EVOLUTION AND DIVERSITY OF STRAMENOPILES

by

Anna Cho

M.Sc, The University of British Columbia, 2019
B.Sc (Hon.), Trent University, 2015
B.Sc., McGill University, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

May 2024

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Evolution and diversity of stramenopiles

submitted by Anna Cho in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

in Botany

Examining Committee:

Patrick Keeling, Professor, Botany, UBC
Supervisor

Brian Leander, Professor, Botany and Zoology, UBC
Supervisory Committee Member

Sean Graham, Professor, Botany, UBC
University Examiner

Dolph Schluter, Professor, Zoology. UBC
University Examiner

Heroen Verbruggen, Professor, School of BioSciences, University of Melbourne
External Examiner

Additional Supervisory Committee Members:

Patrick Martone, Professor, Botany, UBC
Supervisory Committee Member

Abstract

Stramenopiles are a diverse eukaryotic supergroup with considerable genomic information available. Nevertheless, the relationships between major stramenopiles subgroups remain unresolved and incongruent between analyses, in part due to a lack of data from small nanoflagellates that make up much of the genetic diversity of the group, under-represented ochrophyte classes, and a rapid radiation leading to eroded phylogenetic signals in the tree. As a result, assessing genetic diversity and distribution, addressing character evolution, and investigating interactions of these lineages with other organisms are limited. To resolve phylogenomic relationships of stramenopiles, I generated 23 transcriptomes from the most under-sampled subgroups, such as Bigyromonadea, MAST-6, Placididea, and some classes of ochrophytes that had been scarcely represented in phylogenomic data. Of these, 11 are new species of stramenopiles, some of which have helped resolving phylogenomic relationships of Bigyromonadea and the backbone of deep-branching lineages. Some of these species were found to be abundant in sediment sampled across different geographic locations, while others can tolerate a broad range of salinities. I also described behaviours and morphological characters of these species including the ability to form pseudopods and cell-aggregates observed in some bigyromonads. This updated phylogenomic dataset now represents 14 out of 17 classes of ochrophytes and demonstrates robust support for previously contentious or under-tested lineages such as Eustigmatophyceae, Pinguiophyceae, and Olisthodiscophyceae. To address phylogenomic incongruence between multi- or single-gene trees, I explored various gene filtering criteria to identify the phylogenetically informative genes. Selecting genes with long internal phylogenetic branches or removing genes with high levels of phylogenetic noise recovered more topologies that were found in other phylogenomic analyses. Finally, I investigated the only reported case of a prokaryotic endosymbiont found among non-photosynthetic lineages of stramenopiles, in the tiny flagellate *Symbiomonas scintillans*. Instead of endobacteria, I detected multiple giant viruses related to prasinoviruses. This work demonstrates how little we know about symbioses, particularly in nano- or pico-flagellates. Overall, this thesis highlights complex evolutionary histories of stramenopiles inferred from the most up-to-date phylogenomic tree. This work will inform further exploration into trait evolutions related to niche occupation, morphology, and immunity.

Lay Summary

Compared to other major groups of eukaryotes that are neither plants, animals, or fungi (protists), stramenopiles have been relatively well studied due to their economic and ecological significance. Notable examples are oomycetes, which infect many crops species and caused the Great Potato Famine, or the giant kelps and diatoms, which are photosynthetic stramenopiles that provide habitats for marine life and contribute >20% primary production globally. The rest of the stramenopiles are not as relevant to human interests as these groups, and indeed stramenopile studies are highly skewed towards photosynthetic or pathogenic species for this reason. In this thesis, I investigated diversity and abundance of under-studied stramenopiles by describing new species, morphologies, behaviours, symbionts, and distribution. To assess their evolutionary history, I constructed an evolutionary tree by using many genes, and inferred some relationships that were previously controversial or unknown.

Preface

A version of Chapter 2 has been published as a research article in Molecular Phylogenetics and Evolution, of which I was the first co-author. (Cho A, Tikhonenkov DV, Hehenberger E, Karnkowska A, Mylinkov AP, and Keeling PJ. 2022. Monophyly of diverse Bigyromonadea and their impact on phylogenomic relationships within stramenopiles. *Molecular Phylogenetics and Evolution* 171:107468). The project was done in collaboration with former members of the Keeling lab and Denis Tikhonenkov, Institute of Biology of Inland Waters in Russian Academy of Sciences (RAS), Moscow. Denis Tikhonenkov conducted field work, all cell isolations, culture establishment and maintenance, library preparation for transcriptomes. Denis Tikhonenkov also generated micrograph images of the cells with the morphology descriptions. Initial project assessment was done by Anna Karnkowska (Institute of Parasitology, Czech Academy of Science). Elisabeth Hehenberger partly processed transcriptome data for the "approach 2" mentioned in the paper. Alexander Mylinkov (RAS) was a supervisor of Denis Tikhonenkov during the time of field work. I performed all other bioinformatics analyses, determined the scientific goals of the project, and wrote the manuscript during the period when the lab was closed due to COVID restrictions.

A version of Chapter 3 has also been published as a research article (Cho A, Tikhonenkov DV, Lax G, Prokina KI, and Keeling PJ. 2024. *Molecular Phylogenetics and Evolution* 190:107964). This work was done in collaboration with Denis Tikhonenkov (RAS), Kristina Prokina (CNRS, Université Paris-Saclay), and Gordon Lax (a postdoctoral fellow in Patrick Keeling's UBC laboratory). Denis Tikhonenkov and Gordon Lax isolated four new species and micrograph images. Together with Denis Tikhonenkov and Gordon Lax, I wrote species description and taxonomic summaries, and conducted cDNA synthesis and library preparation to generate transcriptome data. Kristina Prokina helped with sampling. I wrote the manuscript and, conceptualized and conducted all other analyses including all the bioinformatics analyses and data curation.

Chapter 4 is a manuscript submitted for a publication (Cho A, Lax G, and Keeling PK. *Under review*. Phylogenomic analyses of ochrophytes (stramenopiles) with an emphasis on neglected lineages). This project was done in collaboration with Gordon Lax, a postdoctoral fellow in the Keeling lab at UBC. Gordon Lax isolated one cell from an environmental sample and performed cDNA synthesis to generate transcriptome data. I maintained cultures, performed RNA extraction, cDNA synthesis, and generated transcriptome for the rest of the data. I conceptualized the data analysis approach, conducted all the bioinformatics analyses, and wrote the manuscript.

A version of Chapter 5 has been published as a research article (Cho A, Lax G, Livingston SJ, Masukagami Y, Naumova M, Millar O, Husnik F, and Keeling PJ. *PLoS Genetics* 20(4):e1011218). This project was done in collaboration between Patrick Keeling's UBC laboratory and Filip Husnik's laboratory in Okinawa Institute of Science and Technology Graduate University, OIST, Japan. Patrick Keeling, Filip Husnik, and I identified the research question and designed the experiment. With some supervision from Gordon Lax, I generated whole genome amplification (WGA) data and Filip Husnik's laboratory generated shotgun metagenomic data (SGM). Sam Livingston, a postdoctoral fellow in the Keeling lab generated negatively stained electron micrograph images. Yumiko Masukagami (OIST) and I performed Fluorescence in situ hybridization (FISH). Mariia Naumova and Olivia Millar, both Filip Husnik's former students, maintained cultures in Japan and involved in laboratory work to generate the SGM data. I maintained the cultures in UBC. I wrote the manuscript and conducted all bioinformatics analyses of WGA and SGM, and discovered and assembled draft genomes of potential viruses associated with the culture maintained at UBC.

Patrick Keeling, the principal investigator and supervisor, was involved in all aspects of conceptualization, supervision, review, and editing of all projects and manuscripts. All the co-authors were involved in reviewing and editing the manuscripts.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

aa Amino acids

AU Approximately unbiased

BB Bacillariophyceae-Bolidophyceae

BC British Columbia

BS Bootstrap supports (synonymous with STB, standard bootstrap)

BLAST Basic Local Alignment Search Tool

BpV *Bathycoccus prasinos virus*

CSS Chrysophyceae-Synurophyceae-Synchromophyceae

DADA2 Divisive Amplicon Denoising Algorithm 2

DAPI 4’,6-diamidino-2-phenylindole

DNA Deoxyribonucleic acid

dsDNA Double-stranded DNA

FISH Fluorescence *in situ* hybridization

GC Guanine-cytosine

HGT Horizontal gene transfer

HMM Hidden Markov Models

MAFFT Multiple Alignment using fast Fourier transform

MAST Marine stramenopile

MCMC Markov Chain Monte Carlo

ML Maximum-likelihood

MOCH Marine ochrophyte

MpV *Micromonas pusilla virus*

NCLDVs Nucleocytoplasmic large DNA viruses

OIST Okinawa Institute of Science and Technology Graduate University

OlV *Ostreococcus lucimarinus virus*

ORFs Open reading frames

PCA Principal Component Analysis

PCR Polymerase chain reaction

PeD Pelagophyceae-Dictyochophyceae

PMSF Posterior mean site frequencies

RNA Ribonucleic acid

RPX Raphidophyceae-Phaeophyceae-Xanthophyceae

SAR Stramenopiles, Alveolates, Rhizaria

SEM Scanning electron microscopy

SGM Shotgun metagenomic

SSU rDNA Small subunit ribosomal DNA

SSU rRNA Small subunit ribosomal RNA

SsV *Symbiomonas scintillans* virus

STB Standard bootstrap

TEM Transmission electron microscopy

UA Uranyl acetate

UBC University of British Columbia

UFB Ultrafast bootstrap

vMAGs Viral metagenomically assembled genomes

VLP Virus-like particle

VOG Virus orthologous groups

WGA Whole-genome amplification

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Patrick Keeling, for all his supports, insightful advice, and encouragement throughout my PhD. His openness to new scientific approaches and discoveries was especially encouraging and inspiring for me to explore broad areas and, led to a serendipitous finding. I was also extremely lucky to be working with the past and present members of the Keeling Lab, despite the COVID pandemic lockdown. I was able to work with Dr. Denis Tikhonenkov on his brilliant collection of cells and learnt so much from Dr. Elisabeth Hehenberger and, my mentor, Dr. Anna Karnkowska about data processing, writing, and academic paths. In the lab, Dr. Liz Cooney, Dr. Gordon Lax, Dr. Vittorio Boscaro, Dr. Corey Halt, Dr. Filip Husnik, and Dr. Waldan Kwong were always there to help and discuss, without any hesitation. I want to thank all other present and past members who made the experience so much fun and made me truly enjoy science (Dr. Emma George, Dr. Sam Livingston, Dr. Noriko Okamoto, Dr. Juan Saldarriaga, Ina Na, Mahara Mtawali, Victoria Jacko-Reynolds, and so many others). I also want to thank Dr. Naomi Fast and her past lab members, Dr. Donald Wong, and Dr. Thomas Whelan, for their support, warm encouragements, and advice throughout the process. I would like to thank everyone for keeping the coffee room filled with sugar and carbs throughout.

I was also very lucky to collaborate and work with many brilliant scientists, including Dr. Liz Cooney, Dr. Gordon Lax, Dr. Denis Tikhonenkov, Dr. Kristina Prokina, Dr. Elisabeth Hehenberger, Dr. Anna Karnkowska, and Dr. Filip Husnik. I would like to thank my committee members, Dr. Brian Leander, and Dr. Patrick Martone for all the helpful feedback and their time. I also want to thank Dr. Brett Couch who made my TA experience so much more rewarding and fun. Thank you, Alice Lou, for being so patient and kind in answering all the questions throughout the program. I'd like to extend my thanks to UBC Sequencing and Bioinformatics Consortium and UBC Bioimaging Facility for the fast turn-around. Many thanks to Hakai Institute for making the sampling trips and writing retreats possible. Additionally, I thank many members of Hakai Institute (Matt Whelan, Derek Van Maanen, Brenton Twist, Krystal Bachen, Gillian Sadlier-Brown), and Camosun Bog (Gerry Mignault) for providing moss samples, although the project never took off.

This work was funded by the National Science and Engineering Resource Council (NSERC), the Gordon and Betty Moore Foundation, NSERC-Canada Graduate Scholarships, and the UBC Botany 4-Year Fellowship.

Finally, I want to thank my partner, family, and friends who helped me get through every step of the process. Thank you to all my dear friends (especially AS) whom I shared all the adventures, late-night drinks, food, skiing, and climbing. My loving partner, Jan Finke ("Dr. Pinky"), thank you for always being there during the lows and the highs, and listening to all my breakdowns and dealing with my gremlin modes. You are my science advisor, bioinformatics saviour, and motivator. Without the love and patience of Jan, friends, and family, I would not been where I am. Lastly, I thank God for always leading me somewhere I least expect and blessing me with never-ending learning, love, and people I am with.

Chapter 1: Introduction

Stramenopiles (= Heterokonts) are generally characterized by having two different lengths of flagella at some point in their life, with the anterior flagellum often bearing tripartite mastigonemes involved in swimming (and, rarely, gliding), and a posterior flagellum involved in beating and feeding. Such characterization does not fully reflect the morphological and genetic diversity of stramenopiles, as they cover broad ranges of sizes, occupied habitats, and nutritional modes. In fact, many stramenopiles have lost a flagellum such as *Symbiomonas scintillans* (Guillou et al., 1999)*,* zoospores of *Hyphochytrium catenoides* (Leonard et al., 2018), or tripartite hairs such as species of *Pelagomonas* (Andersen et al., 1993) and Pinguiophyceae (Kawachi et al., 2002). This supergroup has been proposed (Cavalier-Smith, 1997) to be divided into two major clades: (i) Gyrista, consisting of Ochrophyta, Bigyromonadea, and Oomycota; and (ii) Bigyra, consisting of Sagenista (including Labyrinthulids and Eogyrea (Cavalier-Smith and Scoble, 2013), Bikosia, and Placidozoa (Placididea, Opalinata and Nanomonadea). Phylogenetically, a single taxon *Platysulcus tardus*, represents a new class Platysulcea which is sister to all other stramenopiles (Shiratori et al., 2015). Many studies have focused on the plastid-bearing autotrophs (i.e., ochrophytes) or parasitic stramenopiles (e.g.,oomycetes, labyrinthulids, and opalinids) due to their ecological and economic impacts, and as a result the diversity of the rest of the small heterotrophic stramenopiles are mostly only known through environmental sequences (Massana et al., 2004) or morphological descriptions lacking genomic level data. Characterizing these under-studied stramenopiles will thus not only provide better insights into the evolutionary history of stramenopiles as a whole, but also set a stage to further investigate character evolution that may be relevant for studying diverse niche occupation, roles in food-webs, and virulence.

1.1 The phylogenomic relationship of non-photosynthetic stramenopiles

1.1.1 The unresolved relationship among the stramenopiles

Stramenopiles have been proposed to fall in a variety of places in the tree of eukaryotes, and these suggestions impact how we interpret the biology of the group. For example, they have been proposed to be a member of the Chromalveolata, based on the hypothesis that the last common ancestor of the stramenopiles and various other algal groups possessed a plastid of red-algal origin (Cavalier-Smith, 1999). However, this hypothesis has been challenged and alternative hypotheses have been suggested (Larkum et al., 2007; Dorrell and Smith, 2011), and overall this Chromalveolate Hypothesis remains contentious at best. Understanding phylogenetic relationships of stramenopiles affects how we interpret the evolution of non-photosynthetic stramenopile lineages, especially the way we would model events of plastid gain or loss during the diversification of stramenopiles (Burki et al., 2016; Keeling, 2009; Keeling and Burki, 2019). For instance, whether the ancestor had a plastid or not, dictates how many instances of plastid gain or loss would be required to explain the current diversity of stramenopiles, which in turn would mean non-photosynthetic stramenopiles would be interpreted as having lost a plastid versus never having had one. To assess this type of trait evolution, robust characterization of the plastid-bearing status of many lineages, as well as the construction of a well-supported phylogeny of the stramenopiles, are both necessary. However, no known plastid-associated or plastid-targeted genes, or cryptic plastids akin to the apicoplasts in the Apicomplexa, have been identified from the handful of known heterotrophic stramenopiles, including in the well-studied Oomycota and Opalinata, although the one exception being *SufCB* found in *Blastocystis,* which is hypothesized to have originated from horizontal gene transfer (Tsaousis et al., 2012).

Additionally, the scant amount of genomic-scale molecular data for the already small number of identified heterotrophs is associated with conflicting phylogenomic tree topologies. For example, in phylogenomic trees constructed in Thakur *et al*. (2019) and Derelle *et al*. (2016), both Gyrista *sensu stricto* and Bigyra *sensu stricto* form clades while Bigyra is paraphyletic in other publications (Burki et al., 2016; Leonard et al., 2018; Noguchi et al., 2016) (Fig. 1.1). Additional topological instability exists within Gyrista where the positioning of the Bigyromonadea alternates between ochrophytes and oomycetes depending on the publication and the number of taxa included. Increasing taxon sampling of under-studied stramenopiles will therefore contribute to resolving phylogenomic relationships and can be used to address trait evolution.

1.1.2 The phagoheterotroph clade in Gyrista, the Bigyromonadea, lacks taxon sampling

The phagotrophic heterotrophs, phagoheterotrophs (Mitra et al., 2016), which lack plastid or plastid-associated pathways are found in different lineages of stramenopiles such as Bigyromonadea, Sagenista, and Opalozoa. However, prior to the work in Chapter 2, only one species, a bacterivorous *Developayella elegans* belonging to the subphylum Bigyromonadea, had been characterized at a transcriptome level. Other species such as *Develorapax marinus*, *Pirsonia guinardiae*, and *Mediocremonas mediterraneus* have small subunit ribosomal RNA (SSU rRNA) gene sequences and cellular structures described without genomic level data (Aleoshin et al., 2016; Kühn et al., 2004; Weiler et al., 2021). The rest of the Bigyromonadea are only known through environmental screening using SSU gene markers (Massana et al., 2004). The phagoheterotroph mode of nutrition, lack of known plastid-associated genes, and SSU gene phylogenies were together used to suggest that *D. elegans* is closely related to oomycetes

(Moriya et al., 2002), but subsequently a SSU gene sequence from *D. marinus* suggested instead it is sister to ochrophytes (Aleoshin et al., 2016).

Similar to the unstable phylogenomic affinity to ochrophytes or oomycete of Bigyromonadea, the morphological traits of the flagellar apparatus also seem to vary. The presence of two roots for each of the two flagellar basal bodies in *D. elegans* resemble the ochrophytes while other heterotrophic stramenopiles have an additional ventral root, due to a root being split (Yubuki and Leander, 2013). When examining the flagellar transitional zone of *D. elegans*, a distal portion of a microtubular organizing centre (MTOC = basal body) denoting the boundary between a flagellum and a cell body, it has two connected helices (a double transitional helices) similar to those found in oomycetes and some phagoheterotrophs found in Bigyra (Tong, 1995; Aleoshin et al., 2016).

The semicircular microtubule organization of the ventral flagellar root of *D. marinus* is similar to that of *D. elegans*, in addition to the presence of the double transition helices in the two basal bodies. The SSU gene sequences of the two species are also similar (94% similarity for SSU; 91% for the large subunit ribosomal RNA gene [LSU rRNA]) and together they comprise the order Developea, close to the ochrophytes (Aleoshin et al., 2016). However, unlike *D. elegans*, *D. marinus* can engulf other protists or aggregates of bacteria. This eukaryovory led to a speculation that *D. marinus* might resemble the ochrophyte ancestor, which in turn suggested an independent secondary endosymbiosis of ochrophyte plastid, which is also consistent with the lack of plastid or plastid-associated proteins in all non-ochrophyte stramenopiles (Aleoshin et al., 2016). Recently, a smaller bacteriovorous bigyromonad, *Mediocremonas mediterraneus* (Weiler et al., 2021) has been successfully cultured and its SSU rRNA gene characterized. It branched with the Developea, most closely grouping with an “abyssal” clade of deep-sea environmental data from Sagami Bay sediments (Takishita et al., 2007), southeastern Atlantic plain (Scheckenbach et al., 2010), and Norwegian Sea coral reefs (Jensen et al., 2012). Collectively Developea were sister to both ochrophytes and oomycetes. Although the reliability of SSU gene tree is limited, the assignments of the characterized SSU gene sequences of Bigyromonadea to various environmental sequences point to more taxa yet to be uncovered.

The order Pirsoniales (Cavalier-Smith, 1998) forms a clade in a phylogenetic trees based on SSU gene, and is consistently placed closest to ochrophytes and somewhat less stably to the order Hyphochytriales (Aleoshin et al., 2016; Kühn et al., 2004; Weiler et al., 2021). The first described species within the order was *Pirsonia guinardiae*, a peculiar parasite of diatoms (Schnepf et al., 1990). Unlike other osmotrophic parasites that are phylogenetically placed close to it (e.g., Oomycota and Hyphochytriomycota), *Pirsonia* spp. deploy their pseudopodia to squeeze through the frustule girdles of diatoms, while the main parasitic cell body (auxosome) stays outside of the host. The invading pseudopodium then forms a trophosome which phagocytoses the host cytoplasm, and sometimes chloroplasts as well (Kühn et al., 2004; Schnepf et al., 1990). Whether *Pirsonia* spp. have genes associated with plastid-associated pathways or cryptic plastids has not been thoroughly investigated, although there have been no reports of successful visualization of engulfed plastids that might allude to the existence of a kleptoplast (Schnepf et al., 1990).

With only transcriptomic data from *D. elegans* available and its unstable position in the SSU phylogenetic trees, it was inconclusive whether the Bigyromonadea is monophyletic, comprising the Developea and Pirsoniales (Aleoshin et al., 2016; Weiler et al., 2021).

1.1.3 Ecological diversity and the complex evolutionary history of the phagoheterotrophs of Bigyra

The diversity and distribution of phagoheterotrophs across stramenopile phylogeny have been inferred from environmental screenings of heterotrophic nanoflagellates using ribosomal SSU markers (Kolodziej and Stoeck, 2007; Lin et al., 2012; Massana et al., 2004). These surveys mainly focused on marine environments, and initially discovered several clades consisting of environmental sequences only. Out of these 18 MArine STramenopiles (MAST) clades, only three have been described cellularly and/or phylogenetically at the transcriptome level so far. These include *Pseudophyllomitus vesiculosus* as MAST-6 (Shiratori et al., 2017), *Incisomonas marina* as MAST-3 (Cavalier-Smith and Scoble, 2013), and MAST-4 (Roy et al., 2014). However, recent SSU environmental surveys revealed that what were previously thought to be marine-exclusive MASTs (e.g., MAST-3 and MAST-6) were also present in ecologically diverse freshwater habitats (Simon et al., 2015) and sediments (Logares et al., 2012; Massana et al., 2015; Rodríguez-Martínez et al., 2020). Additional detection of MAST-2 (Simon et al., 2015) and MAST-12 (Kolodziej and Stoeck, 2007) in an estuary demonstrated the ecological diversity of phagoheterotrophs and the biased nature of environmental survey, which are skewed towards marine pelagic environments.

There is much more data from phagoheterophic Bigyra than bigyromonads, including several different strains of MAST-4, MAST-3, Bicosoecids, Placididea, *Cafeteria roenbergensis*, *Cantina marsupialis*, *Wobblia lunata*, and *Pseudophyllomitus vesiculosus* (MAST-6). More recently, genomic level data from MAST-1, -7, -8, -9, and MAST-11 were also described (Labarre et al., 2021). However, even with all these data, the phylogenetic placement of Bigrya remains contradicted, and the phylogenetic incongruence in Bigyra and Gyrista may be related.

Whether conflicts among published phylogenomic trees are due to different numbers or the composition of taxa or genes, quality of data, different statistical indices (e.g., Bayesian posterior probabilities vs. bootstrap percentage), phylogenetic substitution models, or phylogenetic reconstruction methods (e.g., distance vs. likelihood method), it is apparent that the stramenopiles have a complex molecular evolutionary history (e.g., genomic changes such as indels, retrotransposon integration, horizontal gene transfer (HGT), or gene fusion) and phylogenetic signal may have been masked by homoplasy and heterotachy (Delsuc et al., 2005; Lopez et al., 2002).

Platysulcus tardus (Shiratori et al., 2015; Thakur et al., 2019) is the sole member of the Platysulcea, which appears to form a sister clade to the rest of the stramenopiles. At a morphological level, *P. tardus* is an amalgamation of different traits of various phagoheterotrophic stramenopiles. For example, its gliding movement (rare in stramenopiles) is similar to Placididea (e.g., *W. lunata*), but *P. tardus* has a longer posterior flagellum which is similar to another gliding stramenopile belonging to Bikosia (*Caecitellus parvus*), which itself lacks mastigonemes on its anterior flagellum (Shiratori et al., 2015). Furthermore, *P. tardus* lacks any helical structures at the basal bodies, which are present in a double helical form in Placidozoa, oomycetes, and the Bigyromonadea, and in a single helix form in ochrophytes. Instead, *P. tardus* has L-shaped microtubule organization, a similar trait to Bikosia except *P. tardus* has an X-fibre where the S-tubule is absent from R2 of the anterior basal body (Shiratori et al., 2015; Yubuki and Leander, 2013).

1.2 The phylogenomic relationship and systematics of the photosynthetic stramenopiles

All the photosynthetic stramenopiles (excluding mixotrophic labyrinthulids harbouring green algal symbionts) are ochrophytes, but not all ochrophytes are photosynthetic. This is especially the case for many chrysophytes (golden algae) that are mixotrophs or have lost the ability to photosynthesize entirely (Beisser et al., 2017; Graupner et al., 2018). Additionally, ochrophytes have diverse morphologies, ranging from multicellular giant kelps, heliozoan ("sun-like") *Ciliophrys* spp., amoeboid *Chrysamoeba radians* (Hibberd, 1971), and silica-covered diatoms. Despite this morphological and genetic diversity, many studies involving ochrophytes are often focused on two groups, diatoms (Bacillariophyceae) and brown algae (Phaeophyceae), due to their ecological and economical importance. Genome data from most other ochrophyte groups is lacking or they are only represented by one or two species, despite being publicly available cultures. To date, only a single phylogenomic analysis includes comprehensive ochrophyte data (Terpis, 2021, data unpublished).

Phylogenetic analyses including a wide variety of ochrophytes are needed, however, because conflicting phylogenetic trees have been inferred from a handful of SSU rRNA or plastid genes, and initially in combination with morphological traits (Cavalier-Smith and Chao, 2006; Cavalier-Smith and Scoble, 2013). The classification is also in flux, with new classes or families of ochrophytes being erected or existing groups correctly re-classified because of phylogenetic findings. As a result, old names have been re-used or abandoned by some authors (Riisberg et al., 2009; Ševčíková et al., 2015; Derelle et al., 2016), while others have come up with a completely different system (Yang et al., 2012), or a combination of both (Derelle et al., 2016). For example, Yang *et al*., (2012) proposed a new classification composed of three clades (SI, SII, and SIII), in which the SI clade consists of Raphidophyceae-Phaeophyceae-Xanthophyceae (RPX) plus Schizocladiophyceae, "Chrysomerophyceae" which was later identified as being misspelled (Graf et al., 2020), Phaeothamniophyceae, and Aurearenophyceae: the SII clade consists of the Chrysophyceae-Synchromophyceae-Synurophyceae (CSS) plus Pinguiophyceae and

Eustigmatophyceae, and the SIII clade consists of Bolidophyceae and Bacillariophyceae (BB), and Pelagophyceae and Dictyochophyceae (PeD). On the other hand, Derelle *et al*. (2016) proposed the group, Diatomista consisting of Khakista (synonymous to the originally described BB (Cavalier-Smith and Chao, 2006)) and PeD, while reusing Chrysista (Cavalier-Smith and Scoble, 2013) to describe rest of the ochrophytes. Others redefined the original classification as exemplified by Khakista and Phaeista (initially included only PeD) whose redefined groupings include BB and PeD within Khakista (essentially synonymous to Diatomista), and the rest of ochrophytes within Phaeista, essentially synonymous to Chrysista (Riisberg et al., 2009).

Most of these groupings are controversial, as they are not always recovered from different phylogenetic analyses. One example of this is Limnista (Cavalier-Smith and Chao, 2006), which refers to the sister-lineage of Eustigmatophyceae and Chrysophyceae. This grouping is only recovered in phylogenetic trees inferred from plastid genes (Ševčíková et al., 2015; Di Franco et al., 2022) or a single-gene tree (Cavalier-Smith and Scoble, 2013). Pinguiophyceae is recognized to be part of Diatomista in a recent taxonomic revision by Adl et al. (2019) however many phylogenomic analyses recovered the class as being part of Chrysista (Azuma et al., 2022; Burki et al., 2016; Noguchi et al., 2016). The classification is also complicated by misnaming and misidentification [i.e., Chrysomeridophyceae was misnamed as Chrysomerophyceae in (Cavalier-Smith, 1995)], and by the lack of molecular and/or morphological data from type species (i.e., the genus *Chrysomeris*) (Graf et al., 2020).

Even with genome-scale data, the phylogenomic position of some lineages remain unresolved, such as Eustigmatophyceae (Hibberd, 1981), Pinguiophyceae (Kawachi et al., 2002), and *Actinophrys* spp. (Ehrenberg, 1830) (Derelle et al., 2016; Dorrell et al., 2021; Di Franco et al., 2022; Cho et al., 2022). Other lineages are simply not represented in the dataset [except in the recent study (Terpis, 2021)], or represented by a single species (Ševčíková et al., 2019; Amaral et al., 2020; Azuma et al., 2022).

1.2.1 Discrepancies between plastid and nuclear trees

The most obvious controversy in ochrophyte phylogenomics is the incongruent phylogenomic trees inferred from plastid genes versus nuclear genes. This is especially true for the placements of Eustigmatophyceae and Pinguiophyceae. Why they differ and which is correct is unclear, but they appear to vary in phylogenetic signal, perhaps due to different selection pressures.

In phylogenomic analyses inferred from plastid genes, Eustigmatophyceae forms a sister-lineage to CSS (Ševčíková et al., 2015; Di Franco et al., 2022). In contrast, when the phylogenomic tree is inferred from nuclear genes, Eustigmatophyceae forms a sister-lineage to RPX, with weak statistical support (Burki et al., 2016; Noguchi et al., 2016; Cho et al., 2022).

Similar to Eustigmatophyceae, the placement of Pinguiophyceae can differ based on the genes used to infer a tree. It is placed sister to CSS in phylogenomic trees inferred from plastid trees but placed sister to Diatomista in some trees inferred from nuclear genes (Yang et al., 2012; Di Franco et al., 2022) or RPX in the trees inferred from SSU rRNA genes (Kawachi et al., 2002). Despite many statistical tests, such as the approximately unbiased test (Shimodaira, 2002), or comparing the level of phylogenetic signals between nuclear or plastid genes (Di Franco et al., 2022), having only one or two taxa representing each of the groups still leave the placement of these lineages contentious.

1.2.2 Under representative classes of ochrophytes

There are a total of 17 currently recognized classes of ochrophytes described so far (Cavalier-Smith and Chao, 2006; Riisberg et al., 2009; Yang et al., 2012; Graf et al., 2020), with detailed morphological characterization and phylogeny inferred from rRNA and/or plastid genes. In previous phylogenomic analyses, at most 10 classes of ochrophytes were represented, with much of the data collection skewed towards Phaeophyceae (brown algae), Chrysophyceae (golden algae), and Bacillariophyceae (diatoms) (Burki et al., 2016; Derelle et al., 2016; Leonard et al., 2018; Noguchi et al., 2016; Thakur et al., 2019; Cho et al., 2022, 2024). Prior to Chapter 4, four classes were under-represented in phylogenomic analyses: Schizocladiophyceae, Phaeothamniophyceae, Picophagea, and Olisthodiscophyceae. Schizocladiophyceae and Phaeothamniophyceae are positioned within the Phaeophyceae-Xanthophyceae lineage in a phylogenetic tree inferred from a combination of five plastid genes and SSU rRNA genes (Graf et al., 2020), while the position of Picophagea and Olisthodiscophyceae were only explored in single phylogenomic analysis (Terpis, 2021, unpublished data).

1.3 Phylogenomic inference methods and computational burdens

Throughout the thesis, I use phylogenomic analyses to investigate and attempt to resolve phylogenomic incongruencies among reported studies. My analyses involves first finding orthologous genes (up to ~260) in genomic or transcriptomic data of species of interest. Each of these genes is aligned and then concatenated (combined) to construct a supermatrix (Fig. 1.2). The supermatrix is then used to infer a phylogenomic tree using likelihood methods (maximum likelihood and Bayesian) that incorporate models for different character (amino acids or DNA sequences) evolution. Using multiple genes allows for the “genomic” representation of the organisms when inferring a phylogeny, but often limits the number of species that can be used because it requires much more completeness of the data. My approach has been to fill in such gaps to see if newly sampled lineages can be placed, and whether including the new taxa also affects the topology of the tree more generally. Using a single gene such as SSU rRNA, to infer a phylogeny was prevalent before the phylogenomic era, and although these trees lack support, they remain useful for certain purposes, especially for screening environmental data, and this approach is also taken here.

Although using hundreds of genes is a more reliable way to infer a phylogeny, it comes with one of the major hurdles in the method, the computational time and resources. This is especially true for the likelihood methods used in this thesis. Likelihood inference methods utilize complex mathematical models that incorporate various parameters accounting for amino acid site heterogeneity for tree hypothesis testing (Lopez et al., 2002; Delsuc et al., 2005; Lartillot et al., 2007; Quang et al., 2008; Wang et al., 2018). Moreover, testing statistical confidence is perhaps the most computationally demanding process as multiple iterations of tree sampling are required (e.g., non-parametric bootstrap in ML inference). Bayesian inference is another model-based method that incorporates maximum likelihood and was used in conjunction with ML inference to validate the phylogenomic relationship in this thesis. As this inference method relies on a Markov chain Monte Carlo (MCMC) algorithm while measuring the statistical confidence using posterior probabilities, the computational burden can be reduced compared to the ML inference (Holder and Lewis, 2003; Delsuc et al., 2005; Lartillot et al., 2009). However, the MCMC algorithm requires tens of thousands of iterations as opposed to hundreds of bootstraps in ML, and it is difficult to estimate whether the MCMC approximation has reached local maxima or has run long enough. Additionally, Bayesian inferences are sensitive to the misspecification of an prior probabilities, and as a result, multiple chains (four chains in this thesis) are often used to conduct MCMC approximation. In the end, inferring a phylogenomic tree can be a months- to a year-long process depending on the computing resources available. To remediate the computational burden, a "divide-and-conquer" (Delsuc et al., 2005) strategy has been proposed where a dataset is sub-divided by a group and then an optimal tree is generated for each (Strimmer and Von Haeseler, 1996). Then these trees are then combined to generate a supertree.

Another way to reduce computational burden can be to minimize the number of genes used, by selectively using phylogenetically informative genes for the phylogenomic analyses (Salichos and Rokas, 2013; Edwards, 2016; Shen et al., 2016b; Mongiardino Koch, 2021; Di Franco et al., 2022). Although this approach has been tested on more recently diverged eukaryotes (metazoans) or prokaryotes, it is yet to be tested on a specific group of protists.

1.4 Endosymbionts of non-photosynthetic stramenopiles

Many stramenopiles are known to be symbionts of other eukaryotes, either as parasites, or kleptoplasts. The best-known endosymbiotic stramenopiles include the pennate diatom, *Nitzschia frustulum symbiotica* residing in four families of large benthic formaniferans (Lee, 2006), while others only have the plastids sequestered by a host, such as dinoflagellates in the case of dinotoms (e.g., *Durinskia capensis* and *D. kwazulunatalensis* (Yamada et al., 2019)). Of the parasitic stramenopiles, oomycetes are known for their ability to infect a broad range of hosts, including animals, plants, and other protists (Vallet et al., 2019), whereas other parasites are scattered across the stramenopile phylogeny, including the Labyrinthulomycetes (*Labyrinthula zosterae*) (Muehlstein et al., 1991), Opalinata (*Blastocyst hominis*) (Basak et al., 2014), and Bigyromonadea (*Pirsonia* sp.).

Stramenopiles as hosts to endosymbionts, on the other hand, have not been well investigated, and this is true for prokaryotic symbionts of heterotrophic stramenopiles. Based on a recent curation of the published ecological interactions of protists (Bjorbækmo et al., 2020), the supergroup alveolates and rhizaria are the most common hosts of non-parasitic symbionts, which include dinoflagellates, diatoms, and trebouxiophyceans, and together make up 81% of the symbiont-host interaction noted in this analysis. For parasite-host interactions, alveolates again comprise two thirds of both parasite and host examples, while other common hosts were diatoms. Data from the rest of the stramenopile lineages as hosts of parasites or non-parasites is scarce, and the functional role between the pair has not been characterized (denoted as "Unresolved interaction" (Bjorbækmo et al., 2020)). The majority of these "Unresolved interactions" are dominated by protist-bacteria interactions (73%) with rest of the 23% being protist-protist interactions, possibly of under-sampled or unknown protists. The heterotrophs or mixotrophs of stramenopiles, particularly Labyrinthulomycetes and MASTs, were under-represented when compared to SSU-environmental survey (de Vargas et al., 2015), comprising ~0.5% of the interaction entries (Bjorbækmo et al., 2020). This trend can also be observed in a recent review by Husnik et al. (2020) which reports only one case of a prokaryotic endosymbiont among non-photosynthetic stramenopiles (Husnik et al., 2021). This non-photosynthetic stramenopile with reported endobacteria is *Symbiomonas scintillans*, a tiny phagoheterotroph (Guillou et al., 1999).

What is interesting about these two reviews (Bjorbækmo et al., 2020; Husnik et al., 2021) is the number or the proportion of symbiont studies targeting stramenopiles, despite the group being relatively well characterized, if not better described than other eukaryotic supergroups. For example, only 17 cases of stramenopiles are reported to have prokaryotic symbionts (15 of which are from photosynthetic stramenopiles) while 28 cases of symbiont studies are done on rhizaria, a vastly under-studied supergroup.

Although non-phototrophs play vital roles in linking carbon cycling between lower and higher trophic mode of food webs (Sherr and Sherr, 2002), it is surprising how little investigation has been done on symbionts of non-photosynthetic stramenopiles, especially for the nano- or pico-flagellates. Characterizing symbionts of these under-represented stramenopiles may provide some clues into how the complex evolutionary history of stramenopiles took shape. For instance, bacterial endosymbionts may have contributed to the hosts occupying and surviving diverse ecological niches, ranging from shallow freshwater, mosses, or anoxic sediments, with or without being dependent on metabolic network of eukaryotic endosymbionts. It is still not known whether or how symbiotic interactions resulted in diverse morphologies, feeding and motility strategies, further leading to divergence of phylogenomic clades. Uncovering the evidence of the symbionts or a community thereof, may help us understand the divergence of major clades throughout the stramenopile phylogeny.

1.5 Thesis goals and objectives

Compared to other protist supergroups, considerable amounts of genomic and transcriptomic data have been generated for stramenopiles (Sibbald and Archibald, 2017). Despite the amount of the data, the phylogenomic relationships among stramenopiles are largely unresolved, partly due to the complex biodiversity of ochrophytes, and under-sampling the remainder of the group, such that much of their diversity and ecological roles of which are yet to be characterized. To address these problems, the major objectives of my dissertation are:

1. To resolve phylogenomics of stramenopiles by increasing taxon sampling of small non-photosynthetic flagellates.
2. To characterize distribution and abundance of phagoheterotrophic stramenopiles that are associated with sediments and V9-region targeted (SSU) amplicon datasets.
3. To resolve phylogenomics of ochrophytes by sampling under-studied lineages and searching for phylogenetically informative genes.

4. To identify the function and identity of the only reported case of prokaryotic endosymbionts among a heterotrophic stramenopile, *Symbiomonas scintillans*.

In Chapter 2, and 3, I addressed the first objective by generating transcriptomes of 13 heterotrophic stramenopiles, 11 of which are newly described species, belonging to Bigyromonadea, MAST-6, and Placididea. These groups were previously represented by one or two species in a phylogenomic analysis. After inferring phylogenomic trees, I also conducted extensive statistical tests to verify the placement of these taxa. As a result, the inferred phylogenomic tree recovered a robust monophyly of Bigyromonadea and Oomycota and a paraphyletic relationship of Bigyra.

In Chapter 3, I addressed the second objective by searching for SSU rRNA sequences of the newly described sediment-associated MAST-6 and halophilic Placididea species in publicly available environmental amplicon datasets. I focused on sediment amplicon data sets and a dataset that used a V9 SSU rDNA primer pair (preferential amplification of Placididea). As a result, I found that many MAST-6 species related to one of the newly described species are abundant in sediments, while the newly described Placididea species may tolerate a broad range of salinity.

In Chapter 4, I aimed to resolve the phylogenomic relationships of stramenopiles by updating under-represented classes of ochrophytes and searching for phylogenetically informative genes. As a result, I generated 10 new ochrophyte transcriptomes, mostly from public culture collections, and one obtained from a single-cell isolation for an environmental sample. I also added publicly available genomic level data and recovered robust support for some of the previously controversial relationships. To address unresolved relationships of other lineages and explore sub-sampling in a phylogenomic supermatrix, I searched for phylogenetically informative genes and inferred 16 phylogenomic trees consisting of different genes with varying phylogenetic signals. I found no genes that only had phylogenetic signal without noise. However, I did find that inferring phylogenomic trees based on genes with high phylogenetic signal and quality yielded fewer variable topologies, than removing genes with high phylogenetic noise.

Finally, Chapter 5 addresses the fourth objective. This chapter highlights how little we know about symbionts of non-photosynthetic stramenopiles, such as *Symbiomonas scintillans,* whose endosymbionts may be giant viruses and not bacteria, although further experiments are needed. The work done in Chapter 2 and Chapter 3 focused on enriching taxon sampling of these under-studied heterotrophic stramenopiles and provided both morphological descriptions and transcriptome data to provide an updated phylogenomics of stramenopiles. These chapters further provide a platform for better understanding trait evolution such as the evolution of saprotrophy in oomycetes, adaptability in specific niche occupation, horizontal gene transfers, and revisiting plastid evolution.

Overall, my thesis has greatly updated phylogenomic data from stramenopiles by generating 23 transcriptomes, in addition to data curation from publicly available database. The phylogenomic analyses permit a better understanding of the evolutionary history of stramenopiles, while highlighting their diversities in terms of feeding, morphology, and relationship with other organisms.

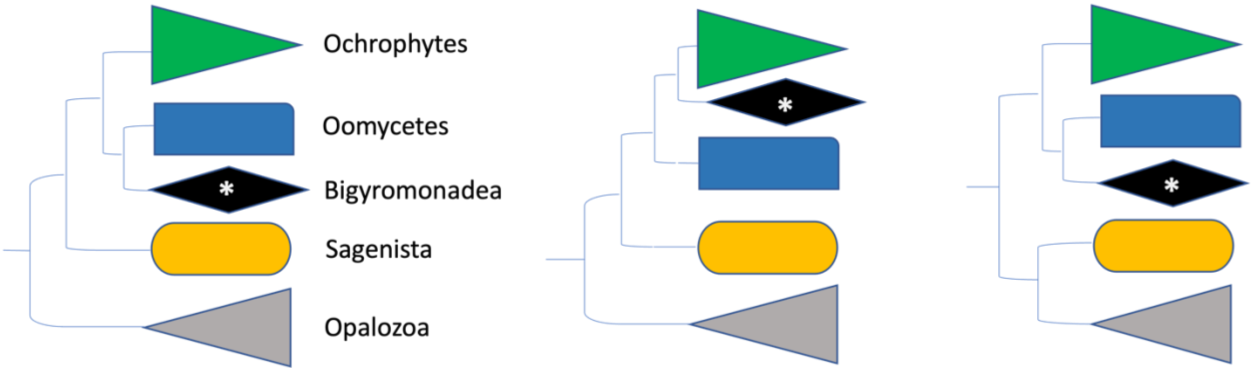


Figure 1.1 Different versions of stramenopile phylogenomic trees recovered in different publications.

Ochrophytes, Oomycetes, and Bigyromonadea have been proposed to form a major group called Gyrista. Sagenista and Opalozoa have been proposed to form a second major group, called Bigyra. Three out of five phylogenomic analyses (Noguchi et al., 2016; Leonard et al., 2018; Thakur et al., 2019) included Bigyromonadea in their analyses, with the clade represented by a single taxon (*). The species that branches sister to the rest of the stramenopiles, *Platysulcus tardus* is omitted.

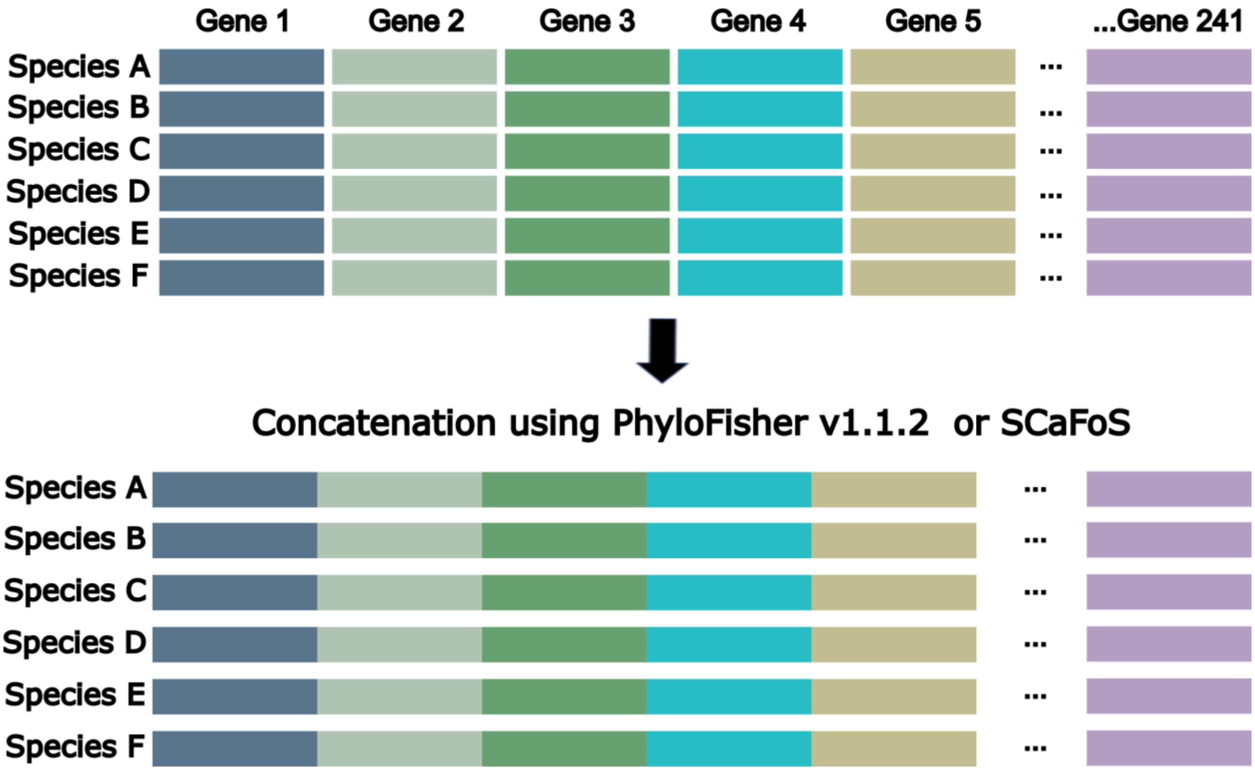


Figure 1.2 Flowchart showing steps to constructing a supermatrix

Concatenation or combination of multiple genes into a single alignment for each species can be done using several programs including PhyloFisher (Tice et al., 2021) or SCaFos (Roure et al., 2007).

Chapter 2: Monophyly of diverse Bigyromonadea and their impact on phylogenomic relationships within stramenopiles

2.1 Introduction

Stramenopiles (= heterokonts) are one of the well-characterized members of the eukaryotic supergroup SAR (**S**tramenopila, **A**lveolate, **R**hizaria) (for a review see Keeling and Burki, 2019). Stramenopiles are very diverse, comprising photoautotrophs (i.e., heterokont algae in ochrophytes), osmotrophic oomycetes and labyrinthulomycetes with a motile zoospore life-stage (e.g., *Phytophthora* sp*.*, *Pythium sp.*, and labyrinthulids) and free-living phagotrophic opalozoans (e.g., *Cafeteria roenbergensis*, *Cantina marsupialis*) that occupy a broad range of environments (Cavalier-Smith and Chao, 2006; Kolodziej and Stoeck, 2007; Stiller et al., 2009; Tsui et al., 2009; Cavalier-Smith and Scoble, 2013). Stramenopiles can be largely classified into two major groups: Gyrista consisting of Ochrophyta, Oomycota, and Bigyromonadea; and Bigyra consisting of Sagenista and Opalozoa. A single species, *Platysulcus tardus*, has also recently been shown to be a basal stramenopile (Thakur et al., 2019). While there is a lot of genomic data from stramenopiles, only a handful comes from phagoheterotrophs (Mitra et al., 2016), despite them representing much of the diversity as well as being key outstanding problems in resolving controversies in stramenopiles phylogeny (Burki et al., 2016; Derelle et al., 2016; Leonard et al., 2018; Shiratori et al., 2017, 2015).

One such clade is the subphylum Bigyromonadea (Cavalier-Smith, 1998), which was proposed to include the class Developea (Aleoshin et al., 2016) and order Pirsoniales (Cavalier-Smith, 1998). The monophyly of the Bigyromonadea is essentially untested, since only small subunit rRNA (SSU) data are known from all but a single species (the exception being "*Developayella elegans*", for which a transcriptome is available), and the two groups never branch together in SSU phylogenies (Aleoshin et al., 2016; Cavalier-Smith and Chao, 2006; Kühn et al., 2004; Weiler et al., 2021).

Developea are marine bacterivores, including *Developayella elegans* (Tong, 1995; Leipe et al., 1996) and *Mediocremonas mediterraneus* (Weiler et al., 2021), and the marine eukaryovore *Develorapax marinus* (Aleoshin et al., 2016). Pirsoniales are parasites of other microbes, including *Pirsonia guinardiae* (Schnepf et al., 1990) and *P. punctigera* (Schweikert and Schnepf, 1997). These parasites deploy a pseudopod to squeeze through the frustule girdles of their diatom host, while the main cell body (auxosome) stays outside of the host. The invading pseudopod then phagocytoses the host cytoplasm or chloroplasts forming a trophosome (food vacuole), which is then transported out to the auxosome (Kühn et al., 2004; Schnepf et al., 1990). The relationship of both groups to other stramenopiles is uncertain, and both have led to hypotheses about the evolution of other related groups. For example, the eukaryovory of *D. marinus* and its placement in rRNA trees has led to the hypothesis that it represents a model for a phagoheterotrophic ochrophyte ancestor (Aleoshin et al., 2016), however its position in the tree varies between grouping with ochrophytes (Leonard et al., 2018) or oomycetes (Noguchi et al., 2016; Thakur et al., 2019). Pirsoniales have also been found as the sister group of ochrophytes based on SSU rRNA trees (Aleoshin et al., 2016; Kühn et al., 2004), although once again not consistently and without strong support.

