–818. https://doi.org/10.1093/molbev/msh075.
Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K.M., Struck, T.H., 2011. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. BMC Evol. Biol. 11, 369. https://doi.org/10.1186/1471-2148-11-369.
Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598. https://doi.org/10.1080/10635150290102339.