To test for the monopoly of bigyromonads and more thoroughly examine their relationship to other stramenopiles, together with the collaborators, I substantially increased the diversity of genomic data from the group by adding transcriptomes from seven newly discovered species belonging to Pirsoniales (*Pirsonia chemainus* nom. prov., *Koktebelia satura* nom. prov., and *Feodosia pseudopoda* nom. prov.) and Developea (*Develocanicus komovi* n. gen. n. sp., *Develocanicus vyazemskyi* n. sp., *Develocauda condao* n. gen. n. sp., and *Cubaremonas variflagellatum* n. gen. n. sp.). The inferred 247-gene phylogenomic tree, reconstructed with various methods, recovered for the first time the monophyly of the Bigyromonadea. Maximum likelihood (ML) recovered a robustly supported monophyly of Bigyromonadea and oomycetes, while Bayesian inference and statistical tests of alternative tree hypothesis were inconclusive. I describe several new features of the seven bigyromonads, and noted their resemblance with oomycete zoospores, and report the first observation of eukaryovory in the flagellated stages of Pirsoniales. Overall, these findings indicate bigyromonada and ooymcetes are most likely sister groups, and suggest potential ancestral state of the oomycetes resembling bigyromonada, including their ability to form auto-aggregates (=self-aggregates) (Hickman, 1970; Ko and Chase, 1973; Galiana et al., 2008) and phagoheterotrophy.

2.2 Materials and Methods

2.2.1 Sample collection, identification, and library preparation

Strain Colp-23 (*Develocanicus komovi*) was obtained from the black volcanic sand on the littoral zone of Maria Jimenez Beach (Playa Maria Jiménez), Puerto de la Cruz, Tenerife, Spain, October 20, 2014. Strains Colp-30 (*Develocanicus vyazemskyi*) and Chromo-1 (*Koktebelia satura*) were isolated from the near shore sediments on the littoral zone near T.I. Vyazemsky Karadag Scientific Station, Crimea, May 2015. Strain Chromo-2 (*Feodosia pseudopoda*) was obtained from the near shore sand on the littoral zone of the beach in the settlement Beregovoye, Feodosiya, Crimea, June 24, 2017. Strain Colp-29c (*Develocauda condao*) was isolated from the near shore sediments on the north-east part of Con Dao Island, South Vietnam, May 4, 2015. Strains '*Pirsonia*-like' (*Pirsonia chemainus*) and Dev-1 (*Cubaremonas variflagellatum*) were obtained from seawater samples taken in the Strait of Georgia, British Columbia, Canada (123°

28'50" W, 49°10'366" N) at 70 m and 220 m depths, respectively using a Niskin bottle, June 13, 2017.

The samples were examined on the third, sixth and ninth day of incubation in accordance with methods described previously (Tikhonenkov et al., 2008). *Procryptobia sorokini* strain B-69 (IBIW RAS), feeding on *Pseudomonas fluorescens*, was cultivated in Schmaltz-Pratt's medium at a final salinity of 20‰, and used as a prey for clones Colp-23, Colp-29c, Colp-30, Chromo-1, Chromo-2, and '*Pirsonia*-like' (Tikhonenkov et al., 2014). Bacterivorous strain Dev-1 was propagated on the *Pseudomonas fluorescens*, which was grown in Schmaltz-Pratt's medium. Strains Colp-23, Colp-29c, and Dev-1 are currently being stored in a collection of live protozoan cultures at the Institute for Biology of Inland Waters, Russian Academy of Sciences. However, strains Colp-30, Chromo-1, Chromo-2, and '*Pirsonia*-like' perished after several months to one year of cultivation.

Studied isolates were identified using a combination of microscopic and molecular approaches. Light microscopy observations were made using a Zeiss AxioScope A.1 equipped with a DIC water immersion objective (63x) and an AVT HORN MC-1009/S analog video camera. The SSU rRNA genes (GenBank accession numbers: OL630092 to OL630098) were amplified by polymerase chain reaction (PCR) using the general eukaryotic primers EukA-EukB (for strains Colp-23, Colp-30, '*Pirsonia*-like'), PF1-FAD4 (Chromo-1), 18SFU-18SRU (Chromo-2, Dev-1), 25F-1801R (Colp-29c) (Medlin et al., 1988; Keeling, 2002; Cavalier-Smith et al., 2009; Tikhonenkov et al., 2016). PCR products were subsequently cloned (Colp-23, Colp-30, Chromo-2, '*Pirsonia*-like') or sequenced directly (Chromo-1, Dev-1, Colp-29c) using Sanger dideoxy sequencing.

For cDNA preparation, cells grown in clonal laboratory cultures were harvested when the cells had reached peak abundance (strains Colp-23, Col-30, Colp29c, Chromo-1, Dev-1) and after the majority of the prey had been eaten (for eukaryovorous strains Colp-23, Col-30, Colp29c, Chromo-1). Cells were collected by centrifugation (1000 x *g*, room temperature) onto the 0.8 µm membrane of a Vivaclear mini column (Sartorium Stedim Biotech Gmng, Cat. No. VK01P042). Total RNA was then extracted using a RNAqueous-Micro Kit (Invitrogen, Cat. No. AM1931) and reverse transcribed into cDNA using the Smart-Seq2 protocol (Picelli et al., 2014), which uses poly-A selection to enrich mRNA. Additionally, cDNA of Colp-29c was obtained from 20 single cells using the Smart-Seq2 protocol (cells were manually picked from the culture using a glass micropipette and transferred to a 0.2 mL thin-walled PCR tube containing 2 µL of cell lysis buffer – 0.2% Triton X-100 and RNase inhibitor (Invitrogen)). The same 'single cell' transcriptomic approach was applied for strains Chromo-2 and '*Pirsonia*-like', which never consumed the prey completely. Sequencing libraries were prepared using NexteraXT protocol and sequencing was performed on an Illumina MiSeq using 300 bp paired-end reads. Additionally, Chromo-1 transcriptome sequencing was performed on the Illumina HiSeq platform (UCLA Clinical Microarray Core) with read lengths of 100 bp using the KAPA stranded RNA-seq kit (Roche) to construct paired-end libraries. Raw reads are available in the NCBI Short Read Archive (SRA) (BioProject number: PRJNA782193, SRR17035338 to SRR17035344).

2.2.2 Small-subunit phylogenetic tree reconstruction

SSU rRNA sequences were identified from the seven new assembled transcriptomes using Barrnap v0.9 (Seemann, 2007) and compared with the SSU sequences obtained with Sanger sequencing, and the longer sequences were used for further analysis.

After an initial BLASTn search of the SSU rRNA sequences against the non-redundant NCBI database to confirm stramenopile identity, the SSU sequences were aligned using MAFFT v7.222 (Katoh and Standley, 2013) with previously compiled SSU datasets (Aleoshin et al., 2016; Yubuki et al., 2015). Additionally, SSU sequences of the other stramenopile taxa that were included in the multi-gene phylogenomic dataset and other closely related taxa were included (see Results). Furthermore, to show the diversity of uncultured Gyrista and provide possible directions for future sampling efforts, environmental sequences of stramenopiles that are closely related to Pirsoniales and Developea were added. The environmental sequences were manually retrieved from NCBI database and PR2 based on previously published alignments (Massana et al., 2004; Weiler et al., 2021; Yubuki et al., 2015). After trimming using trimAl v.1.2rev59 (-gt 0.3, -st 0.001) (Capella-Gutiérrez et al., 2009), the two SSU phylogenetic trees were reconstructed based on 1650 sites and 92 taxa, and 1665 sites and 107 taxa, using IQ-TREE v1.6.12 (Nguyen et al., 2015) 1000 ultrafast bootstrap (UFB) under Bayesian information criterion (BIC): using the TIM2+R6 model selected by ModelFinder (Kalyaanamoorthy et al., 2017) implemented in IQ-TREE.

2.2.3 Transcriptome processing, assembly, and decontamination

Raw sequencing reads were assessed for quality using FastQC v0.11.5 (Andrews, 2010) and remnant transposase-inserts from the library preparation were removed. The reads were assembled using Trinity-v2.4.0 with *–trimmomatic* option to remove NexteraXT adaptors, Smart-Seq2 IS-primer, and low quality leading and trailing ends (quality threshold cut-off:5) (Grabherr et al., 2011; Bolger et al., 2014). To identify contaminants, assembled reads were searched against the NCBI nucleotide database using megaBLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990), followed by diamond BLASTX against a UniProt reference proteome (The UniProt Consortium et al., 2021). To visualize the contig sizes, coverage, and remove bacterial, archaeal, and metazoan contaminants, BlobTools v1.0 (Laetsch and Blaxter, 2017) was used. PhyloFlash v3.3b2 (Gruber-Vodicka et al., 2020) was used in parallel to confirm identified contaminants and coverage based on SILVA v138 SSU database (Quast et al., 2012). To remove sequences from the prey, *Procryptobia sorokini*, which was used in the cultures of *Pirsonia chemainus*, *Koktebelia satura*, *Feodosia pseudopoda*, *Develocanicus komovi*, *D. vyazemskyi*, and *Develocauda condao*, the assembled reads were searched against the *P. sorokini* transcriptome using BLASTn in which the contigs with ≥95% sequence identity were removed from the assembled reads. To predict open reading frames (ORFs) and coding genes, TransDecoder v5.5.0 (Haas et al., 2013) was used and the longest ORFs were annotated using BLASTP search against UniProt database. To estimate the completeness of each of the assembly, BUSCO v4.0.5 (Simão et al., 2015) with eukaryotic database was used.

2.2.4 Phylogenomic matrix construction and ortholog identification

To better represent each stramenopile (sub)group in the phylogenomic reconstruction, recently published and publicly available (Broad Institute and Japan Agency for Marine-Earth Science and Technology; JAMSTEC) additional 27 stramenopile genomic or transcriptome data (de Vargas et al., 2015; Hackl et al., 2020; Keeling et al., 2014; Leonard et al., 2018; Noguchi et al., 2016; Seeleuthner et al., 2018; Thakur et al., 2019; Wawrzyniak et al., 2015) were obtained and analyzed along with the seven new transcriptomes (Appendix A). The updated stramenopile dataset including all the newly added transcriptomes in this study were compiled to the existing gene-set described below. Using BLASTP, the predicted coding genes from each transcriptome were searched against 263 gene-sets (orthologs), each consisting of compiled genes from major supergroups of protists, fungi, and holozoans (Burki et al., 2016; Hehenberger et al., 2017). The blast outputs contained up to four non-redundant (nr) sequences for each gene and were filtered with an e-value threshold of 1e-20 with >50% query coverage. To ensure there is no extension for each of the newly identified genes that might hinder downstream 263 gene-set analysis, the new gene-sets were used as a query for BLASTP search against the UniProt database, followed by removing poorly aligned regions. Each gene-set was aligned using MAFFT-L-INS-i v.7222 and trimmed using trimAL v1.2rev59 (-gt 0.8). To infer orthologs among nr sequences from the newly added transcriptomes aligned to the corresponding 263 gene-sets, 263 gene-trees were built using maximum-likelihood (ML) estimation with IQ-TREE v1.6.12 under the LG+I+G4 model and 1000 ultrafast bootstrap (UFB). Then, each gene tree was manually screened in FigTree v1.4.4 for paralogs and contaminants (e.g., long branching sequences or sequences nested within other distantly related taxa), which were subsequently confirmed using BLASTP search against the nr database. These paralogs and contaminants were removed from each gene-set alignment. To increase ortholog coverage from the added transcriptomes, fragmented orthologs were manually merged. To minimize the creation of artifacts, we followed several criteria for merging ortholog fragments. Up to two fragments were merged and considered fragments of the same ortholog; (1) if the fragments came from the same transcriptome; (2) if they were positioned within the same node in a given gene tree; (3) if they covered different regions of a gene with or without an overlapping region; (4) and if there was an overlapping region present among fragments aligned to a given gene, up to two mismatches were permitted. Out of the 263 gene-sets, 110 gene-sets include manually merged orthologs with up to two taxa per gene-set. The 263 gene-sets containing the selected orthologs of the newly added transcriptomes and 27 newly published stramenopile data were aligned using two approaches and compared by reconstructing two phylogenomic trees. In the first approach (**approach 1**), the sequences were aligned by using MAFFT L-INS-i v.7.222 and trimmed via trimAL v1.2rev59 (-gt 0.8). In the second approach (**approach 2**), the sequences were filtered using PREQUAL (Whelan et al., 2018) to remove non-homologous regions generated due to poor transcriptome quality or assembly errors. The filtered sequences were then aligned using MAFFT G-INS-i (--allowshift and --unalignlevel 0.6 option) and processed for further filtering using Divvier (-mincol 4 and -divvygap option) (Ali et al., 2019) to identify statistically robust pairwise homology characters. The filtered gene-sets were then soft-trimmed using trimAL (-gt 0.1). The two dataset generated by two different filtering and alignment methods were separately processed using SCaFoS v1.2.5 (Roure et al., 2007), by removing gene-sets that have ≥40% missing amino acid positions in the alignment. The resulting 247 gene-set was concatenated into a phylogenomic matrix comprising 75,798 amino acid (aa) sites from 76 taxa for approach 1. For the PREQUAL/Divvier processed data (approach 2), the same 247 gene-sets were concatenated in a phylogenomic matrix comprising 101,314 aa sites from the same 76 taxa.

2.2.5 Phylogenomic tree reconstruction, fast-evolving site removal, and topology test

The ML tree for the concatenated phylogenomic matrix was inferred using IQ-TREE v.1.6.12 under the empirical profile mixture model, LG+C60+F+G4 (Quang et al., 2008). The best tree under this model was used as a guide tree to estimate the "posterior mean site frequencies" (PMSF). The PMSF method allows the conduction of non-parametric bootstrap analyses under complex models on large data matrices and has been shown to mitigate long-branch attraction artifacts (Wang et al., 2018). This LG+C60+F+G-PMSF model was then used to re-estimate the ML tree and for a non-parametric bootstrap analysis with 100 replicates. For Bayesian inference, CAT-GTR mixture model with four gamma rate categories was used with PhyloBayes-MPI v.20180420 (Lartillot and Philippe, 2004; Lartillot et al., 2009), only for the dataset processed with approach 1. To estimate posterior probabilities, four independent Markov Chain Monte Carlo (MCMC) chains were run simultaneously for minimum 10,000 cycles. After discarding the first 2000 burn-in points, consensus posterior probabilities for each branch were computed by subsampling every second tree. Convergence of the four chains were tested by calculating differences in bipartition frequencies (bpcomp) with a threshold maxdiff however, no chains converged (maxdiff=1).

Site-specific substitution rates were inferred using the -wsr option as implemented in IQ-TREE, under the LG+C60+F+G4 substitution model. Increments of the top 5% fastest evolving sites were removed from the phylogenomic matrix until exhaustion, defined as the point when the bootstrap support value significantly began to drop and the topology became unstable (50%; 37,899 sites). Each incremental phylogenomic matrix was analyzed using IQ-TREE for ML estimation using LG+C60+F+G4 and 1000 UFB. All fast-evolving species removal and sites tests were conducted on the dataset processed with approach 1.

Approximately unbiased (AU) tests (Shimodaira, 2002; Nguyen et al., 2015) were performed on set of phylogenomic trees constructed based on the 247 gene-sets generated by the first approach (i.e., MAFFT L-INS-i and trimAL with -gt 0.8) and the second approach (i.e., PREQUAL/Divvier), separately. The set of trees includes the two ML trees generated under LG+C60+F+G4(+PMSF) with 1000UFB (100STB), four consensus trees of MCMC chains, and other hypothetical constrained trees as listed as “Chain modified” in Table 2.1.

2.3 Results

2.3.1 Multi-gene phylogenomic analysis

The concatenated phylogenomic matrix was composed of 68 stramenopiles and eight alveolates (outgroup) with 247 aligned genes totaling 75,798 positions for approach 1, and

101,314 positions for approach 2. The average missing sites and genes were 22% and 19%, respectively (Fig. 2.1). The amount of missing data varied among the seven new transcriptomes. Chromo-1 had nearly complete data (5% missing sites and 6% missing genes) while Colp-29c had 21% missing sites and 12% missing genes. Colp-23 and Chromo-2 had the highest amount of missing data (75% missing sites and 57% genes for Chromo-2 and, 83% and 76% for Colp-23). The ML phylogenomic tree generated under LG+C60+F+G4+PMSF with STB estimation from the two approaches is shown in Figure 2.1, with the tree topology representing the dataset generated from approach 1 (i.e., MAFFT L-INS-i and trimAL with -gt 0.8). The tree topology representing the dataset generated from approach 2 (i.e., Prequal/Divvier) is shown in Appendix A. The tree topology is almost identical between the two, except the position of sub-clades in ochrophytes; for example, the positions of Chrysophyceae + Synurophyceae and Raphidophyceae + Phaeophyceae + Xanthophyceae + Eustigmatophaceae are swapped in the two trees (Fig. 2.1 and Appendix B).

The newly added transcriptomes of the seven new species formed the robust monophyletic bigyromonada with either dataset (approach 1 and approach 2; Figure 2.1 and Appendix A): *Develocanicus komovi*, *D. vyazemskyi*, *Develocauda condao*, and *Cubaremonas variflagellatum* forming a Developea clade (100% STB), while Pirsoniales is composed of *Pirsonia chemainus*, *Koktebelia satura*, and *Feodosia pseudopoda* (100% STB). The ML tree also recovered monophyly of the bigyromonada and oomycetes with 100% STB support (Fig. 2.1). The monophyly of Gyrista was strongly supported, with Sagenista (Labyrinthulomycetes and Eogyrea) forming a sister clade to it, resulting in a paraphyletic Bigyra. Platysulcea formed a sister clade to rest of the stramenopiles with a moderate support (91%/95% STB) (Fig. 2.1; Appendix B).

Bayesian analyses recovered a conflicting topology for the bigyromonada, which formed a sister-clade to ochrophytes in all four consensus trees generated (Appendix C). Additionally, the topology within ochrophytes was conflicting, contributing to the lack of convergence. However, the monophyly of bigyromonada + ochrophytes was rejected by approximately unbiased (AU) tests in three of the four consensus trees. AU test failed to reject the chain 1 consensus tree at a confidence interval of 95% (p-AU ≥ 0.05) (Appendix B). Interestingly, the sub-clade topology of ochrophytes in chain 1 is the same as in the ML phylogenomic tree generated using the approach 1 (Figure 2.1; Appendix C). When the AU tests were repeated on hypothetically constrained trees where bigyromonada + oomycetes were monophyletic but the rest of the topology was unchanged for each of the MCMC chains, the tests failed to reject the monophyly of bigyromonada + oomycetes (Table 2.1). Rejection of bigyromonada + ochrophytes was also observed in constrained trees when the AU test was repeated on the dataset processed with approach 2 (Appendix D). To evaluate the effect of fast-evolving sites, bootstrap support and topology were compared among the ML trees that were reconstructed with increments of 5% fast-evolving sites removed from the dataset processed with approach 1. The topologies of the phylogenomic tree were maintained while the UFB support for Platysulcea increased up to 97% (Fig. 2.2). To account for possible artefacts due to long-branching attraction of fast-evolving species, tree reconstruction was repeated after removing *Cafeteria roenbergensis,* the two *Blastocystis* species, and *Cantina marsupialis*. The monophyly of bigyromonada and oomycetes was recovered with 85% UFB. However, the topology of Bigyra became unresolved with weak support for its monophyly (Appendix E).

2.3.2 Small-subunit ribosomal RNA gene tree reveals two different species assigned as *Developayella*

As shown previously, the SSU rRNA phylogenetic tree recovered the bigyromonada as a paraphyletic group, with the Pirsoniales (*Pirsonia chemainus*, *Koktebelia satura*, and *Feodosia pseudopoda*) forming a sister clade to ochrophytes (92% UFB) while the Developea clade was recovered as sister to oomycetes (Fig. 2.3). Within the Developea clade, in addition to the SSU rRNA sequences obtained from *Cubaremonas variflagellatum* and the JAMSTEC *Developayella elegans* transcriptome, I included three publicly available SSU rRNA sequences assigned as *Developayella* spp.: Accession ID U37107 (Tong, 1995; Leipe et al., 1996), MT355111.1 (Unpublished) and JX272636.1 (Del Campo et al., 2013). Note: although JX272636.1 is assigned as "Cf. *Developayella* sp." in GenBank, it was recently re-assigned as *Mediocremonas mediterraneus* (Weiler et al., 2021). The SSU rRNA sequences of the four "*Developayella*" fell into two separate groups, indicating two different species (and genera) were assigned as *Developayella elegans*, sub-clade I consisted of *Developayella elegans* U37107, *Developayella* sp. MT355111.1, *Develocanicus komovi*, *D. vyazemskyi*, and *Develocauda condao,* while sub-clade II consisted of *M. mediterraneus* (JX272636.1 and MT918788.1), JAMSTEC *Developayella elegans*, and *Cubaremonas variflagellatum* (Fig. 2.3). The SSU rRNA sequence similarity between the two sub-clade I *Developayella* species (U37107 and MT355111.1) is 98.987%, between the two species (JAMSTEC *D. elegans* and *Cubaremonas variflagellatum*) in sub-clade II 97.528% and between the originally described *Developayella elegans* U37107 and *Cubaremonas variflagellatum* 91.143%.

2.3.3 Morphology of the novel species

Developea Karpov et Aleoshin 2016

***Develocanicus vyazemskyi* (Fig. 2.4A, B) and *Develocanicus komovi* (Fig. 2.4C–M)**

Free-swimming naked eukaryovorous heterokont flagellates. The shape of the cell is irregularly flattened ellipse, where the dorsal side is more convex, and the ventral side is flatter. Two species differ in size, *Develocanicus vyazemskyi* (Colp-30) is larger and rounder, 7.4 –12.5 µm long, 4.8 – 9.2 µm wide, typical dimension ranging 9.2 x 7.0 µm. *Develocanicus komovi* (Colp-23) is slightly smaller, with the length 5.4 – 10 µm, width 3.8 – 7.4 µm and a typical dimension of 7.1 x 5.1 µm.

Cell possesses two non-acronematic heterodynamic flagella of unequal lengths (Fig. 2.4A-D, F, I, J). The posterior flagellum is two times longer than the cell, the anterior flagellum is approximately 1 – 1.5 times longer. Flagella emerge from a prominent ventral depression (Fig. 2.4A-D) which passes into a shallow wide groove (Fig. 2.4E) along the entire length of the cell. Cells predominantly exhibit active and quick swimming without rotation. During swimming, the posterior flagellum is directed backward and straight, running along the ventral depression of the cell. The anterior flagellum beats rapidly and is directed forward while slightly curved. In non-motile cells, both flagella are directed backward, beating in a slow sinusoidal wave (Fig. 2.4G, J).

The medial nucleus is located closer to the dorsal side of the cell (Fig. 2.4H). A large digestive vacuole is situated at the posterior part of the cell (Fig. 2.4I, J). As it is digested, the posterior end of the cell becomes thinner. The cells can form aggregations and attach to each other (Fig. 2.4K), sometimes forming pseudopodia (Fig. 2.4L). Transverse binary fission (Fig. 2.4M).

***Develocauda condao* (Fig. 2.4N–W)**

Free-swimming eukaryovorous heterokont flagellates (Colp-29c). The cells are slightly flattened, usually elongated-oval, less often narrow-oval or almost rod-shaped (Fig. 2.4Q). The anterior end is more rounded, the posterior end of the cell can be pointed, forming a characteristic “tail” found in starving cells (Fig. 2.4R, S). Cell length 5.14 – 12 µm, width 2.8 – 5.42 µm typically ranging 7.14 x 4.28 µm in dimension. The caudal extension is about 4.57 x 1.42 µm in size.

The cells have two heterodynamic flagella of an almost equal length with a posterior flagellum compared to the cell body. Flagella emerge from a pronounced deep ventral depression (Fig. 2.4N, O), which almost extends to the dorsal side of the cell. Depression transforms into a shallow groove (Fig. 2.4P) spanning along the entire cell, in which the posterior flagellum can fit.

The cells swim very quickly without rotating along the longitudinal axis. The posterior flagellum is straight and directed backwards. The anterior flagellum is directed forward, beats actively, and is only slightly curved. Rarely, the cells lie at the bottom with both flagella directed backward while making a slow sinusoidal movement, or the posterior flagellum beating actively.

The aggregated (Fig. 2.4U), partially fused cells (Fig. 2.4W) that form clusters were observed in culture. The medial nucleus is located closer to the dorsal side of the cell. Sated cells do not have a tail; at the posterior end of their cells there is a large digestive vacuole (Fig. 2.4T). Transverse binary fission (Fig. 2.4V).

***Cubaremonas variflagellatum* (Fig. 2.4X–AE)**

Cells (clone Dev-1) are naked and solitary bacteriovores with a length of 3.7 – 8 µm, a width 2.6 – 5.4 µm, and a typical dimension of 5.0 x 3.7 µm. The cell shape varies from elongated oval, oviform to rounder form (Fig. 2.4X-AA). Typically, the shape is irregularly ovoid, with the convex dorsal side and the flatter ventral side. The shape and size vary depending on feeding conditions. Starving cells have a small rostrum at the anterior end (Fig. 2.4AD). Cells are larger before division.

The cells possess two heterodynamic flagella of unequal length, emerging from a conspicuous ventral depression (Fig. 2.4Z, AB). Ventral depression starts from the anterior tip and continues ventrally to the middle of the cell. The anterior flagellum is approximately equal to the cell length or slightly longer, while the posterior flagellum is 1.5 – 1.8 times longer than the cell. Digestive vacuole is situated at the cell posterior. An observed cell division produces two or four cells (Fig. 2.4AE).

In culture condition, the cells predominantly lie at the bottom unattached with both flagella directed backward. The posterior flagellum runs along the ventral surface of the cell and beats rapidly with sinusoidal pattern to draw water through the depression. The anterior flagellum is hook-shaped and sweeps slowly down behind the posterior flagellum.

Although less common, when the cells swim, the curved anterior flagellum beats actively, pulling the cell forward. It is almost invisible due to its fast beating. The posterior flagellum extends behind the cell and is likely used as a rudder. The cells swim quickly, only occasionally rotating about the axis of motion. Cells can sharply change the direction of movement.

Pirsoniales Cavalier-Smith 1998 emend. 2006

***Feodosia pseudopoda* (Fig. 2.4AF, AG, AJ–AS), *Koktebelia satura* (Fig. 2.4AH), and *Pirsonia chemainus* (Fig. 2,4AI).**

Free-swimming naked, solitary and eukaryovorous heterokont flagellates. Cells are shaped as a flattened oval, with slightly pointed ends with the size 10.5 – 14 µm in length, 6 – 9.1 µm in width, and typically having the dimension of 12 x 8.2 µm. The flagellated stages of three studied Pirsoniales were almost morphologically identical except for *Feodosia pseudopoda* (Chromo-2) which possesses a small notch at the anterior part of the cell (Fig. 2.4AF, AG). Rarely, *F.*

pseudopoda can produce pseudopodia (Fig. 2.4AM, AN), which are up to 10 μm long and sometimes branched.

Two long heterodynamic flagella originate from the pit located in the anterio-medial part of the cell (Fig. 2.4AL, AM, AO). The length of the anterior flagellum is as long as the cell, while the posterior one is 2.5 times longer.

The cells swim fast in a straight line, without rotating along the longitudinal axis. The anterior flagellum is directed anteriorly, always bent towards the ventral surface. The posterior flagellum propels the cell and beats at a high speed, which can be seen as multiple posterior flagella (Fig. 2.4AI). In stationary cells, the flagella take the form of a sinusoid (Fig. 2.4AJ, AK).

The nucleus is located in the middle of the cell (Fig. 2.4AJ). The cytoplasm contains many refractive granules as observed in previously described *Pirsonia* species (Schweikert and Schnepf, 1997). Non-flagellated cells were also observed with slightly amoeboid and round shape (Fig. 2.4AP–AR). The satiated cells have a large digestive vacuole at the posterior end (Fig. 2.4AS). The eukaryovory of the biflagellates seems to be facultative as they mostly did not actively pursue the prey but only *Koktebelia satura* (clone Chromo-1) consumed all the prey cells in culture.

2.4 Discussion

2.4.1 Monophyly and phylogenetic position of the Bigyromonadea

Of the known subdivisions of stramenopiles, the Bigyromonadea stand out for their lack of data and contentious phylogenetic position (even the newly discovered *Platysulcus tardus* is represented by transcriptomic data and consistently branches at the base of the tree). From the five recent phylogenomic analyses of stramenopiles (Noguchi et al., 2016; Burki et al., 2016; Derelle et al., 2016; Leonard et al., 2018; Thakur et al., 2019), only three included a single bigyromonada representative (*Developayella elegans* JAMSTEC), none tested the monophyly of the group, and they recovered inconsistent positions. Using transcriptomes of seven new species belonging to the Bigyromonadea representing both the Developea and Pirsoniales subgroups, I tested the monophyly of the group and its position relative to other stramenopiles.

Previously, only SSU rRNA phylogenies could be used to test the monophyly of the Bigyromonadea, and such analyses consistently failed to support the monophyly, typically showing Developea with oomycetes and Pirsoniales with ochrophytes. In contrast, phylogenomic data consistently and strongly supports the monophyly of these two groups, and shows each to include multiple distinct genera.

The position of Bigyromonadea within stramenopiles as a whole is also contentious, with some analyses showing the previously available transcriptome from *D. elegans* branching with oomycetes (Noguchi et al., 2016; Thakur et al., 2019), and based on internode consistency analyses (Kobert et al., 2016; Leonard et al., 2018, with ochrophytes). This discrepancy is not entirely eliminated by the addition of new taxa, because ML phylogenomic trees with the expanded representation recovered monophyly of the bigyromonada and oomycetes with robust support, but Bayesian analyses support a clade comprising of bigyromonada+ochrophytes, and AU tests rejected most but not all topologies with this relationship (Table 2.1; Appendix D).

The discrepancy between the ML and Bayesian analyses may be due to two groups (Chrysista and Bigyromonadea) that do not fit the same model for tree reconstruction. Although it is not the aim of this study to resolve the phylogeny of ochrophytes, further examination of ochrophyte phylogeny may reveal whether the discrepancy stems from the unreconciled model used in the two groups, the different data processing approaches used, or insufficient data in one or both groups.

These results change how we interpret these lineages and their biological characteristics within the wider evolution of stramenopiles. For example, the phylogenetic position of Pirsoniales inferred from ribosomal genes implies they share a recent common ancestor with the ochrophytes, which naturally affected the interpretation of the ancestral state of ochrophytes and the role of phagoheterotrophy in their evolution (Aleoshin et al., 2016; Shiratori et al., 2017). However, the phylogenomic tree points instead to a phagoheterotrophic origin of the Pseudofungi. Parallels between this and recent suggestions on the origin of fungi are noteworthy, since *Paraphelidium tribonemae,* a phagoheterotrophic parasite belonging to phylum Aphelida, has recently been found to be sister to the osmotrophic "core" fungi by phylogenomics (Torruella et al., 2018). Close similarities in metabolism and a phagotrophy-related proteome profile of *P. tribonemae* and the osmotrophic "core" fungi suggested the "core" fungi have evolved from a phagoheterotrophic aphelid-like ancestor. Further information on the metabolism and feeding mechanisms of the new species should shed light on whether the origins of fungi and pseudofungi have more parallels and on the possible phagoheterotrophic ancestral state of Gyrista more widely.

Of course, trait evolution is also dependent on conclusively determining the position of Bigyromonadea. Substantial advances in phylogenetic methods have been made, but challenges stemming from systematic errors, compositional bias, or long branch attraction, incomplete or contaminated data, and models that do not account for heterotachy in large datasets (Delsuc et al., 2005; Kapli et al., 2020; Zhou et al., 2007) remain. Similarly, advances in single-cell sequencing have vastly increased the taxonomic scope of phylogenomics, but the severely limited starting material and the fact that they are by definition a snapshot of gene expression in one cell remain important hurdles. Here, the removal of fast-evolving sites (Fig. 2.2), species

(Appendix E), extensive AU tests (Table 2.1; Appendix A-C) and two different data processing approaches collectively tip the scale in favour of the monophyly of bigyromonada and oomycetes over the alternative position of bigyromonada with ochrophytes. However, the conflicting results of Bayesian inferences show that the lack of a robust phylogenomic tree is not just due to lack of taxonomic diversity. Continued sampling efforts in phagoheterotrophic stramenopiles will expand the phylogenetic diversity of the Bigyromonadea (and environmental SSU rRNA data already show there are more new taxa to be found) (Appendix F), but other advances in data generation and analyses will also be required.

2.4.2 Morphology, evolutionary implications, and taxonomic description of the novel phagoheterotrophic Bigyromonadea

2.4.2.1 Newly observed morphological and behavioural features in bigyromonads: cell-aggregation to fusion, pseudopod-formation, and facultative phagotrophy in motile zoospores

Before I compare morphological features, I need to clarify that the JAMSTEC strain of *Developayella elegans* has been mis-named and is a distinct species in a different genus. According to the SSU rRNA gene tree (Fig. 2.3), the originally described *D. elegans* U37107 (Tong, 1995) is placed in a distinct sub-clade of Developea (sub-clade I) whereas, *D. elegans* JAMSTEC is placed within sub-clade II with its most closely related species being *C. variflagellatum*. Renaming *D. elegans* JAMSTEC will be necessary in the future: its close relatedness to *Cubaremonas* is sufficient to indicate that it is mis-named, but rectifying this should take into account morphological information, which is currently unavailable. Overall, however, the novel developeans have similar morphological traits as previously described species. For example, *C. variflagellatum* falls in the same sub-clade as *Mediocremonas mediterraneus* (Del Campo et al., 2013; Weiler et al., 2021) (Fig. 2.3), and both have similar morphology. *C. variflagellatum* is slightly larger, but measurements for *M. mediterraneus* (2.0 – 4.0 µm in length and 1.2 – 3.7 µm in width) were most likely based on scanning electron microscopy (SEM) images and cells tend to shrink in SEM fixatives (Weiler et al., 2021). The cell size, flagella length and swimming movement of *C. variflagellatum* exhibited close similarity to *D. elegans* U37107, which was named after its characteristic "developpé" movement of the anterior flagellum during stationary feeding (Tong, 1995). However, no thread-like substances were observed, which *D. elegans* uses to attach to substrate.

The remaining novel Developea species, *Develocanicus vyazemskyi, D. komovi*, and *Develocauda condao,* differed from *D. elegans* JAMSTEC and *C. variflagellatum* by having a proportionately longer posterior flagellum, forward propulsion without rotating its axis, a eukaryovorous diet [like *Develorapax marinus* (Aleoshin et al., 2016)], and the presence of a "tail" in *D. condao*. Notably, the ability of the cells to form aggregates (Fig. 2.4K, U), pseudopodia (Fig. 2.4L), and to undergo partial cell fusion (Fig. 2.4W) has not been reported in this clade previously. The above-mentioned differences between *D. vyazemskyi D. komovi*, *Develocauda condao*, and *C. variflagellatum* are also phylogenetically reflected in the division of these species into two sub-clades (Fig. 2.1 and Fig. 2.3).

The three novel Pirsoniales, *Feodosia pseudopoda, Koktebelia satura*, and *Pirsonia chemainus*, described here as *nomen provisorium*, most likely represent a motile zoospore stage of unknown algal parasites. The novel Pirsoniales species did not actively pursue the provided prey and only partially consumed their prey (except *K. satura* which consumed all the prey provided), all the cultures died after a few months to one year of cultivation. Although there has been extensive description of auxosome and trophosome formation during the parasitic stage of known Pirsoniales (Schnepf et al., 1990; Schweikert and Schnepf, 1997), the ability of motile zoospores to acquire effective eukaryovory has not been described so far. The observed eukaryovory of the zoospore-like Pirsoniales is likely facultative, as the cells were cultured without potential hosts and the cells with larger food vacuoles became non-flagellated and rounded, a structure akin to an auxosome. However, further culture experimentations with their natural hosts are required to verify their ability to form parasitic auxosomes and trophosomes from motile phagotrophic zoospores.

I postulate that the facultative eukaryovory at the motile zoospore stage provides a significantly increased survival rate and thus extension of the motile stage during their dispersal until a suitable host is found. This ability can be particularly advantageous before the onset of seasonal algal bloom, where the zoospores can efficiently infect multiple hosts without resource competition. Therefore, the sustained survival of the zoospores via facultative eukaryovory could be an important factor leading to the evolutionary success of Pirsoniales parasites.

Feodosia pseudopoda differed from rest of the Pirsoniales studied here by an anterior notch (Fig. 2.4AF, AG) and rare occurrences of pseudopodia (Fig. 2.4AM-AO). The two characteristics have been reported in *Pseudopirsonia mucosa*, a cercomonad rhizarian (Kühn et al., 2004), which had been mis-assigned as *Pirsonia* due to the similarities in their parasitic life cycles. In starving and immobile zoospores of *Pirsonia puntigerae*, filopodium-like processes (Schweikert and Schnepf, 1997) have been described. However, pseudopodia in motile zoospores of Pirsoniales have not been observed previously.

The presence of pseudopodia, and the ability to form aggregated cells in the newly described sub-clade I of Developea and previously reported publications of Pirsoniales may be synapomorphic traits of Bigyromonadea. It will be important for future studies to compare ultrastructure and genes putatively associated with cell-aggregation or fusion among the species of bigyromonada, thus potentially addressing the evolution of an osmotrophic nutritional strategy in stramenopiles.

2.4.2.2 Similarities among Oomycetes motile zoospores, Labyrinthulomycetes, and Bigyromonadea

Morphologically, the novel Developea species have similar features to motile zoospores of previously studied oomycetes, such as the general cell dimension, the ratio of anterior and posterior flagellum, and two laterally oriented flagella (with a tinsellate anterior flagellum) emerging from a ventral groove (Dick, 2000), which resembles the ventral depression observed in the novel species. Behaviourally, the swimming pattern (e.g., direction of flagella, sinusoid form) is comparable (Ho and Hickman, 1967; Hickman, 1970). Another striking similarity between the two groups is their ability to self-aggregate, which is observed in oomycete zoospores as a distinct form of self-aggregation related to aggregation towards host-plant tissues (Ko and Chase, 1973; Bassani et al., 2020). Similarly, cell aggregation observed in this study was not a result of attraction to food as this behaviour was observed rarely, and feeding of these predatory flagellates is associated with active mobile eukaryotic prey hunting. Additionally, cells attaching to each other were distinguishable from the intermediate stage of transverse binary cell division. The mechanism underlying self-aggregation in oomycetes has not been fully resolved. However, recent studies suggest that a combination of chemotaxis (Judelson and Blanco, 2005; Zheng and Mackrill, 2016; Bassani et al., 2020) and bioconvection (Savory et al., 2014), is involved in the process. The exact role of the self-aggregation in oomycete pathogenesis is still unclear. However, the fact that a similar observation was made in its sister-clade, the Bigyromonadea, indicates that self-aggregation may have been present in the ancestor of Pseudofungi, before the osmotrophic parasitism of oomycetes evolved. Cell aggregation is also observed in *Sorodiplophrys* (Dykstra and Olive, 1975), a species belonging to another osmotrophic group of stramenopiles, the labyrinthulomycetes. Cell aggregation has convergently evolved multiple times across many other supergroups (Parfrey and Lahr, 2013), such as Opisthokonta (Brown et al., 2009), Discoba (Brown et al., 2012; He et al., 2014), Amoebozoa (Du et al., 2015), Rhizaria (Brown et al., 2012), and ciliates (Sugimoto and Endoh, 2006), and whether cell aggregation within stramenopiles arose convergently or divergently should be further investigated.

As mentioned previously, some species described in this study formed pseudopodia (Fig. 2.4L,4AM,4AN) and partially fused cells (Fig. 2.4W) resembling amoeboid forms. Labyrinthulomycetes also form filose pseudopodia (Gomaa et al., 2013) akin to pseudopodia observed in this study (Fig. 2.4AM, AN). These are found in Amphitremidae, during an amoeboid stage of *Diplophrys* (Anderson and Cavalier-Smith, 2012), and other labyrinthulids (Raghukumar, 1992), implying this trait either evolved convergently or was present earlier than the divergence of Pseudofungi.

Another notable similarity between oomycetes and the novel bigyromonada is their potential marine origin, as all known bigyromonads are exclusively marine. Molecular clock analyses indicate the Silurian period as the time of oomycete origins (Matari and Blair, 2014), while the earliest fossil evidence points to the Devonian period (Krings et al., 2011). The fossil evidence of the “deep-branching” genera have shown them to be marine parasites of seaweed or of crustaceans based on molecular studies (Beakes and Sekimoto, 2009), both suggesting a marine origin of oomycetes as a facultative parasitic osmotroph (Beakes et al., 2012, 2014; Beakes and Thines, 2017).

The origin and evolution of major stramenopile subgroups is coming into sharper focus with the increase in phylogenomic data from diverse species. The new taxa described here, together with future descriptions of the still-substantial diversity of bigyromonada that has not been well-characterized, can potentially shed more light on this and the origins of oomycetes in particular. I propose that the ancestor of oomycetes was a phagoheterotrophic amoeboid, as postulated in the evolution of true fungi, and that this transition might be better understood through a detailed functional examination of the novel species. Just as the highly successful analyses of choanoflagellates and unicellular opisthokonts changed our understanding of the origin of animals (Sebé-Pedrós et al., 2013; Zmitrovich, 2018; Chow et al., 2019), a similar analysis of the distribution of genes involved in Pseudofungi cell-aggregation or pseudopodia formation across the diversity of bigyromonads could be a future direction to understand the evolution of these unique phagoheterotrophs and oomycetes.

2.4.3 Taxonomic summary

Taxonomy: Eukaryota; SAR Burki et al. 2008, emend. Adl et al. 2012; Stramenopiles Patterson 1989, emend. Adl et al. 2005; Gyrista Cavalier-Smith 1998; Bigyromonadea Cavalier-Smith, T. 1998; Developea Karpov et Aleoshin 2016

Cubaremonas n. gen. Tikhonenkov, Cho, and Keeling

Diagnosis: naked and solitary bacteriovorous protist. Cell shape is irregularly ovoid, with the convex dorsal side and the flatter ventral side. Cells possess two heterodynamic flagella emerging from a conspicuous ventral depression, which starts from the anterior end and continues ventrally to the middle of the cell. In culture condition, the cells predominantly lie at the bottom unattached with both flagella directed backward.

Etymology: from lat. cubare – to lie, to be lying down and monas (lat.) – unicellular organism.

Zoobank Registration. urn:lsid:zoobank.org:act: 169A2385-5669-4FB2-A728-AC5AD74B5076

Type species. *Cubaremonas variflagellatum*

Cubaremonas variflagellatum n. sp. Tikhonenkov, Cho, and Keeling

Diagnosis: cells length 3.7 - 8 μm, cell width 2.6 - 5.4 μm. Flagella of unequal length, the anterior one is approximately equal to the cell length while the posterior flagellum is 1.5 - 1.8 times longer than the cell. At lying cells, posterior flagellum runs along the ventral surface of the cell and beats rapidly with sinusoidal pattern to draw water through the depression. The anterior flagellum is hook-shaped and sweeps slowly down behind the posterior flagellum. Starving cells have a small rostrum at the anterior end. Digestive vacuole is situated at the cell posterior. An observed cell division produces two or four cells.

Type Figure: Fig. 2.4X illustrates a live cell of strain Dev-1.

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OL630098.

Type locality: water column of Strait of Georgia, British Columbia, Canada

Etymology: the species name means “unequal flagella”, lat.

Zoobank Registration: urn:lsid:zoobank.org:act: 2152FF4A-BFC8-4064-A197-74FE6BEE2EC8

Develocanicus n. gen. Tikhonenkov, Cho, Mylnikov, and Keeling

Diagnosis: Free-swimming naked eukaryovorous heterokont flagellates with two non-acronematic heterodynamic flagella of unequal lengths. The shape of the cell is irregularly flattened ellipse, where the dorsal side is more convex, and the ventral side is flatter. Flagella emerge from a prominent ventral depression which passes into a shallow wide groove along the entire length of the cell.

Etymology: from développé (fr.) – characteristic ballet movement and volcanicus (lat.) (found near volcanos in Kanary island and Crimea).

Zoobank Registration. urn:lsid:zoobank.org:act: 74F9B793-53AD-4F4C-8A71-3F29D9F97B9E

Type species. *Develocanicus komovi*

Develocanicus komovi n. sp. Tikhonenkov, Cho, Mylnikov, and Keeling

Diagnosis: cell length 5.4 - 10 μm, cell width 3.8 - 7.4 μm. The posterior flagellum is two times longer than the cell, the anterior flagellum is approximately 1 - 1.5 times longer. Cells swim without rotation. At that, posterior flagellum is directed backward and straight, running along the ventral cell of the cell. Anterior flagellum beats rapidly and is directed forward while slightly curved. Medial nucleus is located closer to the dorsal side of the cell. Large digestive vacuole is situated at the posterior part of the cell. Cells can form pseudopodia and aggregations and attach to each other. Transverse binary fission.

Type Figure: Fig. 2.4C illustrates a live cell of strain Colp-23.

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OL630096.

Type locality: black volcanic sand on the littoral of Maria Jimenez Beach (Playa Maria Jiménez), Puerto de la Cruz, Tenerife, Spain

Etymology: named after Prof., Dr. Viktor T. Komov, Russian ecotoxicologist, who carried out fieldwork and collect samples, where new species was discovered.

Zoobank Registration: urn:lsid:zoobank.org:act: 6C543426-FAFB-4DBD-AEB3-3CA648FD53D5

Develocanicus vyazemskyi n. sp. Tikhonenkov, Cho, Mylnikov, and Keeling

Diagnosis: cell 7.4 - 12.5 μm long, 4.8 - 9.2 μm wide. The posterior flagellum is two times longer than the cell, the anterior flagellum is approximately 1 - 1.5 times longer. Cells swim without rotation. At that, posterior flagellum is directed backward and straight, running along the ventral cell of the cell. Anterior flagellum beats rapidly and is directed forward while slightly curved. In non-motile cells, both flagella are directed backward, beating in a slow sinusoidal wave. Medial nucleus is located closer to the dorsal side of the cell. Large digestive vacuole is situated at the posterior part of the cell. Transverse binary fission.

Type Figure: Fig. 2.4A illustrates a live cell of strain Colp-30.

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OL630097.

Type locality: near shore sediments on the littoral near T.I. Vyazemsky Karadag Scientific Station, Crimea

Etymology: named after Dr. T.I. Vyazemsky, founder and first director of Karadag Scientific Station, Crimea

Zoobank Registration: urn:lsid:zoobank.org:act: 6A2D2D31-E16A-470F-9ED9-26546944A96C

Develocauda n. gen. Tikhonenkov, Cho, and Keeling

Diagnosis: Free-swimming eukaryovorous heterokont flagellates with slightly flattened elongated-oval cells and two heterodynamic flagella. The anterior end is more rounded, the posterior end of the cell can be pointed, forming a characteristic "tail" in starving cells. Flagella emerge from a pronounced deep ventral depression, which almost extends to the dorsal side of the cell. Depression transforms into a shallow groove spanning along the entire cell, in which the posterior flagellum can fit.

Etymology: from développé (fr.) – characteristic ballet movement and cauda (lat.) – tail.

Zoobank Registration. urn:lsid:zoobank.org:act: 5BA3D9B6-0A50-45A5-83D3-7474EA31F13C

Type species. *Develocauda condao*

Develocauda condao n. sp. Tikhonenkov, Cho, and Keeling

Cell length 5.14 - 12 µm, width 2.8 - 5.42 µm. The caudal extension is about 4.57 x 1.42 µm in size. Flagella of almost equal length. The cells swim very quickly without rotating along the longitudinal axis. The posterior flagellum is straight and directed backwards. The anterior flagellum is directed forward, beats actively, and is only slightly curved. Cells can be partially fused and aggregated. Medial nucleus is located closer to the dorsal side of the cell. Transverse binary fission.

Type Figure: Fig. 2.4N illustrates a live cell of strain Colp-29c.

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OL630094.

Type locality: near shore sediments on the littoral of north-east part of Con Dao Island, South Vietnam

Etymology: named after Con Dao Island, South Vietnam, where species was discovered.

Zoobank Registration: urn:lsid:zoobank.org:act: FA73444D-79A5-498C-BB7F-139E9D82C0BA

Pirsoniales Cavalier-Smith 1998, emend. 2006

Studied pirsoniales most likely represent a motile zoospore stages of unknown algal parasites. Since data on the stage of the parasitic trophonts (auxosome and a trophosome) are not available, it is premature to formulate taxonomic diagnoses. But we provide provisional names (nom. prov.) which can be used for future research.

Pirsonia chemainus nom. prov. Tikhonenkov, Cho, and Keeling

Etymology: species epithet is after the Stz'uminus First Nation traditional territory (Strait of Georgia area) claimed by the Chemainus First Nation

Type locality: water column of the Strait of Georgia, British Columbia, Canada

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OL630095.

Koktebelia satura nom. prov. Tikhonenkov, Cho, and Keeling

Etymology: genus epithet reflects the place of finding, Koktebel bay, Crimea; species epithet – from satur (lat.), well-fed.

Type locality: near shore sediments on the littoral near T.I. Vyazemsky Karadag Scientific Station, Crimea

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OL630093.

Feodosia pseudopoda nom. prov. Tikhonenkov, Cho, and Keeling

Etymology: genus epithet reflects the place of finding, the settlement Beregovoye, Feodosiya, Crimea; species epithet reflects the ability to produce pseudopodia.

Type locality: near shore sand on the littoral of the beach in the settlement Beregovoye, Feodosiya, Crimea

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OL630092.

Table 2.1 Approximately unbiased (AU) tests on tree constraints based on approach 1 dataset.

Approach 1 (MAFFT L-INS-i and trimAl -g 0.8)			
Constrained Tree	p-AU	logL	ΔlogL
Unconstrained ML tree	0.78	-3935691.338	0
ML tree	0.759	-3935691.338	0.00089261
Chain 1 (C+S+Pi),(R+P+X+E)	0.0541	-3935828.083	136.75
Chain 1 Modified (Bigyromonada+oomycetes)	0.267	-3935763.845	72.508
Chain 2 (C+S+Pi+E),(R+P+X)	**0.0297**	-3935859.39	168.05
Chain 2 Modified (Bigyromonada+oomycetes)	0.0924	-7871604.549	108.01
Chain 3 (R+P+X+E),(C+S)	**0.0119**	-3935874.998	183.66
Chain 3 Modified (Bigyromonada+oomycetes)	0.0717	-3935805.205	113.87
Chain 4 (C+S+E),(R+P+X)	**0.0186**	-3935860.003	168.67
Chain 4 Modified (Bigyromonada+oomycetes)	0.108	-3935765.741	74.404

Except for the unconstrained ML tree, each tree was constrained under LG+C60+F+G4 using IQ-TREE with the approach 1 dataset. Chain 1 to chain 4 are generated from Bayesian analyses and contain (bigyromonada+ochrophytes). “Chain 1 Modified” to “Chain4 Modified” contain a hypothetical clade (bigyromonada+oomycetes) with the rest of topology remaining the same with their corresponding chains. Each unmodified chain is listed with different topology of Chyrisista as represented in Appendix B. The unconstrained tree is based on ML tree reconstructed under LG+C60+F+G4+PMSF as presented in Fig. 1. The p-AU values were calculated using the AU test with 10,000 RELL bootstrap replicates, implemented in IQ-TREE. The maximum log likelihoods (logL) of each constrained and their differences (ΔlogL) compared to the unstrained tree are listed. Constraints with P-values lower than 0.05 are rejected, indicating confidence interval below 95% (marked bold). Raphidophyceae (R), Eustigmatophyceae (E), Chrysophyceae (C), Synurophyceae (S), Phaeophyceae (P), Pinguiophyceae (Pi), and Xanthophyceae (X).

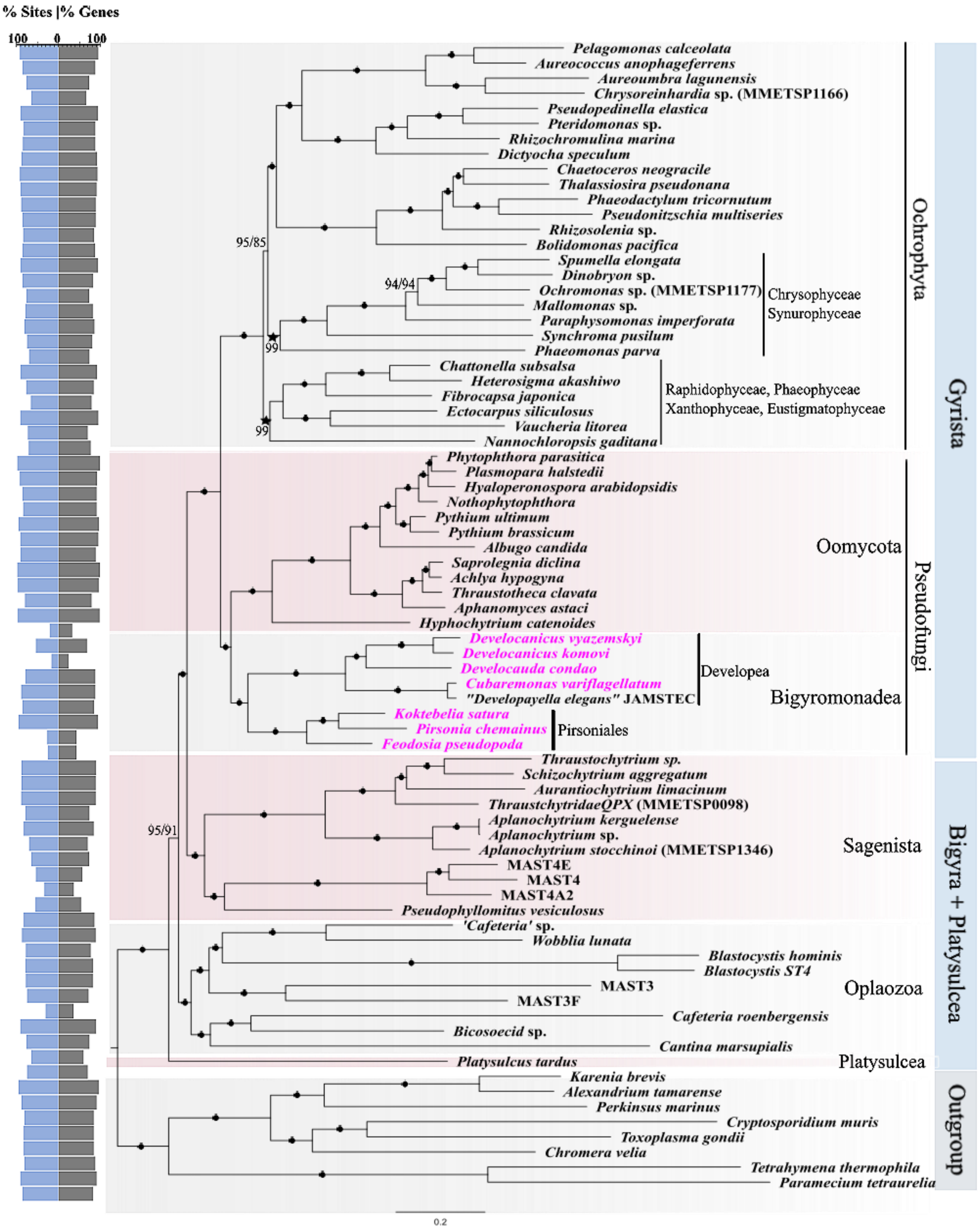


Figure 2.1 Phylogenomic tree of stramenopiles with the seven new Bigyromonadea

Multi-gene tree of stramenopiles with the seven new transcriptomes (pink) are added to Gyrista consisting of the concatenated alignments of 247 genes of 76 taxa. The tree was reconstructed using a maximum-likelihood (ML) analysis, under the site-heterogenous model, LG+C60+F+G4+PMSF, implemented in IQ-Tree. Branch support was calculated using non-parametric PMSF 100 standard bootstrap (STB). Branches with ≥99% STB for both approaches are marked with black bullets while others are labelled as "Approach 1 STB/Approach 2 STB". The topology of the trees generated from the two approaches were the same except for the positions of Raphidophyceae, Phaeophyceae, Xanthophyceae, and Eustigmatophyceae and, Chrysophyceae and Synurophyceae, which were swapped in the tree reconstructed based on the dataset processed using approach 2 (i.e., Prequal/Divvier method); denoted by star symbols (Appendix B). The percent sites (blue) and genes (grey) present for each transcriptome is depicted on the back-to-back bar plot on the left.

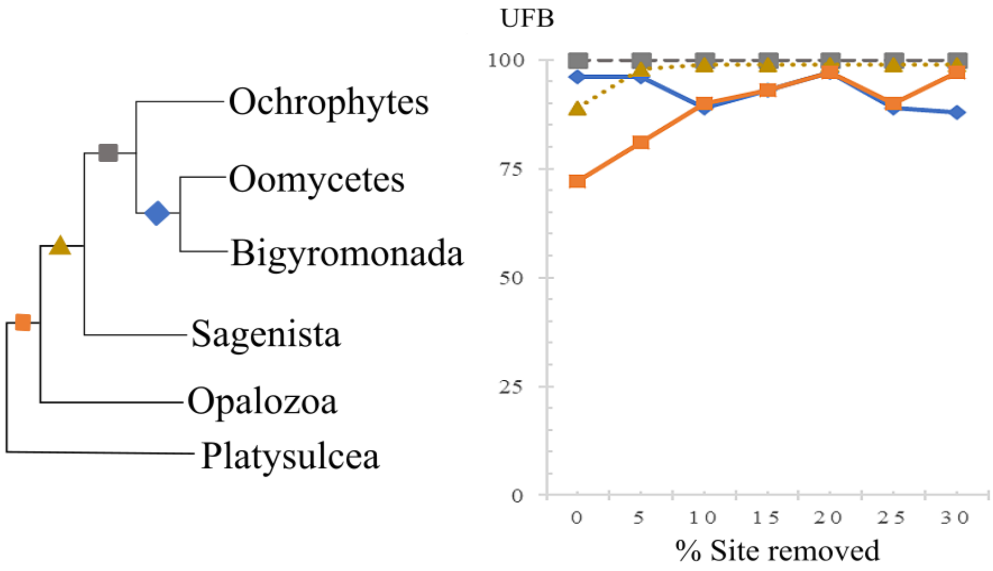


Figure 2.2 Summary of ultrafast bootstrap with fast-evolving sites removed

Summary of ultrafast bootstrap (UFB) with incremental removal of fast-evolving sites, based on the dataset processed with approach 1. Schematic representation the stramenopiles ML tree (left) with each branch marked with different shapes and colours. The line plot (right) showing the change in UFB for each branch when fast-evolving sites were incrementally removed by 5%. The monophyly of Gyrista shows full support throughout while the UFB increases incrementally for 'Sagenista' and 'Platysulcea'.

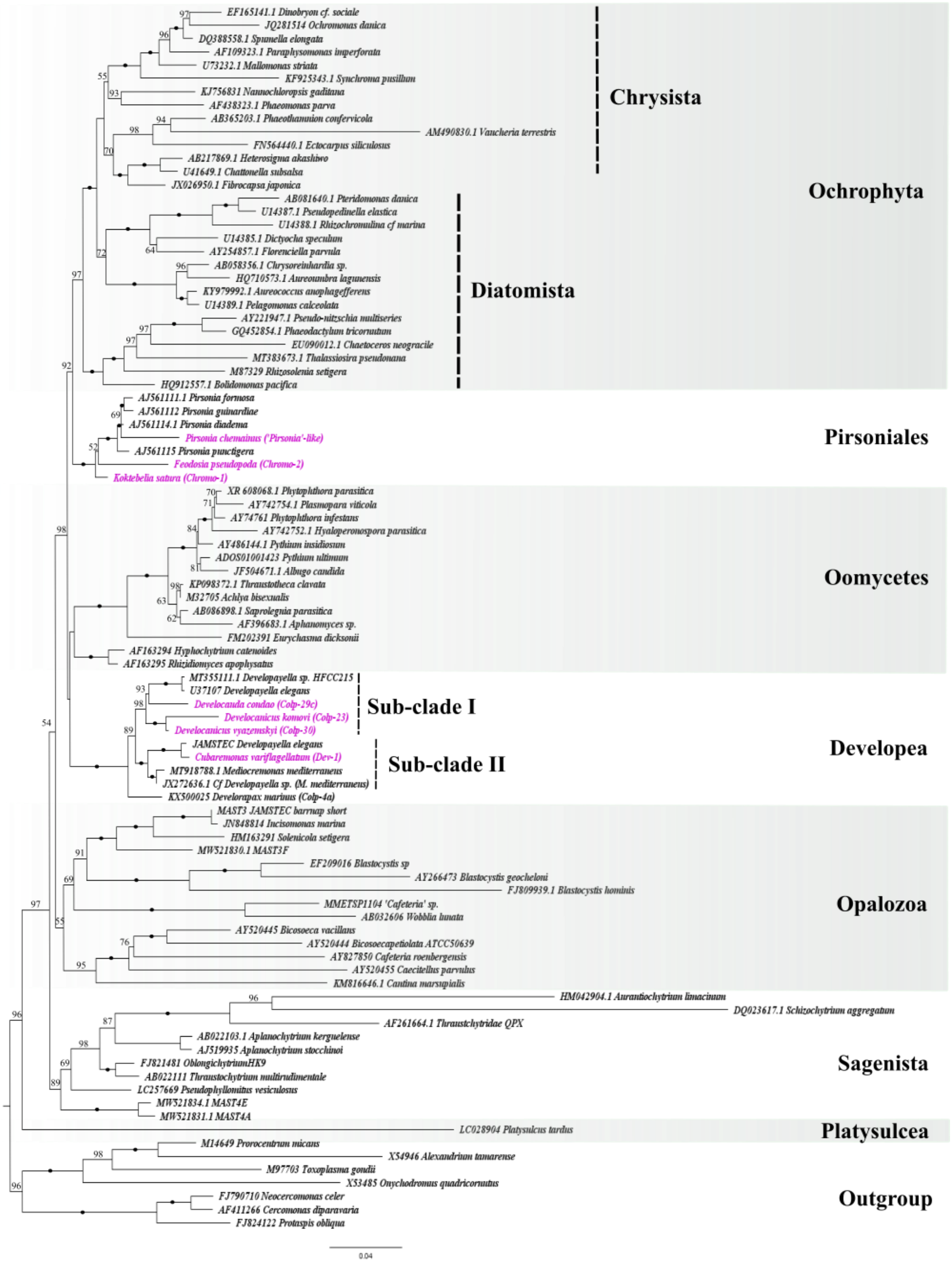

EF165141.1 Dinobryon cf. sociale
JQ281514 Ochromonas danica
DQ388558.1 Spumella elongata
AF109323.1 Paraphysomonas imperforata
U73232.1 Mallomonas striata
KF925343.1 Synchroma pusillum
KJ756831 Nannochloropsis gaditana
AF438323.1 Phaeomonas parva
AB365203.1 Phaeothamnion confervicola
AM490830.1 Vaucheria terrestris
FN564440.1 Ectocarpus siliculosus
AB217869.1 Heterosigma akashiwo
U41649.1 Chattonella subsalsa
JX026950.1 Fibrocapsa japonica
Chrysista
AB081640.1 Pteridomonas danica
U14387.1 Pseudopedinella elastica
U14388.1 Rhizochromulina cf marina
U14385.1 Dictyocha speculum
AY254857.1 Florenciella parvula
AB058356.1 Chrysoreinhardia sp.
HQ710573.1 Aureoumbra lagunensis
KY979992.1 Aureococcus anophagefferens
U14389.1 Pelagomonas calceolata
AY221947.1 Pseudo-nitzschia multiseries
GQ452854.1 Phaeodactylum tricornutum
EU090012.1 Chaetoceros neogracile
MT383673.1 Thalassiosira pseudonana
M87329 Rhizosolenia setigera
HQ912557.1 Bolidomonas pacifica
Diatomista
Ochrophyta
AJ561111.1 Pirsonia formosa
AJ561112 Pirsonia guinardiae
AJ561114.1 Pirsonia diadema
Pirsonia chemainus ('Pirsonia'-like)
AJ561115 Pirsonia punctigera
Feodosia pseudopoda (Chromo-2)
Koktebelia satura (Chromo-1)
Pirsoniales
XR 608068.1 Phytophthora parasitica
AY742754.1 Plasmopara viticola
AY74761 Phytophthora infestans
AY742752.1 Hyaloperonospora parasitica
AY486144.1 Pythium insidiosum
ADOS01001423 Pythium ultimum
JF504671.1 Albugo candida
KP098372.1 Thraustotheca clavata
M32705 Achlya bisexualis
AB086898.1 Saprolegnia parasitica
AF396683.1 Aphanomyces sp.
FM202391 Eurychasma dicksonii
AF163294 Hyphochytrium catenoides
AF163295 Rhizidiomyces apophysatus
Oomycetes
MT355111.1 Developayella sp. HFCC215
U37107 Developayella elegans
Develocauda condao (Colp-29c)
Develocanicus komovi (Colp-23)
Develocanicus vyazemskyi (Colp-30)
Sub-clade I
JAMSTEC Developayella elegans
Cubaremonas variflagellatum (Dev-1)
MT918788.1 Mediocremonas mediterraneus
JX272636.1 Cf Developayella sp. (M. mediterraneus)
KX500025 Develorapax marinus (Colp-4a)
Sub-clade II
Developea
MAST3 JAMSTEC barrnap short
JN848814 Incisomonas marina
HM163291 Solenicola setigera
MW521830.1 MAST3F
EF209016 Blastocystis sp
AY266473 Blastocystis geocheloni
FJ809939.1 Blastocystis hominis
MMETSP1104 'Cafeteria' sp.
AB032606 Wobblia lunata
AY520445 Bicosoeca vacillans
AY520444 Bicosoecapetiolata ATCC50639
AY827850 Cafeteria roenbergensis
AY520455 Caecitellus parvulus
KM816646.1 Cantina marsupialis
Opalozoa
HM042904.1 Aurantiochytrium limacinum
DQ023617.1 Schizochytrium aggregatum
AF261664.1 Thraustchytridae QPX
AB022103.1 Aplanochytrium kerguelense
AJ519935 Aplanochytrium stocchinoi
FJ821481 OblongichytriumHK9
AB022111 Thraustochytrium multirudimentale
LC257669 Pseudophyllomitus vesiculosus
MW521834.1 MAST4E
MW521831.1 MAST4A
Sagenista
LC028904 Platysulcus tardus
Platysulcea
M14649 Prorocentrum micans
X54946 Alexandrium tamarense
M97703 Toxoplasma gondii
X53485 Onychodromus quadricornutus
FJ790710 Neocercomonas celer
AF411266 Cercomonas diparavaria
FJ824122 Protaspis obliqua
Outgroup
0.04

Figure 2.3 ML tree of stramenopiles using a 18S rRNA gene alignment

ML tree reconstructed from a 18S rRNA gene alignment of 92 taxa (1650 sites), under BIC: TIM2+R6 with 1000 UFB. Branch support with ≥99% UFB is marked with black bullets while the values less than 50% are not shown. The seven new species described in this study are marked as pink: Pirsoniales forming a sisterhood with Ochrophytes and Developea forming a sister clade to Oomycetes. Within Developea, two previously assigned *Developayella* species (JAMSTEC transcriptome and the U37107 SSU rRNA sequence) are split into two sub-clades, in which the four novel Developea species are positioned.

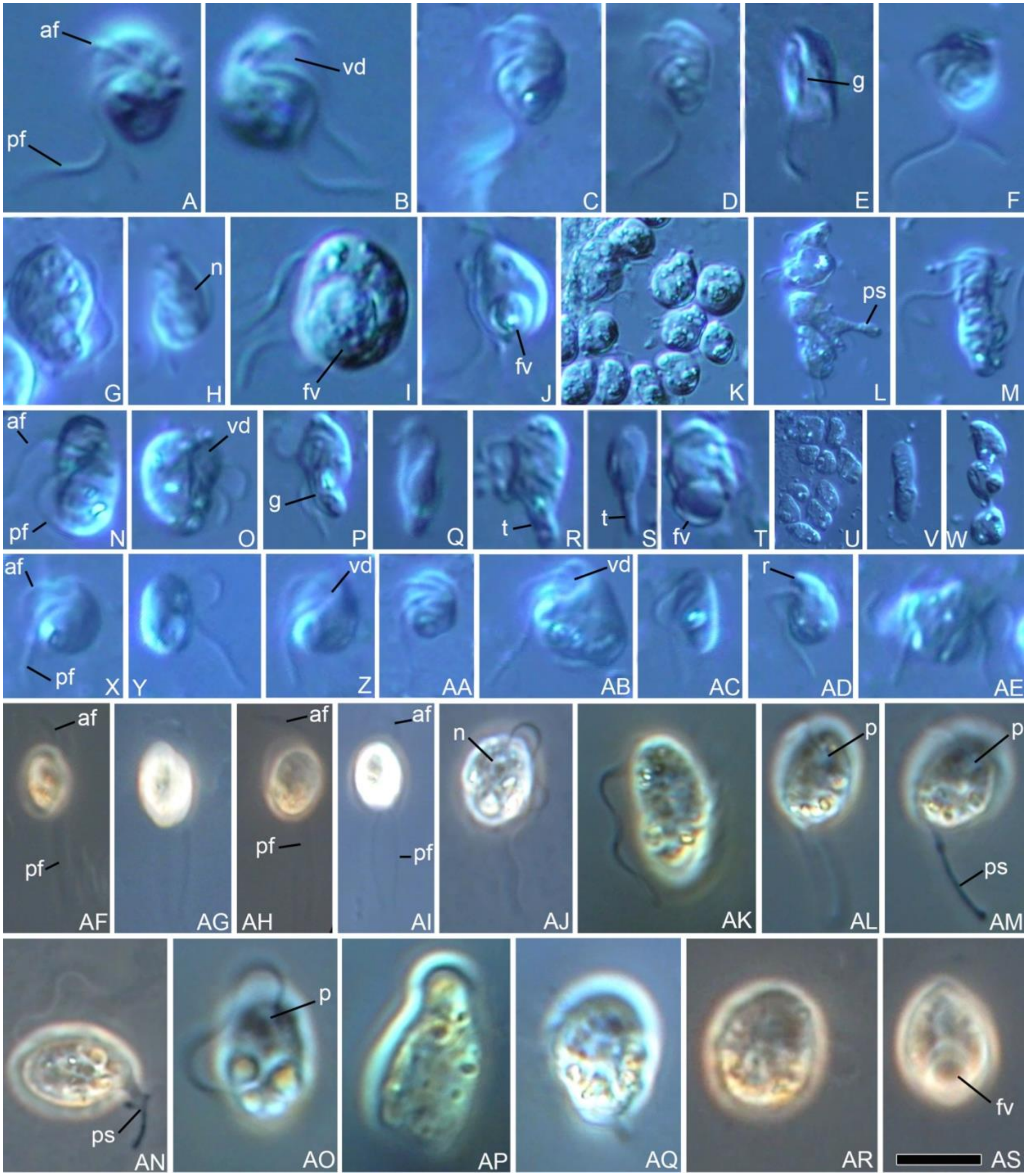


Figure 2.4 Morphology of the seven new phagoheterotrophic Bigyromonadea

A, B. *Develocanicus vyazemskyi,* general cell view with flagella (anterior flagellum [af] and posterior flagellum [pf]) and ventral depression [vd]. **C–M.** *Develocanicus komovi,* C–F – general cell view with flagella and ventral depression, shallow wide groove [g] is visible in (E), G – lying cell with posterior flagellum [pf] beating with a slow sinusoidal wave, H–J – cells with medial nucleus [n] (H) and large food vacuoles [fv] (I, J), K – cell aggregation, L – aggregated cells with pseudopodia [ps], M – transverse binary fission. **N–W.** *Develocauda condao,* N–P – general cell view with two flagella and ventral depression, Q – rod-shaped cell, R,S – cells with pointed 'tail-like' [t] posterior end, T – cells with large food vacuole, U – cell aggregation, V – transverse binary fission, W - partially fused cells. **X–AE.** *Cubaremonas variflagellatum,* X–AA – general cell view with flagella, AB – cell with conspicuous ventral depression, AC, AD – starving cells with small rostrum [r] (AD), AE – division into 4 cells. **AF, AG, AJ – AS.** *Feodosia pseudopoda,* AF, AG – typical fast swimming cell with two flagella, AJ, AK, AO – lying cells with sinusoid shaped flagella, AL–AN – cells with pseudopodia and anterior pit [p] (AL, AM), AP–AR – metabolic cells, AS – cell with large food vacuole. **AH.** *Koktebelia satura*, typical fast swimming cell with two flagella. **AI.** *Pirsonia chemainus*, typical fast swimming cell with two flagella. **Scale bar –** the scale changes in different images with respect to the scale bar in the AS image: A, B, R, AK, AP, AQ, AS – 8 µm; C–H, J, N, P,Q, X–AA, AC, AD – 7 µm; I, O, T, AB – 5 µm; K, L, V, AF–AI – 15 µm; M, S, AE, AJ, AL–AO, AR – 10 µm; U – 25 µm; W – 20 µm.

Chapter 3: Phylogenomic position of genetically diverse phagotrophic stramenopile flagellates in the sediment-associated MAST-6 lineage and a potentially halotolerant placididea

3.1 Introduction

Stramenopiles are a diverse group of eukaryotes, in terms of molecular sequences, size, trophic mode, and morphology. The best known are within one subgroup, the Ochrophyta, which includes diatoms with diverse frustule shapes, microscopic phagotrophic flagellates that have lost photosynthesis (Dorrell et al., 2019; Kayama et al., 2020), and macroscopic multicellular brown algae like kelps. The diversity is less obvious at the morphological level in some "deep-branching" groups of stramenopiles, but their molecular diversity is nonetheless significant. This is most obvious in the Bigyra Cavalier-Smith, 1998, which is a large assemblage composed of Sagenista Cavalier-Smith, 1995 and Opalozoa Cavalier-Smith, 1993, and includes the most deep-branching stramenopile, *Platysulcus tardus* (Shiratori et al., 2015). Other than saprotrophic Labyrinthulea (Sagenista), epiphytic *Solenicola setigera* (MAST-3), and symbiotic Opalinata (Opalozoa), rest of the species of Bigyra are marine phagotrophic flagellates, generally in the size range of 2–10 mm (Gómez et al., 2011; Guillou et al., 1999; Lee, 2002; Moriya et al., 2002, 2000; Schoenle et al., 2022; Shiratori et al., 2017, 2015; Yubuki et al., 2015, 2010). These small and inconspicuous flagellates had been historically mistaken for cercozoans or discobans in light microscopy surveys (Larsen and Patterson, 1990; Patterson et al., 1993; Lee, 2002). Without detailed morphological examination and molecular surveys, it is difficult to discern among flagellated species of Bigyra, or indeed even between members of the major subdivisions, leading to under-sampling and under-estimation of their diversity.

The diversity and abundance of Bigyra has accordingly been determined by molecular surveys, and these have revealed a number lineages without any morphological identity simply referred to as MArine Stramenopile (MAST), a term that was originally coined to include 18 uncharacterized lineages, most of which are still only known from small subunit (SSU) ribosomal RNA (rRNA) sequences identified from environmental sampling efforts (Logares et al., 2012; Massana et al., 2014, 2009, 2006, 2004). These surveys also showed that differences in community composition among different environments, particularly between benthic and pelagic samples (Massana et al., 2015; Forster et al., 2016). Notably, MAST-6 (along with MAST-9, and -12) are common in sediments but rare in pelagic samples (Logares et al., 2012; Massana et al., 2015; Rodríguez-Martínez et al., 2020). Extensive ultrastructural and cellular examination of the first cultured MAST-6 lineage, *Pseudophyllomitus vesiculosus* (family Pseudophyllomitidae Shiratori et al., 2017) describe it as a relatively large algivore (dimensions up to 18.3 x 12.4 mm) with the characteristic flagellar apparatus ultrastructure of deep-branching stramenopiles (e.g., no x-fiber, R2 flagellar root with 13 microtubules) (Lee and Patterson, 2002; Moestrup, Ø, 1976; Shiratori et al., 2017; Yubuki et al., 2010). The genus *Pseudophyllomitus* (Lee, 2002) was erected to describe *Phyllomitus*-like taxa without two adhering flagella. This resulted in re-designation of four new species (i.e., *Pseudophyllomitus apiculatus, P. granulatus, P. salinus*, and *P. vesiculosus*). However, with limited molecular and ultrastructural data available to support the monophyly of the genus. Later, a new family Pseudophyllomitidae (Shiratori et al., 2017) was erected, apparently corresponding to a MAST-6 clade. However, the phylogenetic position of the type species, *P. granulatus* is unknown. As a result, we refrain from using Pseudophyllomitidae in replacement of MAST-6 in this study.

In multi-gene analyses, MAST-6 are closely related to many ecologically important groups such as MAST-4 (Cho et al., 2022; Thakur et al., 2019), which is one of the most common heterotrophic flagellate groups in coastal ecosystems, substantially affecting microbial food webs (Massana et al., 2006; Rodríguez-Martínez et al., 2009; Logares et al., 2012). MAST-

4 and -6, in turn form a sister group to Labyrinthulea, a detrital decomposer that is also abundant in sediments (Collado-Mercado et al., 2010; Massana et al., 2015; Nakai and Naganuma, 2015; Rodríguez-Martínez et al., 2020). The diversity of MAST-6 has been demonstrated by SSU rRNA amplicon sequencing of various sediment studies from surface to deep-sea push core sediments (Massana et al., 2015; Rodríguez-Martínez et al., 2020; Schoenle et al., 2021) and the high relative abundance in sediments estimated by up to 46 Operational Tasonomic Units (OTUs) and >1000 sequencing reads in a recent study (Rodríguez-Martínez et al., 2020). Two morphotypes of MAST-6 were observed in plankton samples that differ in sizes and seasonal abundances (Piwosz and Pernthaler, 2010), with the larger morphoptype (9.9-22 mm) showing rapid increase in abundance only for a week. This study showed MAST-6 is not only phylogenetically diverse, but also that community composition can respond quickly to fluctuating environments and food availability. Despite these advances, however, only a single transcriptome is available for the MAST-6 lineage (Shiratori et al., 2017; Thakur et al., 2019), limiting our understanding of its biology and its relationship to other stramenopiles, and character evolution of Sagenista.

Another major but under-sampled subgroup of Bigyra is the Placididea, a class within Opalozoa. Like Sagenista, the phylogenetic diversity of Placididea is largely represented by SSU rRNA genes, with -omic data available for only two species: *Wobblia lunata* (Moriya et al., 2000; Thakur et al., 2019) and *Placididea* sp. (the CaronLab strain formerly mis-labelled as '*Cafeteria*') (Keeling et al., 2014). Placidideans are closely related to MAST-3, another abundant and highly diverse heterotrophic flagellate group that play an important role in marine food webs and are found in all coasts and open oceans around the world (Gómez et al., 2011; Logares et al., 2012; Massana et al., 2004). Unlike the MAST clades, SSU sequences of Placididea (Moriya et al., 2002) do not amplify well with V4-targeting primers (Lee et al., 2022). Consequently, the diversity of this group was reported either using V9-targeting SSU primers (Lee et al., 2022) or by individually isolating placidideans (Park et al., 2006; Park and Simpson, 2010; Rybarski et al., 2021). Isolated placidideans are often from hypersaline environments (>40‰), however, many characterized halophilic placidideans can tolerate lower salinity (Park and Simpson, 2010; Rybarski et al., 2021), raising the possibility that they are also present in non-hypersaline environments.

Here, I describe three new MAST-6 species including two new genera, and one new Placididea species, providing microscopic observation, transcriptomic data, and SSU rRNA sequence comparisons with previously generated environmental amplicon data. Strain PhM-7 (Placididea, *Haloplacidia sinai*) and Colp-33 (MAST-6, *Vomastramonas tehuelche*) were maintained in culture for a year, but subsequently lost. Two other MAST-6 species (*Mastreximonas tlaamin* and *Pseudophyllomitus* sp. BSC2) were obtained by single cell isolations. I also describe transcriptomes from two cultured species, *Symbiomonas scintillans* RCC257 (Guillou et al., 1999) and *Caecitellus* sp. RCC1078 (O'Kelly and Nerad, 1998), to further fill out the diversity of deep-branching stramenopiles for phylogenomic analyses. I report the relative abundance and diversity of the four new species of MAST-6 and Placididea in publicly available environmental sequence surveys, and re-examine stramenopile phylogeny, particularly with the aim to resolve relationships within the Bigyra using a multi-gene approach based on my new transcriptome data.

3.2 Materials and Methods

3.2.1 Sample collection and imaging

Strain PhM-7 (*Haloplacidia sinai* sp. nov.) was isolated from the Red Sea (average salinity 36-41 ‰), Sharm El Sheikh, Egypt (27°50’50.5" N, 34°18’59.4’’ E), scraped from coral at 75 m depth in April 2015. Strain Colp-33 (*Vomastramonas tehuelche* gen. et sp. nov.) was isolated from nearshore bottom sediments, Chile, Punta Arenas (53°37'49" S, 70°56’58’’ W, T=9.4 °C, Salinity 24 ‰) in November 2015. These strains were propagated in a predator-prey culture with the bodonid *Procryptobia sorokini* as a steady food source but both perished after a year of cultivation. Light microscopy observations for PhM-7 and Colp-33 were made using a Zeiss AxioScopeA.1 equipped with phase contrast and DIC water immersion objectives (63x) and an AVT HORN MC-1009/S analog video camera. For scanning electron microscopy (SEM) imaging of PhM-7, cells from exponential growth phase were fixed at 22 °C for 10 min in a cocktail of 0.6% glutaraldehyde and 2% OsO_4 (final concentration) prepared using a 0.1 M cacodylate buffer (pH 7.2), and gently drawn onto a polycarbonate filter (diameter 24 mm, pores 0.8 µm). Following filtration, the specimen was taken through a graded ethanol dehydration and acetone, and critical-point dried. The dry filters with fixed specimens were mounted on aluminum stubs, coated with gold-palladium, and observed with a JSM-6510LV scanning electron microscope (JEOL, Japan).

Two uncultured single cells, PRC5 (*Mastreximonas tlaamin* gen. et sp. nov.) and BSC2 (*Pseudophyllomitus* sp. BSC2), were isolated from oxic marine intertidal sediment. Sediment for PRC5 was collected from Powell River, British Columbia, Canada (49°50'42" N, 124°31'60" W) in August 2020; whereas the BSC2 sample was collected from Boka Santa Cruz, Curaçao (12°18'24" N, 69°8'44" W) in April 2022. Both cells were manually isolated using a drawn-out glass micropipette under a Leica DLIM inverted microscope and imaged with a Sony α7rIII camera. The cells were rinsed twice in filtered sea water and transferred into a 0.2 mL PCR tube containing lysis buffer (Picelli et al., 2014) and stored in -80°C until cDNA synthesis.

Cultures of *Symbiomonas scintillans* strain RCC257 and *Caecitellus* sp. strain RCC1078 were obtained from the Roscoff culture collection (France) in March 2022. The cultures were grown in 30 mL of 0.22 µm filtered f/2 medium (30 ‰) and autoclaved seawater (30 ‰), respectively, both with an autoclaved rice grain added. The cultures were kept in a 20°C incubator with a 12 hour:12 hour light:dark cycle and sub-cultured every two weeks.

3.2.2 cDNA synthesis, library preparation and sequencing

Cells of PhM-7 (*H. sinai*) and Colp-33 (*V. tehuelche*) grown in clonal cultures were harvested when the cells had reached peak abundance and after most of the prey had been eaten. The cells were collected by centrifugation (2,000 x g for PhM-7 and 1,000 x g for Colp-33, both at room temperature) onto the 0.8 µm membrane of a Vivaclear mini column (Sartorium Stedim Biotech Gmng, Cat. No. VK01P042). Total RNA was then extracted using a RNAqueous-Micro Kit (Invitrogen, Cat. No. AM1931). In addition to the RNA extraction from the Colp-33 clonal cultures, 20 single cells were manually picked from its culture using a glass micropipette and transferred into a 0.2 mL PCR tube containing the cell lysis buffer for an additional Smart-Seq2 cDNA synthesis and library preparation.

For cultures obtained from the Roscoff Culture Collection (RCC257 and RCC1078), TRIzol™ LS Reagent was used to extract total RNA, following the manufacturer's instructions with a modification at the aqueous-organic layer separation step. Briefly, 100 mL of each culture was centrifuged at 3220 x g for 20 min at 4°C to pellet cells at the bottom of the centrifuge tubes. After carefully discarding the media, 1 mL of TRIzol™ LS was added to the pelleted cells. For an easier transfer of the aqueous phase containing the RNA without an interphase contaminant, the aqueous-organic layer separation by chloroform was done in Phasemaker™ (Invitrogen) tubes. The quality and quantity of the RNA yield was determined using a NanoDrop 1000 Spectophotometer v3.8.1 (Thermo Fisher Scientific). Additionally, using glass micropipettes, approximately 20 cells were manually isolated from each culture and processed in the same manner as the single-cell isolation method used for Colp-33 (*V. tehuelche*), PRC5, and BSC2.

For cDNA synthesis, the poly-A selection based Smart-Seq2 protocol was used (Picelli et al., 2014). For manually isolated single cells in the lysis buffer, 2-3 rounds of freeze-thaw steps were included prior to the cDNA synthesis (Onsbring et al., 2020). For RNA extracts, 4µL of the extract was used for cDNA synthesis. The rest of the library preparation and sequencing steps (tagmentation, quality control, and adaptor ligation) for PRC5, BSC2, RCC257 and RCC1078 were carried out by the Sequencing and Bioinformatics Consortium (University of British Columbia, BC Canada), using the Illumina Nextera^{TM} DNA Flex Library Preparation Kit. The sequencing was performed on a NextSeq (mid-output) platform with 150 bp paired-end library constructs. For PhM-7 and Colp-33, the libraries were prepared using Nextera^{TM} XT DNA Library Preparation Kit (Illumina, Inc., Cat. # FC-131-1024) followed by Illumina Miseq 300 bp paired-end sequencing at GenoSeq, Sequencing & Genotyping Core (University of California Los Angeles, CA USA) for PhM-7, and Sequencing and Bioinformatics Consortium (University of British Columbia, BC Canada) for Colp-33. All the raw reads of the transcriptomes are deposited in the NCBI Short Read Archive (SRA) under the BioProject number PRJNA961826 (SRR24392492 to SRR24392501).

3.2.3 Transcriptome processing, assembly, and decontamination

Along with the six newly generated transcriptomes in this study, recently published transcriptomes of *Actinophrys sol* (Azuma et al., 2022) and its prey, *Chlorogonium capillatum,* were processed as follows. The quality of the raw sequencing reads was assessed using FastQC v0.11.9 (Andrews, 2010). To correct random sequencing errors of the short Illumina RNA-seq reads, *k*-mer based Rcorrector (version 3) was used on the raw reads (Song and Florea, 2015). The error-corrected reads were then trimmed using Trimmomatic v0.39 (Bolger et al., 2014) to remove remnant transposase-inserts from the library preparation, Nextera^{TM} DNA Flex adaptors, low quality reads (-phred33), and Smart-Seq2 IS-primers with the leading and trailing cut-off at 3, SLIDINGWINDOW:4:15, and MINLEN:36. Processed forward, reverse, and unpaired transcripts were assembled using the *de novo* transcriptome assembler rnaSPAdes v3.15.1 (Bushmanova et al., 2019). Additionally, for species with two libraries prepared from both RNA extract and single cell isolations (i.e., Colp-33, RCC257, and RCC1078), the resulting transcripts were co-assembled. BlobTools v2.3.3 (Challis et al., 2010; Laetsch and Blaxter, 2017) was used to identify contaminants and visualize contig coverage. In short, megaBLAST was used to search assembled transcripts against the NCBI nucleotide database followed by a diamond BLASTX (Altschul et al., 1990; Buchfink et al., 2015) protein search against the UniProt reference database (Buchfink et al., 2015; The UniProt Consortium et al., 2021). Both searches were performed with an e-value cut-off 1e-25. Bacterial, Viriplantae, obazoan, and archaeal reads were removed from all transcripts. To remove prey contaminants from PhM-7, Colp-33, and *A. sol*, the assembled transcripts were first searched against the transcriptome of the respective prey (*Procryptobia sorokini* for PhM-7 and Colp-33; and *C. capillatum* for *A. sol*) using BLASTn, followed by the removal of contigs with ≥ 95% sequence identity. TransDecoder v5.5.0 (Haas et al., 2013) was used to predict open reading frames (ORFs) and the longest ORFs were annotated using a BLASTP search against UniProt database with the e-value cut-off 1e-5. BUSCO v5.2.2 (Simão et al., 2015) with 'stramenopiles_odb10' database was used to assess the completeness of each transcriptome.

3.2.4 Small subunit sequences and amplicon processing using QIIME 2

Small subunit (SSU) rRNA sequences were extracted from PRC5 and BSC2 transcriptomes using barrnap v0.9 (Seemann, 2007). For *S. scintillans* and *Caecitellus* sp., SSU rRNA sequences were generated by polymerase chain reaction (PCR) amplification of cDNA using 18SFU and 18SRU eukaryotic primers (Tikhonenkov et al., 2016), followed by Sanger dideoxy sequencing. Although the SSU sequences for *S. scintillans* RCC257 and *Caecitellus* sp. and RCC1078 are available in GenBank, I did SSU PCR to confirm species identity and to obtain longer sequences as the published *S. scintillans* RCC257 (accession KT861043) SSU is 760 bp. For all the downstream analyses, I included SSU sequences from this study for the two cultured bikosia.

To obtain SSU rRNA sequences of Colp-33 and PhM-7, the cells were first harvested when the cultures had reached peak abundance and after the prey had been eaten (confirmed with light microscopy), followed by centrifugation (7,000 x g, room temperature) onto an 0.8 µm membrane of a Vivaclear mini column (Sartorius Stedim Biotech Gmng, Cat. No. VK01P042). Total DNA was extracted from the filters using the MasterPure Complete DNA and RNA Purification Kit (Epicentre, Cat. No. MC85200). The SSU rRNA genes were PCR-amplified using the general eukaryotic primers EukA-EukB for strain Colp-33 (Medlin et al., 1988), and GGF-GGR for strain PhM-7 (Tikhonenkov et al., 2022). PCR products were subsequently cloned prior to sequencing (PhM-7) or sequenced directly (Colp-33), using Sanger dideoxy sequencing with two additional internal sequencing primers 18SintF and 18SintR (Tikhonenkov et al., 2022). All the SSU rRNA sequences from the four newly described species and two culture strains are deposited in GenBank with the accession OQ909082-OQ909087.

To compare SSU rRNA sequences of newly identified speciess to previously reported studies, five sediment datasets were obtained via European Nucleotide Archive (ENA). The datasets are designated as follows: BioMarKs (Dunthorn et al., 2014; Massana et al., 2015), SouthChina (Wu and Huang, 2019), Norway (unpublished BioProjects PRJEB24876; PRJEB24158; PRJEB24888), Deepsea (Schoenle et al., 2021), and ISME2020 (Rodríguez-Martínez et al., 2020) (Table 3.1). For the sixth dataset (designated as ESBig), I obtained ten SSU rRNA sequences (ESBig130-139) assigned to Placididea directly from the authors (Lee et al., 2022) (Table 3.1). These studies examined sediments from different bodies of water across the US, Europe, and Asia, including the South China Sea, North Atlantic Ocean, Mariana Basin, Philippine Basin, Bunnefjorden (Norway), Pacific Ocean, and a freshwater lake. The depths of the sample sites vary from 20 m to 5497 m, and cover diverse marine, brackish and freshwater environments such as push-cores or surface sediments of seafloors, fjords, abyssal plains, and continental rises. Except for ESBig, all datasets were processed using QIIME 2 (q2cli v2020.11.1) (Bolyen et al., 2018). For the 454 pyrosequencing data (BioMarKs and SouthChina), the raw reads were imported and demultiplexed with '--type SampleData[SequencesWithQuality]' and '--input-format SingleEndFastqManifestPhred33' options. After trimming the raw reads with respective primer-pair sequences, both 454 pyrosequencing and Illumina sequencing data were filtered with a DADA2 denoising step (Callahan et al., 2016). To remove chimeric sequences, denoised sequences were further processed with 'uchime-denovo' (Rognes et al., 2016). For taxonomic classification of amplicon sequence variants (ASVs), a QIIME 2 compatible PR2 v4.14.0 dataset was obtained (Del Campo et al., 2018; Guillou et al., 2012) and modified by manually adding the SSU rRNA sequences of the new species described from this study and relevant sequences from the recent PR2 database and GenBank (Park and Simpson, 2010; Guillou et al., 2012; Rybarski et al., 2021). The modified PR2 dataset was used to pre-train the QIIME 2 classifier using 'qiime feature-classifier fit-classifier-naïve-bayes' (Pedregosa et al., 2011). The trained classifier was then used to assign taxonomy to filtered representative amplicon sequence variants (ASVs). Amplicon sequence variants assigned to MAST-6 and Placididea were extracted and added to a stramenopile SSU rDNA alignment consisting of partial to nearly full-length sequences (Cho et al., 2022; Yubuki et al., 2015). Additionally, relevant environmental sequences from the PR2 database, GenBank, and 10 placididean-associated operational taxonomic units (OTUs) from ESBig (Lee et al., 2022) were added. The extracted feature sequences were further subjected to CD-HIT to remove duplicates (Li and Godzik, 2006).

To check presence and visualize relative abundance of newly acquired MAST-6 and Placididea species in the amplicon dataset (Table 3.1), feature tables from QIIME2 were exported and processed in RStudio (R v4.2.0) with ggplot2 (Wickham, 2016).

3.2.5 Small-subunit (SSU) rRNA gene tree construction

The compiled SSU rRNA sequences were aligned with MAFFT v7.481 (Katoh and Standley, 2013) resulting 8,771 sites, followed by maximum likelihood inference using RAxML v8.2.12 (Stamatakis, 2014) under the GTRGAMMA model with 1000 ultrafast bootstrap replicates (UFB). To further evaluate the phylogenetic placement of short amplicon sequences from the amplicon datasets (Table 3.1), additional phylogenetic supports were estimated using the Evolutionary Placement Algorithm (EPA) (Berger et al., 2011) with EPA-ng v0.3.8 (Barbera et al., 2018). This method used the reference ML tree constructed with the same conditions as above with partial to nearly full-length SSU rRNA sequences. To determine the placement probability of each amplicon sequence variant (ASVs) assigned to MAST-6 or Placididea, a likelihood weight ratio (LWR) was determined using GAPPA (Czech et al., 2020). The ASVs with an LWR value higher than 95% were inspected for chimerism using BLASTn and passing sequences were considered to be accurate with high confidence (Dunthorn et al., 2014). The SSU rRNA tree with EPA analysis is hereafter referred as SSU-EPA tree.

To evaluate phylogenetic relationships of newly added MAST-6, placididean, and other species of Bigyra, another SSU rRNA phylogenetic tree was constructed without short amplicon sequences, hereafter referred as the SSU-tree. A total of 224 SSU rRNA sequences ≥ 900bp consisting of previously compiled datasets and new sequences (Aleoshin et al., 2016; Cho et al., 2022; Rybarski et al., 2021; Yubuki et al., 2015) were aligned using MAFFT v7.481 (Katoh and Standley, 2013), followed by trimming using trimAl 1.2rev59 (-gt 0.3, -st 0.001) (Capella-Gutiérrez et al., 2009). The phylogenetic tree was then constructed based on 1649 sites using IQ-TREE v2.1.0 (Minh et al., 2020) under TIM2+F+R6, the optimal model determined with ModelFinder (Kalyaanamoorthy et al., 2017) and 1000 UFB.

3.2.6 Phylogenomic matrix construction using PhyloFisher

The phylogenomic matrix including the predicted proteins of the newly produced transcriptomes were generated using PhyloFisher v1.1.2 (Tice et al., 2021). Briefly, annotated ORFs from the newly generated transcriptomes were searched against the 241 gene set embedded in PhyloFisher and the resulting homologs were then added to each of the gene alignments. For each of the updated 241 gene alignments, a single-gene tree was constructed using IQ-TREE v1.6.12 (Nguyen et al., 2015) under the L+G4+X model and 1000 UFB. Each single-gene tree was manually screened using ParaSorter v1.0.4 to ensure orthologs were inferred from the newly added proteins. Predicted orthologs of recently published or relevant stramenopiles (Azuma et al., 2022; Cho et al., 2022; Keeling et al., 2014; Richter et al., 2022; Thakur et al., 2019) were kept. To generate a final concatenated phylogenomic matrix, 98 taxa (including 15 taxa for an outgroup) were selected, resulting in a 240 gene set with 76,516 amino acid (aa) sites. Beside the main concatenated matrix, two additional concatenated matrices were generated to evaluate the effects of ortholog completeness in determining the phylogeny: one that included only orthologs found in ≥39% of taxa (233 orthologs with 74,531 aa sites, referred as 39per-matrix), and another that included orthologs found in ≥59% of taxa (215 orthologs with 67,630 aa sites; referred as 59per-matrix). Additionally, I generated another matrix with the most recent genomic data of other MAST lineages (MAST-1, MAST-7, MAST-8, MAST-9, and MAST-11) (Labarre et al., 2021; Richter et al., 2022) with 74,898 aa sites (234 orthologs) composed of 104 taxa (hereafter, referred as MASTer-matrix).

3.2.7 Phylogenomic tree reconstruction, removal of fast-evolving sites, and recoding

The initial maximum likelihood (ML) tree of the main concatenated phylogenomic matrix was inferred using IQ-TREE v2.1.2 under the empirical profile mixture model, LG+C60+F+G4 (Quang et al., 2008) with 1000 UFB. The resulting ML tree was used as a guide to estimate posterior mean site frequencies (PMSF) (Wang et al., 2018), which was then used to re-estimate a final ML-PMSF tree with 100 non-parametric standard bootstraps under the same model. The construction of the ML-PMSF phylogenomic tree was repeated with the 39per- and 59per-matrices. To consider the effect of fast-evolving sites on tree topology, the main concatenated matrix was further subjected to a stepwise 10,000 aa site removal using PhyloFisher (fast_site_removal.py) followed by construction of ML-PMSF trees. To account for potential amino acid composition bias in the dataset, web-based Composition Profiler (Vacic et al., 2007) was used with default settings to compare relative abundances of GARP vs. FYMINK amino acids with "SwissProt 51" (Bairoch, 2004) as a background, in addition to examining a distance matrix tree output generated by 'aa_comp_calculator.py' in PhyloFisher. To remove potential amino acid composition bias, the main concatenated matrix was recoded with the Dayhoff 18 (Dayhoff et al., 1978; Wang et al., 2018; Hernandez and Ryan, 2021) option using PhyloFisher v1.2.4 (aa_recode.py) followed by a tree reconstruction under the MULTI18_GTR+FO and 100 replicates of standard bootstrap with RAxML-NG v.1.1.0 (Kozlov et al., 2019).

To infer a phylogenomic tree using Bayesian estimation, the CAT-GTR mixture model was used with the -dgm 4 option in PhyloBayes-MPI v4.0.3 (Lartillot and Philippe, 2004; Lartillot et al., 2009). Four independent Markov Chain Monte Carlo (MCMC) chains were run in parallel for at least 10,000 generations. The consensus posterior probability and topology were estimated after discarding first 20% as burn-in and subsampling every second tree. Convergence of the four chains was tested with bpcomp.

3.3 Results

3.3.1 Phylogenomic tree of stramenopiles

The final phylogenomic matrix used for constructing the main phylogenomic tree is a concatenated alignment of 240 genes (76,517 sites) and 98 taxa (including 15 taxa belonging to an outgroup). The average percentage of genes present for each included transcriptome is 71.6%, with 76.4% of sites covered. These values are lower in the newly added transcriptomes: 8.5% genes and 16.3% sites for *Pseudophyllomitus* sp. BSC2; 21.1% genes and 25% sites for *Mastreximonas tlaamin*; 38.6% genes and 52.5% sites for *Vomastramonas tehuelche*; 38.2%

genes and 47.5% sites for *Haloplacidia sinai*; 21.42% genes and 35.42% sites for *Caecitellus* sp.; 42.7% genes and 58.3% sites for *Symbiomonas scintillans* (Fig. 3.1). The BUSCO scores showed a similar pattern where *Pseudophyllomitus* sp. BSC2 and *M. tlaamin* had the lowest values (4%:4% and 8%:2% completed:fragmented) while *V. tehuelche* and *H. sinai* had 28%:19% and 26%:16% and, *S. scintillans* and *Caecitellus* sp. 43%:10% and 11%:7%, respectively.

Based on the main phylogenomic tree inferred from ML analysis under LG+C60+F+G4+PMSF, Gyrista was monophyletic and the Bigyra was paraphyletic (Fig. 3.1). Within Gyrista, Ochrophytes and Pseudofungi are monophyletic with strong support. In Ochrophyta, the Raphidophyceae, Phaeophyceae, and Xanthophyceae (RPX) clade formed a monophyly with the Chrysophyceae, Synurophyceae, and Synchromophyceae + Pinguiophyceae clade (CSS + Pi) with moderate bootstrap support (84%). Bacillariophyceae + Bolidophyceae + Pelagophyceae and Dictyochophyceae (BBDPe) formed a fully supported clade. The monophyly of RPX, CCS + Pi, and BBDPe was moderately supported (84%). However, phylogenetically unstable Eustigmatophyceae formed a weakly supported (71%) clade with *Actinophrys sol*, a non-photosynthetic heliozoan stramenopile. The clade comprising Eustigmatophyceae + *A. sol* clade branched sister to the rest of the Ochrophyta. In the ML-PMSF trees reconstructed from on MASTer, 39per- and 59per-matices, *A. sol* was sister to CSS + Pi while Eustigmatophyceae was sister to RPX with moderate support (81% to 94%, and 76% to 89%, respectively) (Appendix G-H).

Within Bigyra, the three new MAST-6 species, *M. tlaamin*, *Pseudophyllomitus* sp. BSC2 and *V. tehuelche* formed a clade with *Pseudophyllomitus vesiculosus,* with *Pseudophyllomitus* sp. BSC2 being the immediate sister lineage to *P. vesiculosus*. MAST-6, Eogyrea (MAST-4), and Labyrinthulea all formed a monophyletic group, Sagenista. In the tree reconstructed with

MASTer-matrix (Appendix H), MAST-7 and MAST-11 were robustly supported as monophyletic, which then was sister to Eogyrea. MAST-8 and MAST-9 formed close relationship to the grouping consisted of Eogyrea. MAST-7, and MAST-11. MAST-6 formed robust monophyly with this grouping composed of Eogyrea, MAST-7, -8, -9, and -11. The new Placididea species, *H. sinai*, is closest to *Placididea* sp. (Caron Lab) and, together with *Wobblia lunata*, comprise the Placididea clade. Placididea formed a sister lineage to the rest of the Placidozoa (MAST-3 and *Blastocystis* sp.). However, the support value for the Nanomonadea (MAST-3) and Opalinata (*Blastocystis* sp.) clade was weak (70% bootstrap). Placidozoa and Bikosia in turn comprise a robust monophyletic group, the Opalozoa, which is the sister lineage to the rest of the stramenopiles, except for *Platysulcus tardus*. *Symbiomonas scintillans* RCC257 is sister to *Cafeteria burkhardae* (Fenchel and Patterson, 1988; Schoenle et al., 2020) and this clade is a well-supported sister lineage to a clade composed of *Caecitellu*s sp. RCC1078 and *Halocafeteria seosinensis* (Park et al., 2006) (Fig. 3.1).

When fast-evolving sites were removed from the concatenated matrix to assess the effects of long-branch attraction, monophyly of each of the Ochrophyta, Gyrista, Sagenista, and Opalozoa were well supported up to 65% site removal (50,000 aa; Fig 3.2A). The monophyly of pseudofungi and the relationship between Gyrista and paraphyletic Bigyra were well supported up to 39% sites removed (30,000 aa; Fig. 3.2A). However, the groups with weak to moderate supports (70-84%) in Fig. 3.1 continue to show unstable relationships when fast-evolving sites are removed (Fig. 3.2B). Particularly, the placement of *A. sol*, *Microchloropsis gadidata* (Eustigmatophyceae) and sub-groups of opalozoans is different. An alternative placement for *A. sol* is as a sister lineage to rest of the ochrophytes when 20,000 aa sites (26%) and 40,000 sites (52%) are removed (Fig. 3.2B). For *M. gadidata*, it formed a sister lineage with Pinguiophyceae or the rest of the ochrophytes except *A. sol*. Although the paraphyly of Bigyra was always supported with the progression of fast-evolving site removal, the relationships among the sub-groups kept changing with weak support (~70%) (Fig. 3.2B).

To evaluate the effect of amino acid composition bias within sub-groups of opalozoans (namely Placidozoa), I inspected GC% of each transcriptome. All taxa belonging to Placididea are enriched in GARP amino acids compared to the background dataset, whereas all Opalinata are enriched in FYMINK, as is Nanomonadea with the exception of *Incisomonas marina*. Additionally, the amino acid composition of Placididea is more similar to Bikosia than the rest of the Placidozoa (Appendix I). However, when a phylogenomic tree is reconstructed using the recoded main matrix, the topology of the Placidozoa remains the same as Fig. 3.1, while the placement of *A. sol* and *M. gadidata* changes, with *A. sol* being recovered as the sister lineage to Pinguiophyceae and *M. gadidata* as the sister lineage to RPX (Appendix J).

For the Bayesian analysis, the chains did not converge (maxdiff=1), with all chains conflicting with one another. When a consensus tree from each chain was compared, all the trees had the same topology of Sagenista and Bikosia that was also seen in the ML-PMSF inferred trees (Fig. 3.1; Appendix K). For Placidozoa, all the consensus trees had Nanomonadea branching sister to a clade composed of Opalinata and Placididea, a different topology from the ML-PMSF analysis, except the one constructed with the MASTer-matrix (Fig. 3.1; Appendix H and K). All chains had different Ochrophyta topologies, although the sub-clade relationship of BBPe was the same as the ML-PMSF inferred trees. Compared to the ML-PMSF tree (Fig. 3.1), the same topology was observed for the monophyletic CSS and the monophyletic RPX. The placement of *M. gadidata*, *A. sol*, and Pinguiophyceae were the most inconsistent across chains.

In all chains, Bigyromonadea is a sister lineage to Ochrophyta and two out of four chains recovered Bigyra as monophyletic (excluding *P. tardus*) (Appendix K).

3.3.2 New species represent phylogenetically diverse MAST-6 group in SSU rRNA analysis

To determine the genetic diversity of MAST-6 in publicly available sediment datasets, a SSU rRNA tree was constructed with the extracted amplicon sequence variants (ASVs) trained with the modified PR2 reference database, including the SSU rRNA sequences of the newly described species in this study. All the SSU rRNA sequences obtained from the newly described MAST-6 species (>1800bp), *H. sinai* (>1800bp) and two bikosia (>1600bp) species are nearly full length. In total, 12 unique ASVs from BioMarKs were assigned to MAST-6 species; 9 for SouthChina; 16 for Norway; 6 for Deepsea; and 61 for the ISME2020 dataset. In general, studies that targeted the V4 region had sequence lengths between 183 to 460 bp; the V1-2 region 397 to 429 bp; and the V9 region 128-154 bp (Table 3.1). Shorter sequence lengths (~180 bp) from V4 targeted amplicon data are unpaired reads where low quality reverse reads were dropped.

The SSU rRNA analysis of environmental data revealed substantial diversity of the MAST-6 group, which were largely grouped into four sub-groups. The new MAST-6 species and previously cultured species were found within the three sub-groups (Fig. 3.3); *M. tlaamin* in sub-group I, *Pseudophyllomitus* sp. BSC2 in sub-group II, and *V. tehuelche* in sub-group III. However, their phylogenetic relationship needs additional examination as the branch support values were weak (Appendix L). To evaluate the prevalence of these new MAST-6 species in the amplicon studies in sediment samples, relative abundance was plotted against other environmental MAST-6 ASVs (Fig. 3.4). All sediment datasets had relatively high abundance of MAST-6, particularly BioMarKs (65% of all MASTs and 2.24% of all ASVs) and ISME2020 (37% of all MASTs and 13.9% of all ASVs) studies. Amplicon sequence variants assigned to *M.*

tlaamin (PRC5) were dominant MAST-6 groups in Deepsea (67% of all MAST-6), ISME2020 (44%) and Norway (20%) datasets, while no sequences assigned to the new MAST-6 species were present in the BioMarKs study (Fig. 3.4B). It is important to note that not all ASVs assigned to *M. tlaamin* correspond exactly to the same species. However, they were assigned based on which most closely related MAST-6 species was available in the training dataset. This assignment may change as new MAST-6 transcriptomes representing each sub-group are added to the updated SSU reference data. I did not find shared MAST-6 ASVs across the four studies, which may be due to different sequencing technologies with different coverage, or presence of biological sequence variants by different sampling sites and time. Within sub-group I (Fig. 3.3), the ASVs from the ISME2020 and Norway datasets are placed closest to *M. tlaamin*, while the ASVs from the Deepsea dataset are more distantly related. This indicates that MAST-6 species closely related to *M. tlaamin* are not only genetically diverse and abundant, but the present in various sediment samples across different depths and geological locations. Additionally, within the MAST-6 sub-group I, *M. tlaamin* and the environmental sequence “SA2_3F7” are the only two with nearly full length SSU rRNA sequences, compared to sub-group II, which includes more close-to-full length SSU rRNA sequences. The addition of the *M. tlaamin* SSU rRNA sequence in the taxonomic assignment has markedly improved phylogenetic resolution among the MAST-6 lineages. A similar trend was observed in sub-group III where *V. tehuelche* is placed. Amplicon sequence variants from the BioMarKs dataset that were assigned to MAST-6, however, were mostly placed across the different sub-groups, except sub-group II. Along with many ASVs from ISME2020, six out of 12 unique MAST-6 ASVs of BioMarKs are placed within sub-group IV, which have no sequences from cell isolates with genomic data. When I visualized the abundance of different sub-groups across different datasets, Sub-group I was the dominant group in all cases (Appendix L). More sub-groups were present in ISME2020 and Norway and this is likely due to sequencing techniques (i.e., pyrosequencing in BioMarKs) and limited universality of V9 primer used in Deepsea dataset (Appendix M). As QIIME2 generates ASVs, we interpreted the data without clustering. However, clustering the ASVs by ≥98% sequence similarity resulted in 10 and 37 ASVs assigned to MAST-6 in BioMarKs and ISME2020, respectively. For other MAST lineages, ASVs assigned to MAST-1, -3, -9, and -12 were present in all studies. Depending on the dataset, the relative abundance these MAST lineages were high although values fluctuated depending on the sample within the study.

Two ASVs assigned to *V. tehuelche* were present in the SouthChina study, and no ASVs were assigned to *Pseudophyllomitus* sp. BSC2 (Appendix M). However, based on initial phylogenetic evaluation of assigned MAST-6 sequences from the SouthChina study, blastn searches, and the EPA analysis (low LWR values with the equal likelihood of alternative placements), the sequences were excluded from main the SSU-EPA tree (Appendix L-M). Additionally, one Deepsea ASVs assigned to MAST-6 was excluded from the downstream analysis based on the initial phylogenetic tree, and blastn search places it close to MAST-8 (Appendix L). Aside from *M. tlaamin*, other MAST-6 sequences from cell isolates (*P. vesiculosus* and NY13S_181 clone) were found in Deepsea (1.5%) and Norway (0.8%), although in low relative abundance. The rest of the MAST-6 sequences were assigned to environmental “MAST-6_X” and “SA2_3F7” from the PR2 dataset and “MAST-6”, a potentially new MAST-6 variant (Fig. 3.4B).

As an additional measure to quantify the confidence of the extracted SSU rRNA sequence placements, sequences with LWR values ≥95% verified with blastn searches are highlighted in red in SSU-EPA tree (Fig. 3.3) and considered to be of highly confident (Berger et al., 2011; Dunthorn et al., 2014). No Deepsea_MAST6 ASVs had LWR values ≥ 95%, with many of them having equally likely alternative placements (blue lines in nodes in Fig. 3.3).

3.3.3 The new Placididea may be rare in sediments

For the Deepsea study, the only ASVs with high LWR values were the ones assigned to Placididea species. Although there was a total of 15 ASVs assigned to Placididea, none were assigned to *H. sinai*. When the SSU sRNA tree was constructed including the 10 Placididean OTU sequences of ESBig, *H. sinai* formed a sister lineage with ESBig133, which were found in water samples with salinities of 78, 124 and 380‰ (Lee et al., 2022) and, *Placididea* sp. (Caron Lab), cultured in 36‰ (Caron, 2000; Keeling et al., 2014) (Fig. 3.3). This clade formed a sister-lineage to "Group-D" containing *Haloplacidia cosmopolita* (described in Park and Simpson, 2010; Rybarski et al., 2021), which can tolerate 15–175‰ salinity. Additionally, ESBig sequences and Deepsea placididean sequences were placed across the major sub-groups of Placididea, despite being isolated from different geographical locations and a broad range of salinities (36‰ for Deepsea and 76–380‰ for ESBig) (Fig. 3.3). The confidence of the extracted SSU rRNA sequences placement within the partial-to-full length SSU sequences was inferred from LWR values ≥95%. Seven out of 15 Deepsea and four out of 10 ESBig placididean ASVs showed high confidence (red nodes in SSU-EPA tree in Fig. 3.3).

3.3.4 Morphological description and new name designation

3.3.4.1 An undescribed Pseudophyllomitus sp. BSC2

The cell is a biflagellated, naked, and free-living single-celled protist. The outline of the cell is oblong and slightly concave at the middle, measuring 22 µm in length and 7 µm in width (Fig. 3.5A-E). Both flagella emerge subapically from a gullet which continues for two-thirds (approximately 5 µm) of the cell width. The anterior flagellum is ~1x cell length and directed forward. The posterior flagellum is 0.5x cell length and inserts to the left of the anterior flagellum. When the cell was stationary, the anterior flagellum beats rapidly in a sinusoidal wave, often sweeping to the right. The posterior flagellum is anchored sideways, likely attached to the surface, and occasionally trailing behind when changing direction. The two flagella are clearly visible and do not adhere to each other, the morphological trait that separates *Pseudophyllomitus* from *Phyllomitus* species (Lee, 2002). Some refractile granules are visible at the cell surface. Although no feeding was observed at the time of sampling, the cell is likely be a phagotroph. The shape of the cell is comparable to *P. salinus* (Lackey, 1940) in its oblong shape however, it is distinguishable by the longer anterior flagellum and the shorter posterior flagellum, and the presence of refractile granules on the cell surface (Lee, 2002). The cell is also similar to *P. granulatus* (Larsen and Patterson, 1990; Lee and Patterson, 2002) in terms of length and movement of both flagella and presence of the vesicles on the cell surface. However, its oblong shape is distinguished from sac-shaped *P. granulatus*.

3.3.4.2 New genera and species designation

***Mastreximonas* gen. nov. Lax, Cho, and Keeling**

Taxonomy: Eukaryota; SAR Burki et al. 2008, emend. Sar Adl et al. 2012; Stramenopiles Patterson 1989, emend. Adl et al. 2005; Bigyra Cavalier-Smith 1998 emend. 2006; Sagenista Cavalier-Smith 1995; Eogyrea Cavalier-Smith 2013.

Diagnosis: Flagellated, naked, and single-celled protist. Cell outline is elongated sac-shape with a slightly flattened anterior end. Thick anterior flagellum emerging apically, posterior flagellum may be very short and trailing under the cell, or absent.

Etymology: Acronym for *ma*rine *stra*menopile, *éxi* (Greek έξι, number 6), and *monas* (Greek, fem.), commonly used for unicellular organisms.

Zoobank Registration. LSID for this publication: urn:lsid:zoobank.org:pub:583E6EDF-B1A2-4220-96D7-C4CF47DA9A6C. LSID for the new genus: urn:lsid:zoobank.org:act:960070EF-0259-4A31-936F-A372FED9B7FE

Type species. *Mastreximonas tlaamin*

Mastreximonas tlaamin sp. nov. Lax, Cho, and Keeling

Diagnosis: The cell measures 15.6 µm in length and 4.8-6.4 µm in width. The prominent anterior flagellum is markedly thicker than the posterior flagellum and roughly two-thirds of the cell length (13 µm), directed forward, and emerges apically from a gullet. The posterior flagellum was not observed. Many large vesicles (approximately 1.5-2.5 µm in diameter) are present in the cytoplasm and two similarly sized golden vacuoles (2.4 µm) are present at the posterior end. The cell swims in a circular motion with the anterior flagellum beating in a sine wave. The nucleus is located just below the base of the anterior flagellum and is 3.5-4.0 µm in diameter. Although no feeding was observed at the time of sampling, the cell is likely a phagotroph.

Type Figure: Fig. 3.5F.

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OQ909084.

Type material: The specimen shown in Fig. 3.5F–J is the holotype. The actual specimen (single cell) was destroyed in the process of single-cell genome sequencing by necessity (see International Code of Zoological Nomenclature, Art. 72.5.6, Declaration 45).

Type locality: Oxic marine intertidal sediment of the Powell River, British Columbia, Canada (49°50'42" N, 124°31'60" W)

Etymology: The species epithet ‘tlaamin’ is derived from the Tla’amin Nation, an indigenous First Nation in Powell River, BC. It means ‘our people’ in Tla’amin language.

Zoobank Registration: LSID for this publication: urn:lsid:zoobank.org:pub:583E6EDF-B1A2-4220-96D7-C4CF47DA9A6C. LSID for the new species:urn:lsid:zoobank.org:act:8B0835A7-679C-441A-AFFC-18D8596201BC

***Vomastramonas* gen. nov. Tikhonenkov, Prokina, Cho, and Keeling**

Taxonomy: Eukaryota; SAR Burki et al. 2008, emend. Sar Adl et al. 2012; Stramenopiles Patterson 1989, emend. Adl et al. 2005; Bigyra Cavalier-Smith 1998 emend. 2006; Sagenista Cavalier-Smith 1995; Eogyrea Cavalier-Smith 2013.

Diagnosis: Biflagellate, naked, and solitary eukaryovorous protist. Cells are slightly flattened and ovoid, with a slightly narrowed posterior end and a notch at the anterior end. Both flagella are acronematic, emerging apically from a notch at the anterior end of the cell.

Etymology: Acronym for *vo*racious, *ma*rine *stra*menopile, and *monas* (Greek, fem.) – commonly used for unicellular organisms.

Zoobank Registration. LSID for this publication: urn:lsid:zoobank.org:pub:583E6EDF-B1A2-4220-96D7-C4CF47DA9A6C. LSID for the new genus:urn:lsid:zoobank.org:act:610C621F-9C18-49E5-983A-6194BE4F97CB

Type species. *Vomastramonas tehuelche*

Vomastramonas tehuelche sp. nov. Tikhonenkov, Prokina, Cho, and Keeling

Diagnosis: cell body is 11.5-13 µm in length and 7.5-10 µm in width. Anterior flagellum is approximately equal to the cell length, posterior flagellum is 1.2-1.5 times longer than the cell. Anterior flagellum is markedly thicker than the posterior flagellum and clearly visible, directed forward and sideways, curved in form of an arc, vibrates very rapidly with a short wavelength but doesn’t change its position during cell movement. Posterior flagellum is barely visible during cell movement, directed backwards. Cells swim close to the substrate in a circle, pushing off with the posterior flagellum, without rotation around its longitudinal axis and without changing the direction of movement. The anterior flagellum is directed towards the outer side of the circle when cell moves. When cell stops, posterior flagellum is directed sideways and curved in arc towards the anterior flagellum, so the flagella seem to stretch towards each other. Cells also can swim relatively straight, with small jerks. Numerous light-refracting granules and digestive vacuoles are present in the posterior half of the cell. No cysts.

Remarks: this species differs from the other member of MAST-6 clade, *Pseudophyllomitus vesiculosus* Shiratori et al., 2017 because the cells are not flexible and lack the rod or bar laid against the anterior side of the nucleus (Shiratori et al., 2017).

Type material: The specimen shown in Fig. 3.5K is the holotype (see International Code of Zoological Nomenclature, Art. 72.5.6, Declaration 45).

Type Figure: Fig. 5K.

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OQ909086.

Type locality: Nearshore bottom sediments of the Strait of Magellan, Punta Arenas, Chile.

Etymology: Tehuelche is the collective name (in Araucanian) of the indigenous peoples of Patagonia.

Zoobank Registration: LSID for this publication: urn:lsid:zoobank.org:pub:583E6EDF-B1A2-4220-96D7-C4CF47DA9A6C. LSID for the new species:urn:lsid:zoobank.org:act:9C52101E-1B3E-45F2-9DB0-A6DFAB1349D1

***Haloplacidia sinai* sp. nov. Tikhonenkov, Cho, and Keeling**

Taxonomy: Eukaryota; SAR Burki et al. 2008, emend. Sar Adl et al. 2012; Stramenopiles Patterson 1989, emend. Adl et al. 2005; Bigyra Cavalier-Smith 1998 emend. 2006; Opalozoa Cavalier Smith 1991 emend. 2006; Placidozoa Cavalier-Smith 2013; Placididea Moriya, Nakayama & Inouye 2002; *Haloplacidia* Rybarski, Nitsche & Arndt 2021.

Diagnosis: Cells are oval, roundish or irregularly ovoid, with the convex dorsal side and the flatter ventral side. Cell body is 5.4-8.3 µm in length and 3.4-6.6 µm in width. Anterior flagellum is approximately 1.5 times longer than the cell, posterior flagellum is approximately equal to the cell length. Posterior flagellum is acronematic and both flagella emerge from a shallow groove at the central part of the ventral side of the cell and oriented in the opposite directions. Anterior flagellum bears mastigonemes. Cells are often attached to the substrate with a posterior flagellum and produce very fast trembling movements. No cysts.

Remarks: This species differs from the other member of the genus, *H. cosmopolita* Rybarski, Nitsche & Arndt 2021, by having a slightly different shape of the cell without pronounced kidney-like morphology, and by the absence of cysts, even under starvation conditions (Rybarski et al., 2021).

Type material: The specimen shown in Fig. 3.5U is the holotype (see International Code of Zoological Nomenclature, Art. 72.5.6, Declaration 45).

Type Figure: Fig. 3.5U.

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number OQ909082.

Type locality: surface of corals in the Red Sea, Sharm El Sheikh, Egypt.

Etymology: named after the place it was found in the Mount Sinai region, where the Ten Commandments were given to Moses by God, according to the Book of Exodus in the Hebrew Bible. The English name Sinai came from Latin, ultimately from Hebrew סִינַי, pronounced /siˈnái/.

Zoobank Registration: LSID for this publication: urn:lsid:zoobank.org:pub:583E6EDF-B1A2-4220-96D7-C4CF47DA9A6C. LSID for the new species:urn:lsid:zoobank.org:act:6EF3B31E-BAFB-4E5C-8010-DF3BBE3DCE43

3.5 Discussion

3.5.1 Updated taxon sampling and phylogeny of MAST-6

MAST-6 has been shown to be both abundant and diverse through various amplicon sequencing studies in sediment samples (Rodríguez-Martínez et al., 2009; Massana et al., 2015; Schoenle et al., 2021) (Table 3.1). Despite the known abundance and distribution across various sediment sites, inferring the diversity of MAST-6 species has been limited to a reference database composed of handful of SSU rRNA sequences. Moreover, only a single taxon for which genomic-level data are available (i.e., *Pseudophyllomitus vesiculosus*) has represented the MAST-6 clade in phylogenomic analyses. In this chapter, together with collaborators, I generated transcriptomes of three new MAST-6 taxa: *Mastreximonas tlaamin*, *Vomastramonas tehuelche*, and *Pseudophyllomitus* sp. BSC2, and updated the deep phylogeny of stramenopiles. These three new MAST-6 species in turn reflect broader genetic diversity by representing different sub-groups of the MAST-6 lineage.

As with previously described *P. vesiculosus*, all new MAST-6 species described here were found in sediments, and have relatively large and numerous vesicles or granules underlying the cell surface. The new *Pseudophyllomitus* sp. BSC2 was the most closely related to previously described *P. vesiculosus* and one of the longest *Pseudophyllomitus* species described so far (22 µm) (Lee and Patterson, 2002). The overall morphological characteristics are most similar to *P. granulatus* and, in a lesser extent to *P. salinus*. However, due to not observing feeding behaviour, I refrained from establishing a new species for this cell. *Mastreximonas tlaamin* had a similar oblong shape to *Pseudophyllomitus* sp. BSC2 and is sister to the two *Pseudophyllomitus* species. *Vomastramonas tehuelche*, on the other hand, has a more circular shape and is a sister lineage to the rest of the MAST-6 species in the phylogenomic tree.

3.5.2 The new MAST-6 species broaden the genetic diversity

This chapter showed that *M. tlaamin*-related ASVs (sub-group I) are the most abundant MAST-6 across different sediment amplicon studies (Fig. 3.4), representing a largest MAST-6 sub-group consisting of ASVs from various sediment locations and depths (Fig. 3.3). Amplicon sequence variants assigned to *M. tlaamin* were absent in other studies (e.g., BioMarKs and SouthChina). This can be due to pyrosequencing, which is prone to non-homopolymer errors and has less sequencing coverage (Luo et al., 2012), or low abundance of *M. tlaamin* at the time of sampling. None of the new MAST-6 species from this study is found within sub-group IV despite its high relative abundance in the ISME2020 (Appendix M). Future efforts in isolating and describing cells of the subgroup IV may not only confirm phylogenetic diversity but help us better understand biology behind the sediment associated MAST-6 species. Based on the absence of *Pseudophyllomitus* sp. BSC2-and *V. tehuelche*-assigned ASVs, and low relative abundance of ASVs assigned to *P. vesiculosus* in sediment studies, these MAST-6 species may be rare. Additionally, sample timing may have played a role in lack of detection of some MAST-6 species, as the cell abundance has been reported to be affected by seasonality and salinity (Piwosz and Pernthaler, 2010). For example, both small and large morphotypes of MAST-6 are observed to have short-lived peaks at mid-May to early-June in the Gulf of Gdansk shortly after freshwater inflow, followed by a substantial decline in relative abundance (Piwosz and Pernthaler, 2010). All datasets except BioMarKs were sampled mostly in September, while some sampled in August and July (sampling months for BioMarKs from February to October). These months were the time when the number of sub-group II associated MAST-6 were reported to be very low (Piwosz and Pernthaler, 2010). Although the work by Piwosz and Pernthaler, 2010 was done on plankton samples, the rapid and short-lived seasonal fluctuation of MAST-6 abundance revealed that this group may respond quickly to changing environment, including the ones in sediments.

3.5.3 Rare and potentially halotolerant *Haploplacidia sinai* and its implication in trait evolution

Haloplacidia sinai is the fourth new species reported here. *Haloplacidia sinai* belongs to Placididea, another major clade of Bigyra that was represented by two transcriptomes before this chapter. As with some of the previously described species of Placididea (Park and Simpson, 2010), *H. sinai* was found in a relatively high salinity environment. Although I did not detect any ASVs assigned to *H. sinai*, based on its relationship (Fig. 3.3) with other isolated cells cultured in broad range of salinity, *H. sinai* might also be found in non-hypersaline environments. Absence of ASVs assigned to *H. sinai* may be due to the choices of sampling habitats in the datasets examined, as collaborators isolated the cell from coral scrapes. Three "Deepsea_Placididea" sequences formed a clade with the two *Suigetsumonas* spp. isolated from brackish lakes in Japan and Kenya (Okamura and Kondo, 2015; Rybarski et al., 2021) (Fig. 3.3), further demonstrating the broad range of salinity in which species of Placdidiea can be found.

The halophilic trait is not just limited to Placididea but can also be found in Bikosia. The extremely halophilic *Halocafeteria seosinensis* (Park et al., 2006; Park and Simpson, 2010) can survive between 75 to 363 ‰ (Lee and Patterson, 2002; Park et al., 2006). Furthermore, several traits including differential gene expressions involved in anti-oxidization, membrane fluidity, O-linked glycosylation, and gene-duplication were linked to high salt adaptability of *H. seiosinesis* (Harding et al., 2017). Exploring the evolution of halotolerancy in these deep-branching stramenopiles may lead to a better understanding of the ancestral state of the stramenopiles, determining whether the trait evolved separately in Placididea and Bikosia or arose in the last common ancestor of the two groups involving transition between different salinity barrier (Dunthorn et al., 2014; Jamy et al., 2022).

3.5.4 Phylogenomics of stramenopiles with a twist

In this chapter, *H. seosinensis* is a sister lineage to *Caecitellus* sp., and this atypical mastigoneme-lacking group (O'Kelly and Nerad, 1998; Park et al., 2006) in turn formed a robust sister lineage to the clade composed of *S. scintillans* and *Cafeteria burkhardae* (Fig. 3.1; Appendix G and I). When I added the most recent genomic data of MAST-1, MAST-7, MAST-8, MAST-9, and MAST-11, the relationship remained the same (Appendix H). The bikosian phylogenomic relationship in this study (Fig. 3.1) is consistent with previous SSU phylogenetic trees (Cavalier-Smith and Chao, 2006; Cavalier-Smith and Scoble, 2013; Guillou et al., 1999; Park et al., 2006; Shiratori et al., 2017, 2015). However, an alternative SSU rRNA phylogeny showed *H. seosinensis* forming a sister lineage to a clade composed of *Cafeteria* spp. and *Caecitellus* spp. (Yubuki et al., 2015), similar to the SSU-tree generated in this chapter (Appendix N). This could be due to the fast-evolving nature of many bikosian SSU rRNA genes, as indicated by the long branch length of *S. scintillans* and *C. burkhardae* (Appendix N).

Additionally, the topology of Bikosia in the phylogenomic tree may be prone to future change as there are far more bikosia that are not represented in transcriptomic or genomic datasets, such as diverse *Bicosoeca* spp. (Karpov et al., 1998), *Pseudobodo* spp. (Griessmann, 1913), freshwater or soil bikosians, including *Siluania monomastiga* (Karpov et al., 1998), *Nerada mexicana* (Cavalier-Smith and Chao, 2006), *Adriamonas peritocrescens* (Verhagen et al., 1994), and *Paramonas globosa* (Cavalier-Smith and Chao, 2006; Saville-Kent, 1880) (Appendix N).

The paraphyly of Bigyra has been repeatedly demonstrated in recent publications as more genomic data across different lineages of stramenopiles have become available (Burki et al., 2016; Noguchi et al., 2016; Azuma et al., 2022; Cho et al., 2022), including the ML-PMSF tree in my study (Fig. 3.1). However, the Bigyra are monophyletic in some other studies (Derelle et al., 2016; Thakur et al., 2019), as well as the two consensus trees obtained from MCMC chains in this study (Appendix K). As these studies all have differing numbers of taxa (as well as different taxa) and orthologs, and use different methods for data processing, it is difficult what might be causing topological incongruencies across these analyses.

In contrast to previously published work (Azuma et al., 2022), the placement of *A. sol* is not sister to the rest of the Ochrophyta. Rather, it forms a weakly-supported clade with *Microchloropsis gadidata* (Eustigmatophyceae). As a single transcriptome represents each of Eustigmatophyceae and Actinophrydae, and I argue that this is the result of long branch attraction artefacts (LBA) caused by eroded phylogenetic signals (class II LBA), rather than parallel substitutions (class III) or saturation (Fig. 3.1) (Wägele and Mayer, 2007). I infer the probable existence of LBA in trees reconstructed from fast-evolving-site removal (Fig. 3.2B), Bayesian analysis (Appendix K) and, 39per-, 59per-, and recoded matrix (Appendix G and J), where the placement is chaotic rather than showing a pattern. The Ochrophyta phylogeny was further complicated by other unstable relationships of Eustigmatophyceae, Pinguiophyceae, and among CCS, RPX, and BBPe. Phylogenomic discrepancies found in the Ochrophyta nuclear dataset should be addressed by more taxon sampling to break the long branches (e.g., Marine OCHrophytes (MOCH) (Massana et al., 2014) and Olisthodiscophyceae (Barcytė et al., 2021) and developing new phylogenomic models that can resolve short internal branches within early ochrophyte divergence (Philippe et al., 2011; Ševčíková et al., 2015; Di Franco et al., 2022). A similar discrepancy between maximum likelihood (ML) and Bayesian analyses was also observed in Cho et al. (2022) where all four chains yielded different topologies compared to the one from ML analysis. In my consensus trees also differed from the ML analysis in recovering Bigyromonadea as the sister lineage to Ochrophyta, as was observed in previous study (Cho et al., 2022). However, constrained AU tests (Shimodaira, 2002; Nguyen et al., 2015) failed to reject the monophyly of Bigyromonadea, together with Oomycetes (Winter 1897) and Hyphochytriomycetes (Dick 1983), forming a sister lineage to Ochrophyta in all four consensus trees (Cho et al., 2022).

Within the monophyletic Placidozoa (Placididea+Nanomonadea+*Blastocystis*), the relationship among the sub-groups is not strongly supported, in contrast to three other studies (Azuma et al., 2022; Cho et al., 2022; Thakur et al., 2019) and my Bayesian analysis (Appendix K). Based on the amino acid composition of the placidozoan data used in this study, the topology appears to result from LBA due to enriched GARP aa in this group (Fig. 3.1; -Appendix I). However, repeating the ML-PMSF analysis without *H. sinai* (data not shown) recovered the same Placidozoa topology as previous studies (Azuma et al., 2022; Cho et al., 2022; Thakur et al., 2019) with weak support (76%). Despite the suspected LBA due to amino acid composition bias, the recoded tree analysis did not change the topology of Placidozoa, although the bootstrap support was weak. The present placidozoan topology is likely unstable in my dataset due to a combination of long branches leading to Placididea and Opalinata, and low taxon sampling in Bigyra. As shown by the Opalinata + MAST-12 clade and diverse placidideans shown in the SSU-tree (Appendix N), future efforts in increasing taxon sampling will likely help stabilize the placidozoan topology, in addition to deploying a phylogenetic model that can resolve LBA amongst stramenopiles.

3.6 Conclusion

The first impression of phagotrophic Bigyra to most observers may be a jumble of heterotrophic flagellates with few distinguishing features. It was only through SSU rRNA-amplicon sequencing that their identities and phylogenetic diversities were revealed. Even then the reference-dependent taxonomy assignment and usage of a single SSU primer-set often led to an under-detection of their diversity. Placididea on the other hand were initially discovered through cell isolates, but an assessment of their environmental distribution was limited due to its preferential amplification with a V9-targeting primer set (Lee et al., 2021; Rybarski et al., 2021). Despite the group's diversity and ability to survive in a broad range of salinity, only very limited transcriptome or genome data had been available prior to this study. After adding another transcriptome of a placididean (*H. sinai*), we observed a topology change in Placidozoa that conflicts with previous studies. Based on the long unbroken branch leading to Placididea and alternative tree construction methods, the topology from the current study may be an artefactual relationship caused by long branch attraction (Felsenstein, 1978; Hendy and Penny, 1989; Delsuc et al., 2005b; Philippe et al., 2005). Combined with a lack of taxon sampling, the presence of highly divergent species, such as symbiotic Opalinata, *Incisomonas marina,* and their long-branching sister lineage, Bikosia, it is likely that currently available models cannot resolve the true relationship of Placidozoa. Although the phylogenomics of Ochrophyta are beyond the scope of the present study, I note that it remained unresolved with conflicting ML and Bayesian analyses in this and previous studies (Azuma et al., 2022; Cho et al., 2022), which suggests more data will be required. Adding three new MAST-6 transcriptomes to my phylogenomic tree resulted in robust monophyly of MAST-6 and MAST-4, a relationship only recently revealed in phylogenetic studies (Shiratori et al., 2017; Thakur et al., 2019; Cho et al., 2022). Along with the new MAST-6 species, I also showed phylogenomic relationship among Sagenista with recently published genomic data of MAST-7, -8, -9, and -11 for the first time. Newly described MAST-6 species improved the detection of considerable phylogenetic diversity of sediment-associated MAST-6 species from various sample sites, and demonstrated a higher diversity compared to that of the most abundant MAST-4 group (Logares et al., 2012; Rodríguez-Martínez et al., 2012). One of the abundantly detected MAST-6 is closely related to the newly described *M. tlaamin* (PRC5), while few or no ASVs were detected for *V. tehuelche* and *Pseudophyllomitus* sp. BSC2. This indicates different MAST-6 species may be rare and have different seasonal dynamics.

Table 3.1 List of selected amplicon sequencing datasets from the European Nucleotide Archive (ENA).

Dataset designation	Sample environment	Sequencing technology	18S rRNA region	Length (bp)	Number of ASVs		Sample number	BioProject
					Placididea	MAST-6		
BioMarKs*	Seafloor sediment	454 GS FLX Titanium	V4	380-384	0	12	24* (Run accessions: ERR861806-ERR861811, ERR861839, ERR861843, ERR861849, ERR861853, ERR861860, ERR861870, ERR861884, ERR861885, ERR861894, ERR861895, ERR861900, ERR861901, ERR861905, ERR861910, ERR861911, ERR861915- ERR861917)	PRJEB9133 (Dunthorn et al., 2014; Massana et al., 2015)
SouthChina	Seafloor sediment	454 GS FLX Titanium	V1-V2	396-429	0	9†	6	PRJNA341446 (Wu and Huang, 2019)
Norway	Marine and brackish sediment	Illumina MiSeq paired-end	V4	426-429	0	16	24	PRJEB24876, PRJEB24158, PRJEB24888
Deepsea	Abyssal seafloor sediment	Illumina Genome Analyzer II paired-end	V9	134-138	15	6†	20	PRJNA635512 (Schoenle et al., 2021)
ISME2020	Seafloor sediment	Illumina MiSeq paired-end	V4	182-425	0	61	49	PRJNA521526 (Rodríguez-Martínez et al., 2020)
ESBig**	Solar saltern	Illumina MiSeq paired-end	V9**	128-154	10**	0	**Accession number: MZ297173, MZ297191, MZ299824, MZ299825, MZ299969, MZ300048, MZ300314, MZ300350, MZ300439, MZ300768	**PRJNA732544 (Lee et al., 2021)

Accession numbers are included only if selected samples of a given BioProject were analyzed. For example, out of 139 samples for the BioMarks dataset, only sediment samples (24) were processed to access the diversity of the newly identified MAST-6 species (*). Extracted length indicates the length of the amplicon sequence variants (ASVs) assigned to MAST-6 or Placididea lineages. The ESBig dataset was not processed in this study, but the sequences assigned to placidideans were directly obtained from the authors of the BioProject (**). All SouthChina ASVs and one Deepsea_MAST6 ASV were excluded from the main figures based on LWR-values and manual blastn searches (†).

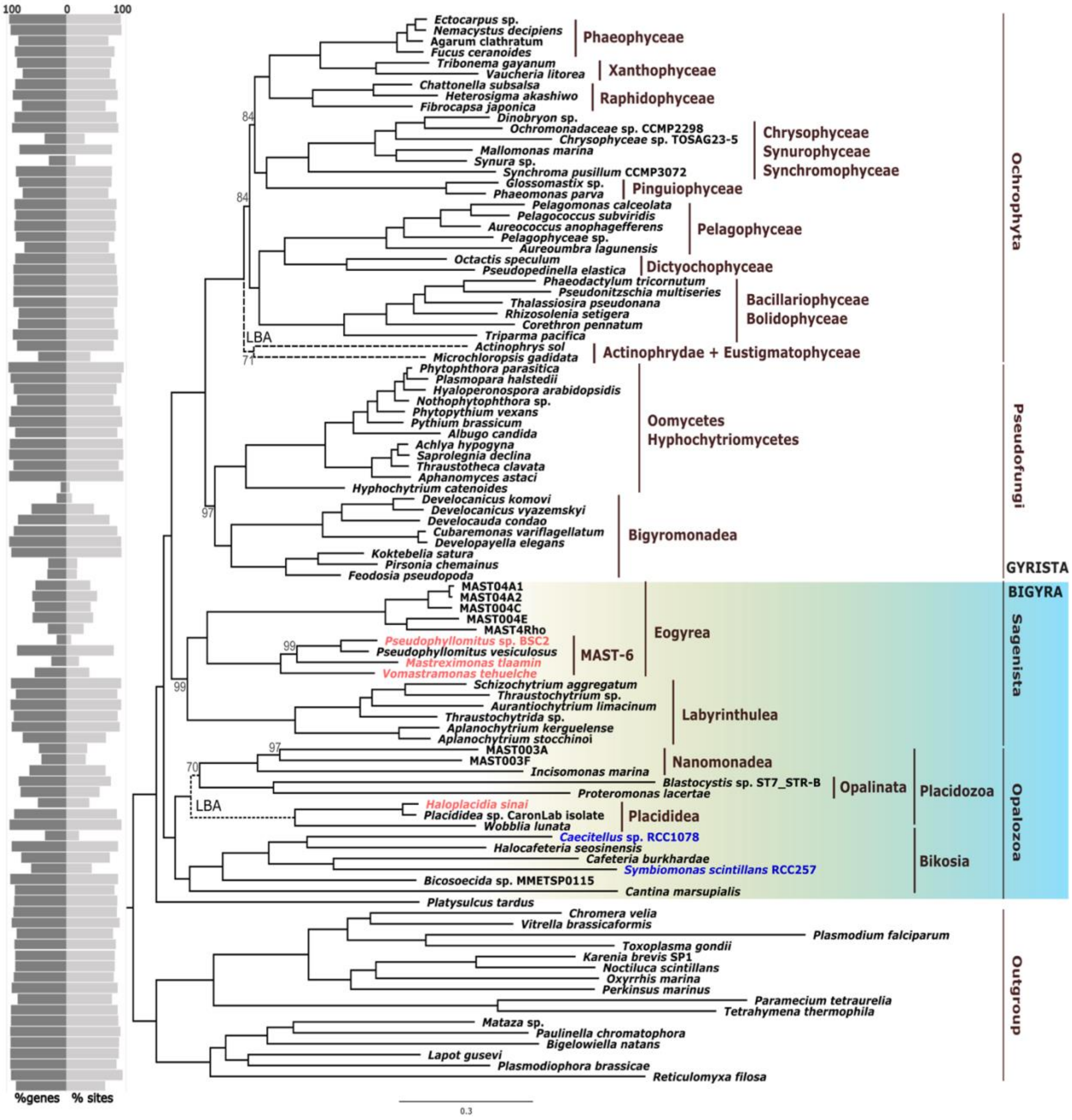


Figure 3.1 Phylogenomic tree of stramenopiles

Maximum-likelihood (ML) multi-gene tree of stramenopiles, including six new transcriptomes; four from newly described Bigyra in this study (light red), and two from culture collections (blue). The tree was constructed from concatenated alignments of 240 genes from 98 taxa (76,516 aa sites) under the site-heterogeneous model LG+C60+F+G4+PMSF with 100 standard bootstraps. Only nodes with ≤99% support values are labelled, with unlabelled nodes indicating 100% bootstrap support. Dashed branches indicate potential long branch attraction artefacts (LBA). The % genes (dark grey) and sites occupied (light grey) for each taxon are shown on the mirrored bar plot on the left.

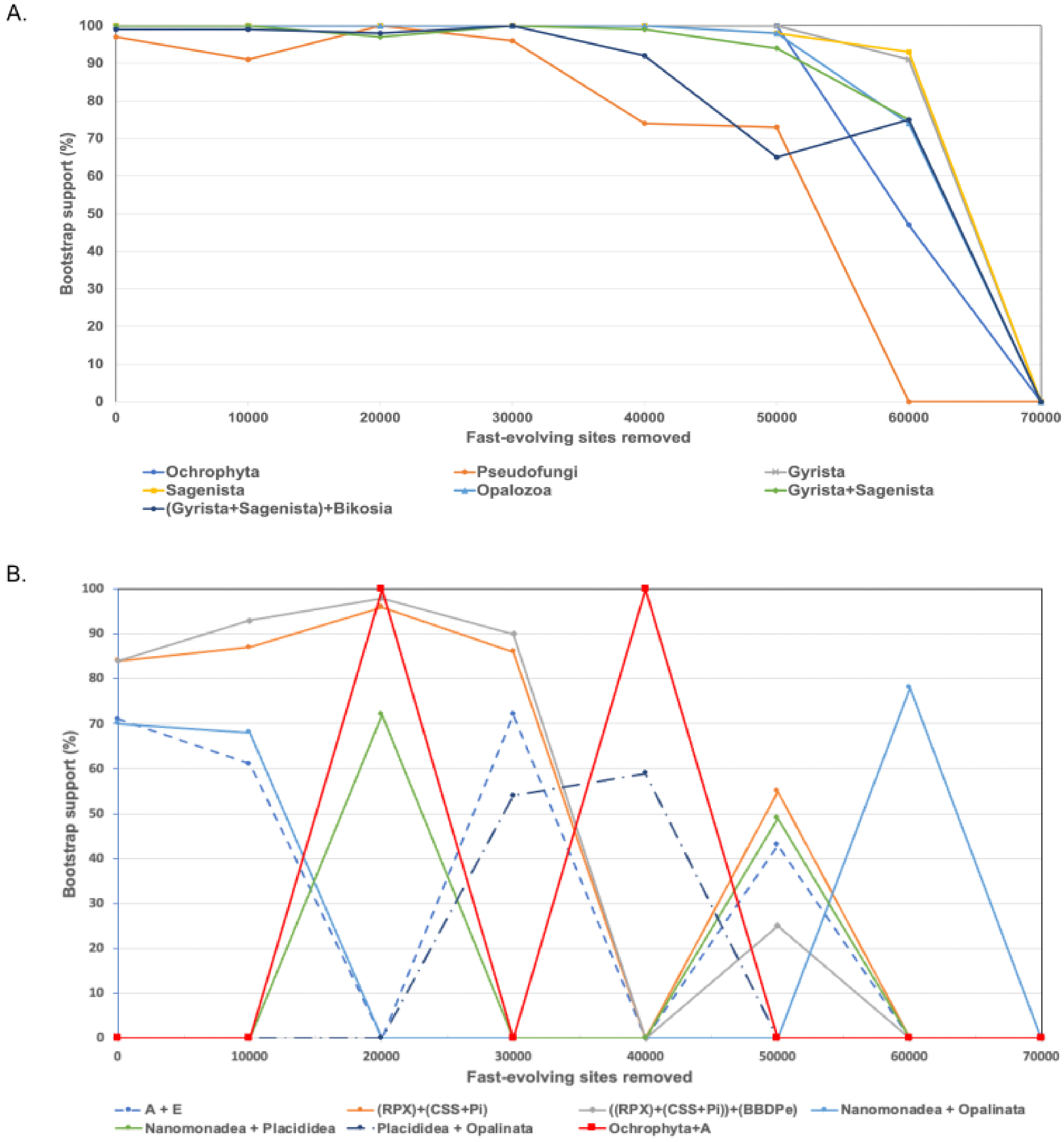


Figure 3.2 Removal of fast-evolving sites

Change in bootstrap support with the incremental removal of fast-evolving sites (10,000 sites removed at each step) for the monophyly of major stramenopile groups (A) and minor unstable groups (B). **A**. Monophyly of major stramenopile groups show strong bootstrap support up to 30,000 sites removed. Paraphyly of Bigyra, represented by "(Gyrista+Sagenista)+Bikosia" and "Gyrista+Sagenista". **B.** Monophyly of unstable groups showing fluctuation in bootstrap support. Bootstrap supports with zero values indicate alternative topology (not shown here) with weak support (22-55%). Topologies within Opalozoa (Nanomonadea, Placididea and Oplinata) were unstable and weakly supported. Topologies within the Ochrophyta were also largely unstable, especially for Eustigmatophyceae ("E") and *Actinophrys sol* ("A"). "RPX" =

Raphidophyceae+Phaeophyceae+Xanthophyceae; “CSS+Pi” = Chyrsophyceae+Synurophyceae+Synchromophyceae; “BBDPe” = Bacillariophyceae+Bolidophyceae+Dictyophyceae+Pelagophyceae.

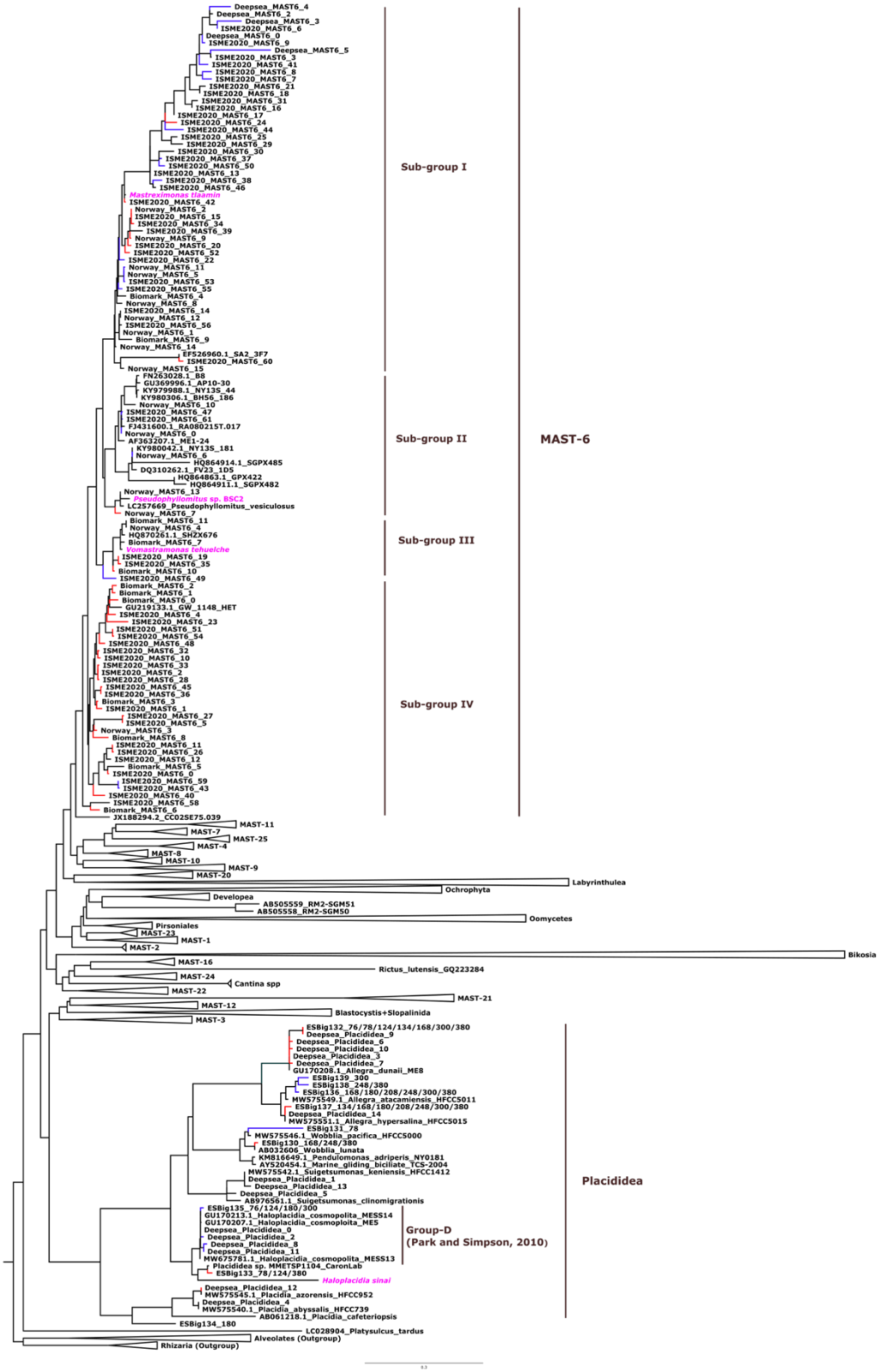


Figure 3.3 SSU-EPA tree of stramenopiles

A RAxML SSU rRNA phylogenetic tree (SSU-EPA tree) of stramenopiles. The tree was constructed under the GTR+GAMMA model with 1000 rapid bootstrap replicates, using an alignment of 527 stramenopile sequences and seven outgroup sequences (8,771 sites): 109 extracted ASVs assigned to MAST-6 or Placididea from the amplicon dataset, and 10 placididean OTU sequences from ESBig study. The four new Bigyra species are coloured in pink. The likelihood weight ratio (LWR) values calculated from our EPA analysis are coloured in red for high confidence (LWR ≥95%), and in blue for low confidence (LWR <95%), indicating equally likelihood of alternative placements. The label structure for the ASVs is "Dataset_MAST6/Placididea_count". Clades other than MAST-6 and Placididea are collapsed. For bootstrap supports, see Appendix M.

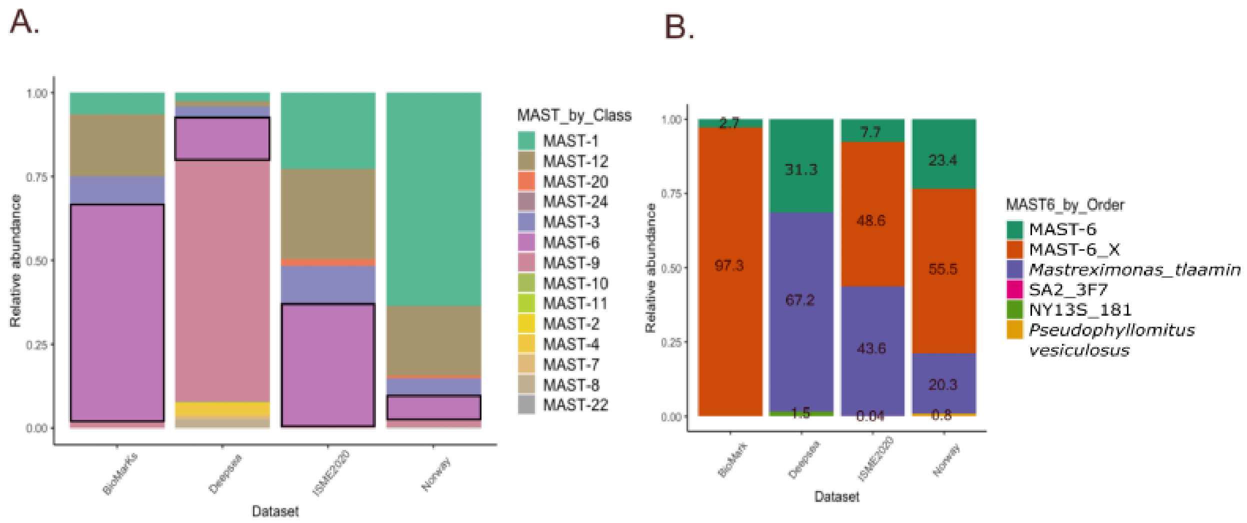


Figure 3.4 Relative abundance bar plots of MASTs

Stacked bar plots of the relative abundance of unique ASVs assigned to main MAST groups (**A**) and MAST-6 (**B**) from four sediment datasets: BioMarKs, Deepsea, ISME2020, and Norway. Deepsea is the only study with a SSU rRNA gene primer targeting the V9 region. **A.** Composition of different MASTs from each dataset grouped by class level. Black frames indicate the relative abundance of MAST-6. **B.** Composition of MAST-6 lineages from each dataset grouped by order to further show higher taxonomic assignment. "MAST-6_X" represents an unknown MAST-6 lineages classified from the PR2 database, and "MAST-6" represents a potentially new MAST-6 lineage based on the updated taxonomic training database. "*Mastreximonas tlaamin*" is one of the new MAST-6 species descried in the current study. "*Pseudophyllomitus vesiculosus*" and "NY13S_181" are previously reported cultures and "SA2_3F7" is an environmental sequence.

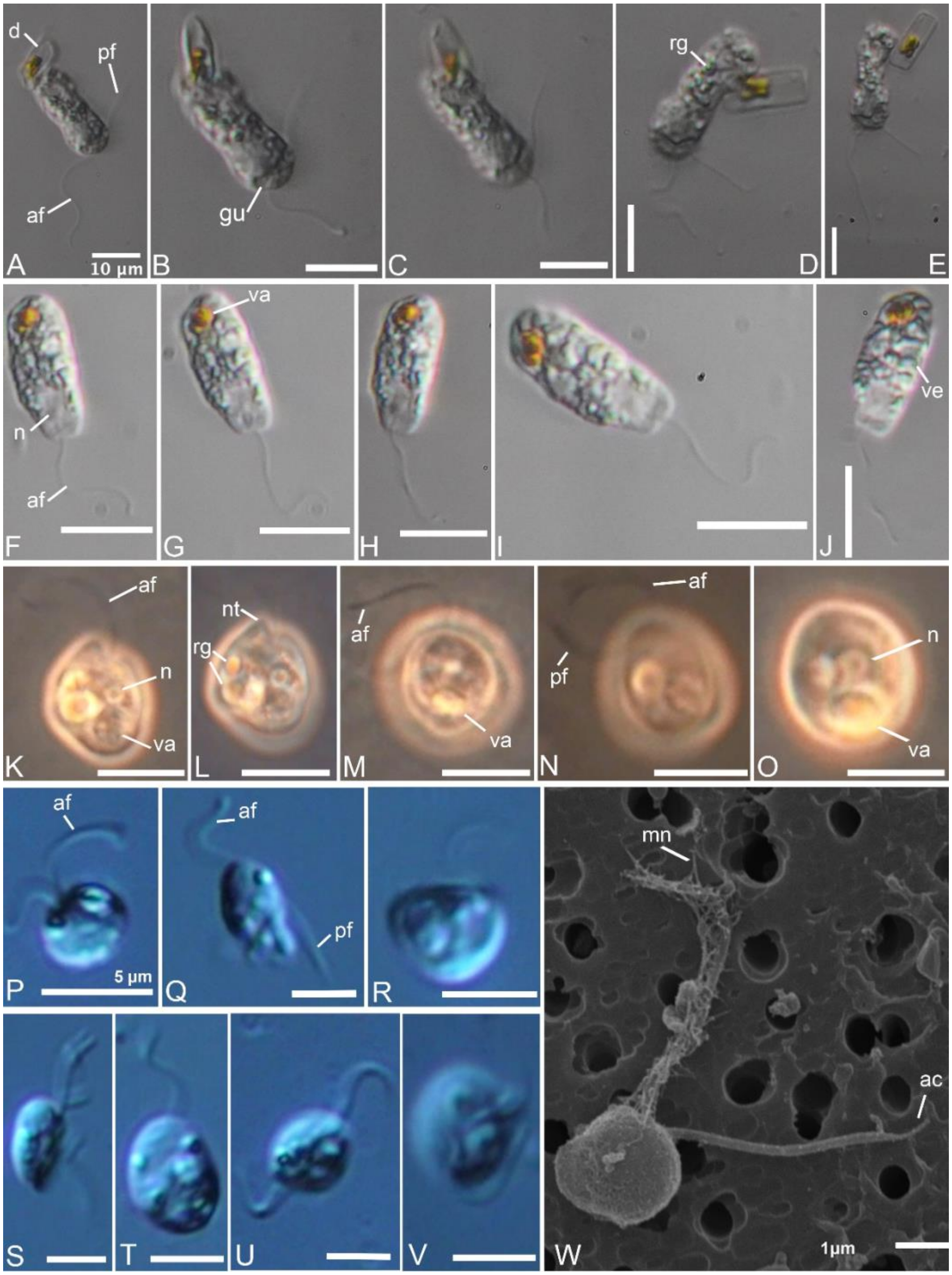


Figure 3.5 Morphology of four new Bigyra

A-E. *Pseudophyllomitus* sp. BSC2. General view of the cell including anterior flagellum [af] and posterior flagellum [pf]. Both flagella emerge from a horizontal gullet [gu] and some refractile granules are visible on the cell surface [rg]. A diatom [d] is attached at the posterior end. **F-J**. *Mastreximonas tlaamin*, general view of the cell with an anterior flagellum. The nucleus [n] is visible just below the base of the anterior flagellum. Some vesicles [ve] and golden vacuoles [va] are present from the mid to posterior end of the cell. **K-O**. *Vomastramonas tehuelche*. General view of the cell with clearly visible anterior flagellum. A notch [nt] is present at the anterior end. Refractile granules and food vacuoles are present. **P-V.** *Haloplacidia sinai*, general view of the cell with two flagella. **W**. *H. sinai* in scanning electron micrograph, showing mastigonemes [mn] on the anterior flagellum and acroneme [ac] on the posterior flagellum. **Scale bars** are 10 µm for A-O, 5 µm for P-V, and 1 µm for W.

Chapter 4: Phylogenomic analyses of ochrophytes (stramenopiles) with an emphasis on neglected lineages

4.1 Introduction

Ochrophyta is a group of protists that are often used as an example of the vast molecular and morphological diversity of stramenopiles. Ochrophytes 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 relationship. Yet, despite this attention, phylogenomic analyses of ochrophytes remain incongruent with one another (Burki et al., 2016; Derelle et al., 2016; Noguchi et al., 2016; Thakur et al., 2019; Di Franco et al., 2022; Cho et al., 2022; Azuma et al., 2022), especially between the trees reconstructed from nuclear and plastid genes (Ševčíková et al., 2015; Barcytė et al., 2021; Dorrell et al., 2021; Di Franco et al., 2022). 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 under-represented and has been somewhat static (Driskell et al., 2004; Burki et al., 2016; Derelle et al., 2016; Noguchi et al., 2016; Thakur et al., 2019; Azuma et al., 2022; Cho et al., 2022;) (for an exception, see Terpis, 2021).

Current ochrophyte phylogenomic analyses all differ in dataset composition and size, processing approaches, and phylogenetic inference methods. Although there is some consensus around the backbones of the ochrophyte phylogeny (Derelle et al., 2016; Azuma et al., 2022; Cho et al., 2022), 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 that can carry limited phylogenetic signals (Whitfield and Lockhart, 2007; Di Franco et al., 2022; Pardo-De La Hoz et al., 2023). 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 selection (Stiller et al., 2003; Edwards, 2009; Dorrell et al., 2019), incomplete lineage sorting (ILS), introgression via hybridization, and horizontal gene transfers (Maddison, 1997; Nichols, 2001; Dorrell et al., 2021; Dong et al., 2022).

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 (Bapteste et al., 2007; Pick et al., 2010; Superson and Battistuzzi, 2022). 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 radiations (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 signals than nuclear genes with comparable numbers of sites. However, plastid datasets are not suitable for inferring evolutionary history of stramenopiles as a whole, as many stramenopiles lack plastid or its associated genes.

In this chapter, I aimed 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, I added ten new transcriptomes from ochrophytes some of which had not been 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; Riisberg et al., 2009; Yang et al., 2012; Graf et al., 2020) 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). I 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, I explored different gene filtering criteria. I used a previously established method (Mongiardino Koch, 2021; Mongiardino Koch and Thompson, 2021), which calculates phylogenetic signal, noise, and data quality. Overall, I 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, I observed that using genes with high phylogenetic signal and quality resulted in the most stable tree topologies, as opposed to selecting genes with low phylogenetic noise or removing the ones with high noise.

4.2 Materials and Methods

4.2.1 Ochrophyte sample collection and processing

Nine cultures of under-represented ochrophytes were obtained from various culture collections (Table 4.1). Except for *Actinosphaerium* sp. (which was processed immediately and the culture not maintained), I sub-cultured all cultures every two weeks in 30 mL and kept at 20°C with a 12 hour:12 hour light:dark cycle. Both *Olisthodiscus luteus* and *O. tomasii* were kept in TL30 media; *Schizocladia ischiensis* was maintained in L1-Si (Guillard and Ryther, 1962; Guillard, 1975); *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.

I extracted RNA with TRIzol^{TM} 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^{TM} 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^{TM} 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^{TM} (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, I 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^{TM} 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., I 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, my collaborator 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^{TM} LS and CTAB-based methods were assessed using a NanoDrop 1000 Spectrophotometer v3.8.1 (Thermo Fisher Scientific) and Qubit^{TM} RNA High Sensitivity Assay Kits (Thermo Fisher Scientific).

4.2.2 cDNA synthesis, library preparation and sequencing

I 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^{TM} dsDNA HS Assay Kits (Thermo Fisher Scientific). To confirm taxonomic identities, I performed small subunit ribosomal DNA (SSU rDNA) polymerase chain reaction (PCR) on each cDNA sample (except *V. globosus*), using 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.

4.2.3 Transcriptome processing and phylogenomic matrix construction

Along with the ten newly generated transcriptomes, I 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) geomic level data were obtained from previous publications (Azuma et al., 2022; Cho et al., 2024, 2022; 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^{TM} 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), I used BlobTools v2.3.3 (Laetsch and Blaxter, 2017; Challis et al., 2020). 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’.

4.2.4 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, I 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 (UFB), using IQ-TREE v1.6.12 (Nguyen et al., 2015). To ensure correct orthologs were identified for each gene from each transcriptome, I manually screened 241 single-gene trees using ParaSorter v1.0.4. To generate a concatenated supermatrix, I 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').

4.2.4.1 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'). I also randomly removed genes in 20% increments to compare with trees recovered from different gene filtering criteria ('randGene-supermatrix'). The average BS values of phylogenomic trees from each of randSite- and randGene-supermatrices were calculated and used to determine minimum data size (i.e. amino acid sites) required to reduce the effect of small data size and distinguish from the effect of different gene-filtering criteria (see below). With the condition of recovering paraphyletic Bigyra and well-recognized relationship of ochrophyte lineages (e.g., Chrysista or Diatomista), I decided the cut-off BS values to be >89%. Based on the cut-off, I determined that approximately 22,000 sites are the minimum sites required.

4,2.4.2 Conceptual design for phylogenomic gene filtering

To identify phylogenetically informative genes and investigate incongruence among different phylogenomic analyses, I 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 finding the most informative set of genes. Furthermore, I 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.

4.2.4.3 Filter by phylogenetic biases, signals, and other data qualities

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) (Struck, 2014; Mongiardino Koch and Thompson, 2021), variance of root-to-tip distances (root_tip_var, a proxy for inferring deviation from clock-like evolution) (Smith et al., 2018; Mongiardino Koch and Thompson, 2021), saturation (Nosenko et al., 2013; Kocot et al., 2016), and relative composition frequency variability (RCFV, a proxy for amino acid compositional heterogeneity) (Zhong et al., 2011; Whelan et al., 2015; Shen et al., 2016b), 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), I 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).

I 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, I 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, calculate correlation coefficients and significance I generated Pearson correlation graphs using R-packages ‘corrr’, ‘ggcorrplot’, ‘GGally’, ‘ggfortify’, and ‘FactoMineR’ R-packages. Based on the correlation analyses, PC loadings of each properties, I subsampled genes by 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, I extracted shared genes that are found in the properties of a given criterion by searching the top 40 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 33 genes were present in the top 40 lowest values for each properties in criterion C (criterion C40). I also combined subsampled 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, I also subsampled genes that are not well represented by any of the two PCA axes (i.e., genes with low cos2 values) (criterion N). A size of different supermatrices generated from each criterion is summarized in Table 4.2. For each of the gene sets that were filtered by different criterion or a combination thereof, I generated supermatrices as described in 4.2.4.

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

For all the supermatrices generated above, I 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, I verified that the F-class values are < 0.11 (Baños et al., 2023), and repeated the tree reconstruction under the LG+C60+G4 model.

For the '231-supermatrix', I 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, I checked for convergence using bpcomp, and estimated the consensus posterior probability and topology by subsampling every second tree. Finally, I 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 (Wang et al., 2018; Szantho et al., 2023). 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) 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.

4.3. Results and Discussion

4.3.1 The phylogenomic tree of stramenopiles

4.3.1.1 Updating ochrophytes dataset with under-represented classes

I 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 4.1). 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. 4.1). When I 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 change the topology of the rest of the stramenopiles, except the placement of Nanomonadea and Placididea (Fig. 4.1). The phylogenomic trees inferred from these two supermatrices are summarized in Figure 4.1.

In both trees of '231-supermatrix', C60-PMSF and CAT-PMSF (Figs. 4.1 and 4.2), 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, I recovered Chrysophyceae + Synurophyceae + Synchromophyceae (CSS) + Picophagea (Pico) as monophyletic in all trees examined, as previously reported in (Guillou et al., 1999; Barcytė et al., 2021) (Fig. 4.1; Table 4.3). This relationship was also observed in the only other phylogenomic analysis with a comprehensive ochrophyte dataset (Terpis, 2021). Schizocladiophyceae is sister to Phaeophyceae, while Phaeothamniophyceae is a sister-lineage to

Phaeophyceae-Xanthophyceae-Schizocladiophyceae (Fig. 4.1). This placement of Schizocladiophyceae is found in previous studies (Yang et al., 2012; Graf et al., 2020; Barcytė et al., 2021). However, the placement of Phaeothamniophyceae showed more inconsistency within Raphidophyceae-Phaeophyceae-Xanthophyceae (RPX) clades. As I found here, Phaeothamniophyceae falls sister to PX-Schi zocladiophyceae 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).

My 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 multi-gene phylogenetic analyses, along with Raphidophyceae (Yang 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, 2021). The absence of these classes therefore, account for the low BS values for PX in the phylogenomic analyses (53% BS in '231-supermatrix' C60-PMSF; 95% in CAT-PMSF) (Fig. 4.1).

The two Actinophrydae taxa are sister to CSS+Pico, although with a modest BS support of 83% (Fig. 4.1). 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 my dataset. The newly generated transcriptome of *Vicicitus globosus* was nested within the Dictyochophyceae with 100% BS support.

The *Vicicitus globosus* is known to produce fast-acting cytotoxin (Chang, 2015) and its transcriptome was included in the analyses due to its availability at the time.

4.3.1.2 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, 2021). 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, I added newly generated and publicly available transcriptomes belonging to different Eustigmatophyceae sub-groups and Pinguiophyceae.

I recovered a robust monophyly of RPX and Eustigmatophyceae (RPX+Eustig) in a majority of the trees (Figs. 4.1 and 4.2; Table 4.3), 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, 2021). Eustigmatophyceae is the sister lineage to CSS in a phylogenomic tree inferred from plastid genes (Ševčíková et al., 2015; Di Franco et al., 2022), while it is sister to RPX in a nuclear phylogeny (Burki et al., 2016; Derelle et al., 2016; Noguchi et al., 2016; Thakur et al., 2019; Terpis, 2021; Azuma et al., 2022; Di Franco et al., 2022; Cho et al., 2024, 2022). However, the latter studies only included a single Eustigmatophyceae taxon, likely contributing to with weak bootstrap supports. Two chains of the Bayesian analysis did recover the Eustigmatophyceae grouping close to CSS, along with Olisthodiscophyceae and Actinophrydae (Appendix O), 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. I 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 4.3).

Although I replicated the similar placement of Eustigmatophyceae that would be observed in trees inferred from plastid genes, I suspect that these groupings are the result of small data size (C60) and/or LBA artefact (Eustig+Actino in F140), rather than replicating evolutionary or artefactual processes of plastid genes. Instead, it is likely the incongruence observed in nuclear versus plastid trees is the result of molecular convergence arising from non-neutral selection force. Molecular convergence arising from neutral or random mutations (e.g., homoplasy) can be remediated by current phylogenomic mixture models (Lartillot and Philippe, 2004; Wang et al., 2008, 2018). However, non-neutral force on plastids such as balancing selection that selects similar sets of plastid genes across eukaryotes (Maier et al., 2013; Dorrell et al., 2019) can result strong phylogenetic signal in these genes (Stiller et al., 2003; Edwards, 2009). Further investigation on the effects on non-neutral force on plastid and nuclear genes may help understanding the incongruence between the two datasets (Stiller et al., 2003; Castoe et al., 2009). Additionally, it may be worthwhile examining gene properties of plastid genes and compare with nuclear genes.

I observed a clade comprising Pinguiophyceae+Olisthodiscophyceae (Olis+Ping) in almost all trees examined, including the ones with fast-evolving sites, random sites and genes removed (Fig. 4.1-4.2; Table 4.2; Appendix P). This clade was the sister group of CSS, often with strong branch supports (Fig. 4.1; Table 4.3) and was also observed in a previous phylogenomic study (Terpis, 2021). The close relatedness between Pinguiophyceae and CSS has been demonstrated in other studies including the ones using plastid genes however these only used a single taxon representing Pinguiophyceae or recovered lower bootstrap supports (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 (Appendix O) had different placements of Pinguiophyceae (branching as a 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', most with higher BS supports (Fig. 4.1). I observed a clade comprising Bigyromonadea and Ochrophyta in the CAT-PMSF tree (Fig. 4.2; Table 4.3) with up to 88% BS.

4.3.1.3 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 (Appendix O), while the topology of chain 3 and 4 had the same topology in Gyrista topology (Appendix O). The topology of the ochrophytes were almost the same (except for the placement of *Attheya septentrionalis*; Bacillariophyceae) between the chains 1-4 and the C60-PMSF tree inferred from the '231-supermatrix' (Figs. 4.1 and 4.2A). This conflicting placement of *A. septentrionalis* can also be found in previous studies (Theriot et al., 2010, 2015; Parks et al., 2018; Dorrell et al., 2021) where different set sizes of genes were sampled; small subunit ribosomal genes and plastid genes (Theriot et al., 2010, 2015), high occupancy orthologs (58,294 sites) found in diatoms (Parks et al., 2018) or ochrophytes (26,399 sites) (Dorrell et al., 2021).

For Bigyra, I found paraphyly similar to that observed by Cho et al. (2024) in addition to the unstable groupings within Placidozoa (Fig. 4.1; Appendix O). 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 (Kolodziej and Stoeck, 2007; Okamura and Kondo, 2015; Cho et al., 2024).

4.3.2 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, I 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 a far more complex relationship. As a result, it was challenging to devise a suitable filtering criteria that could discern genes by the “good” or the “bad” gene properties (Fig. 4.3A; Appendix Q). 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), my 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 in an additional PC axis (Appendix Q). This made the delineation of the “good” or the “bad” genes further challenging. All values of the 13 properties are summarised in Appendix R.

I observed that the majority of noise (e.g., saturation, av_patristic, and root_tip_var - coloured in red Fig. 4.3A) 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. 4.3; Appendix Q). The rest of the noise, RCFV (coloured in red in Fig. 4.3A), 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 the phylogenetic signal (e.g., average_BS_support, robinson_sim – coloured in blue in Fig. 4.3A), although in an opposing direction (i.e., negative correlation). The two properties, the treeness and occupancy were explained by PC1 but negatively correlated with the noise and data quality (Fig. 4.3; Appendix Q). Consequently, I included various filtering criteria (criteria A-D) by PC loadings and their correlations (Appendix Q and S) among different properties regardless of the nature (e.g., noise, signal, or data quality) of the gene properties. Additionally, not all the gene properties of the same nature showed strong positive correlations (Fig. 4.3; Appendix S). I 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 4.3; Appendix S). Presence of many recalcitrant nodes, 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 radiation are likely the cause of such difference between stramenopile dataset and the dataset analysed by the initial research that established this method (Mongiardino Koch, 2021).

4.3.2.1 Evolutionary rate provides phylogenetic signals 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) (Appendix Q). 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, I treated them as noise in my analyses (Fig. 4.3B). Similarly, I 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) (Appendix Q).

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 (Philippe et al., 2005; Bapteste et al., 2007; Edwards, 2016; Superson and Battistuzzi, 2022). However, despite their correlation with other noise in this study (Fig. 4.3B), 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 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.

4.3.3 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 see the effects of gene properties on phylogenomic analysis (Table 4.3; Appendix R). To minimize the effect of the small data size (i.e., number of amino acid sites) on the phylogenomic analyses, I compared the average BS support of all trees reconstructed from random site or gene removal to the C60-PMSF tree reconstructed from the 231-supermatrix (Fig. 4.1). Based on the change of backbone topologies and their average BS supports, supermatrices with the average BS less than 89% was deemed too small to sufficiently distinguish from the effects of different gene-filtering criteria and small data size. Therefore, I 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. 4.2; Table 4.2 and 4.3).

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. 4.2A and D; Table 4.3).

To investigate the effects of signal, noise, and data quality alone, I included additional filtering criteria (criteria S, E, F, and Q) to compare the trees with those reconstructed from supermatrices A-D and ABC (Fig. 4.2). When I 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. 4.2A and J). These criteria all selected for high signal while the criteria A and B distinguished the signal associated PC axis in addition to other highly correlated gene properties (i.e. data quality and information).

The trees reconstructed from high data quality (criterion Q) had the most stable topologies (Fig. 4.2K), all of which were identical to the ’231-supermatrix’ C60-PMSF, except the placement of Actinophrydae. For the 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. 4.2D, E, J, K). Similarly, the trees reconstructed from supermatrix D120-160 (Fig. 4.2F) showed unstable topology of Pseudofungi and the placement of Actinophrydae (Fig. 4.2F).

When I examined the trees reconstructed from supermatrices E and F (selecting genes with high noise), the placement of *Platysulcus tardus* became unstable, no longer was sister to the rest of the ochrophytes (Fig. 4.2H and I; Table 4.3). 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 a data size and phylogenetic noise, likely due to having the phylogenetic signal present is smaller set of genes compared 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. 4.2H and I; Table 4.3). The latter topology observed in plastid multi-gene trees (Ševčíková et al., 2019; Barcytė et al., 2021; Di Franco et al., 2022).

The majority of the Actinophrydae (Actino) placement was observed to be sister to CSS+Pico or Olis+Ping, each relationship with the same frequency (14 occurrences) (Fig. 4.2; Table 4.3). 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 (Appendix T). However, Actino+[CSS+Pico] was also recovered the '231-supermatrix' C60-PMSF (Fig. 4.1). I 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 4.3) 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.

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, rest of the topologies were the same relationship as the ones found in '231-supermatrix' C60-PMSF (Fig. 4.2A and G).

The placement of Bigyromonadea being sister to ochrophytes was observed in different criteria that had relatively small data size (e.g. N, A and B60-80, C80-120, D140-160, ABC60-100) and for the one that select for genes with high noise (F120-140) (Table 4.2 and 4.3). Thus, the groupings of Bigyromonadea+Ochrophytes, Sagenista+Opalozoa found in other trees generated with different filtering criteria may have been the result of lack of phylogenetic signal arising from small data size or the effect of compositional bias (Fig. 4.2I). When I incrementally removed fast-evolving sites, I observed monophyly of Pseudofungi (oomycetes, hyphochytriomycetes, and Bigyromonadea) in trees with up to 67% aa sites removed (Appendix P: A). Even when I randomly removed aa sites, bigyromonads formed a monophyly with oomycetes, and Platysulcidae remained sister to rest of the stramenopiles in most cases (Appendix P). When I 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 (Appendix P: C).

4.3.3.1 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 (Koshi and Goldstein, 1995; Jimenez et al., 2018; 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 (Koshi and Goldstein, 1995; Boussau et al., 2008; Jimenez et al., 2018; 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. 4.2B). When I compared the trees inferred from supermatrices E and F, I observed that the monophyly of Bigyromondea and Oomycetes were no longer stable in trees inferred from supermatrices F (selecting for genes with all the high noise, including RCFV) (Fig. 4.2H 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 various selecting criteria (Fig. 2), we speculate that paraphylectic relationship of Bigyomonadea and Oomycetes is an artefact of across-taxa aa compositional bias (i.e., RCFV).

4.4 Conclusion

To resolve the placement of several contentious lineages of stramenopiles, I updated stramenopile supermatrix and conducted phylogenomic analyses using various inference methods. I recovered robust relationships of previously phylogenomically unavailable or contentious lineages such as Eustigmatophyceae, Olisthodiscophyceae, and Pinguiophyceae. Additionally, based on 13 proxies for phylogenetic noise, signal, and quality for each gene, I constructed numerous supermatrices based on different criteria selecting for genes with high signal or low noise. I found the tree topologies were more stable when I selected for genes with high signal and data quality. Selecting the most conserved, the 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.

Table 4.1 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

Table 4.2 Summary of supermatrices generated using different filtering criteria

The number values are size of the amino acid and the brackets indicate the number of genes. ‘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’; 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)	—	—	—	—		
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)	11,353 (43)	72,932 (231)
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)		

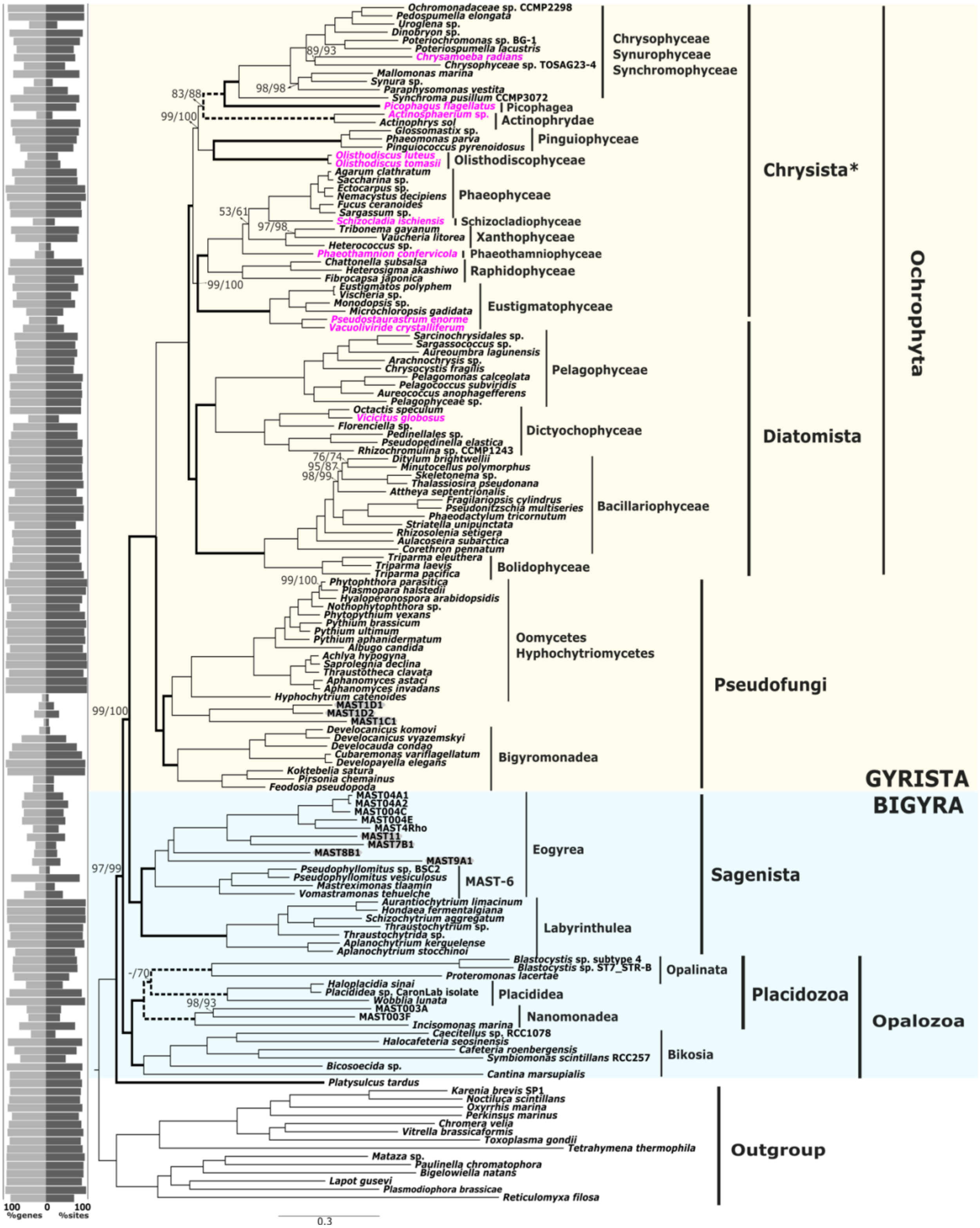


Figure 4.1 Phylogenomic of stramenopiles with 10 new ochrophyte transcriptomes

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 and inference methods. 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. 4.2 and Table 4.3). 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.

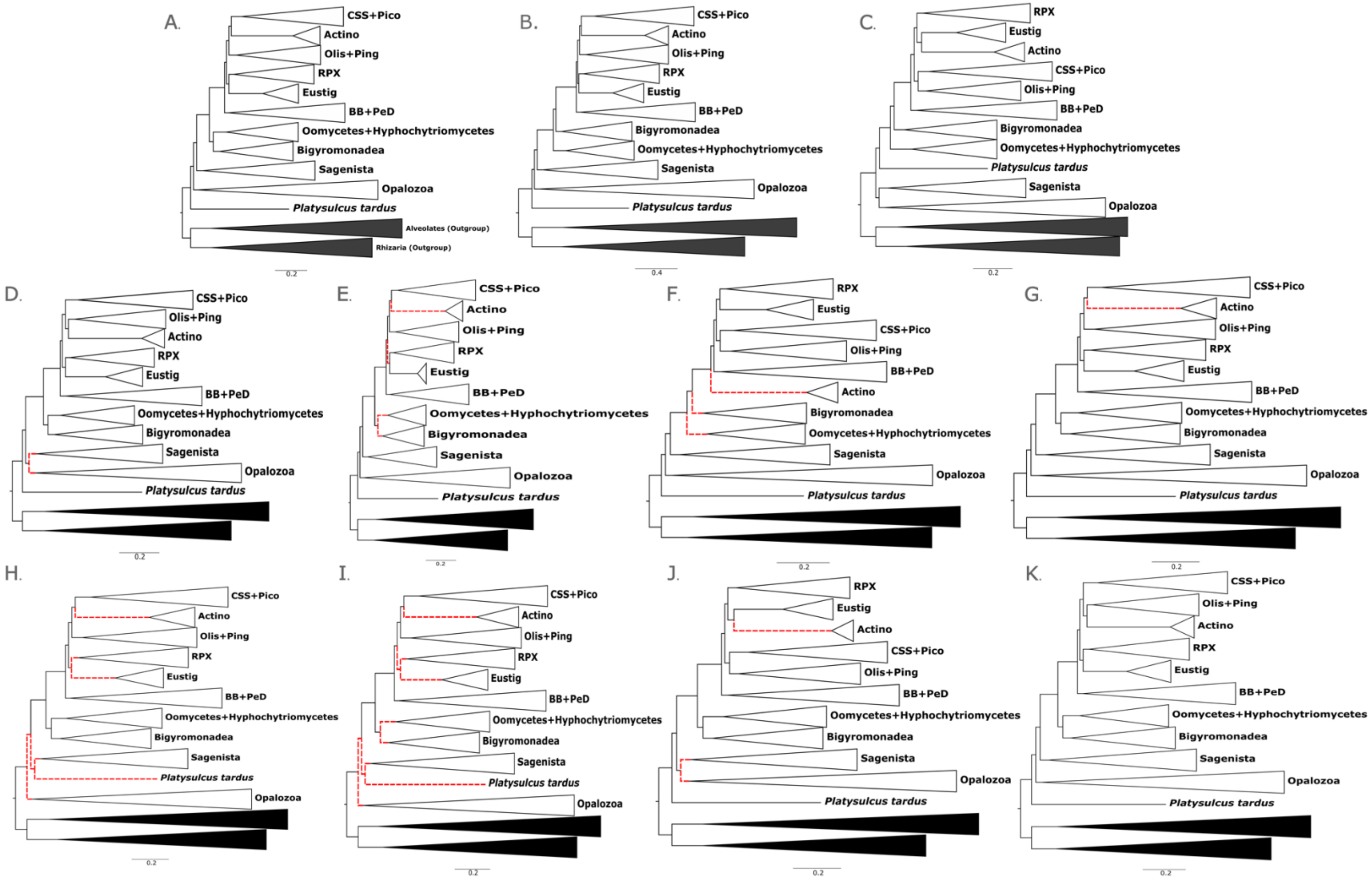


Figure 4.2 Schematic representation of major stramenopile topologies

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.

Table 4.3 Summary of bootstrap support for all the supermatrices

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	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

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	
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	—	—	—	—	—	—	—	—	—	—	—	—
[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

Groupings	Criterion E				Criterion F				Criterion S				Criterion Q				Criterion ABC						231-supermatrix	CAT-PMSF
	120	140	160	180	120	140	160	180	120	140	160	180	120	140	160	180	60	80	100	120	140	160		
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)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
[PeD+BB]+[RPX+Eustig] (2)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	49	—	—	—	—	—	—	—
[CSS+Pico]+[Eustig+Actino] (1)	—	61	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

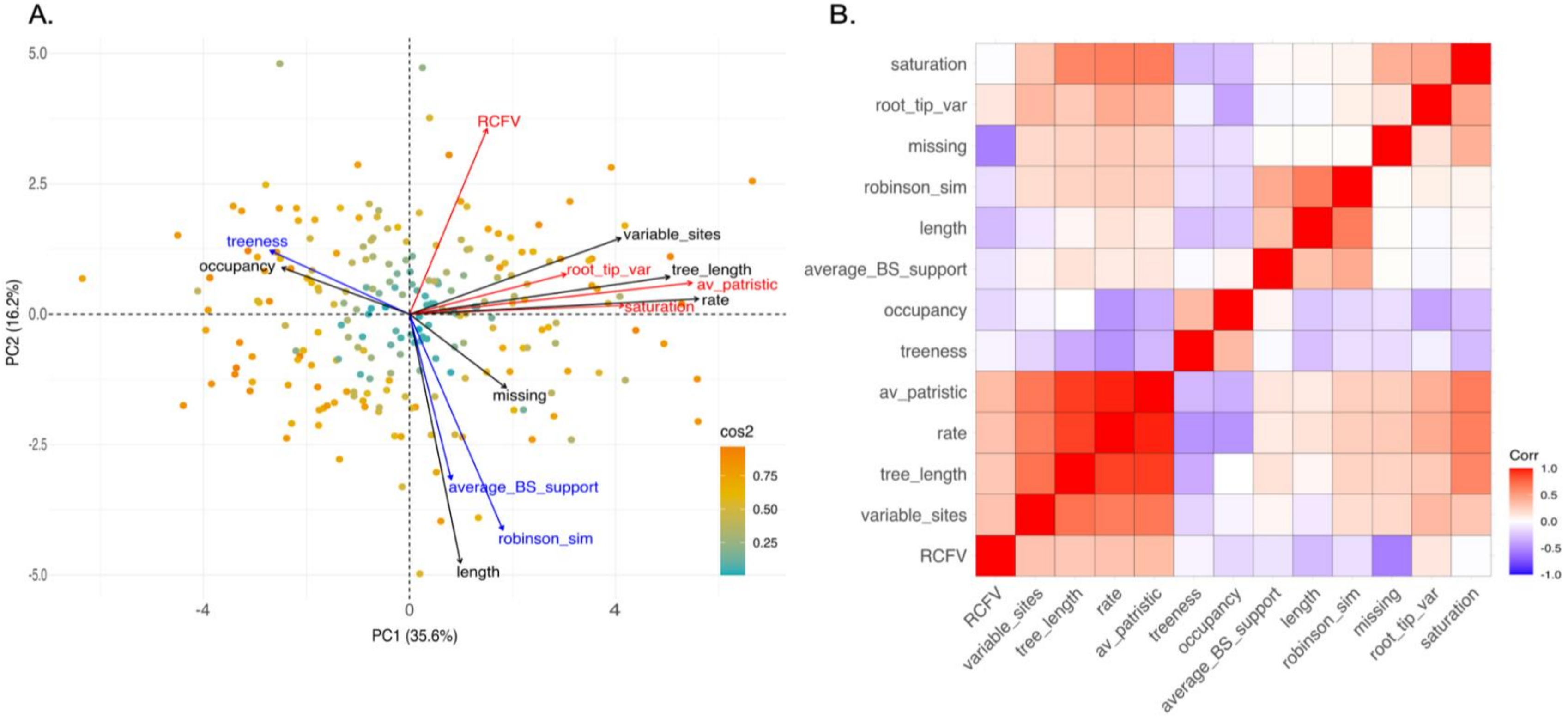


Figure 4.3 Gene properties on a PCA plot

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 biases (red), signals (blue), and data quality (black). (B) Correlation matrix with hierarchical clustering of 13 gene properties. Positive correlation is indicated by red and negative by blue.

Chapter 5: Genomic analyses of *Symbiomonas scintillans* show no evidence for endosymbiotic bacteria but does reveal the presence of giant viruses

5.1 Introduction

Understanding the evolutionary history of eukaryotes is inherently linked to understanding their symbiotic relationships with prokaryotes, whether it is in the form of genetically integrated organelles or the multitude of short-term endosymbioses with bacteria or archaea. Most of our understanding about the effects of endosymbiosis on eukaryotic evolution relates to the origin of mitochondria and plastids, and their involvement in eukaryotic diversification (Gray, 1999; Gray and Doolittle, 1982; Keeling, 2010, 2009). However, the impact of prokaryotic symbioses goes far beyond these rare organellogenesis events, given the diverse nature of symbioses affecting hosts in different ways (reviewed in Husnik and Keeling, 2019; Nowack and Melkonian, 2010). Thanks to genome sequencing, prokaryotic symbionts have been found to be associated with all major eukaryotic supergroups, involved in a myriad of functions such as metabolism (Fenchel and Finlay, 1991; Kneip et al., 2008; Nowack and Melkonian, 2010; Seah et al., 2019), defense (Ishida et al., 2014), parasitism (Corsaro et al., 2010, 2013), and motility (Okude et al., 2012; Ishida et al., 2014). Additionally, some bacterial lineages have evolved to be "professional symbionts" (Husnik and Keeling, 2019) such as Chlamydiae, Rickettsiales, and Holosporales, consisting entirely of obligate endosymbionts or intracellular parasites of eukaryotic hosts (Montagna et al., 2013; Boscaro et al., 2019; Husnik and Keeling, 2019; Giannotti et al., 2022).

Despite these findings, most prokaryotic symbionts of eukaryotes are poorly studied, generally only observed with microscopy, and left unidentified and uncharacterized. For example, the only known case of prokaryotic endosymbiosis in a non-phototrophic stramenopiles is found in the tiny (~1.4 µm) bikosia, *Symbiomonas scintillans*, where two geographically distinct strains were reported to harbour up to six endobacteria, and which served as the inspiration for its genus name (Guillou et al., 1999). The location of these endobacteria within the endoplasmic reticulum was of particular interest, as this is where plastids of phototrophic stramenopiles are located (Cavalier-Smith, 1989; Guillou et al., 1999). However, the identity and functional role of these apparent endobacteria has not been further investigated. To identify the endobacterium and its role in such a small protist, we conducted Fluorescent *in situ* hybridization (FISH) targeting various groups of bacteria and generated amplification-free shotgun metagenomics and whole-genome amplification sequencing data of two strains of *S. scintillans*. This showed the absence of endobacteria of known endosymbiotic lineages. Instead, we observed a viral-like particle by transmission electron microscopy (TEM) and recovered three draft viral genomes related to prasinovirus, nucleocytoplasmic large DNA viruses (NCLDVs) belonging to a member of the Phycodnaviridae family (Van Etten et al., 2002; Wilson et al., 2009). During the course of this work, one strain apparently lost the virus, while the other strain perished, so I was unable to conduct further experiments to verify the nature of the viral association. This chapter underscores how much is still unknown about endosymbioses, particularly in small heterotrophic protists. I expect that viral association is especially relevant to nano- or pico-eukaryotes, as there may simply not be enough space for endobacteria, and predict more such findings in the future.

5.2 Materials and Methods

5.2.1 Culture collection and maintenance

All strains of *S. scintillans* used in this study are summarized in Appendix U, with the initial isolation dates and locations, sequencing methods, dates, and locations, and the culture collection centres. Two *Symbiomonas scintillans* culture strains RCC257 and RCC24 were obtained from the Roscoff culture collection (RCC, France) on March 7th, 2022. The cultures were grown and maintained in 0.22 µm filtered and autoclaved marine f/2 media (30 PSU) with an autoclaved rice grain at the University of British Columbia (UBC), Canada. The cultures were kept in a 20°C incubator with a 12:12 h light:dark cycle and sub-cultured every two weeks in 30 mL. Using glass micropipettes, approximately 50 to 100 cells from each strain were collected and stored in 5 µL PCR-grade water after two rounds of rinsing in PCR-grade water on April 6th, 2022. The isolated cells were immediately subjected to three rounds of freeze-thaw cycles to promote lysis and stored at -80°C until whole genome amplification (WGA). Upon receiving the two strains, they were slow to grow (low culture density and no noticeable movement) and within two months of receipt, the strain RCC24 showed reduced viability and was eventually lost. This was also observed in the RCC, as their cultures perished with no identifiable cause at a similar time (M. Gachenot, assistant engineer/curator of RCC, personal communication, Oct 12th, 2022). In contrast, the strain RCC257 became denser and more active between the first round of cell collection in April 2022 and the second round of cell collection on June 28th, 2022. I suspected this boost of culture viability can be due to resistant cells or loss of viruses (see below). As a result, I also collected 50 cells from strain RCC257 on June 28th, 2022, for an additional WGA (hereafter, referred to as RCC257-late).

Independently at OIST, Okinawa (Japan), the culture strains RCC257 (which I refer as RCC257-jp) and NIES-2589 (strain synonymous to RCC24) were obtained from the RCC in December 2022, and the Microbial Culture Collection at the National Institute for Environmental Studies (NIES Collection, Tsukuba, Japan) in March 2021. Strain NIES-2589 will be hereafter referred to as RCC24-jp. RCC24-jp was cryopreserved at -160°C and was thawed in f/2 medium with an added rice grain. The RCC24-jp cultures were maintained in the same condition as above, except with a 10:14 h light:dark cycle, and were further processed for amplification-free shotgun metagenomics (AF-SMG; see Library preparation and sequencing). Strain RCC257-jp was grown in 20 µm filtered and autoclaved seawater with rice. All cultures were sub-cultured every four weeks.

5.2.2 Library preparation and sequencing

Two strains of *S. scintillans* (RCC24 and RCC257) maintained at UBC were subject to WGA sequencing and one strain RCC24-jp, maintained at OIST was subject to amplification-free shotgun metagenomic (AF-SMG) sequencing. To prepare a WGA library of the isolated cells, a 4BB^{TM} TruePrime® Single Cell WGA Kit was used following a manufacturer's protocol with 12 h incubation at 30°C for the amplification reaction step. The amplified product was then cleaned with AMPure XP beads (Beckman Coulter, US), following a protocol described in the Nanopore Ligation Sequencing Kit protocol (SQK-LSK110, Oxford Nanopore Technologies, UK). Library preparation for WGA sequencing followed the Illumina DNA Preparation kit (Illumina, US) which uses a Bead-linked Transposome complex, resulting in ~350 bp library constructs. The WGA sequencing was performed on a NextSeq (mid-output) platform with 150 bp paired-end library constructs at the UBC Sequencing and Bioinformatics Consortium (Vancouver, Canada). Whole genome amplification sequencing was repeated twice using the same library constructs. For downstream analysis, the transcriptome of RCC257 (NCBI SRA accession number SRR24392496) was also used, which was prepared from approximately 20 isolated cells from the same sub-culture, described in Cho et al. (2024). To minimize culture-associated bacterial reads, only single-cell isolated transcriptomes were used, as opposed to cDNA prepared from whole-culture RNA extract.

For shotgun metagenomics, 10 mL of RCC24-jp culture was filtered through a 5 µm syringe filter for enrichment (removal of large bacteria) followed by DNA extraction using the MasterPure Complete DNA and RNA Purification kit (Lucigen, US). The DNA extractions were prepared from multiple subsequent subcultures (in March, May, June, and October 2022). PCR-free shotgun metagenomic libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB, US) and sequenced by the OIST Sequencing Centre using the Illumina MiSeq platform with 300 bp paired-end reads.

The strain information and sequencing methods are summarized in Appendix U. The raw genomic data for this study is deposited in the NCBI Sequence Read Archive (SRA) with the accession numbers SRR26451788-SRR26451790, SRR26412500-SRR26412501, and SRR26943481, under the BioProject PRJNA1029166.

5.2.3 Sequence processing: assemblies and sub-assemblies of viral reads

The quality of raw sequencing reads for amplification-free shotgun metagenome, WGA, and transcriptome data were all examined using FastQC v0.11.9 (Andrews, 2010). The transcriptomic data were processed as described in Cho et al. (2024). Briefly, to correct random sequencing errors of the raw data, *k-mer* based Rcorrector (v3) (Song and Florea, 2015) was used followed by Trimmomatic v0.39 (Bolger et al., 2014) to remove transposase-inserts, SmartSeq2 primers, adaptors, IS-primers from library preparation and, low-quality reads (-phred33 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). Error-corrected and trimmed forward, reverse and unpaired transcriptome reads were then *de novo* assembled using rnaSPAdes v3.15.1 (Bushmanova et al., 2019). After removing bacterial (non-endosymbiotic lineages) and metazoan reads, open reading frames (ORFs) were predicted with TransDecoder v5.5.0 (Haas et al., 2013). The raw shotgun metagenome and WGA sequencing data were trimmed as described above without Rcorrector step, with corresponding adaptors and primers removed. The trimmed WGA reads from the three rounds of sequencing runs were then co-assembled using SPAdes v3.15.1 (Prjibelski et al., 2014, 2020; Vasilinetc et al., 2015) with --sc and --phredoffset33 options. The same assembly parameters were used for the shotgun metagenome reads. For initial taxonomic and coverage screenings of the assembled transcriptomes, shotgun metagenome, and WGA assemblies, and in particular to search for reported endobacteria, BlobTools v2.3.3 (Laetsch and Blaxter, 2017; Challis et al., 2020) was used to visualize search results of assemblies against NCBI nucleotide (nt) (using megaBLAST) and UniProt reference databases (using diamond BLASTX), both with e-value cut-offs at 1e^{-25} (--taxrule bestsumorder). After failing to detect any obvious taxonomic signatures of endobacterial origin in both transcriptome and genomic data, a subset of WGA reads was reassembled by filtering reads with GC content below 40% (a common range for endosymbionts) and coverage above 10^{25} using SeqKit v2.3.0 (Shen et al., 2016), and assembling those with both SPAdes and Unicycler v0.5.0 (Wick et al., 2017; Prjibelski et al., 2020).

With the initial BlobTools screening indicating the presence of prasinovirus taxonomic assignments in WGA sequencing data, trimmed WGA contigs were searched against the Reference Viral Database (RVDB) (Goodcare et al., 2018) using blastn (e-value cut-off 1e^{-10}) followed by protein domain searches using hmmsearch (HMMER3.3) (hmmer.org) against virus orthologous groups (VOGs) (vogdb.org), Pfam, and giant VOG (GVOG) hidden Markov models (HMM) databases compiled in ViralRecall (Aylward and Moniruzzaman, 2021). The open reading frames (ORFs) were predicted using Prodigal-gv (Hyatt et al., 2010). Contigs with final ViralRecall scores above 10 were considered of viral in origin. For those with the final ViralRecall scores less than 10, if the number of VOG hits were higher than 3 and Pfam hits at the same time, we also considered these contigs to be viral. Additionally, all the contigs mapping to 16 prasinovirus genomes (BIIV, BpVs, OlVs, OtVs, OmVs, and MpVs) (Appendix V) using DNAdiff v1.3 (Kurtz et al., 2004) were kept. Contigs with viral hits from NCBI nt, clustered RVDB (RVDB-c), UniProt blast searches were kept, excluding circular elements. These select contig results were cross-validated with blastx and blastn searches among VOG, RVDB, and diamond viral databases. The same searches were repeated on amplification-free shotgun metagenomic data, WGA data from RCC257-late, NIES-2589 (RCC24-jp), and ORF-predicted transcriptome data. However, neither prasinovirus nor NCLDV reads were detected. Aside from microscopic observation, to confirm the absence of green algae contamination in the culture, I searched for small subunits of ribosomal RNA (SSU rRNA) in all sequencing data using barrnap v0.9 (Seemann, 2007). We also carefully screened eukaryotic reads from the initial BlobTools results and found no evidence of green algae or other eukaryotic protist contamination.

An extracted subset of WGA viral contigs (707 contigs out of 69,958) was reassembled using SPAdes v3.15.1 with –sc, --careful and –phredoffset33 options, resulting in 748 scaffolds. Scaffolds with lengths under 100 bp were removed. Additionally, blastn searches were repeated against NCBI-nt and RVDB databases to remove bacteriophage reads, resulting in 543 scaffolds with the total length of 469,314 bp with 37.31% GC content. These filtered subset assemblies are hereafter referred as “viral-subset-scaffolds”. The viral-subset-scaffolds were then further scaffolded using 16 prasinovirus genomes as a guide with homology-based RagTag v2.1.0 (Alonge et al., 2019, 2022). This reference-guided assembly method does not alter the scaffold sequences but reorients and reorders them by aligning to a reference genome, creating a single scaffold or a pseudomolecule. The pseudomolecule or the single scaffold of viral metagenomically assembled genomes will be hereafter referred to as vMAG. Out of 543 viral-subset-scaffolds, 279 were recruited for the assembly of 16 vMAGs. The remainder of the 264 scaffolds were not recruited to any reference genomes despite having 194 out 264 scaffolds with ≥80% similarity to known sequence identities (ID) and e-value < $1e^{-25}$ hit to prasinovirus (the rest of the scaffolds had lower % ID or no hits to the database). This is due to the majority of the scaffolds (215/264) being shorter than 500 bp, which were filtered out due to the small alignment length threshold (1000 bp). Additionally, a pre-defined *k*-mer and window size (19 bp) in read mapping to the reference genomes may have affected correct scaffold placements of sequence variants in these potentially new vMAGs.

The completeness of each reference-guided assembly was assessed using CheckV v0.8.1 with CheckV-db v1.5 (Nayfach et al., 2021) (Appendix V). The assembled vMAGs with the highest completeness and the corresponding reference genomes are circularized for visualization with BLAST Ring Image Generator (BRIG) v3 (Alikhan et al., 2011).

5.2.4 Draft viral genome annotation and gene content comparison

The ORFs for each reference guided vMAGs (=pseudomolecules) were predicted with Prodigal-gv and further annotated with Prokka and ViralRecall (scores >=10). The annotation was repeated with the 16 prasinovirus genomes (Appendix V). To compare shared orthologs among vMAGs and published viral genomes, all the ORFs were used in all-versus-all blastx search (Burns et al., 2018). The blastx (e-value=$1e^{-5}$ and query-cover=50) result was then clustered first by 95% similarity using CD-hit v4.8.1 (Fu et al., 2012) followed by MCL algorithm (inflation=2). Only clusters with hits from a minimum of three different genomes (including vMAGs) were retained (432 clusters). Amino acid sequences of each cluster were then aligned using MAFFT v7.481 (Katoh and Standley, 2013) and trimmed using trimAl v1.2rev59 (Capella-Gutiérrez et al., 2009), which was then used to build 432 HMMs. The resulting HMMs were then searched against individual reference and draft genomes using hmmsearch HMMER v3.3 (e-value 1e^{-10} and domain e-value 1e^{-8}) to confirm the presence of the protein clusters in the genomes. The outcome of shared protein clustering hits for each genome was summarized in an upset plot. All BpV- and BIIV-vMAGs were combined as “BV-vMAGs”, all OlV-, OmV-, and OtV-vMAGs were combined as “OV-vMAGs”. Similar grouping was done for published genomes that were used as a reference-guide to assemble vMAGs.

5.2.5 Prasinovirus hallmark gene search and phylogeny construction

To construct a phylogenetic tree of prasinoviruses, we searched for 22 prasinovirus hallmark genes (Rozenberg et al., 2020; Bachy et al., 2021) in the predicted ORFs of WGA viral-subset-scaffolds using blastp (e-value 0.001) and hmmsearch (–E 1e-3 –domE 1e-3 –incE 1e-3 –incdomE 1e-3). Candidate genes from the predicted ORFs were then concatenated with the corresponding alignments and realigned with MAFFT (--auto) and trimmed with trimAl (-gt 0.3 and -st 0.001). I constructed a single gene tree for each of the prasinoviral hallmark genes using IQ-TREE v2.1.0 under the LG+G4 amino acid model with 1000 ultrafast bootstrap pseudoreplicates. Each single-gene tree and corresponding alignment were manually examined to discern viral paralogs and orthologs from cellular proteins. For prasinovirus some single-gene alignments, the candidate genes were manually merged if the gene fragments had overlapping regions and were positioned within the same clade. The 22 cleaned prasinovirus hallmark single-gene trees were then concatenated, realigned with MAFFT, trimmed with trimAl, and a multi-gene phylogenetic tree was inferred using IQ-TREE v2.1.0 under the LG+G4+F model and 1000 ultrafast bootstrap pseudoreplicates. We searched for the same prasinovirus hallmark genes in predicted ORFs from my transcriptome data in the same manner. However, no hits were found.

The hallmark gene-alignments, relevant intermediate files, gene-tree files, vMAG genome and protein sequences are uploaded on Dryad (10.5061/dryad.mw6m90644).

5.2.6 Transmission Electron Microscopy (TEM)

To visualize a virus-like particle (VLP) in unfiltered strain RCC24, 5 µL of the culture was deposited onto glow-discharged (60 sec at 50 mA; Leica EM ACE200) formvar/carbon-coated 400 mesh copper TEM grids. Samples were stained with 2% uranyl acetate for 60 s. Excess UA was removed by gently placing a filter paper at the edge of the grids and subsequently transferred to a FEI Tecnai Spirit TEM (Thermo Fisher, USA) operating at 80 kV acceleration voltage. Images were captured with a DVC1500M camera and AMT Image Capture Engine V601 software (MA, USA). VLP diameter was measured with the AMT built-in measurement tool. All sample processing and TEM imaging were carried out in a sterile environment where no other viral experiments were done prior to the imaging.

5.2.7 Fluorescence *in situ* hybridization (FISH)

For the RCC257 strain grown at UBC, Canada, 10 mL of culture was spun down in 15 mL centrifuge tubes at 3000 rpm, at 4°C for 10 min. The centrifuged cells were collected from the bottom of the tubes and transferred into 1.5 mL microcentrifuge tubes. Approximately 7 µL of the collected cells were placed on Poly-D-Lysine-coated glass slides (Sigma-Aldrich, US) and demarcated with a LiquidBlocker (Electron Microscopy Sciences, US). An equal amount of 4% paraformaldehyde (in water) was added to the slides. After all the liquid evaporated, 95% ethanol was added to the marked spot on the slides and incubated until complete drying. The slides were dipped in 50%, 80%, and 100% ethanol for 10 min each. The slides were then incubated overnight in a dark humidity chamber at 46°C with 10 µM of probe EUB338-Green prepared in a hybridization solution (1 M, pH 8.0 Tris HCl; 5 M NaCl, 1.3% SDS). The slides were gently rinsed twice in 48°C hybridization solution for 10 min, followed by 15 min rinse in water at room temperature. After completely drying liquid, 20 µL of SlowFade Gold with DAPI (Life Technologies, US) were added and visualized with an Olympus BX53 at the UBC Bioimaging Facility, Canada.

To verify the lack of endobacteria in sub-cultures grown in Japan, a separate FISH protocol was done on the RCC24-jp and RCC257-jp strains. Each of the 10 mL of culture were fixed with 3.2% formaldehyde at 20 °C for 20 min and spun down at 4000 rpm at 4 °C for 15 min. The centrifuged cells were washed with 1x PBS and seeded onto a 0.1% polyethyleneimine-coated 18 mm round coverslip (Matsunami Glass Ind., Ltd, Japan) in a 12-well plate. To allow attachment to the coverslip, the fixed cells were incubated for 3 h in 1x PBS. The attached cells were then washed three times each for 5 min in 1x PBS, 0.3% 1x PBS-Tx (0.3% Triton X-100, pH7.4), then in hybridization buffer (20 mM Tris-HCl; 30% formamide; 0.01% SDS). The fixed cells were hybridized with probes EUB338-Alexa488, EUB338-Alexa647 (Eubacteria), and CF319a-Alexa647 (Bacteroidetes) [0.1 µM] (Manz et al., 1992) (ThermoFisher Scientific, Japan) with DAPI [0.01 ug/mL] (Roche, Germany) and incubated overnight in a 42°C humidity chamber. For RCC24-jp, additional probes targeting Planctomycetes (PLA46) (Neef et al., 1998), alpha- (ALF969), and gamma-Proteobacteria (GAM42a) (Neef, 1997) were hybridized. To remove unbound probes and DAPI, the coverslip was gently rinsed three times in 0.3% 1x PBS-Tx solution for 5 min and twice in 1x PBS. After drying, the coverslip was mounted onto a glass slide with ProLong Diamond Antifade Mountant (ThermoFisher Scientific, Japan) and incubated at room temperature overnight in the dark. The hybridized sample was kept at 4 °C in the dark until visualization on Leica TCS SP8 Inverted Confocal Microscope at the OIST

Imaging Facility (Okinawa, Japan). The brightness and contrast of all images were adjusted using ImageJ v1.53 and sharpness with Inkscape v1.2.1.

5.3 Results and Discussion

5.3.1 No bacterial sequences from known clades of common endosymbionts

To identify the symbionts of *Symbiomonas scintillans,* I sequenced two geographically distinct strains (RCC24 isolated from Pacific Ocean and RCC257 from the Atlantic Ocean) maintained under culture conditions. In most of the sequencing data, a large representation of the host sequence was found as expected. The exception to this is the WGA data from RCC24, where no host sequences could be identified (see below). As the original description of this taxon suggested these symbionts were bacterial, we first searched for bacterial reads assigned to well-known endosymbiotic lineages such as Rickettsiales, Holosporales, or Chlamydiae in all the analyzed genomic and transcriptomic data. No such putative symbiont reads were found, and instead the bacterial reads were largely assigned to common environmental, or culture-associated Alphaproteobacteria, Gammaproteobacteria and Balneolia such as *Marinobacter* spp., *Epibacterium* spp., *Hyphomonas* spp., *Zhongshania* spp., *Balneola* spp., and *Labrenzia* spp., (Appendix W). When sequences that had no taxonomic affiliation in WGA data were removed, a scaffold assigned to *Marinobacte*r *salinus* had the third highest coverage up to x95,000 (N50=116K), after the ones assigned to Oomycota (N50=276) and *Cafeteria roenbergensis,* (up to x102,851 coverage with N60=63K), a species closely related to *S. scintillans* (Cho et al., 2024). Notably, *Marinobacter* spp., *Labrenzia* spp., and *Hyphomonas* spp. were all reported to be common in cultures of *Ostreoccocus tauri*, Symbiodiniaceae, *Alexandrium* spp., and discobids (Alavi et al., 2001; Seibold et al., 2001; Jasti et al., 2005; Lupette et al., 2016; Bolch et al., 2017; Aponte et al., 2021; Maire et al., 2021). Accounting for this overwhelming representation of culture-associated bacteria, a subset of whole-genome amplification (WGA) data was selected and reassembled based on lower GC content, which is usually associated with endosymbionts. However, no sequences assigned to endosymbiotic bacterial lineages were detected. To account for unequal genomic amplification of WGA causing loss of AT-rich and local repeat regions, and secondary structures (Karlsson et al., 2015), I also searched bacterial reads in amplification-free shotgun metagenomic data. Many bacterial lineages with high-coverage in WGA were also found in the shotgun metagenomic data (e.g., *Marinobacter* spp., *Hyphomonas* spp., *Balneola* spp.) in addition to *Marinovum algicola* and a member of Phycisphaeraceae, but no known endosymbiotic lineages nor any draft bacterial genomes with “symbiotic features” such as small genome size, AT-rich content, or rapid sequence evolution could be identified in any of these data.

The absence of endosymbiotic bacteria in all the sequencing data was further supported by the absence of a bacterial signal using fluorescence *in situ* hybridization (FISH) of all sub-cultures of *S. scintillans* grown in Canada (RCC257) and Japan (RCC257-jp) (Fig. 5.1). My collaborators observed the same trend in RCC24-jp (Appendix X) using additional probes targeting Planctomycetes, Bacteroidetes, Alphaproteobacteria, and Gammaproteobacteria. In all my assembled WGA data, no sequences were assigned to Archaea while the amplification-free shotgun metagenome data had some Archaea contigs with low coverage (x1-7).

5.3.2 Multiple prasinovirus-like vMAGs are associated with the strain RCC257 and RCC24

Instead of endobacteria, we detected contigs assigned to prasinovirus with up to x200 coverage (Appendix W). When viral-subset-scaffolds were re-assembled and further scaffolded using a reference-guide approach, we recovered three viral metagenomically assembled genomes (vMAGs) related to the prasinovirus genera, *Bathycoccus prasinos* virus 2 (BpV2), *Ostreococcus* *lucimarinus* virus 1 (OlV1), and *Micromonas pusilla* virus Pl1 (MpV_Pl1). The completeness of vMAGs were the highest for the BpV2-guided assembly (BpV2-vMAG), with 100% completeness. Among OV-guided and MpV-guided vMAGs, OlV-1-guided assembly (OlV1-vMAG) and MpV_Pl1-guided assembly (MpVPl1-vMAG) had the most completeness with 54% and 18%, respectively (Appendix V).

I compared the number of shared scaffolds and gene contents among BV-, OV-, and MpV-vMAGs to verify the presence of multiple different virus genomes. Only up to two recruited scaffolds were shared between vMAGs of BVs, OVs and MpVs (Appendix Y). When the shared orthologs were examined among all vMAGs using 16 reference genomes, I observed the same trend (Appendix Y). Multiple copies of single-copy-genes (e.g., DNA polB, DNA helicase, and mRNA capping enzyme) (Clerissi et al., 2014; Moniruzzaman et al., 2020) were detected in viral-subset-scaffolds, each corresponding to three groups of prasinoviruses (Moniruzzaman et al., 2020). All 22 genes were placed within a BV clade, nine genes in an OV clade, and four in a MpV clade (Fig. 5.2), and similar trends were observed in RCC24 (Appendix Y). These results support the presence of multiple giant viruses, altogether referred as *S. scintillans* virus (SsVs), rather than a single genome mapping to multiple reference genomes. In RCC24 we found no evidence of host reads (see above), but also found evidence for three giant viruses very similar to those found in RCC257 (Appendix Z and AA). No prasinovirus reads were detected in RCC24-jp.

The presence of multiple giant viral species within a single host species is rare in protists. However, multiple viral species were detected in three different species of Ectocarpales, a group of brown algal stramenopiles (Muller and Parodi, 1993; McKeown et al., 2018). In these host species, up to two major capsid protein (MCP) genes of different Phaeoviruses

(Phycodnaviridae) subgroups were found. One of these phaeoviruses (EfasV), can infect different genera of Ectocarpales (Muller et al., 1996). Although prasinoviruses are reported to have a narrow host range at the strain or species level (Bellec et al., 2014; Baudoux et al., 2015; Derelle et al., 2015; Bachy et al., 2018, 2021), the close relationship to phaeoviruses might indicate wider host range is also possible for these new prasinoviral vMAGs. Additionally, the name "prasinoviruses" likely reflects a sampling bias in the first reports, as is the case for many viruses. Notably, both *Monkeypox* (MPXV) (Von Magnus et al., 1959) and *Cucumber mosaic viruses* (CMV) (Price, 1934) were named after their first isolation from *Macaca fascicularis* (macaque monkeys) and *Cucumis sativus* (cucumbers), respectively, but MPXV was subsequently reported to infect other hosts including humans and squirrels (for MPXV) (reviewed in Ullah et al., 2023), and CMV in legumes and ornamental plants (Heo et al., 2020).

5.3.3 Genome characteristics of vMAGs

While many genes and ORFs were predicted on all vMAGs, only BpV-vMAGs were fully annotated (Fig. 5.3; Appendix AB and AC). For BpV-vMAG, 297 ORFs were predicted, including homologues of Hsp70 [a known protein in BpVs with a green algal host origin (Moreau et al., 2010)], DNA methyltransferase, and multiple MCPs were identified (Appendix AC). For the OlV1-vMAG, up to 149 ORFs were predicted with 146 genes while MpVPl1-vMAG had 47 ORFs predicted with 40 genes (Appendix V). OlV- and MpV-vMAGs from RCC24 had more complete assemblies (Appendix V and AB).

Compared to published prasinovirus genomes with 3-5 tRNAs (three for BpVs), only two tRNAs in RCC257 BpV-vMAG were predicted (Fig. 5.3A): tRNA-Leu and tRNA-Asn. Similar to chloroviruses, four tRNAs were predicted in RCC24 BpV-vMAG, two of them being tRNA-Asn (Appendix AB) (Moreau et al., 2010). I detected five and six MCPs in RCC24 and RCC257

BpV-vMAGs, respectively, as was the case for BVII1-3 (Fig. 5.3; Appendix AC) (Bachy et al., 2021). A high number of MCPs (up to nine) is unique to Phycodnaviridae, however, its implications in host entry or capsid assembly are currently poorly understood (Moreau et al., 2010; Weynberg et al., 2011; Moniruzzaman et al., 2020). Along with other common prasinovirus proteins involved in carbohydrate synthesis (i.e., dTDP-4-dehydrorhamnose reductase, and five glycosyltransferases), I also detected ribulose-phosphate 3-epimerase in RCC257 BpV-vMAG (Appendix AC), which was unique to BIIV-2 and -3 among prasinoviruses (Bachy et al., 2021).

To evaluate unique gene contents in BV-vMAGs, I generated protein clusters and compared them between 16 vMAGs and reference genomes. I observed that 26 protein clusters were unique to BV-vMAG (including BpVs- and BIIVs-vMAGs) (Appendix Y and AD). Although most of the annotation indicated HMM hits to hypothetical proteins of prasinoviruses, I detected a protein cluster assigned to 4-hydroxy-2-oxopentanoic acid aldolase. In prasinoviruses, this enzyme was only found in MpVs and is involved in biosynthesis of isoleucine, leucine, and valine which might be important in capsid formation (Moreau et al., 2010; Weynberg et al., 2017). Additionally, in both RCC24 and RCC257 BpV-vMAGs, we detected the IceA gene (“induced by contact with epithelium” endonuclease) gene, a putative virulence gene in *Helicobacter pylori* (Peek et al., 1998) which is also in the *Chrysochromulina ericina* virus (Mimiviridae; NCLDVs) (Gallot-Lavallée and Blanc, 2017).

5.3.4 SsV vMAGs are associated with *S. scintillans*

As prasinoviruses are known to be host-specific and have not yet been described in other hosts, I wanted to rule out the unlikely possibility that these new viruses came from a cryptic prasinophyte in the culture. I detected no green algal SSU sequences or signals indicative of green algal contaminants in any of the microscopic observation and sequencing data. In my WGA data, there were 15 scaffolds assigned mitochondrial genes of various Chlorophyta species (Appendix Z), with their counts ranging from 1 to 331. A close inspection of these scaffolds showed that these hits are likely not green algal contamination, as the taxonomic assignments were based on short read lengths. Additionally, some of the blastp hits of the same scaffolds indicated a stramenopile origin (Bikosia, ochrophytes, and oomycetes), suggesting these regions of the scaffolds are likely from the host and represent conserved homologs found in mitochondria across different eukaryotes. I observed similar patterns with scaffolds taxonomically identified as belonging to Rhodophyta.

The possibility that prasinoviruses contaminated the culture media is also highly unlikely, given both the sterilizing protocol (autoclaving, filtering, and UV sterilization) and single-cell isolation. These methods could hardly result in near-complete BpV-vMAGs from contaminant viruses, which require a minimum of 10^5 VLP to reach the observed read depth (Illingworth et al., 2017). Due to the loss of viral signals in RCC257-late and the complete loss of the RCC24 strain, I could not conduct an infection assay or purify lysates. However, given the sequence coverage of prasinoviral reads, lack of evidence for green algae in the cultures, and sample processing method, I argue that the SsV vMAGs are indeed directly associated with *S. scintillans*. This is further supported by the fact that the two similar but distinct strains of *S. scintillans* contained two similar but distinct sets of giant virus genomes.

5.3.5 TEM observation of a VLP

A virus-like particle (VLP) from RCC24 was visualized with negative stain TEM (Fig. 5.2A-B). The VLP exhibited an icosahedral shape with a diameter of 192 nm, which is unusually large compared to previously characterized prasinoviruses (Weynberg et al., 2017). However, it fell within the size range (180-240 nm) of the endobacteria described by Guillou et al in 1999 (Guillou et al., 1999). Indeed, the morphology of the “endobacteria” in the original description [Figure 1D in 13] closely matches that of the VLP in Figure 5.2A-B. I did not find VLPs in the actively growing RCC257 strain, as expected as the NCLDV reads were no longer detected in RCC257-late.

The *S. scintillans* “endobacteria” were also described to be located within the endoplasmic reticulum (ER), which continues as perinuclear space of a nuclear envelope (Guillou et al., 1999). This location was emphasized to be potentially relevant for the origin of plastids in deep-branching lineages of stramenopiles and compared to the location of plastids found in photosynthetic lineage of stramenopiles (Hibberd, 1971; Husnik et al., 2021). However, the ER is also a site for viral protein glycosylation (Agarkova et al., 2006), membrane protein folding (Doms, Robert W. et al., 1993), genome replication, and pre-capsid assembly (Inoue and Tsai, 2013; Romero-Brey and Bartenschlager, 2016). Within the Phycodnaviridae, the development of a Phaeovirus infecting *Hinckisa hinckisae* has been observed within the ER, in which viral capsids are derived from the ER membrane (Wolf et al., 1998; Van Etten et al., 2002).

5.3.6 Possible nature of associations: endobacteria, SsVs, and *S. scintillans*

Two decades have passed since the original description of *S. scintillans*, and the present analysis, raising many questions about how to connect data from the original description with data currently at hand. There is no direct evidence to verify the exact nature of the association between SsVs and *S. scintillans* and similarly, there is no way to equate the SsVs to the intracellular inclusions described in 1999. Because the experimental design was to identify endobacteria and because there is no longer any living host-virus pair in culture, experiments such as infection assays, virus-targeted FIHS or PCRs, or thin-section TEM to show virus particles within the cells are not possible. At the same time because there was no sequence data associated with the original genus or endobacteria descriptions, I cannot compare the current data directly with any data from the original description.

There are several possible explanations that formally account for the data, and I will review several in turn here. First, it is possible that inclusions originally described are endobacteria that are still present, but were not detectable in genomic analyses, or belong to one of the normally free-living lineages we did detect. This is not readily consistent with the FISH data, however, and is also not consistent with the genomic observations from most other bacteria endosymbionts of protists (Husnik et al., 2019).

Second, it is also possible the endobacteria were lost and the viruses were acquired later. The idea that the endobacteria may have been lost is not without precedent, since this has been observed in previous cultures (Boscaro et al., 2013), but how the viruses could have been gained is a much more difficult problem. The read-depth in the vMAG assemblies suggest viral DNA was highly represented in these cultures, and by extension these viruses were replicated in the cultures. Since no other eukaryotes were in the cultures, it also suggests the viurses were most likely replicating in *S. scintillans* (since the viruses need some host and no other eukaryote is evidence). Therefore, for the viruses to have been gained after the original description, the two cultures would have to have been exposed to two unrelated but distinct sets of viruses that could each infect and replicate *in S. scintillans*.

Third, it is also possible the viruses have been endogenized within the host genome. This is not obviously consistent with absence of viral reads in some of the sequencing data (Appendix U) or the TEM evidence for viral particles. I also examined this possibility using ViralRecall

(Moniruzzaman et al., 2020; Aylward and Moniruzzaman, 2021; Bellas et al., 2023), which did not detect viral regions with potential host sequences flanking the contig.

Lastly, it is possible that the initially reported endobacteria are actually giant viruses. This possibility is consistent with all the sequencing and FISH analyses, but contrary to the identification of the inclusions made in the original description based on thin section TEM. However, when this was observed the field giant viruses was relatively young, so the only logical identification of a large inclusion in the ER would be a bacterium. In retrospect, many of these TEMs actually resemble giant virus particles, and I observed an extracellular VLP that falls within a similar size range and resembles a shape of the reported endobacteria [compare Fig. 5.2A and B with Figure 1D in Guillou et al., 1999]. However, as noted above since the cultures are now gone and the data are generally non-overlapping, this possibility cannot obviously be verified either.

Another complication with the last possibility is how to explain the long-term persistence of viruses in these cultures, in particular as it must have been followed after 20 years by a sudden loss of viruses (RCC257-late) or the death of the strain (RCC24). One *O. mediterraneus* culture with a decade-long co-existence with OmV2, was found to be a co-culture of resistant (R) and susceptible (S) strains, where the host showed two reversible phenotype phases that are thought to explain the long-term stability of the system (Yau et al., 2016, 2020). It was hypothesized that the RS-switching may be a common long-term strategy for other NCLDVs-affected hosts, and persistent infection is a known strategy for phaeoviruses, a close relative of prasinoviruses (Delaroque et al., 1999; Van Etten et al., 2002; Stevens et al., 2014). Some resistant hosts have been observed to produce infective viruses without typical lytic events (Thomas et al., 2011; Yau et al., 2016), reminiscent of the fact that no prasinovirus reads and hallmark genes were detected in the RCC257 transcriptome data [also to Herpesvirales (Goodrum and McWeeney, 2018), another dsDNA virus distant related to NCLDVs]. When susceptible and different types of resistant cells (R^P vs. R^{NP}: viral-producing vs. non-producing) were cloned and co-cultivated, the viruses were eventually eliminated in the co-cultivated R^P and R^{NP} culture while, susceptible cells became dominant in the S and R^{NP} co-cultivated culture (Thomas et al., 2011).

To examine the possibility that virophages are involved in the host-virus dynamic, I searched for virophage genes or virophage-like elements (VLEs) (Yutin et al., 2013; Blanc et al., 2015) in the initial assembly without taxonomically filtering out scaffolds, due to some virophage genes being recombinant, horizontally transferred, or homologs that are shared with cellular organisms or transposable elements (i.e., polintons), and NCLDVs (Yutin et al., 2013). I detected OLV2 (an uncharacterized protein) only in RCC257 WGA, forming a sister lineage to Yellowstone Lake virophage 1 (YSLV1) (Gong et al., 2016). Although this result is insufficient to conclude the involvement of virophages or VLEs in this chapter, deeper sequencing and assembly of the *S. scintillans* genome could potentially verify the presence and nature of virophages or VLEs association.

I suspect that the lack of prasinovirus reads in RCC24-jp is due to long-term cryopreservation. For example, in the *Paramecium bursaria chlorella* virus (PBCV-1), the strength of infectivity decreased upon cryopreservation and more so if the samples were frozen shortly after post-infection (Nagasaki and Yamaguchi, 1999; Coy et al., 2019). Whether this observation is based on differences in host strains or SsVs, or a combination of both, characterization of host genomes along with further searches of prasinovirus in non-Mamiellophycean hosts should provide insights into the dynamics of persistent infection.

The current observations echo the first discovery of the mimivirus, which was initially described as “Chlamydia-like obligate parasites” in an amoeba (Birtles et al., 1997). It took six years to correctly characterize the parasites as Mimivirus (Scola et al., 2003). Conversely, the bacterium *Chromulinavorax destructans* (Deeg et al., 2019) was recently been described as a bacterial parasite of *Spumella elongata* (a photosynthetic stramenopile), but it was initially studied as a putative giant virus due to a replicating morphology resembling a viral factory of some giant viruses. Both these cases illustrate how difficult it can be to identify the nature of an intracellular symbiont, suggesting that more studies on the diversity of symbioses in heterotrophic nano- or pico-flagellates should yield more such surprises and taxonomic re-assignments of many symbionts will also follow.

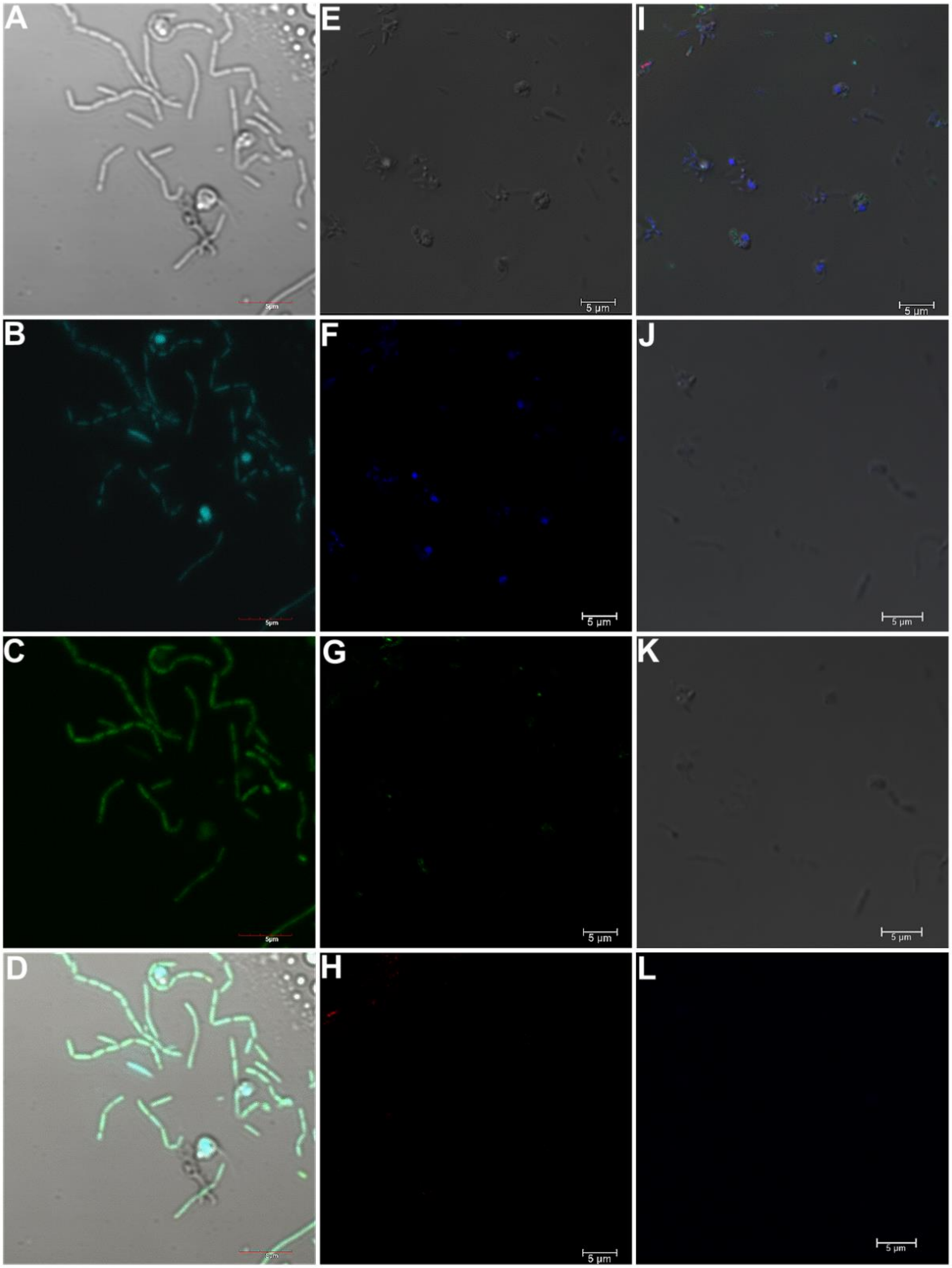


Figure 5.1 FISH analyses on *S. scintillans* cultures

Symbiomonas scintillans RCC257 (A-D) and RCC257-jp (E-L) showing no endobacterial signals. (A), (E) and (J) Brightfield; (B) and (F) DAPI; (C) and (G) EUB388 probe under 473 and 488 nm; (D) merged image of (A-C); (H) CF319 probe under 647 nm; (I) merged image of (E-H); (K) merged image of unstained DAPI, CF319 and (L) EUB388 images. Scale bars = 5 μm.

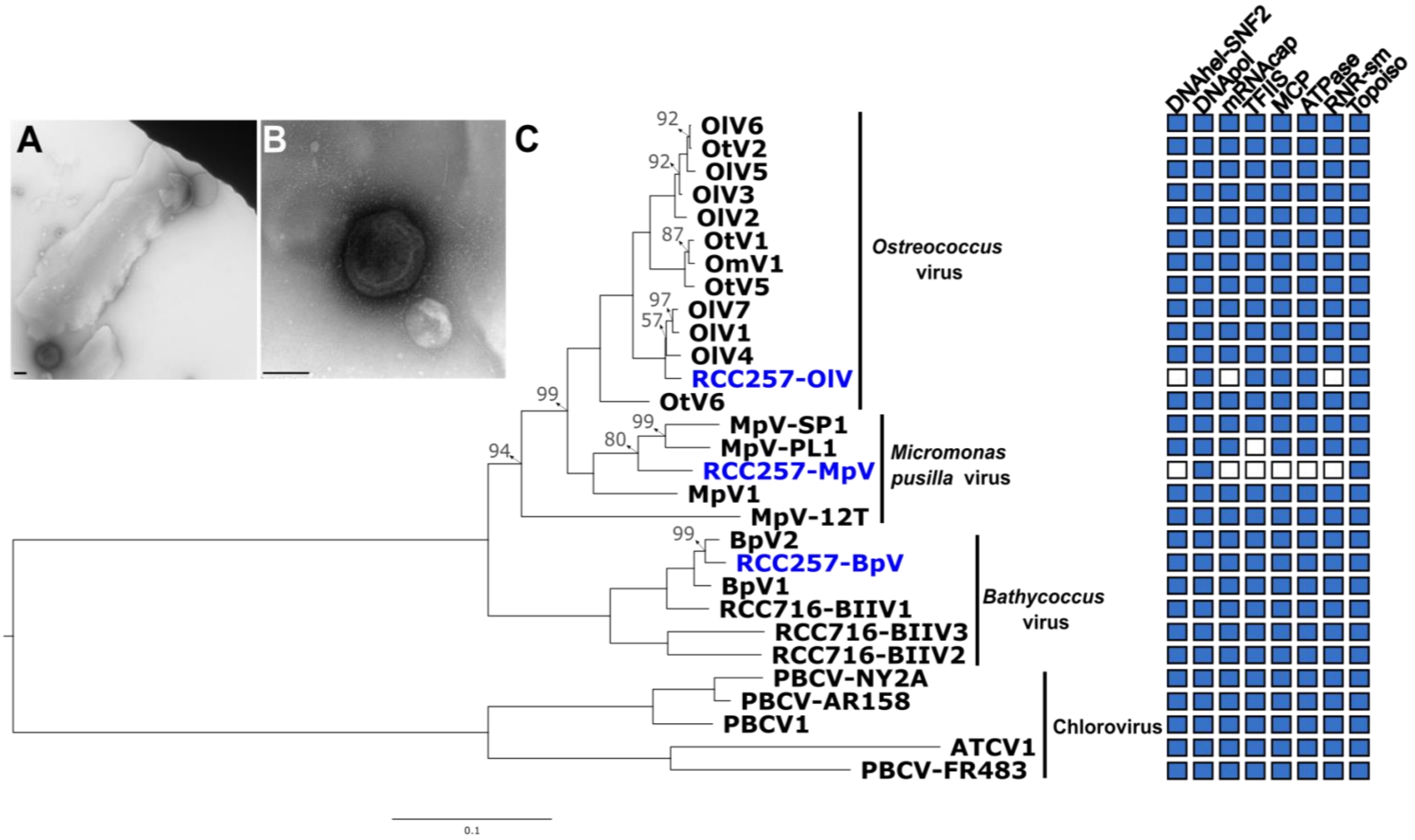


Figure 5.2 Virus-like particle and a multi-gene phylogeny of prasinoviruses

(A-B) Detection of a virus-like particle (VLP) in negatively stained RCC24. (B) Close up of the VLP in (A). The diameter of the VLP is 192 nm. Scale bars = 100 nm. (C) A multi-gene prasinovirus phylogeny reconstructed from 22 core genes (5,213 sites) using IQ-TREE2 LG+F+G4 model. The right panel shows presence-absence of select core genes. Single-copy genes are DNApol (DNA polymerase B), DNAhel-SNF2 (SNF2 helicase), mRNAcap (mRNA capping enzyme), ATPase, and RNR-sm (RNR small subunit). The tree is rooted with Chlorovirus (PBCVs and ATCV) for visualization. Only nodes <100% ultrafast bootstrap supports are labelled.

OlV=*Ostreococcus lucimarinus* virus; OtV=*Ostreococcus tauri* virus; OmV=*Ostreococcus mediterraneus* virus; MpV=*Micromonas pusilla* virus; BpV=*Bathycoccus prasino* virus; BIIV=*Bathycoccus* sp. virus clade BII. PBCV=*Paramecium bursaria chlorella* virus; ATCV=*Acanthocystis turfaceae chlorella* virus.

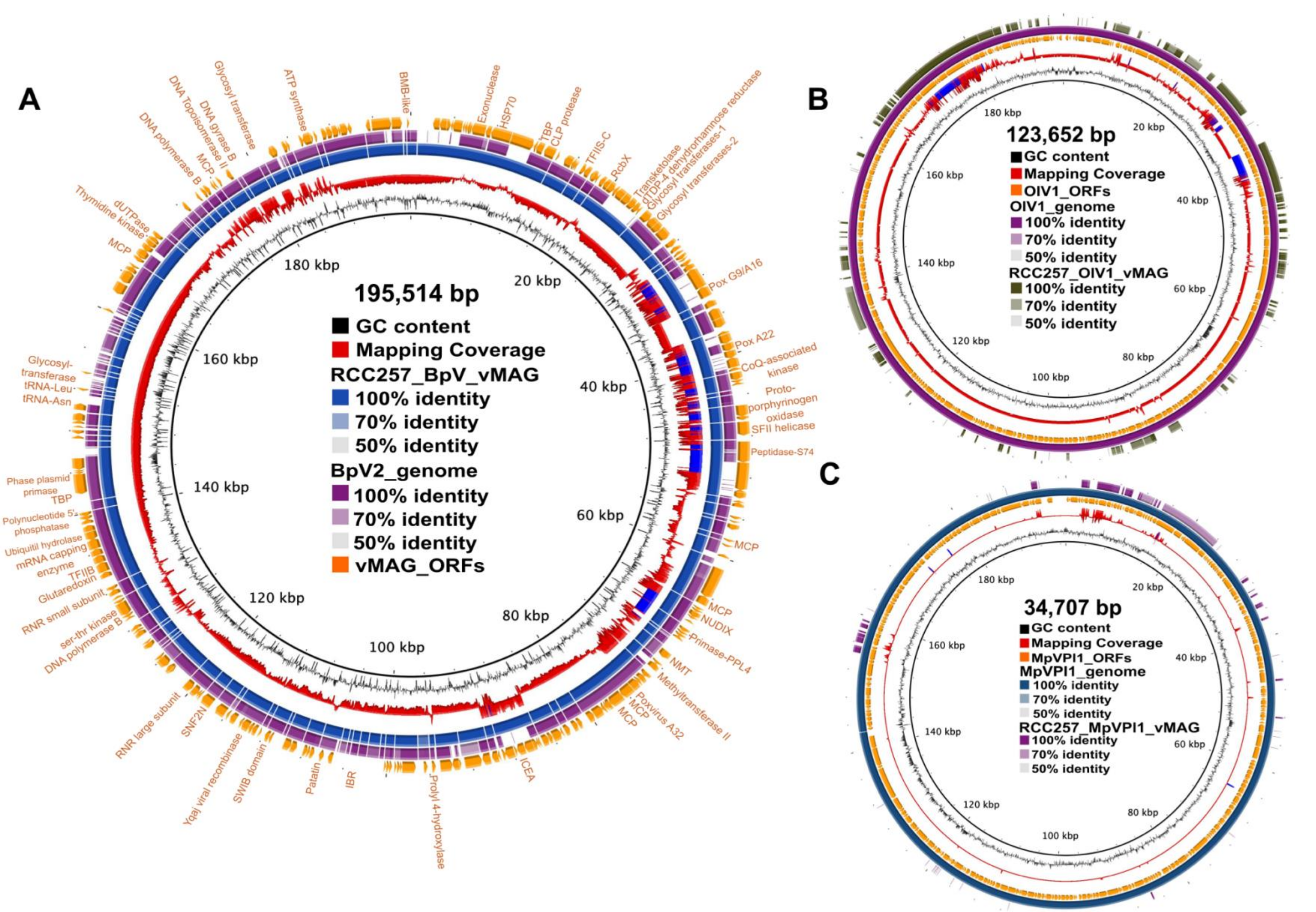

A
195,514 bp
GC content
Mapping Coverage
RCC257_BpV_vMAG
100% identity
70% identity
50% identity
BpV2_genome
100% identity
70% identity
50% identity
vMAG_ORFs
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
180 kbp
BMB-like
Exonuclease
HSP70
TBP
CLP protease
TFIIS-C
RcbX
Transketolase
dTDP-4 dehydrorhamnose reductase
Glycosyl transferases-1
Glycosyl transferases-2
Pox G9/A16
Pox A22
CoQ-associated kinase
Proto-porphyrinogen oxidase
SFII helicase
Peptidase-S74
MCP
MCP
NUDIX
Primase-PPL4
NMT
Methyltransferase II
Poxvirus A32
MCP
MCP
ICEA
Prolyl 4-hydroxylase
IBR
Patatin
SWIB domain
Yqaj viral recombinase
SNF2N
RNR large subunit
DNA polymerase B
ser-thr kinase
RNR small subunit
Glutaredoxin
TFIIB
mRNA capping enzyme
Ubiquitil hydrolase
Polynucleotide 5' phosphatase
TBP
Phase plasmid primase
tRNA-Asn
tRNA-Leu
Glycosyl-transferase
MCP
Thymidine kinase
dUTPase
DNA polymerase B
MCP
DNA Topoisomerase II
DNA gyrase B
Glycosyl transferase
ATP synthase
B
123,652 bp
GC content
Mapping Coverage
OlV1_ORFs
OlV1_genome
100% identity
70% identity
50% identity
RCC257_OlV1_vMAG
100% identity
70% identity
50% identity
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
180 kbp
C
34,707 bp
GC content
Mapping Coverage
MpVPl1_ORFs
MpVPl1_genome
100% identity
70% identity
50% identity
RCC257_MpVPl1_vMAG
100% identity
70% identity
50% identity
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
180 kbp

Figure 5.3 Circularized draft vMAGs overview

Genome overview and comparison of the most complete BpV-, OlV-, and MpV-vMAGs to corresponding reference genomes, with the size of vMAGs labelled in the centre. (A) Circularized representation of (A) RCC257 BpV-vMAG compared to BpV2 genome; (B) OlV1 genome compared to RCC257 OlV1-vMAG; (C) MpV_Pl1 genome compared to MpVPl1_vMAG, in an ordered set of coding sequences, represented by blocks shaded by similarity. (A) Mapping coverage is based on RCC257 BpV-vMAG mapped to viral-subset-scaffolds and regions with the coverage more than one standard deviation [62.1] from the mean coverage [50.8] are shown in blue spikes. The outermost ring represents predicted ORFs of the vMAG with manually annotated protein from Prodigal-gv and Viralrecall. (B) Mapping coverage is based on OlV1 genome mapped to viral-subset-scaffolds and regions with the coverage more than one standard deviation [8.4] from the mean coverage [3.1] shown in blue spikes. Only ORFs from the reference OlV1 genome is shown and the partial RCC257 OlV1-vMAG CDS are shown in the outermost ring. (C) Mapping coverage is based on MpV_Pl1 genome mapped to viral-subset-scaffolds and regions with the coverage more than one standard deviation [7.8] from the mean coverage [0.6] shown in blue spikes. Only ORFs from the reference OlV1 genome is shown and the partial RCC257 OlV1-vMAG CDS are shown in the outermost ring. See Table S2 for annotation in a tabular format.

Chapter 6: Conclusion

6.1 Major findings and significance

In Chapter 2, I processed transcriptomes of seven new species belonging to Bigyromonadea, a poorly-understood group which was previously represented by a single taxon in phylogenomic analyses (Kühn et al., 2004; Leonard et al., 2018; Noguchi et al., 2016; Thakur et al., 2019; Susan M. Tong, 1995). Along with the new transcriptomes, I updated phylogenomic dataset for stramenopiles by curating publicly available transcriptome and genome data (Appendix D). Using this dataset, I inferred a phylogenomic tree that recovered well-supported sister-group relationships between the two Bigyromonadea subgroups, Developea and Pirsoniales. These relationships had never been observed in phylogenetic trees inferred from a handful of genes, which did not support this clade (Aleoshin et al., 2016; Kühn et al., 2004). Bigyromonadea in turn, are a sister-lineage to oomycetes. Additionally, together with collaborators, I described morphologies and behaviours of the seven new bigyromonads, in which some of the Developea species were able to form cell-aggregates that occasionally fused. Some of the new bigyromonads were also able to form pseudopods, while the zoospores of the newly described Pirsoniales were able to actively feed on smaller eukaryotic prey. Together with the phylogenomic data and morphological and the behavioural observations, I hypothesized that the last common ancestor of the oomycetes may have looked more similar to Bigyromonadea, and were likely phagotrophic amoeboids.

In Chapter 3, I further updated the phylogenomic data of stramenopiles by processing and generating transcriptomes of six species belonging to Sagenista and Opalozoa, including one new species of Placididea and three new benthic species of MAST-6 (Sagenista). MAST-6 is one of the few MArine STramenopiles (MASTs) with an ultrastructure description and available genomic level data (Shiratori et al., 2017). This group is associated with sediments, but has unknown phylogenetic diversity (Massana et al., 2015). By searching for SSU rRNA gene sequences of the newly characterized MAST-6 species in several sediment amplicon datasets, I found that they are not only abundant in sediments but are phylogenetically more diverse than their sister lineage, MAST-4, one of the most abundant groups in the open ocean (Rodriguez-Martinez et al., 2012; Rodríguez-Martínez et al., 2009; Thakur et al., 2019). Additionally, I observed high relative abundance of SSU sequences similar to one of the newly described MAST-6 species, *Mastreximonas tlaamin* in most of the sediment datasets. For phylogenetic trees including SSU sequences of the new Placididea species, *Haloplacidia sinai,* it is depicted as a sister-lineage to a previously described clade ("Group-D") that consists of a number of species that can tolerate a wide range of salinities (Park and Simpson, 2010). Based on this relationship and the location of its isolation in the Red Sea, *H. sinai* also likely tolerates a broad range of salinities. Chapter 3 also provides an updated stramenopile phylogenomic tree by adding publicly available MAST-1, -7, -8, -9, and MAST-11 in addition to the new transcriptomes generated in this study. Together with phylogenomic data and the distribution of some of the Sagenista described in this study, I conclude that Bigyra is indeed paraphyletic and some clades showed phylogenetic parallelism with niche occupation.

In Chapter 4, I focused on the ochrophytes phylogenomic dataset (supermatrix) and conducted phylogenomic analyses using various inference methods and gene selection criteria. This was done by obtaining under-represented ochrophyte classes from publicly available culture collections and generating new transcriptome data. I additionally added publicly available ochrophyte genomic or transcriptomic data to the supermatrix to "break" long branches of many clades previously represented by one or two taxa in phylogenomic analyses. The resulting phylogenomic tree represented four classes that had only been included in a single phylogenomic analysis (Terpis, 2021, unpublished data). The inferred phylogenomic tree recovered robust relationships of Eustigmatophyceae + RPX and Pinguiophyceae + Olisthodiscophyceae, whose relationships were either contentious or had rarely been phylogenomically analysed prior to this study. To further investigate and resolve other lineages that were incongruent between the Bayesian and the maximum-likelihood tree, I quantified various gene properties considered to represent phylogenetic signal or noise using a previously established method (Mongiardino Koch, 2021; Mongiardino Koch and Thompson, 2021). When alternative supermatrices were concatenated with different combinations of genes using various gene properties as filtering criteria, selecting genes with high signal and data quality resulted in the similar topology as the initial phylogenomic tree, but selecting genes with low noise resulted more unstable lineages. This study not only provides the most up-to-date stramenopiles phylogenomic tree while resolving some contentious relationships, but also hints at a potential way to sub-sample a supermatrix to the least number of genes to reduce the computational resources and time – both major hurdles in phylogenomic analyses.

While stramenopiles endosymbionts of eukaryotes are well documented, this group has barely been investigated for being potential endosymbiont hosts, particularly among free-living heterotrophic lineages. In Chapter 5, I investigated the only reported case of putative endobacteria among free-living heterotrophic stramenopiles, *S. scintillans*. However, even with extensive genomic sequencing from several strains I found no evidence of a bacterial endosymbiont. The absence of endobacteria was further confirmed with multiple eubacterial FISH probes. Instead, I recovered and assembled up to three viral genomes, and detected a VLP with similar size and appearance to the “endobacteria” described by Guillou *et al.* (1999). Based on this finding, I proposed persistent infection of SsVs in *S. scintillans*, although the viruses were later lost in all my cultures, either because the strain lost the virus or the susceptible strain itself was lost. Host genome assemblies and further search on SsVs in a broader host range, particularly focusing on pico- or nano-flagellates, will not only help us understand persistent infection of NCLDVs, but also broaden our view on endosymbiosis in free-living heterotrophs.

6.2 Future direction and outstanding questions

6.2.1 Character evolution and niche occupation of stramenopiles

An extensive update on the stramenopile phylogenomics in this thesis resulted in 23 new transcriptomes, which were mostly generated from newly described species and previously under-represented groups. In addition to updating our knowledge on the stramenopile phylogeny, the work from my thesis can be used to address the functional evolution of stramenopiles. For example, an amoeboid morphology and saprotrophic mode of feeding can be found across different groups of stramenopiles. Two ochrophyte species, *Leukarachnion* sp. (Hibberd, 1971; Jaške et al., 2022) and *Chrysamoeba radians,* are amoeboids that can form net-like pseudopods, both traits resembling oomycetes and labyrinthulids. Some of the newly described Bigyromonadea in this study were also able to fuse with each other and form pseudopodia. Other groups of stramenopiles seem to have evolved adaptations to specific habitats where they are found. For example, all the newly described MAST-6 species from this study were benthic and shown to be abundant in sediment samples. Many species belonging to Placididea on the other hand, have been cultured in a wide range of salinities, many isolated from hypersaline conditions.

Identifying genes that are linked to the amoeboid morphology, pseudopodia formation, or their habitats in stramenopiles would enhance our current understanding of the character evolution of stramenopiles. Exploring the gene expression patterns from each group with these features can bring us closer to elucidating the characteristics of the the last common ancestor of stramenopiles. Using transcriptome analyses, this type of approach has been used to infer characteristics of the last common ancestor of fungi as a phagotroph (Torruella et al., 2018), and to identify genes that were highly responsive during highly saline conditions in *Halocafeteria seosinensis* (Opalozoa; Bigyra).

6.2.2 Minimum number of genes for a phylogenomic tree

Phylogenomic analyses have revolutionized our understanding of eukaryotic phylogeny and evolutionary history (Delsuc et al., 2005; Burki et al., 2016; Keeling and Burki, 2019). However, the problem with a heavy computational burden will only increase as I generate data from more and more of the phylogenetic diversity of protists. To reduce the computational burden, I searched for phylogenetically informative genes that can be selected to build a smaller supermatrix with decreased phylogenetic noise. Likely due to the rapid ancient diversification of stramenopiles, I was unable to identify phylogenetically informative genes that would be sufficient to replace the supermatrix consisting of hundreds of genes. However, this work provided several findings that can bring us a step closer to subsampling a supermatrix: (1) selecting for the slowest evolving genes omits significant phylogenetic information and this can especially impact deep-branching lineages; (2) selecting genes with long internal branches provide less variable topologies among multiple trees and can be informative for recently and/or rapidly diverged group; (3) removing genes with the most apparent biases yielded the same topology as the phylogenomic tree inferred from an unfiltered supermatrix. Future studies can expand on this work to investigate rate-driven phylogenetic information and establish the maximum number of phylogenetically “noisy” genes that can be removed to lessen the computational burden for likelihood-based tree inference methods.

6.2.3 Verifying persistent infection in *Symbiomonas scintillans*

Finally, I investigated the only reported case of endobacteria within a non-photosynthetic lineage of stramenopile, *Symbiomonas scintillans*, and found no evidence of endobacteria but instead three giant viruses. I observed this in two different strains of *S. scintillans* with different outcomes. Strain RCC257 appears to have lost the viruses, since the culture grew faster over time, and eventually no viral contigs could be found in the WGA data. Conversely, strain RCC24 appears to have perished, leaving only viral sequences. Due to fact that the initial experimental design was to search for endobacteria, I was not able to purify viral lysates and isolate viruses. Therefore, future work should address this by designing an experiment that involves infecting the culture with closely related giant viruses (e.g., three different groups of prasinoviruses) and investigating whether the virus can reside within the cell without actively replicating (i.e., persistent infection). To consider the possibility of the resistant and susceptible strategy employed by the host, and an involvement of a virophage in *S. scintillans*, a close monitoring of culture strains upon re-infection should be done, accompanied by electron micrograph imaging, transcriptome, and host genome analyses. If successful, *S. scintillans* can be suitable a model organism for studying persistent infection and immunity in protists.

References

Agarkova, I. V., D. D. Dunigan, and J. L. Van Etten. 2006. Virion-Associated Restriction Endonucleases of Chloroviruses. *Journal of Virology* 80: 8114–8123.

Alavi, M., T. Miller, K. Erlandson, R. Schneider, and R. Belas. 2001. Bacterial community associated with Pfiesteria-like dinoflagellate cultures. *Environmental Microbiology* 3: 380–396.

Aleoshin, V. V., A. P. Mylnikov, G. S. Mirzaeva, K. V. Mikhailov, and S. A. Karpov. 2016. Heterokont predator Develorapax marinus gen. et sp. nov. - A model of the ochrophyte ancestor. *Frontiers in Microbiology* 7: 1–14.

Ali, R. H., M. Bogusz, and S. Whelan. 2019. Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments K. Tamura [ed.],. *Molecular Biology and Evolution* 36: 2340–2351.

Alikhan, N.-F., N. K. Petty, N. L. Ben Zakour, and S. A. Beatson. 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12: 402.

Alonge, M., L. Lebeigle, M. Kirsche, K. Jenike, S. Ou, S. Aganezov, X. Wang, et al. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* 23: 258.

Alonge, M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin, F. J. Sedlazeck, Z. B. Lippman, and M. C. Schatz. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology* 20: 224.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215: 403–410.

Amaral, R., K. P. Fawley, Y. Němcová, T. Ševčíková, A. Lukešová, M. W. Fawley, L. M. A. Santos, and M. Eliáš. 2020. Toward Modern Classification of Eustigmatophytes, Including the Description of Neomonodaceae Fam. Nov. and Three New Genera [1] O. De Clerck [ed.],. *Journal of Phycology* 56: 630–648.

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

Andersen, R. A., D. Potter, R. R. Bidigare, M. Latasa, K. Rowan, and C. J. O'Kelly. 1998. Characterization and phylogenetic position of the enigmatic golden alga Phaeothamnion confervicola: ultrastructure, pigment composition and partial SSU rDNA sequence. *Journal of Phycology* 34: 286–298.

Andersen, R. A., G. W. Saunders, M. P. Paskind, and J. P. Sexton. 1993. Ultrastructure and 18s rRNA gene sequence for Pelagomonas calceolata gen. et sp. nov. and the description of a new alagal class, the Pelagophyceae class nov. *Journal of Phycology* 29: 701–715.

Anderson, O. R., and T. Cavalier-Smith. 2012. Ultrastructure of Diplophrys parva, a New Small Freshwater Species, and a Revised Analysis of Labyrinthulea (Heterokonta). *Acta Protozoologica* 51: 291–304.

Andrews, S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online].

Aponte, A., Y. Gyaltshen, J. A. Burns, A. A. Heiss, E. Kim, and S. D. Warring. 2021. The Bacterial Diversity Lurking in Protist Cell Cultures. *American Museum Novitates* 2021.

Apt, K. E., S. K. Clendennen, D. A. Powers, and A. R. Grossman. 1995. The gene family encoding the ucoxanthin chlorophyll proteins from the brown alga Macrocystis pyrifera. *Molecular and General Genetics MGG* 246: 455–464.

Aylward, F. O., and M. Moniruzzaman. 2021. ViralRecall—A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. *Viruses* 13: 150.

Azuma, T., T. Pánek, A. K. Tice, M. Kayama, M. Kobayashi, H. Miyashita, T. Suzaki, et al. 2022. An Enigmatic Stramenopile Sheds Light on Early Evolution in Ochrophyta Plastid Organellogenesis H. Hendrickson [ed.],. *Molecular Biology and Evolution* 39: msac065.

Bachy, C., C. J. Charlesworth, A. M. Chan, J. F. Finke, C. Wong, C. Wei, S. Sudek, et al. 2018. Transcriptional responses of the marine green alga Micromonas pusilla and an infecting prasinovirus under different phosphate conditions. *Environmental Microbiology* 20: 2898–2912.

Bachy, C., C. C. M. Yung, D. M. Needham, M. C. Gazitúa, S. Roux, A. J. Limardo, C. J. Choi, et al. 2021. Viruses infecting a warm water picoeukaryote shed light on spatial co-occurrence dynamics of marine viruses and their hosts. *The ISME Journal* 15: 3129–3147.

Bairoch, A. 2004. The Universal Protein Resource (UniProt). *Nucleic Acids Research* 33: D154–D159.

Baños, H., E. Susko, and A. J. Roger. 2023. Is Over-parameterization a Problem for Profile Mixture Models? *Systematic Biology* syad063.

Bapteste, E., E. Susko, J. Leigh, I. Ruiz-Trillo, J. Bucknam, and W. F. Doolittle. 2007. Alternative Methods for Concatenation of Core Genes Indicate a Lack of Resolution in Deep Nodes of the Prokaryotic Phylogeny. *Molecular Biology and Evolution* 25: 83–91.

Barbera, P., A. M. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri, and A. Stamatakis. 2018. Data from: EPA-ng: massively parallel evolutionary placement of genetic sequences. 16579930 bytes.

Barcytė, D., W. Eikrem, A. Engesmo, S. Seoane, J. Wohlmann, A. Horák, T. Yurchenko, and M. Eliáš. 2021. *Olisthodiscus* represents a new class of Ochrophyta O. De Clerck [ed.],. *Journal of Phycology* 57: 1094–1118.

Basak, S., M. N. Rajurkar, and S. K. Mallick. 2014. Detection of Blastocystis hominis: A controversial human pathogen. *Parasitology Research* 113: 261–265.

Bassani, I., C. Rancurel, S. Pagnotta, F. Orange, N. Pons, K. Lebrigand, F. Panabières, et al. 2020. Transcriptomic and Ultrastructural Signatures of K+-Induced Aggregation in Phytophthora parasitica Zoospores. *Microorganisms* 8: 1012.

Baudoux, A.-C., H. Lebredonchel, H. Dehmer, M. Latimier, R. Edern, F. Rigaut-Jalabert, P. Ge, et al. 2015. Interplay between the genetic clades of Micromonas and their viruses in the Western English Channel. *Environmental Microbiology Reports* 7: 765–773.

Beakes, G. W., S. L. Glockling, and S. Sekimoto. 2012. The evolutionary phylogeny of the oomycete "fungi". *Protoplasma* 249: 3–19.

Beakes, G. W., D. Honda, and M. Thines. 2014. Systematics of the Straminipila: Labyrinthulomycota, Hyphochytriomycota, and Oomycota. The Mycota: Fungal Taxonomy and Systematics, 39–97. Springer Verlag, Germany.

Beakes, G. W., and S. Sekimoto. 2009. The Evolutionary Phylogeny of Oomycetes—Insights Gained from Studies of Holocarpic Parasites of Algae and Invertebrates. *In* K. Lamour, and S. Kamoun [eds.], Oomycete Genetics and Genomics, 1–24. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Beakes, G. W., and M. Thines. 2017. Hyphochytriomycota and Oomycota. Handbook of the Protists., 435–505. Springer Verlag, Cham, Switzerland.

Beisser, D., N. Graupner, C. Bock, S. Wodniok, L. Grossmann, M. Vos, B. Sures, et al. 2017. Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. *PeerJ* 5: e2832.

Bellas, C., T. Hackl, M.-S. Plakolb, A. Koslová, M. G. Fischer, and R. Sommaruga. 2023. Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses. *Proceedings of the National Academy of Sciences* 120: e2300465120.

Bellec, L., C. Clerissi, R. Edern, E. Foulon, N. Simon, N. Grimsley, and Y. Desdevises. 2014. Cophylogenetic interactions between marine viruses and eukaryotic picophytoplankton. *BMC Evolutionary Biology* 14: 59.

Berger, S. A., D. Krompass, and A. Stamatakis. 2011. Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology* 60: 291–302.

Birtles, R., T. Rowbotham, C. Storey, T. Marrie, and D. Raoult. 1997. Chlamydia-like obligate parasite of free-living amoebae. *The Lancet* 349: 925–926.

Bjorbækmo, M. F. M., A. Evenstad, L. L. Røsæg, A. K. Krabberød, and R. Logares. 2020. The planktonic protist interactome: where do we stand after a century of research? *ISME Journal* 14: 544–559.

Blanc, G., L. Gallot-Lavallée, and F. Maumus. 2015. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proceedings of the National Academy of Sciences* 112.

Bolch, C. J. S., T. A. Bejoy, and D. H. Green. 2017. Bacterial Associates Modify Growth Dynamics of the Dinoflagellate Gymnodinium catenatum. *Frontiers in Microbiology* 8.

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

Bolyen, E., J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. Abnet, G. A. Al-Ghalith, H. Alexander, et al. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints.

Boscaro, V., F. Husnik, C. Vannini, and P. J. Keeling. 2019. Symbionts of the ciliate *Euplotes* : diversity, patterns and potential as models for bacteria–eukaryote endosymbioses. *Proceedings of the Royal Society B: Biological Sciences* 286: 20190693.

Boscaro, V., M. Schrallhammer, K. A. Benken, S. Krenek, F. Szokoli, T. U. Berendonk, M. Schweikert, et al. 2013. Rediscovering the genus Lyticum, multiflagellated symbionts of the order Rickettsiales. *Scientific Reports* 3: 3305.

Boussau, B., S. Blanquart, A. Necsulea, N. Lartillot, and M. Gouy. 2008. Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456: 942–945.

Brown, J. W., and U. Sorhannus. 2010. A Molecular Genetic Timescale for the Diversification of Autotrophic Stramenopiles (Ochrophyta): Substantive Underestimation of Putative Fossil Ages M. T. P. Gilbert [ed.],. *PLoS ONE* 5: e12759.

Brown, M. W., M. Kolisko, J. D. Silberman, and A. J. Roger. 2012. Aggregative Multicellularity Evolved Independently in the Eukaryotic Supergroup Rhizaria. *Current Biology* 22: 1123–1127.

Brown, M. W., J. D. Silberman, and F. W. Spiegel. 2012b. A contemporary evaluation of the acrasids (Acrasidae, Heterolobosea, Excavata). *European Journal of Protistology* 48: 103–123.

Brown, M. W., F. W. Spiegel, and J. D. Silberman. 2009. Phylogeny of the ‘forgotten’ cellular slime mold, Fonticula alba, reveals a key evolutionary branch within Opisthokonta. *Molecular Biology and Evolution* 26: 2699–2709.

Buchfink, B., C. Xie, and D. H. Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12: 59–60.

Burki, F., M. Kaplan, D. V. Tikhonenkov, V. Zlatogursky, B. Q. Minh, L. V. Radaykina, A. Smirnov, et al. 2016. Untangling the early diversification of eukaryotes: A phylogenomic study of the evolutionary origins of centrohelida, haptophyta and cryptista. *Proceedings of the Royal Society B: Biological Sciences* 283.

Burki, F., K. Shalchian-Tabrizi, and J. Pawlowski. 2008. Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biology Letters* 4: 366–369.

Burns, J. A., A. A. Pittis, and E. Kim. 2018. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nature Ecology & Evolution* 2: 697–704.

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

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13: 581–583.

Capella-Gutiérrez, S., J. M. Silla-Martínez, and T. Gabaldón. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.

Caron, D. A. 2000. Symbiosis and mixotrophy among pelagic microorganisms. Microbial Ecology of the Oceans, 495–523. John Wiley and Sons, New York.

Castoe, T. A., A. P. J. De Koning, H.-M. Kim, W. Gu, B. P. Noonan, G. Naylor, Z. J. Jiang, et al. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences* 106: 8986–8991.

Cavalier-Smith, T. 1998. A revised six-kingdom system of life. *Biological Reviews of the Cambridge Philosophical Society* 73: 203–266.

Cavalier-Smith, T. 1995. Evolutionary Protistology Comes of Age: Biodiversity and Molecular Cell Biology. *Archiv für Protistenkunde* 145: 145–154.

Cavalier-Smith, T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *Journal of Eukaryotic Microbiology* 46: 347–366.

Cavalier-Smith, T. 1997. Sagenista and bigyra, two phyla of heterotrophic heterokont chromists. *Archiv für Protistenkunde* 148: 253–267.

Cavalier-Smith, T. 1989. The kingdom Chromista. The Chromophyte Algae: Problems and Perspectives, Systematics Association special volume, 381–407. Oxford Science Publications, New York.

Cavalier-Smith, T., and E. E. Y. Chao. 2006. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). *Journal of Molecular Evolution* 62: 388–420.

Cavalier-Smith, T., R. Lewis, E. E. Chao, B. Oates, and D. Bass. 2009. Helkesimastix marina n. sp. (Cercozoa: Sainouroidea superfam. n.) a Gliding Zooflagellate of Novel Ultrastructure and Unusual Ciliary Behaviour. *Protist* 160: 452–479.

Cavalier-Smith, T., and J. M. Scoble. 2013. Phylogeny of Heterokonta: Incisomonas marina, a uniciliate gliding opalozoan related to Solenicola (Nanomonadea), and evidence that Actinophryida evolved from raphidophytes. *European Journal of Protistology* 49: 328–353.

Challis, R., E. Richards, J. Rajan, G. Cochrane, and M. Blaxter. 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. *Journal of Marine Science and Engineering* 3: 401–411.

Cho, A., D. V. Tikhonenkov, E. Hehenberger, A. Karnkowska, A. P. Mylnikov, and P. J. Keeling. 2022. Monophyly of diverse Bigyromonadea and their impact on phylogenomic relationships within stramenopiles. *Molecular Phylogenetics and Evolution* 171: 107468.

Cho, A., D. V. Tikhonenkov, G. Lax, K. I. Prokina, and P. J. Keeling. 2024. Phylogenomic position of genetically diverse phagotrophic stramenopile flagellates in the sediment-associated MAST-6 lineage and a potentially halotolerant placididean. *Molecular Phylogenetics and Evolution* 190: 107964.

Choi, S.-W., L. Graf, J. W. Choi, J. Jo, G. H. Boo, H. Kawai, C. G. Choi, et al. 2024. Ordovician origin and subsequent diversification of the brown algae. *Current Biology* 34: 740-754.e4.

Chow, J., H. M. Dionne, A. Prabhakar, A. Mehrotra, J. Somboonthum, B. Gonzalez, M. Edgerton, and P. J. Cullen. 2019. Aggregate Filamentous Growth Responses in Yeast Y.-S. Bahn [ed.],. *mSphere* 4: e00702-18.

Clerissi, C., N. Grimsley, H. Ogata, P. Hingamp, J. Poulain, and Y. Desdevises. 2014. Unveiling of the Diversity of Prasinoviruses (Phycodnaviridae) in Marine Samples by Using High-Throughput Sequencing Analyses of PCR-Amplified DNA Polymerase and Major Capsid Protein Genes K. E. Wommack [ed.],. *Applied and Environmental Microbiology* 80: 3150–3160.

Collado-Mercado, E., J. Radway, and J. Collier. 2010. Novel uncultivated labyrinthulomycetes revealed by 18S rDNA sequences from seawater and sediment samples. *Aquatic Microbial Ecology* 58: 215–228.

Corsaro, D., R. Michel, J. Walochnik, K.-D. Müller, and G. Greub. 2010. Saccamoeba lacustris, sp. nov. (Amoebozoa: Lobosea: Hartmannellidae), a new lobose amoeba, parasitized by the novel chlamydia 'Candidatus Metachlamydia lacustris' (Chlamydiae: Parachlamydiaceae). *European Journal of Protistology* 46: 86–95.

Corsaro, D., K.-D. Müller, J. Wingender, and R. Michel. 2013. "Candidatus Mesochlamydia elodeae" (Chlamydiae: Parachlamydiaceae), a novel chlamydia parasite of free-living amoebae. *Parasitology Research* 112: 829–838.

Coy, S. R., A. N. Alsante, J. L. Van Etten, and S. W. Wilhelm. 2019. Cryopreservation of Paramecium bursaria Chlorella Virus-1 during an active infection cycle of its host S. A. Wood [ed.],. *PLOS ONE* 14: e0211755.

Czech, L., P. Barbera, and A. Stamatakis. 2020. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data R. Schwartz [ed.],. *Bioinformatics* 36: 3263–3265.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure, 5, 345–352. National Biomedical Research Foundation, Washington DC.

Deeg, C. M., M. M. Zimmer, E. E. George, F. Husnik, P. J. Keeling, and C. A. Suttle. 2019. Chromulinavorax destructans, a pathogen of microzooplankton that provides a window into the enigmatic candidate phylum Dependentiae E. A. McGraw [ed.],. *PLOS Pathogens* 15: e1007801.

Del Campo, J., F. Not, I. Forn, M. E. Sieracki, and R. Massana. 2013. Taming the smallest predators of the oceans. *The ISME Journal* 7: 351–358.

Delaroque, N., I. Maier, R. Knippers, and D. G. M√°ller. 1999. Persistent virus integration into the genome of its algal host, Ectocarpus siliculosus (Phaeophyceae). *Journal of General Virology* 80: 1367–1370.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* 6: 361–375.

Derelle, E., A. Monier, R. Cooke, A. Z. Worden, N. H. Grimsley, and H. Moreau. 2015. Diversity of Viruses Infecting the Green Microalga Ostreococcus lucimarinus R. M. Sandri-Goldin [ed.],. *Journal of Virology* 89: 5812–5821.

Derelle, R., P. López-García, H. Timpano, and D. Moreira. 2016. A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Molecular Biology and Evolution* 33: 2890–2898.

Di Franco, A., D. Baurain, G. Glöckner, M. Melkonian, and H. Philippe. 2022. Lower statistical support with larger datasets: insights from the Ochrophyta radiation. *Molecular Biology and Evolution* 39: msab300.

Dick, M. W. 2000. Straminipilous fungi: systematics of the peronosporomycetes, including accounts of the marine straminipilous portests, the plasmodiophorids, and similar organisms. 1st ed. Spring Netherlands, Netherlands.

Doms, Robert W., Lamb, Robert A., Rose, John K., and Helenius, Ari. 1993. Minireview: Folding and assembly of viral membrane proteins. *Virology* 193: 545–562.

Dong, S., Y. Wang, N. Xia, Y. Liu, M. Liu, L. Lian, N. Li, et al. 2022. Plastid and nuclear phylogenomic incongruences and biogeographic implications of *Magnolia* s.l. (Magnoliaceae). *Journal of Systematics and Evolution* 60: 1–15.

Dorrell, R. G., T. Azuma, M. Nomura, G. Audren De Kerdrel, L. Paoli, S. Yang, C. Bowler, et al. 2019. Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. *Proceedings of the National Academy of Sciences* 116: 6914–6923.

Dorrell, R. G., and A. G. Smith. 2011. Do Red and Green Make Brown?: Perspectives on Plastid Acquisitions within Chromalveolates. *Eukaryotic Cell* 10: 856–868.

Dorrell, R. G., A. Villain, B. Perez-Lamarque, G. Audren De Kerdrel, G. McCallum, A. K. Watson, O. Ait-Mohamed, et al. 2021. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. *Proceedings of the National Academy of Sciences* 118: e2009974118.

Driskell, A. C., C. Ané, J. G. Burleigh, M. M. McMahon, B. C. O'Meara, and M. J. Sanderson. 2004. Prospects for Building the Tree of Life from Large Sequence Databases. *Science* 306: 1172–1174.

Du, Q., Y. Kawabe, C. Schilde, Z. Chen, and P. Schaap. 2015. The Evolution of Aggregative Multicellularity and Cell–Cell Communication in the Dictyostelia. *Journal of Molecular Biology* 427: 3722–3733.

Dunthorn, M., J. Otto, S. A. Berger, A. Stamatakis, F. Mahé, S. Romac, C. De Vargas, et al. 2014. Placing Environmental Next-Generation Sequencing Amplicons from Microbial Eukaryotes into a Phylogenetic Context. *Molecular Biology and Evolution* 31: 993–1009.

Dykstra, M. J., and L. S. Olive. 1975. Sorodiplophrys: An Unusual Sorocarp-Producing Protist. *Mycologia* 67: 873–879.

Edwards, S. V. 2009. Natural selection and phylogenetic analysis. *Proceedings of the National Academy of Sciences* 106: 8799–8800.

Edwards, S. V. 2016. Phylogenomic subsampling: a brief review. *Zoologica Scripta* 45: 63–74.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systemic Biology* 27: 401–410.

Fenchel, T., and B. J. Finlay. 1991. Endosymbiotic Methanogenic Bacteria In Anaerobic Ciliates: Significance For the Growth Efficiency of the Host. *The Journal of Protozoology* 38: 18–22.

Finke, J., D. Winget, A. Chan, and C. Suttle. 2017. Variation in the Genetic Repertoire of Viruses Infecting Micromonas pusilla Reflects Horizontal Gene Transfer and Links to Their Environmental Distribution. *Viruses* 9: 116.

Forster, D., M. Dunthorn, F. Mahé, J. R. Dolan, S. Audic, D. Bass, L. Bittner, et al. 2016. Benthic protists: the under-charted majority J. Olson [ed.],. *FEMS Microbiology Ecology* 92: fiw120.

Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150–3152.

Galiana, E., S. Fourr, and G. Engler. 2008. *Phytophthora parasitica* biofilm formation: installation and organization of microcolonies on the surface of a host plant. *Environmental Microbiology* 10: 2164–2171.

Gallot-Lavallée, L., and G. Blanc. 2017. A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. *Viruses* 9: 17.

Giannotti, D., V. Boscaro, F. Husnik, C. Vannini, and P. J. Keeling. 2022. The “Other” *Rickettsiales* : an Overview of the Family “ *Candidatus* Midichloriaceae” K. N. Johnson [ed.],. *Applied and Environmental Microbiology* 88: e02432-21.

Gomaa, F., E. A. D. Mitchell, and E. Lara. 2013. Amphitremida (Poche, 1913) Is a New Major, Ubiquitous Labyrinthulomycete Clade K. A. Crandall [ed.],. *PLoS ONE* 8: e53046.

Gómez, F., D. Moreira, K. Benzerara, and P. López-García. 2011. *Solenicola setigera* is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3: *Solenicola* belongs to uncultured stramenopiles MAST-3. *Environmental Microbiology* 13: 193–202.

Gong, C., W. Zhang, X. Zhou, H. Wang, G. Sun, J. Xiao, Y. Pan, et al. 2016. Novel Virophages Discovered in a Freshwater Lake in China. *Frontiers in Microbiology* 7.

Goodcare, N., A. Aljanahi, S. Nandakumar, M. Mikailov, and A. S. Khan. 2018. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* 3.

Goodrum, F., and S. McWeeney. 2018. A Single-Cell Approach to the Elusive Latent Human Cytomegalovirus Transcriptome. *mBio* 9: e01001-18.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644–652.

Graf, L., E. C. Yang, K. Y. Han, F. C. Küpper, K. M. Benes, J. K. Oyadomari, R. J. H. Herbert, et al. 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.

Graupner, N., M. Jensen, C. Bock, S. Marks, S. Rahmann, D. Beisser, and J. Boenigk. 2018. Evolution of heterotrophy in chrysophytes as reflected by comparative transcriptomics. *FEMS Microbiology Ecology* 94.

Gray, M. W. 1999. Evolution of organellar genomes. *Current Opinion in Genetics & Development* 9: 678–687.

Gray, M. W., and W. F. Doolittle. 1982. Has the endosymbiont hypothesis been proven? *Microbiological Reviews* 46: 1–42.

Gruber-Vodicka, H. R., B. K. B. Seah, and E. Pruesse. 2020. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes M. Arumugam [ed.],. *mSystems* 5: e00920-20.

Guillard, R. R. L. 1975. Culture of phytoplankton for feeding marine invertebrates. Culture of marine invertebrate animals., 29–60. Springer, Boston, MA.

Guillard, R. R. L., and J. H. Ryther. 1962. Studies of marine planktonic diatoms: I. Cyclotella nana Hustedt, and Detonula confervacea (Cleve) Gran. *Canadian Journal of Microbiology* 8: 229–230.

Guillou, L., D. Bachar, S. Audic, D. Bass, C. Berney, L. Bittner, C. Boutte, et al. 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 41: D597–D604.

Guillou, L., M.-J. Chrétiennot-Dinet, S. Boulben, S. Y. Moon-van Der Staay, and D. Vaulot. 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.

Haas, B. J., S. Kamoun, M. C. Zody, R. H. Y. Jiang, R. E. Handsaker, L. M. Cano, M. Grabherr, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. *Nature* 461: 393–398.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* 8: 1494–1512.

Hackl, T., R. Martin, K. Barenhoff, S. Duponchel, D. Heider, and M. G. Fischer. 2020. Four high-quality draft genome assemblies of the marine heterotrophic nanoflagellate Cafeteria roenbergensis. *Scientific Data* 7: 1–9.

Harding, T., A. J. Roger, and A. G. B. Simpson. 2017. Adaptations to High Salt in a Halophilic Protist: Differential Expression and Gene Acquisitions through Duplications and Gene Transfers. *Frontiers in Microbiology* 8: 944.

He, D., O. Fiz-Palacios, C. J. Fu, C. C. Tsai, and S. L. Baldauf. 2014. An alternative root for the eukaryote tree of life. *Current Biology* 24: 465–470.

Hehenberger, E., D. V. Tikhonenkov, M. Kolisko, J. Del Campo, A. S. Esaulov, A. P. Mylnikov, and P. J. Keeling. 2017. Novel Predators Reshape Holozoan Phylogeny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals. *Current Biology* 27: 2043-2050.e6.

Hendy, M. D., and D. Penny. 1989. A Framework for the Quantitative Study of Evolutionary Trees. *Systematic Zoology* 38: 297.

Heo, K.-J., S.-J. Kwon, M.-K. Kim, H.-R. Kwak, S.-J. Han, M.-J. Kwon, A. L. N. Rao, and J.-K. Seo. 2020. Newly emerged resistance-breaking variants of cucumber mosaic virus represent ongoing host-interactive evolution of an RNA virus. *Virus Evolution* 6: veaa070.

Hernandez, A. M., and J. F. Ryan. 2021. Six-State Amino Acid Recoding is not an Effective Strategy to Offset Compositional Heterogeneity and Saturation in Phylogenetic Analyses J. Uyeda [ed.],. *Systematic Biology* 70: 1200–1212.

Hibberd, D. J. 1971. Observations on the cytology and ultrastructure of *Chrysamoeba radians* Klebs (Chrysophyceae). *British Phycological Journal* 6: 207–223.

Hickman, C. J. 1970. Biology of Phytophthora Zoospores. *Phytopathology*.

Ho, H. H., and C. J. Hickman. 1967. Asexual reproduction and behavior of zoospores of Phytophthora megasperma Var. Sojae. *Canadian Journal of Botany* 45: 1963–1981.

Holder, M., and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* 4: 275–284.

Husnik, F., and P. J. Keeling. 2019. The fate of obligate endosymbionts: reduction, integration, or extinction. *Current Opinion in Genetics & Development* 58–59: 1–8.

Husnik, F., D. Tashyreva, V. Boscaro, E. E. George, J. Lukeš, and P. J. Keeling. 2021. Bacterial and archaeal symbioses with protists. *Current Biology* 31: R862–R877.

Hyatt, D., G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11.

Illingworth, C. J. R., S. Roy, M. A. Beale, H. Tutill, R. Williams, and J. Breuer. 2017. On the effective depth of viral sequence data. *Virus Evolution* 3.

Inoue, T., and B. Tsai. 2013. How Viruses Use the Endoplasmic Reticulum for Entry, Replication, and Assembly. *Cold Spring Harbor Perspectives in Biology* 5: a013250–a013250.

Ishida, K., T. Sekizuka, K. Hayashida, J. Matsuo, F. Takeuchi, M. Kuroda, S. Nakamura, et al. 2014. Amoebal Endosymbiont Neochlamydia Genome Sequence Illuminates the Bacterial Role in the Defense of the Host Amoebae against Legionella pneumophila M. Horn [ed.],. *PLoS ONE* 9: e95166.

Jamy, M., C. Biwer, D. Vaulot, A. Obiol, H. Jing, S. Peura, R. Massana, and F. Burki. 2022. Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Nature Ecology & Evolution* 6: 1458–1470.

Jaške, K., D. Barcytė, T. Pánek, T. Ševčíková, A. Eliášová, and M. Eliáš. 2022. The net-like heterotrophic amoeba *Leukarachnion salinum* sp. nov. (Ochrophyta, Stramenopiles) has a cryptic plastid. Evolutionary Biology.

Jasti, S., M. E. Sieracki, N. J. Poulton, M. W. Giewat, and J. N. Rooney-Varga. 2005. Phylogenetic Diversity and Specificity of Bacteria Closely Associated with *Alexandrium* spp. and Other Phytoplankton. *Applied and Environmental Microbiology* 71: 3483–3494.

Jensen, S., D. G. Bourne, M. Hovland, and J. Colin Murrell. 2012. High diversity of microplankton surrounds deep-water coral reef in the Norwegian Sea. *FEMS Microbiology Ecology* 82: 75–89.

Jimenez, M. J., M. Arenas, and U. Bastolla. 2018. Substitution Rates Predicted by Stability-Constrained Models of Protein Evolution Are Not Consistent with Empirical Data. *Molecular Biology and Evolution* 35: 743–755.

Judelson, H. S., and F. A. Blanco. 2005. The spores of Phytophthora: weapons of the plant destroyer. *Nature Reviews Microbiology* 3: 47–58.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. Von Haeseler, and L. S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.

Kapli, P., Z. Yang, and M. J. Telford. 2020. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* 21: 428–444.

Karlsson, K., E. Sahlin, E. Iwarsson, M. Westgren, M. Nordenskjöld, and S. Linnarsson. 2015. Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations. *Genomics* 105: 150–158.

Karpov, S. A., R. Kersanach, and D. M. Williams. 1998. Ultrastructure and 18S rRNA gene sequence of a small Heterotrophic flagellate Siluania monomastiga gen. et sp. nov. (bicosoecida). *European Journal of Protistology* 34: 415–425.

Katoh, K., and D. M. Standley. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30: 772–780.

Kawachi, M., I. Inouye, D. Honda, C. J. O'Kelly, J. C. Bailey, R. R. Bidigare, and R. A. Andersen. 2002. The Pinguiophyceae classis nova, a new class of photosynthetic stramenopiles whose members produce large amounts of omega-3 fatty acids. *Phycological Research* 50: 31–47.

Kawai, H., S. Maeba, H. Sasaki, K. Okuda, and E. C. Henry. 2003. Schizocladia ischiensis: A New Filamentous Marine Chromophyte Belonging to a New Class, Schizocladiophyceae. *Protist* 154: 211–228.

Kayama, M., K. Maciszewski, A. Yabuki, H. Miyashita, A. Karnkowska, and R. Kamikawa. 2020. Highly Reduced Plastid Genomes of the Non-photosynthetic Dictyochophyceans Pteridomonas spp. (Ochrophyta, SAR) Are Retained for tRNA-Glu-Based Organellar Heme Biosynthesis. *Frontiers in Plant Science* 11: 602455.

Keeling, P. J. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *Journal of Eukaryotic Microbiology* 56: 1–8.

Keeling, P. J. 2002. Molecular phylogenetic position of Trichomitopsis termopsidis (Parabasalia) and evidence for the Trichomitopsiinae. *European Journal of Protistology* 38: 279–286.

Keeling, P. J. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 729–748.

Keeling, P. J., and F. Burki. 2019. Progress towards the Tree of Eukaryotes. *Current Biology* 29: R808–R817.

Keeling, P. J., F. Burki, H. M. Wilcox, B. Allam, E. E. Allen, L. A. Amaral-Zettler, E. V. Armbrust, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology* 12: e1001889.

Kneip, C., C. Voβ, P. J. Lockhart, and U. G. Maier. 2008. The cyanobacterial endosymbiont of the unicellular algae Rhopalodia gibba shows reductive genome evolution. *BMC Evolutionary Biology* 8: 30.

Ko, W. H., and L. L. Chase. 1973. Aggregation of Zoospores of Phytophthora palmivora. *Journal of General Microbiology* 78: 79–82.

Kobert, K., L. Salichos, A. Rokas, and A. Stamatakis. 2016. Computing the Internode Certainty and Related Measures from Partial Gene Trees. *Molecular Biology and Evolution* 33: 1606–1617.

Kocot, K. M., T. H. Struck, J. Merkel, D. S. Waits, C. Todt, P. M. Brannock, D. A. Weese, et al. 2016. Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. *Systematic Biology*: syw079.

Kolodziej, K., and T. Stoeck. 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. *Applied and Environmental Microbiology* 73: 2718–2726.

Koshi, J. M., and R. A. Goldstein. 1995. Context-dependent optimal substitution matrices. *Protein Eng.* 8: 641–645.

Kozlov, A. M., D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference J. Wren [ed.],. *Bioinformatics* 35: 4453–4455.

Krings, M., T. N. Taylor, and N. Dotzler. 2011. The fossil record of the Peronosporomycetes (Oomycota). *Mycologia* 103: 445–457.

Kühn, S., L. Medlin, and G. Eller. 2004. Phylogenetic Position of the Parasitoid Nanoflagellate Pirsonia inferred from Nuclear-Encoded Small Subunit Ribosomal DNA and a Description of Pseudopirsonia n. gen. and Pseudopirsonia mucosa (Drebes) comb. nov. *Protist* 155: 143–156.

Kurtz, S., A. Phillipy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. 2004. Versatile and open software for comparing large genomes. *Genome Biology* 5.

Labarre, A., D. López-Escardó, F. Latorre, G. Leonard, F. Bucchini, A. Obiol, C. Cruaud, et al. 2021. Comparative genomics reveals new functional insights in uncultured MAST species. *The ISME Journal* 15: 1767–1781.

Laetsch, D. R., and M. L. Blaxter. 2017. BlobTools: Interrogation of genome assemblies. *F1000Research* 6: 1287.

Lanyon, S. M. 1988. The Stochastic Mode of Molecular Evolution: What Consequences for Systematic Investigations? *The Auk* 105: 565–573.

Larkum, A. W. D., P. J. Lockhart, and C. J. Howe. 2007. Shopping for plastids. *Trends in Plant Science* 12: 189–195.

Larsen, J., and D. J. Patterson. 1990. Some flagellates (Protista) from tropical marine sediments. *Journal of Natural History* 24: 801–937.

Lartillot, N., H. Brinkmann, and H. Philippe. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* 7: S4.

Lartillot, N., T. Lepage, and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25: 2286–2288.

Lartillot, N., and H. Philippe. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution* 21: 1095–1109.

Lee, H. B., D. H. Jeong, B. C. Cho, and J. S. Park. 2022. The Diversity Patterns of Rare to Abundant Microbial Eukaryotes Across a Broad Range of Salinities in a Solar Saltern. *Microbial Ecology* 84: 1103–1121.

Lee, J. J. 2006. Algal symbiosis in larger foraminifera. *Symbiosis* 42: 63–75.

Lee, W. J. 2002. Redescription of the Rare Heterotrophic Flagellate (Protista) - Phyllomitus undulans Stein, 1878, and Erection of a New Genus - Pseudophyllomitus gen. n. *Acta Protozoologica* 41: 375–381.

Lee, W. J., and D. J. Patterson. 2002. Abundance and Biomass of Heterotrophic Flagellates, and Factors Controlling Their Abundance and Distribution in Sediments of Botany Bay. *Microbial Ecology* 43: 467–481.

Lehman, J. T. 1967. Ecological and nutritional studies on Dinobryon Ehrenb.: Seasonal periodicity and the phosphate toxicity problem. *Limnology and Oceanography* 21: 646–658.

Leipe, D. D., S. M. Tong, C. L. Goggin, S. B. Slemenda, N. J. Pieniazek, and M. L. Sogin. 1996. 16S-like rDNA sequences from Developayella elegans, Labyrinthuloides haliotidis, and

Proteromonas lacertae confirm that the stramenopiles are a primarily heterotrophic group. *European Journal of Protistology* 32: 449–458.

Leonard, G., A. Labarre, D. S. Milner, A. Monier, D. Soanes, J. G. Wideman, F. Maguire, et al. 2018. Comparative genomic analysis of the 'pseudofungus' *Hyphochytrium catenoides*. *Open Biology* 8: 170184.

Li, W., and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.

Lin, Y. C., T. Campbell, C. C. Chung, G. C. Gong, K. P. Chiang, and A. Z. Worden. 2012. Distribution patterns and phylogeny of marine stramenopiles in the North Pacific Ocean. *Applied and Environmental Microbiology* 78: 3387–3399.

Logares, R., S. Audic, S. Santini, M. C. Pernice, C. De Vargas, and R. Massana. 2012. Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *The ISME Journal* 6: 1823–1833.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. *Molecular Biology and Evolution* 19: 1–7.

Luo, C., D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis. 2012. Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample F. Rodriguez-Valera [ed.],. *PLoS ONE* 7: e30087.

Lupette, J., R. Lami, M. Krasovec, N. Grimsley, H. Moreau, G. Piganeau, and S. Sanchez-Ferandin. 2016. Marinobacter Dominates the Bacterial Community of the Ostreococcus tauri Phycosphere in Culture. *Frontiers in Microbiology* 7.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–636.

Maier, U.-G., S. Zauner, C. Woehle, K. Bolte, F. Hempel, J. F. Allen, and W. F. Martin. 2013. Massively Convergent Evolution for Ribosomal Protein Gene Content in Plastid and Mitochondrial Genomes. *Genome Biology and Evolution* 5: 2318–2329.

Maire, J., S. K. Girvan, S. E. Barkla, A. Perez-Gonzalez, D. J. Suggett, L. L. Blackall, and M. J. H. Van Oppen. 2021. Intracellular bacteria are common and taxonomically diverse in cultured and in hospite algal endosymbionts of coral reefs. *The ISME Journal* 15: 2028–2042.

Manz, W., R. Amann, W. Ludwig, M. Wagner, and K.-H. Schleifer. 1992. Phylogenetic Oligodeoxynucleotide Probes for the Major Subclasses of Proteobacteria: Problems and Solutions. *Systematic and Applied Microbiology* 15: 593–600.

Massana, R., J. Castresana, V. Balague, L. Guillou, K. Romari, A. Groisillier, K. Valentin, and C. Pedros-Alio. 2004. Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. *Applied and Environmental Microbiology* 70: 3528–3534.

Massana, R., J. Del Campo, M. E. Sieracki, S. Audic, and R. Logares. 2014. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *The ISME Journal* 8: 854–866.

Massana, R., A. Gobet, S. Audic, D. Bass, L. Bittner, C. Boutte, A. Chambouvet, et al. 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing: Protist diversity in European coastal areas. *Environmental Microbiology* 17: 4035–4049.

Massana, R., R. Terrado, I. Forn, C. Lovejoy, and C. Pedros-Alio. 2006. Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. *Environmental Microbiology* 8: 1515–1522.

Massana, R., F. Unrein, R. Rodríguez-Martínez, I. Forn, T. Lefort, J. Pinhassi, and F. Not. 2009. Grazing rates and functional diversity of uncultured heterotrophic flagellates. *The ISME Journal* 3: 588–596.

Matari, N. H., and J. E. Blair. 2014. A multilocus timescale for oomycete evolution estimated under three distinct molecular clock models. *BMC Evolutionary Biology* 14: 101.

McKeown, D., J. Schroeder, K. Stevens, A. Peters, C. Sáez, J. Park, M. Rothman, et al. 2018. Phaeoviral Infections Are Present in Macrocystis, Ecklonia and Undaria (Laminariales) and Are Influenced by Wave Exposure in Ectocarpales. *Viruses* 10: 410.

Medlin, L., H. J. Elwood, S. Stickel, and M. L. Sogin. 1988. The characterization of enzymatically amplified eukaryotic 16!Wke rRNA-coding regions. *Gene* 71: 491–499.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era E. Teeling [ed.],. *Molecular Biology and Evolution* 37: 1530–1534.

Mitra, A., K. J. Flynn, U. Tillmann, J. A. Raven, D. Caron, D. K. Stoecker, F. Not, et al. 2016. Defining Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition: Incorporation of Diverse Mixotrophic Strategies. *Protist* 167: 106–120.

Moestrup, Ø, T., HA. 1976. Fine structural studies on the flagellate genus Bicoeca I. - Bicoeca maris with particular emphasis on the flagellar apparatus. *Protistologica* 12: 101–120.

Mongiardino Koch, N. 2021. Phylogenomic Subsampling and the Search for Phylogenetically Reliable Loci Y. Satta [ed.],. *Molecular Biology and Evolution* 38: 4025–4038.

Mongiardino Koch, N., and J. R. Thompson. 2021. A Total-Evidence Dated Phylogeny of Echinoidea Combining Phylogenomic and Paleontological Data J. Serb [ed.],. *Systematic Biology* 70: 421–439.

Moniruzzaman, M., A. R. Weinheimer, C. A. Martinez-Gutierrez, and F. O. Aylward. 2020. Widespread endogenization of giant viruses shapes genomes of green algae. *Nature* 588: 141–145.

Montagna, M., D. Sassera, S. Epis, C. Bazzocchi, C. Vannini, N. Lo, L. Sacchi, et al. 2013. “Candidatus Midichloriaceae” fam. nov. (Rickettsiales), an Ecologically Widespread Clade of Intracellular Alphaproteobacteria. *Applied and Environmental Microbiology* 79: 3241–3248.

Moreau, H., G. Piganeau, Y. Desdevises, R. Cooke, E. Derelle, and N. Grimsley. 2010. Marine Prasinovirus Genomes Show Low Evolutionary Divergence and Acquisition of Protein Metabolism Genes by Horizontal Gene Transfer. *Journal of Virology* 84: 12555–12563.

Moriya, M., T. Nakayama, and I. Inouye. 2002. A New Class of the Stramenopiles, Placididea Classis nova: Description of Placidia cafeteriopsis gen. et sp. nov. *Protist* 153: 143–156.

Moriya, M., T. Nakayama, and I. Inouye. 2000. Ultrastructure and 18S rDNA Sequence Analysis of Wobblia lunata gen. et sp. nov., a New Heterotrophic Flagellate (Stramenopiles, Incertae Sedis). *Protist* 151: 41–55.

Muehlstein, L. K., D. Porter, F. T. Short, S. Mycologia, and N. M. Apr. 1991. Labyrinthula zosterae sp . nov ., the Causative Agent of Wasting Disease of Eelgrass , Zostera marina Stable URL : http://www.jstor.org/stable/3759933 REFERENCES Linked references are available on JSTOR for this article : You may need to log in to JSTOR t. *Mycologia* 83: 180–191.

Muller, D. G., and E. Parodi. 1993. Transfer of a marine DNA virus from Ectocarpus to Feldmannia (Ectocarpales, Phaeophyceae): aberrant symptoms and restitution of the host. *Protoplasma* 175: 121–125.

Muller, D. G., M. Sengco, S. Wolf, M. Brautigam, C. E. Schmid, M. Kapp, and R. Knippers. 1996. Comparison of two DNA Viruses Infecting the Marine Brown Algae Ectocarpus Siliculosus and E. Fasciculatus. *Journal of General Virology* 77: 2329–2333.

Nagasaki, K., and M. Yamaguchi. 1999. Cryopreservation of a Virus (HaV) Infecting a Harmful Bloom Causing Microalga, *Heterosigma akashiwo* (Raphidophyceae). *Fisheries science* 65: 319–320.

Nakai, R., and T. Naganuma. 2015. Diversity and Ecology of Thraustochytrid Protists in the Marine Environment. *In* S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not [eds.], Marine Protists, 331–346. Springer Japan, Tokyo.

Nayfach, S., A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrpides. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology* 39: 578–585.

Neef, A. 1997. Anwendung der in situ Einzelzell-Identifizierung von Bakterien zur Populationsanalyse in komplexen mikrobiellen Biozonosen. Technische Universitat Munchen, Munich, Germany.

Neef, A., R. Amann, H. Schlesner, and K.-H. Schleifer. 1998. Monitoring a widespread bacterial group: in situ detection of planctomycetes with 16S rRNA-targeted probes. *Microbiology* 144: 3257–3266.

Nguyen, L.-T., H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32: 268–274.

Nichols, R. 2001. Gene trees and species trees are not the same. *Trends in Ecology & Evolution* 16: 358–364.

Noguchi, F., G. Tanifuji, M. W. Brown, K. Fujikura, and K. Takishita. 2016. Complex evolution of two types of cardiolipin synthase in the eukaryotic lineage stramenopiles. *Molecular Phylogenetics and Evolution* 101: 133–141.

Nosenko, T., F. Schreiber, M. Adamska, M. Adamski, M. Eitel, J. Hammel, M. Maldonado, et al. 2013. Deep metazoan phylogeny: When different genes tell different stories. *Molecular Phylogenetics and Evolution* 67: 223–233.

Nowack, E. C. M., and M. Melkonian. 2010. Endosymbiotic associations within protists. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 699–712.

Okamura, T., and R. Kondo. 2015. Suigetsumonas clinomigrationis gen. et sp. nov., a Novel Facultative Anaerobic Nanoflagellate Isolated from the Meromictic Lake Suigetsu, Japan. *Protist* 166: 409–421.

O'Kelly, C. J., and T. A. Nerad. 1998. Kinetid architecture and bicosoecid affinities of the marine heterotrophic nanoflagellate Caecitellus parvulus (Griessmann, 1913) Patterson et al., 1993. *European Journal of Protistology* 34: 369–375.

Okude, M., J. Matsuo, S. Nakamura, K. Kawaguchi, Y. Hayashi, H. Sakai, M. Yoshida, et al. 2012. Environmental Chlamydiae Alter the Growth Speed and Motility of Host Acanthamoebae. *Microbes and Environments* 27: 423–429.

Onsbring, H., A. K. Tice, B. T. Barton, M. W. Brown, and T. J. G. Ettema. 2020. An efficient single-cell transcriptomics workflow for microbial eukaryotes benchmarked on Giardia intestinalis cells. *BMC Genomics* 21: 448.

Pardo-De La Hoz, C. J., N. Magain, B. Piatkowski, L. Cornet, M. Dal Forno, I. Carbone, J. Miadlikowska, and F. Lutzoni. 2023. Ancient Rapid Radiation Explains Most Conflicts Among Gene Trees and Well-Supported Phylogenomic Trees of Nostocalean Cyanobacteria C. Lemus-Solis [ed.],. *Systematic Biology* 72: 694–712.

Parfrey, L. W., and D. J. G. Lahr. 2013. Multicellularity arose several times in the evolution of eukaryotes (Response to DOI 10.1002/bies.201100187). *BioEssays* 35: 339–347.

Park, J. S., B. C. Cho, and A. G. B. Simpson. 2006. Halocafeteria seosinensis gen. et sp. nov. (Bicosoecida), a halophilic bacterivorous nanoflagellate isolated from a solar saltern. *Extremophiles* 10: 493–504.

Park, J. S., and A. G. B. Simpson. 2010. Characterization of halotolerant Bicosoecida and Placididea (Stramenopila) that are distinct from marine forms, and the phylogenetic pattern of salinity preference in heterotrophic stramenopiles: Novel halotolerant heterotrophic stramenopiles. *Environmental Microbiology* 12: 1173–1184.

Parks, M. B., T. Nakov, E. C. Ruck, N. J. Wickett, and A. J. Alverson. 2018. Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *American Journal of Botany* 105: 330–347.

Patterson, D. J., K. Nygaard, G. Steinberg, and C. M. Turley. 1993. Heterotrophic flagellates and other protists associated with oceanic detritus throughout the water column in the mid North Atlantic. *Journal of the Marine Biological Association of the United Kingdom* 73: 67–95.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Peek, R. J., S. Thompson, J. Donahue, K. Tham, J. Atherton, M. Blaser, and G. Miller. 1998. Adherence to gastric epithelial cells induces expression of a Helicobacter pylori gene, iceA, that is associated with clinical outcome. *Proceedings of the Association of American Physicians.* 110: 531–544.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough D. Penny [ed.],. *PLoS Biology* 9: e1000602.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* 5: 50.

Picelli, S., O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* 9: 171–181.

Pick, K. S., H. Philippe, F. Schreiber, D. Erpenbeck, D. J. Jackson, P. Wrede, M. Wiens, et al. 2010. Improved Phylogenomic Taxon Sampling Noticeably Affects Nonbilaterian Relationships. *Molecular Biology and Evolution* 27: 1983–1987.

Piwosz, K., and J. Pernthaler. 2010. Seasonal population dynamics and trophic role of planktonic nanoflagellates in coastal surface waters of the Southern Baltic Sea. *Environmental Microbiology* 12: 364–377.

Price, W. C. 1934. Isolation and study of some yellow strains of cucumber mosaic. *Phytopathology* 24: 743–761.

Prjibelski, A. D., D. Antipov, D. Meleshko, A. Lapidus, and A. Korobeynikov. 2020. Using SPAdes de novo assembler. *Current Protocols in Bioinformatics* 70.

Prjibelski, A. D., I. Vasilinetc, A. Bankevich, A. Gurevich, T. Krivosheeva, S. Nurk, S. Pham, et al. 2014. ExSPANder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics* 30: i293–i301.

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

Quang, L. S., O. Gascuel, and N. Lartillot. 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24: 2317–2323.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41: D590–D596.

Raghukumar, S. 1992. Bacterivory: a novel dual role for thraustochytrids in the sea. *Marine Biology* 113: 165–169.

Richter, D. J., C. Berney, J. F. H. Strassert, Y.-P. Poh, E. K. Herman, S. A. Muñoz-Gómez, J. G. Wideman, et al. 2022. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal* 2: e56.

Riisberg, I., R. J. S. Orr, R. Kluge, K. Shalchian-Tabrizi, H. A. Bowers, V. Patil, B. Edvardsen, and K. S. Jakobsen. 2009. Seven Gene Phylogeny of Heterokonts. *Protist* 160: 191–204.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53: 131–147.

Rodríguez-Martínez, R., M. Labrenz, J. Del Campo, I. Forn, K. Jürgens, and R. Massana. 2009. Distribution of the uncultured protist MAST-4 in the Indian Ocean, Drake Passage and Mediterranean Sea assessed by real-time quantitative PCR. *Environmental Microbiology* 11: 397–408.

Rodríguez-Martínez, R., G. Leonard, D. S. Milner, S. Sudek, M. Conway, K. Moore, T. Hudson, et al. 2020. Controlled sampling of ribosomally active protistan diversity in sediment-surface layers identifies putative players in the marine carbon sink. *The ISME Journal* 14: 984–998.

Rodriguez-Martinez, R., G. Rocap, R. Logares, S. Romac, and R. Massana. 2012. Low Evolutionary Diversification in a Widespread and Abundant Uncultured Protist (MAST-4). *Molecular Biology and Evolution* 29: 1393–1406.

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4: e2584.

Romero-Brey, I., and R. Bartenschlager. 2016. Endoplasmic Reticulum: The Favorite Intracellular Niche for Viral Replication and Assembly. *Viruses* 8: 160.

Roure, B., N. Rodriguez-Ezpeleta, and H. Philippe. 2007. SCaFoS: a tool for Selection, Concatenation and Fusion of Sequences for phylogenomics. *BMC Evolutionary Biology* 7: S2.

Roy, R. S., D. C. Price, A. Schliep, G. Cai, A. Korobeynikov, H. S. Yoon, E. C. Yang, and D. Bhattacharya. 2014. Single cell genome analysis of an uncultured heterotrophic stramenopile. *Scientific Reports* 4: 1–8.

Rozenberg, A., J. Oppermann, J. Wietek, R. G. Fernandez Lahore, R.-A. Sandaa, G. Bratbak, P. Hegemann, and O. Béjà. 2020. Lateral Gene Transfer of Anion-Conducting Channelrhodopsins between Green Algae and Giant Viruses. *Current Biology* 30: 4910-4920.e5.

Rybarski, A. E., F. Nitsche, J. Soo Park, P. Filz, P. Schmidt, R. Kondo, A. Gb Simpson, and H. Arndt. 2021. Revision of the phylogeny of Placididea (Stramenopiles): Molecular and morphological diversity of novel placidid protists from extreme aquatic environments. *European Journal of Protistology* 81: 125809.

Salichos, L., and A. Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497: 327–331.

Saville-Kent, W. 1880. A manual of the Infusoria : including a description of all known flagellate, ciliate, and tentaculiferous protozoa, British and foreign, and an account of the organization and the affinities of the sponges. St. Martin's Place, London.

Savory, A. I. M., L. J. Grenville-Briggs, S. Wawra, P. Van West, and F. A. Davidson. 2014. Auto-aggregation in zoospores of *Phytophthora infestans* : the cooperative roles of bioconvection and chemotaxis. *Journal of The Royal Society Interface* 11: 20140017.

Scheckenbach, F., K. Hausmann, C. Wylezich, M. Weitere, and H. Arndt. 2010. Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. *Proceedings of the National Academy of Sciences of the United States of America* 107: 115–120.

Schnepf, E., G. Drebes, and M. Elbrachter. 1990. Pirsonia guinardiae , gen. et spec. nov.: A parasitic flagellate on the marine diatom Guinardia flaccida with an unusual mode of food uptake. *Helgolander Meeresunters* 44: 275–293.

Schoenle, A., M. Hohlfeld, K. Hermanns, F. Mahé, C. De Vargas, F. Nitsche, and H. Arndt. 2021. High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. *Communications Biology* 4: 501.

Schoenle, A., M. Hohlfeld, A. Rybarski, M. Sachs, E. Freches, K. Wiechmann, F. Nitsche, and H. Arndt. 2022. Cafeteria in extreme environments: Investigations on C. burkhardae and three new species from the Atacama Desert and the deep ocean. *European Journal of Protistology* 85: 125905.

Schweikert, M., and E. Schnepf. 1997. Light and electron microscopical observations on Pirsonia punctigerae spec, nov., a nanoflagellate feeding on the marine centric diatom Thalassiosira punctigera. *European Journal of Protistology* 33: 168–177.

Scola, B. L., S. Audic, C. Robert, L. Jungang, X. De Lamballerie, M. Drancourt, R. Birtles, et al. 2003. A Giant Virus in Amoebae. *Science* 299: 2033–2033.

Seah, B. K. B., C. P. Antony, B. Huettel, J. Zarzycki, L. Schada Von Borzyskowski, T. J. Erb, A. Kouris, et al. 2019. Sulfur-Oxidizing Symbionts without Canonical Genes for Autotrophic CO $_2$ Fixation S. J. Giovannoni [ed.],. *mBio* 10: e01112-19.

Sebé-Pedrós, A., M. Irimia, J. Del Campo, H. Parra-Acero, C. Russ, C. Nusbaum, B. J. Blencowe, and I. Ruiz-Trillo. 2013. Regulated aggregative multicellularity in a close unicellular relative of metazoa. *eLife* 2: e01287.

Seeleuthner, Y., S. Mondy, V. Lombard, Q. Carradec, E. Pelletier, M. Wessner, J. Leconte, et al. 2018. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nature communications* 9: 310.

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

Seibold, A., A. Wichels, and C. Schütt. 2001. Diversity of endocytic bacteria in the dinoflagellate Noctiluca scintillans. *Aquatic Microbial Ecology* 25: 229–235.

Ševčíková, T., A. Horák, V. Klimeš, V. Zbránková, E. Demir-Hilton, S. Sudek, J. Jenkins, et al. 2015. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Scientific Reports* 5: 10134.

Ševčíková, T., T. Yurchenko, K. P. Fawley, R. Amaral, H. Strnad, L. M. A. Santos, M. W. Fawley, and M. Eliáš. 2019. Plastid Genomes and Proteins Illuminate the Evolution of

Eustigmatophyte Algae and Their Bacterial Endosymbionts L. A. Katz [ed.],. *Genome Biology and Evolution* 11: 362–379.

Shen, W., S. Le, Y. Li, and F. Hu. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation Q. Zou [ed.],. *PLOS ONE* 11: e0163962.

Shen, X.-X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* 1: 0126.

Shen, X.-X., L. Salichos, and A. Rokas. 2016b. A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference. *Genome Biology and Evolution* 8: 2565–2580.

Sherr, E. B., and B. F. Sherr. 2002. Significance of predation by protists in aquatic microbial food webs. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* 81: 293–308.

Shimodaira, H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection N. Goldman [ed.],. *Systematic Biology* 51: 492–508.

Shiratori, T., T. Nakayama, and K. Ishida. 2015. A New Deep-branching Stramenopile, Platysulcus tardus gen. nov., sp. nov. *Protist* 166: 337–348.

Shiratori, T., R. Thakur, and K. Ishida. 2017. Pseudophyllomitus vesiculosus (Larsen and Patterson 1990) Lee, 2002, a poorly studied phagotrophic biflagellate is the first characterized member of stramenoile environmental clade MAST-6. *Protist* 168: 439–451.

Sibbald, S. J., and J. M. Archibald. 2017. More protist genomes needed. *Nature Ecology and Evolution* 1: 1–3.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212.

Simon, M., L. Jardillier, P. Deschamps, D. Moreira, G. Restoux, P. Bertolino, and P. López-garcía. 2015. Complex communities of small protists and unexpected occurrence of typical marine lineages in shallow freshwater systems. *Environmental Microbiology* 17: 3610–3627.

Smith, S. A., J. W. Brown, and J. F. Walker. 2018. So many genes, so little time: A practical approach to divergence-time estimation in the genomic era H. Escriva [ed.],. *PLOS ONE* 13: e0197433.

Song, L., and L. Florea. 2015. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* 4: 48.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stevens, K., K. Weynberg, C. Bellas, S. Brown, C. Brownlee, M. T. Brown, and D. C. Schroeder. 2014. A Novel Evolutionary Strategy Revealed in the Phaeoviruses S. J. Martin [ed.],. *PLoS ONE* 9: e86040.

Stiller, J. W., J. Huang, Q. Ding, J. Tian, and C. Goodwillie. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics* 10: 484.

Stiller, J. W., D. C. Reel, and J. C. Johnson. 2003. A single origin of plastids revisited: convergent evolution in organellar genome content. *Journal of Phycology* 39: 95–105.

Strimmer, K., and A. Von Haeseler. 1996. Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution* 13: 964–969.

Struck, T. H. 2014. TreSpEx—Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. *Evolutionary Bioinformatics* 10: EBO.S14239.

Sugimoto, H., and H. Endoh. 2006. Analysis of Fruiting Body Development in the Aggregative Ciliate Sorogena stoianovitchae (Ciliophora, Colpodea). *The Journal of Eukaryotic Microbiology* 53: 96–102.

Superson, A. A., and F. U. Battistuzzi. 2022. Exclusion of fast evolving genes or fast evolving sites produces different archaean phylogenies. *Molecular Phylogenetics and Evolution* 170: 107438.

Szantho, L. L., N. Lartillot, G. L. Szollosi, and D. Schrempf. 2023. Compositionally Constrained Sites Drive Long Branch Attraction. *Systematic Biology* 72: 767–780.

Takishita, K., N. Yubuki, N. Kakizoe, Y. Inagaki, and T. Maruyama. 2007. Diversity of microbial eukaryotes in sediment at a deep-sea methane cold seep: Surveys of ribosomal DNA libraries from raw sediment samples and two enrichment cultures. *Extremophiles* 11: 563–576.

Terpis, K. X. 2021. A phylogenomic approach to explore photosynthetic stramenopile evolution. University of Rhode Island, Kingston, RI.

Thakur, R., T. Shiratori, and K. Ishida. 2019. Taxon-rich Multigene Phylogenetic Analyses Resolve the Phylogenetic Relationship Among Deep-branching Stramenopiles. *Protist* 170: 125682.

The UniProt Consortium, A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, R. Agivetova, S. Ahmad, et al. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* 49: D480–D489.

Theriot, E. C., M. P. Ashworth, T. Nakov, E. Ruck, and R. K. Jansen. 2015. Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Molecular Phylogenetics and Evolution* 89: 28–36.

Theriot, E. C., M. Ashworth, E. Ruck, T. Nakov, and R. K. Jansen. 2010. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution* 143: 278–296.

Thomas, R., N. Grimsley, M. Escande, L. Subirana, E. Derelle, and H. Moreau. 2011. Acquisition and maintenance of resistance to viruses in eukaryotic phytoplankton populations: Viral resistance in *Mamiellales*. *Environmental Microbiology* 13: 1412–1420.

Tice, A. K., L. L. Shadwick, A. M. Fiore-Donno, S. Geisen, S. Kang, G. A. Schuler, F. W. Spiegel, et al. 2016. Expansion of the molecular and morphological diversity of Acanthamoebidae (Centramoebida, Amoebozoa) and identification of a novel life cycle type within the group. *Biology Direct* 11: 69.

Tice, A. K., D. Žihala, T. Pánek, R. E. Jones, E. D. Salomaki, S. Nenarokov, F. Burki, et al. 2021. PhyloFisher: A phylogenomic package for resolving eukaryotic relationships A. Hejnol [ed.],. *PLOS Biology* 19: e3001365.

Tikhonenkov, D. V., J. Janouškovec, P. J. Keeling, and A. P. Mylnikov. 2016. The Morphology, Ultrastructure and SSU rRNA Gene Sequence of a New Freshwater Flagellate, *Neobodo borokensis* n. sp. (Kinetoplastea, Excavata). *Journal of Eukaryotic Microbiology* 63: 220–232.

Tikhonenkov, D. V., J. Janouškovec, A. P. Mylnikov, K. V. Mikhailov, T. G. Simdyanov, V. V. Aleoshin, and P. J. Keeling. 2014. Description of Colponema vietnamica sp.n. and Acavomonas peruviana n. gen. n. sp., Two New Alveolate Phyla (Colponemidia nom. nov. and Acavomonidia nom. nov.) and Their Contributions to Reconstructing the Ancestral State of Alveolates and Eukaryotes S. Gribaldo [ed.],. *PLoS ONE* 9: e95467.

Tikhonenkov, D. V., Iu A. Mazei, and E. A. Embulaeva. 2008. Degradation succession of heterotrophic flagellate communities in microcosms. *Zh Obschch Biol* 69: 57–64.

Tikhonenkov, D. V., K. V. Mikhailov, R. M. R. Gawryluk, A. O. Belyaev, V. Mathur, S. A. Karpov, D. G. Zagumyonnyi, et al. 2022. Microbial predators form a new supergroup of eukaryotes. *Nature* 612: 714–719.

Tong, S. M. 1995a. Developayella elegans nov. gen., nov. spec., a New Type of Heterotrophic Flagellate from Marine Plankton. *European Journal of Protistology* 31: 24–31.

Tong, S. M. 1995b. Developayella elegans nov. gen., nov. spec., a new type of heterotrophic flagellate from marine plankton. *European Journal of Protistology* 31: 24–31.

Torruella, G., X. Grau-Bové, D. Moreira, S. A. Karpov, J. A. Burns, A. Sebé-Pedrós, E. Völcker, and P. López-García. 2018. Global transcriptome analysis of the aphelid Paraphelidium tribonemae supports the phagotrophic origin of fungi. *Communications Biology* 1: 231.

Tsaousis, A. D., S. O. De Choudens, E. Gentekaki, S. Long, D. Gaston, A. Stechmann, D. Vinella, et al. 2012. Evolution of Fe/S cluster biogenesis in the anaerobic parasite Blastocystis. *Proceedings of the National Academy of Sciences of the United States of America* 109: 10426–10431.

Tsui, C. K. M., W. Marshall, R. Yokoyama, D. Honda, J. C. Lippmeier, K. D. Craven, P. D. Peterson, and M. L. Berbee. 2009. Labyrinthulomycetes phylogeny and its implications for the evolutionary loss of chloroplasts and gain of ectoplasmic gliding. *Molecular Phylogenetics and Evolution* 50: 129–140.

Ullah, M., Y. Li, K. Munib, and Z. Zhang. 2023. Epidemiology, host range, and associated risk factors of monkeypox: an emerging global public health threat. *Frontiers in Microbiology* 14.

Vacic, V., V. N. Uversky, A. K. Dunker, and S. Lonardi. 2007. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* 8: 211.

Vallet, M., T. U. H. Baumeister, F. Kaftan, V. Grabe, A. Buaya, M. Thines, A. Svatoš, and G. Pohnert. 2019. The oomycete Lagenisma coscinodisci hijacks host alkaloid synthesis during infection of a marine diatom. *Nature Communications* 10: 1–8.

Van Etten, J. L., M. V. Graves, D. G. Müller, W. Boland, and N. Delaroque. 2002. Phycodnaviridae– large DNA algal viruses. *Archives of Virology* 147: 1479–1516.

de Vargas, C., S. Audic, N. Henry, J. Decelle, F. Mahe, R. Logares, E. Lara, et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348: 1261605–1/11.

Vasilinetc, I., A. D. Prjibelski, A. Gurevich, A. Korobeynikov, and P. A. Pevzner. 2015. Assembling short reads from jumping libraries with large insert sizes. *Bioinformatics* 31: 3262–3268.

Verhagen, F. J. M., M. Zölffel, G. Brugerolle, and D. J. Patterson. 1994. Adriamonas peritocrescens gen. nov., sp. nov., a new free-living soil flagellate (Protista, Pseudodendromonadidae Incertae Sedis). *European Journal of Protistology* 30: 295–308.

Von Magnus, P., E. Andersen, K. Petersen, and A. Birch-Andersen. 1959. A pox-like disease in cynomolgus monkeys. *Acta Pathologica Microbiologica Scandinavica* 46: 156–176.

Wägele, J., and C. Mayer. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *BMC Evolutionary Biology* 7: 147.

Wang, H.-C., B. Q. Minh, E. Susko, and A. J. Roger. 2018. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology* 67: 216–235.

Wang, H.-C., E. Susko, M. Spencer, and A. J. Roger. 2008. Topological Estimation Biases with Covarion Evolution. *Journal of Molecular Evolution* 66: 50–60.

Watanae, M. M., M. Kawachi, M. Hiroki, and F. Kasai. 2000. NIES Collection List of Strains. 6th Ed. NIES, Japan.

Wawrzyniak, I., D. Courtine, M. Osman, C. Hubans-Pierlot, A. Cian, C. Nourrisson, M. Chabe, et al. 2015. Draft genome sequence of the intestinal parasite Blastocystis subtype 4-isolate WR1. *Genomics Data* 4: 22–23.

Weiler, B. A., E. L. Sà, M. E. Sieracki, R. Massana, and J. Campo. 2021. *Mediocremonas mediterraneus* , a New Member within the Developea. *Journal of Eukaryotic Microbiology* 68.

Weiler, B. A., E. L. Sa, M. E. Sieracki, R. Massana, and J. del Campo. 2020. Mediocremonas mediterraneus ,a new member within the Developea. *The Journal of Eukaryotic Microbiology* 1: 0–2.

Wetherbee, R., C. J. Jackson, S. I. Repetti, L. A. Clementson, J. F. Costa, A. Meene, S. Crawford, and H. Verbruggen. 2019. The golden paradox – a new heterokont lineage with chloroplasts surrounded by two membranes L. Graham [ed.],. *Journal of Phycology* 55: 257–278.

Weynberg, K., M. Allen, and W. Wilson. 2017. Marine Prasinoviruses and Their Tiny Plankton Hosts: A Review. *Viruses* 9: 43.

Weynberg, K. D., M. J. Allen, I. C. Gilg, D. J. Scanlan, and W. H. Wilson. 2011. Genome Sequence of Ostreococcus tauri Virus OtV-2 Throws Light on the Role of Picoeukaryote Niche Separation in the Ocean. *Journal of Virology* 85: 4520–4529.

Whelan, N. V., K. M. Kocot, L. L. Moroz, and K. M. Halanych. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences* 112: 5773–5778.

Whelan, S., I. Irisarri, and F. Burki. 2018. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences J. Hancock [ed.],. *Bioinformatics* 34: 3929–3930.

Whitfield, J. B., and P. J. Lockhart. 2007. Deciphering ancient rapid radiations. *Trends in Ecology & Evolution* 22: 258–265.

Wick, R. R., L. M. Judd, C. L. Gorrie, and K. E. Holt. 2017. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology* 13: e1005595.

Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. 2016. Springer International Publishing : Imprint: Springer, Cham.

Wilson, W. H., J. L. Van Etten, and M. J. Allen. 2009. The Phycodnaviridae: The Story of How Tiny Giants Rule the World. *In* J. L. Van Etten [ed.], Lesser Known Large dsDNA Viruses, Current Topics in Microbiology and Immunology, 1–42. Springer Berlin Heidelberg, Berlin, Heidelberg.

Wolf, S., I. Maier, C. Katsaros, and D. G. M�ller. 1998. Virus assembly in Hincksia hincksiae (Ectocarpales, Phaeophyceae): An electron and fluorescence microscopic study. *Protoplasma* 203: 153–167.

Wu, W., and B. Huang. 2019. Protist diversity and community assembly in surface sediments of the South China Sea. *MicrobiologyOpen* 8.

Yamada, N., J. J. Bolton, R. Trobajo, D. G. Mann, P. Dąbek, A. Witkowski, R. Onuma, et al. 2019. Discovery of a kleptoplastic 'dinotom' dinoflagellate and the unique nuclear dynamics of converting kleptoplastids to permanent plastids. *Scientific Reports* 9: 1–13.

Yang, E. C., G. H. Boo, H. J. Kim, S. M. Cho, S. M. Boo, R. A. Andersen, and H. S. Yoon. 2012. Supermatrix Data Highlight the Phylogenetic Relationships of Photosynthetic Stramenopiles. *Protist* 163: 217–231.

Yao, J., W. Fu, X. Wang, and D. Duan. 2009. Improved RNA isolation from Laminaria japonica Aresch (Laminariaceae, Phaeophyta). *Journal of Applied Phycology* 21: 233–238.

Yau, S., C. Hemon, E. Derelle, H. Moreau, G. Piganeau, and N. Grimsley. 2016. A Viral Immunity Chromosome in the Marine Picoeukaryote, Ostreococcus tauri S.-W. Ding [ed.],. *PLOS Pathogens* 12: e1005965.

Yau, S., M. Krasovec, L. F. Benites, S. Rombauts, M. Groussin, E. Vancaester, J.-M. Aury, et al. 2020. Virus-host coexistence in phytoplankton through the genomic lens. *Science Advances*.

Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya. 2004. A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. *Molecular Biology and Evolution* 21: 809–818.

Yubuki, N., and B. S. Leander. 2013. Evolution of microtubule organizing centers across the tree of eukaryotes. *Plant Journal* 75: 230–244.

Yubuki, N., B. S. Leander, and J. D. Silberman. 2010. Ultrastructure and Molecular Phylogenetic Position of a Novel Phagotrophic Stramenopile from Low Oxygen Environments: Rictus lutensis gen. et sp. nov. (Bicosoecida, incertae sedis). *Protist* 161: 264–278.

Yubuki, N., T. Pánek, A. Yabuki, I. Čepička, K. Takishita, Y. Inagaki, and B. S. Leander. 2015. Morphological Identities of Two Different Marine Stramenopile Environmental Sequence Clades: *Bicosoeca kenaiensis* (Hilliard, 1971) and *Cantina marsupialis* (Larsen and Patterson, 1990) gen. nov., comb. nov. *Journal of Eukaryotic Microbiology* 62: 532–542.

Yutin, N., D. Raoult, and E. V. Koonin. 2013. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virology Journal* 10: 158.

Zheng, L., and J. J. Mackrill. 2016. Calcium Signaling in Oomycetes: An Evolutionary Perspective. *Frontiers in Physiology* 7.

Zhong, M., B. Hansen, M. Nesnidal, A. Golombek, K. M. Halanych, and T. H. Struck. 2011. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. *BMC Evolutionary Biology* 11: 369.

Zhou, Y., N. Rodrigue, N. Lartillot, and H. Philippe. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. *BMC Evolutionary Biology* 7: 206.

Zimmerman, A. E., C. Bachy, X. Ma, S. Roux, H. B. Jang, M. B. Sullivan, J. R. Waldbauer, and A. Z. Worden. 2019. Closely related viruses of the marine picoeukaryotic alga *Ostreococcus lucimarinus* exhibit different ecological strategies. *Environmental Microbiology* 21: 2148–2170.

Zmitrovich, I. V. 2018. The Oomycota phenomenon. 188–190. Mycology and lichenology, St. Petersburg.

Appendices

Appendix A: List of published genomic level data of stramenopiles used in this study.

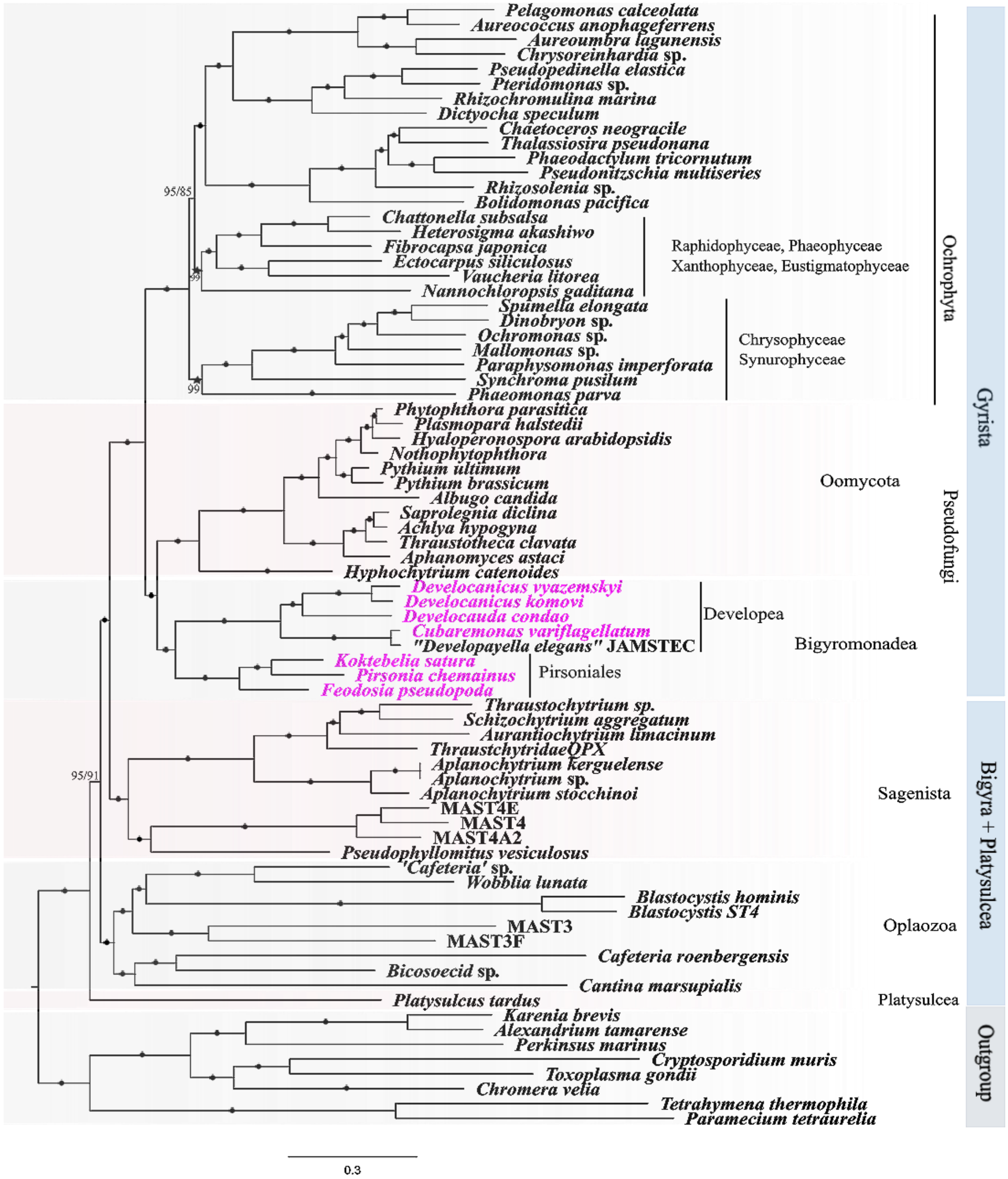
List of 27 recently published Stramenopiles taxa and the corresponding genome or transcriptome data included in this study. 'Peptide reads' were extracted from an annotated genome sequences and publicly available Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) website (Callahan et al., 2016). 'Transcripts' were extracted from an unannotated genome sequences, and protein sequences were predicted as described in Method.

Group	Sample ID	Type	Project	SRA Run	GenBank	Strain
Ochrophytes	*Rhizosolenia setigera* (MMETSP0789)	Peptide reads	PRJNA248394	SRR1296707		CCMP 1694
	Synchroma pusilum (MMETSP1452)	Peptide reads	PRJNA248394	SRR1300531		CCMP 3072
	Aureococcus anophagefferens	Peptide reads	PRJNA13500		GCA_000186865.1	CCMP 1984
Oomycetes	*Phytophthora parasitica*	Peptide reads	PRJNA73155		GCA_000247585.2	INRA-310
	Plasmopara halstedii	Peptide reads	PRJEB6932		GCA_900000015.1	
	Hyaloperonospora arabidopsidis	Peptide reads	PRJNA298674		GCA_001414525.1	Noks1
	Nothophytophthora sp.	Peptide reads	PRJNA328215		GCA_001712635.2	Chile5
	Pythium ultimum	Peptide reads	PRJNA36503		GCA_000143045.1	DAOM BR144
	Pythium brassicum	Peptide reads	PRJNA498716		GCA_008271595.1	P1
	Albugo candida	Peptide reads	PRJNA291031		GCA_001306755.1	Ac 7v
	Saprolegnia diclina	Peptide reads	PRJNA86859		GCA_000281045.1	VS20
	Achlya hypogyna	Peptide reads	PRJNA169234		GCA_002081595.1	ATCC 48635
	Thraustotheca clavata	Peptide reads	PRJNA169235		GCA_002081575.1	ATCC 34112
	Aphanomyces astaci	Peptide reads	PRJNA187372		GCA_000520075.1	APO3
	Hyphochytrium catenoides	Transcripts	PRJEB13950		GCA_900088475.1	

Developea	*Developayella elegans*	Transcriptome	PRJDB4370	DRR049556		CCAP:1917/1
Group	Sample ID	Type	Project	SRA Run	GenBank	Strain
Sagenista	*Pseudophyllomitus vesiculosus*	Transcriptome	PRJDB8568	DRR186658		NIES-4114
	MAST4	Peptide reads	PRJNA244411	SRR1263007		dcp33
	MAST4A2	Transcripts	PRJEB6603		GCA_900128395.1	TOSAG23-2
	MAST4E	Transcripts	PRJEB6603		GCA_900128585.1	TOSAG23-3
Oplozoa	*Wobblia lunata*	Transcriptome	PRJDB4369	DRR049555		NIES-1015
	*Blastocystis*_ST4	Peptide reads	PRJNA257240		GCA_000743755.1	WR1
	MAST3	Transcripts	PRJEB6603	ERR1198953		TOSAG41-2
	MAST3F	Transcripts	PRJEB6603		GCA_900128565.1	TOSAG23-6
	Cafeteria roenbergensis	Peptide reads	PRJNA552725		GCA_008330645.1	BVI
	Bicosoecid sp. (MMETSP0115)	Peptide reads	PRJNA231566	SRR1294380		
	Cantina marsupialis	Transcriptome	PRJDB3523	DRR030401		YPF1205
Platysulcea	*Platysulcus tardus*	Transcriptome	PRJDB8466	DRR186656		NIES-3720

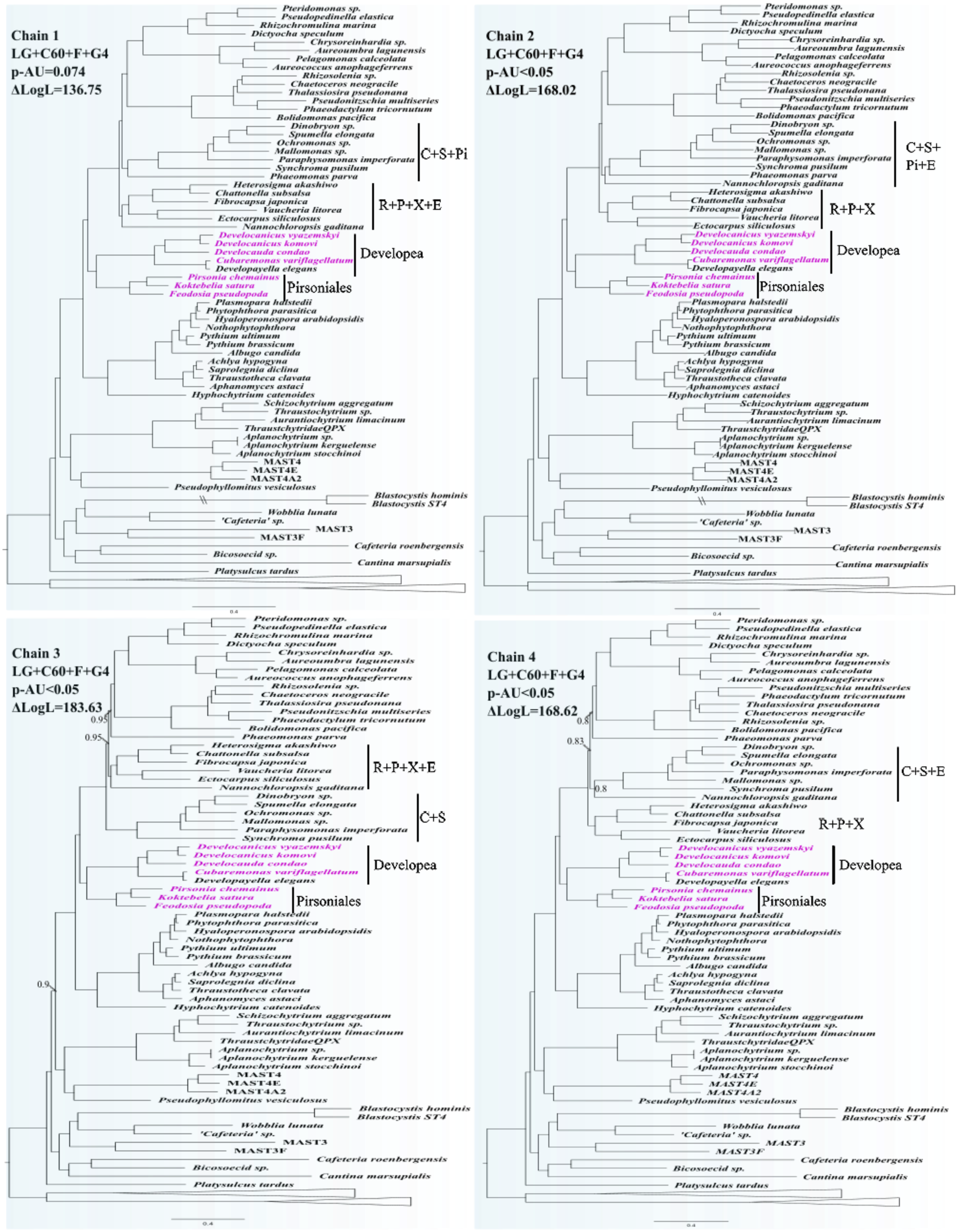
Appendix B: Phylogenomic tree of stramenopiles with the seven new transcriptomes using the approach 2 dataset.

Multi-gene phylogenomic tree of stramenopiles with the seven new transcriptomes (pink) added to Gyrista, consisting of the concatenated alignments of 247 aligned gene-sets. The tree was reconstructed using a maximum-likelihood (ML) analysis, under the site-heterogenous model, LG+C60+F+G4+PMSF implemented in IQ-Tree. The tree topology is based on the tree reconstructed on the dataset process with approach 2. Branch support was calculated separately



Appendix C: Bayesian phylogenomic tree of stramenopiles

Bayesian phylogenomic tree of stramenopiles of the seven new transcriptomes (pink) added to Gyrista. The tree was reconstructed based on the 247 gene-sets of 76 taxa processed with the approach 1 using PhyloBayes under the CAT+GTR+G4 model. No chains converged (maxdiff=1) and all chains have identical tree topologies except the sub-clades of ochrophytes as summarized as different combination of Raphidophyceae (R), Eustigmatophyceae (E), Chrysophyceae (C), Synurophyceae (S), Phaeophyceae (P), Pinguiophyceae (Pi), and Xanthophyceae (X). *Rhizosolenia* sp., also showed inconsistent topology across the chains. P-values were calculated using the approximately unbiased test (p-AU) with 10,000 RELL bootstrap replicates, implemented in IQ-TREE. The difference in maximum log likelihoods (ΔLogL) of each tree was calculated by comparing to the maximum log likelihood of ML tree reconstructed under the LG+C60+F+G4. Except chain1, the topologies of the trees from chain 2-4 were rejected where their p-AU were less than 0.05, indicating confidence interval below 95%. Only the Bayesian posterior probabilities (PP) lower than 1 are marked in the figure. All other nodes have PP=1. The collapsed clades in the figures indicate outgroup (alveolates).


Chain 1
LG+C60+F+G4
p-AU=0.074
ΔLogL=136.75
Pteridomonas sp.
Pseudopedinella elastica
Rhizochromulina marina
Dictyocha speculum
Chrysoreinhardia sp.
Aureoumbra lagunensis
Pelagomonas calceolata
Aureococcus anophageferrens
Rhizosolenia sp.
Chaetoceros neogracile
Thalassiosira pseudonana
Pseudonitzschia multiseries
Phaeodactylum tricornutum
Bolidomonas pacifica
Dinobryon sp.
Spumella elongata
Ochromonas sp.
Mallomonas sp.
Paraphysomonas imperforata
Synchroma pusilum
Phaeomonas parva
C+S+Pi
Heterosigma akashiwo
Chattonella subsalsa
Fibrocapsa japonica
Vaucheria litorea
Ectocarpus siliculosus
Nannochloropsis gaditana
R+P+X+E
Develocanicus vyazemskyi
Develocanicus komovi
Develocauda condao
Cubaremonas variflagellatum
Developayella elegans
Developea
Pirsonia chemainus
Koktebelia satura
Feodosia pseudopoda
Pirsoniales
Plasmopara halstedii
Phytophthora parasitica
Hyaloperonospora arabidopsidis
Nothophytophthora
Pythium ultimum
Pythium brassicum
Albugo candida
Achlya hypogyna
Saprolegnia diclina
Thraustotheca clavata
Aphanomyces astaci
Hyphochytrium catenoides
Schizochytrium aggregatum
Thraustochytrium sp.
Aurantiochytrium limacinum
ThraustchytridaeQPX
Aplanochytrium sp.
Aplanochytrium kerguelense
Aplanochytrium stocchinoi
MAST4
MAST4E
MAST4A2
Pseudophyllomitus vesiculosus
Blastocystis hominis
Blastocystis ST4
Wobblia lunata
'Cafeteria' sp.
MAST3
MAST3F
Cafeteria roenbergensis
Bicosoecid sp.
Cantina marsupialis
Platysulcus tardus
0.4
Chain 2
LG+C60+F+G4
p-AU<0.05
ΔLogL=168.02
Pteridomonas sp.
Pseudopedinella elastica
Rhizochromulina marina
Dictyocha speculum
Chrysoreinhardia sp.
Aureoumbra lagunensis
Pelagomonas calceolata
Aureococcus anophageferrens
Rhizosolenia sp.
Chaetoceros neogracile
Thalassiosira pseudonana
Pseudonitzschia multiseries
Phaeodactylum tricornutum
Bolidomonas pacifica
Dinobryon sp.
Spumella elongata
Ochromonas sp.
Mallomonas sp.
Paraphysomonas imperforata
Synchroma pusilum
Phaeomonas parva
Nannochloropsis gaditana
C+S+
Pi+E
Heterosigma akashiwo
Chattonella subsalsa
Fibrocapsa japonica
Vaucheria litorea
Ectocarpus siliculosus
R+P+X
Develocanicus vyazemskyi
Develocanicus komovi
Develocauda condao
Cubaremonas variflagellatum
Developayella elegans
Developea
Pirsonia chemainus
Koktebelia satura
Feodosia pseudopoda
Pirsoniales
Plasmopara halstedii
Phytophthora parasitica
Hyaloperonospora arabidopsidis
Nothophytophthora
Pythium ultimum
Pythium brassicum
Albugo candida
Achlya hypogyna
Saprolegnia diclina
Thraustotheca clavata
Aphanomyces astaci
Hyphochytrium catenoides
Schizochytrium aggregatum
Thraustochytrium sp.
Aurantiochytrium limacinum
ThraustchytridaeQPX
Aplanochytrium sp.
Aplanochytrium kerguelense
Aplanochytrium stocchinoi
MAST4
MAST4E
MAST4A2
Pseudophyllomitus vesiculosus
Blastocystis hominis
Blastocystis ST4
Wobblia lunata
'Cafeteria' sp.
MAST3
MAST3F
Cafeteria roenbergensis
Bicosoecid sp.
Cantina marsupialis
Platysulcus tardus
0.4
Chain 3
LG+C60+F+G4
p-AU<0.05
ΔLogL=183.63
Pteridomonas sp.
Pseudopedinella elastica
Rhizochromulina marina
Dictyocha speculum
Chrysoreinhardia sp.
Aureoumbra lagunensis
Pelagomonas calceolata
Aureococcus anophageferrens
Rhizosolenia sp.
Chaetoceros neogracile
Thalassiosira pseudonana
Pseudonitzschia multiseries
Phaeodactylum tricornutum
Bolidomonas pacifica
0.95
0.95
Phaeomonas parva
Heterosigma akashiwo
Chattonella subsalsa
Fibrocapsa japonica
Vaucheria litorea
Ectocarpus siliculosus
Nannochloropsis gaditana
R+P+X+E
Dinobryon sp.
Spumella elongata
Ochromonas sp.
Mallomonas sp.
Paraphysomonas imperforata
Synchroma pusilum
C+S
Develocanicus vyazemskyi
Develocanicus komovi
Develocauda condao
Cubaremonas variflagellatum
Developayella elegans
Developea
Pirsonia chemainus
Koktebelia satura
Feodosia pseudopoda
Pirsoniales
Plasmopara halstedii
Phytophthora parasitica
Hyaloperonospora arabidopsidis
Nothophytophthora
Pythium ultimum
Pythium brassicum
Albugo candida
Achlya hypogyna
Saprolegnia diclina
Thraustotheca clavata
Aphanomyces astaci
Hyphochytrium catenoides
0.9
Schizochytrium aggregatum
Thraustochytrium sp.
Aurantiochytrium limacinum
ThraustchytridaeQPX
Aplanochytrium sp.
Aplanochytrium kerguelense
Aplanochytrium stocchinoi
MAST4
MAST4E
MAST4A2
Pseudophyllomitus vesiculosus
Blastocystis hominis
Blastocystis ST4
Wobblia lunata
'Cafeteria' sp.
MAST3
MAST3F
Cafeteria roenbergensis
Bicosoecid sp.
Cantina marsupialis
Platysulcus tardus
0.4
Chain 4
LG+C60+F+G4
p-AU<0.05
ΔLogL=168.62
Pteridomonas sp.
Pseudopedinella elastica
Rhizochromulina marina
Dictyocha speculum
Chrysoreinhardia sp.
Aureoumbra lagunensis
Pelagomonas calceolata
Aureococcus anophageferrens
Pseudonitzschia multiseries
Phaeodactylum tricornutum
Thalassiosira pseudonana
Chaetoceros neogracile
Rhizosolenia sp.
Bolidomonas pacifica
0.8
0.83
Phaeomonas parva
Dinobryon sp.
Spumella elongata
Ochromonas sp.
Paraphysomonas imperforata
Mallomonas sp.
0.8
Synchroma pusilum
Nannochloropsis gaditana
C+S+E
Heterosigma akashiwo
Chattonella subsalsa
Fibrocapsa japonica
Vaucheria litorea
Ectocarpus siliculosus
R+P+X
Develocanicus vyazemskyi
Develocanicus komovi
Develocauda condao
Cubaremonas variflagellatum
Developayella elegans
Developea
Pirsonia chemainus
Koktebelia satura
Feodosia pseudopoda
Pirsoniales
Plasmopara halstedii
Phytophthora parasitica
Hyaloperonospora arabidopsidis
Nothophytophthora
Pythium ultimum
Pythium brassicum
Albugo candida
Achlya hypogyna
Saprolegnia diclina
Thraustotheca clavata
Aphanomyces astaci
Hyphochytrium catenoides
Schizochytrium aggregatum
Thraustochytrium sp.
Aurantiochytrium limacinum
ThraustchytridaeQPX
Aplanochytrium sp.
Aplanochytrium kerguelense
Aplanochytrium stocchinoi
MAST4
MAST4E
MAST4A2
Pseudophyllomitus vesiculosus
Blastocystis hominis
Blastocystis ST4
Wobblia lunata
'Cafeteria' sp.
MAST3
MAST3F
Cafeteria roenbergensis
Bicosoecid sp.
Cantina marsupialis
Platysulcus tardus
0.4

Appendix D: Approximately unbiased test of constrained trees based on approach 2 dataset.

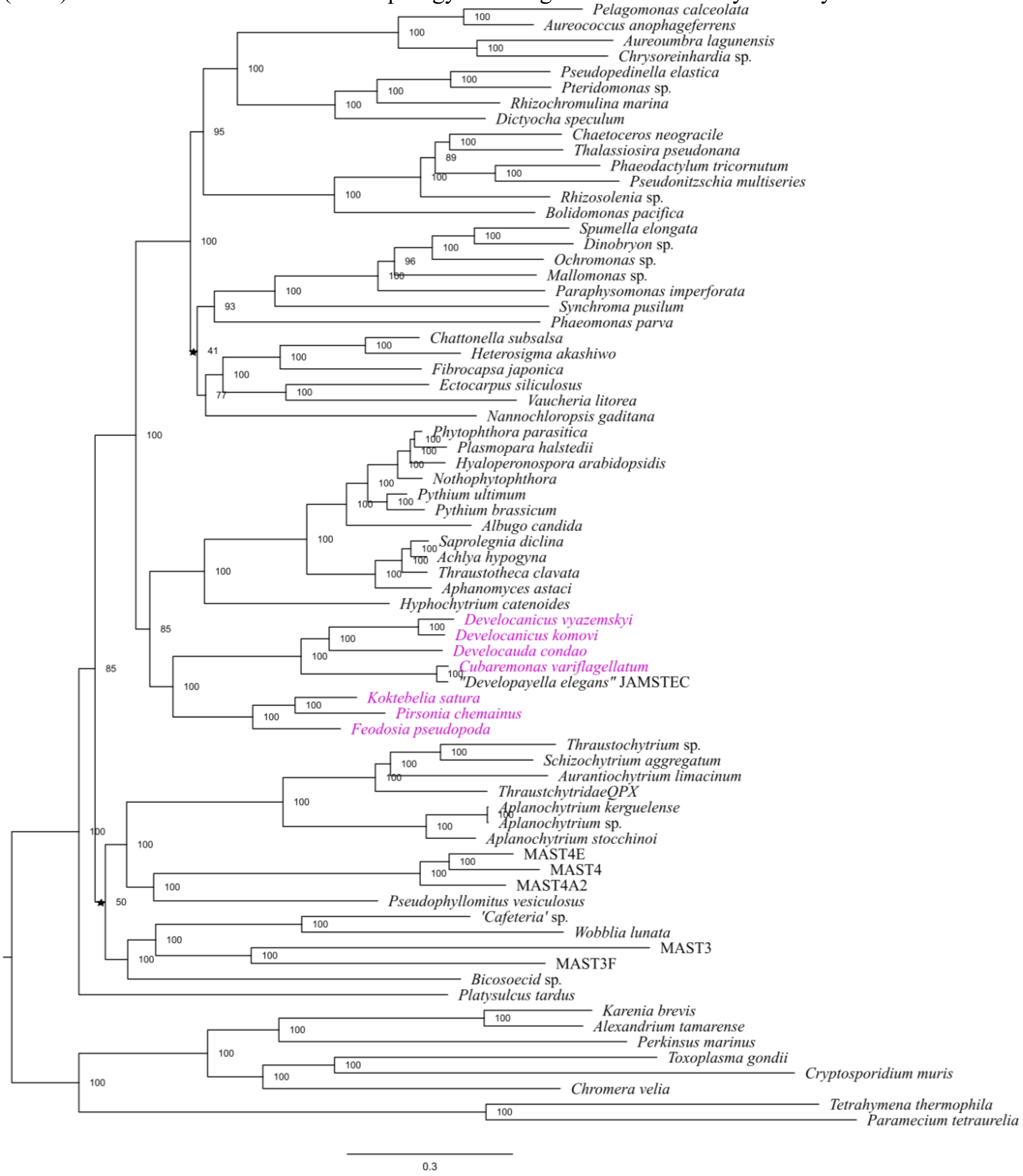
Approach 2 (Prequal/Divvier, MAFFT G-INS-i, -gt 0.1)

Constrained Tree	p-AU	logL	ΔlogL
Unconstrained ML tree	0.602	-4112709.551	0
ML tree under LG+C60+F+G4+PMSF	0.569	-4112709.552	0.00035827
ML tree under LG+C60+F+G4	0.543	-4112709.552	0.00035827
ML tree Modified (Bigyromonada+ochrophytes)	**0.0492**	-4112783.641	74.089

Except for the unconstrained ML tree, each tree was constrained under LG+C60+F+G4 using IQ-TREE with the dataset processed with approach 2. All the ML tree generated in this study (bigyromonada + oomycetes). “ML tree Modified” is a hypothetical tree constraint containing (bigyromonada + ochrophytes) with the rest of topology remaining the same with the unconstrained ML tree. The unconstrained tree is based on ML tree reconstructed under LG+C60+F+G4+PMSF as presented in Appendix B. The p-AU values were calculated using the AU test with 10,000 RELL bootstrap replicates, implemented in IQ-TREE. The maximum log likelihoods (logL) of each constrained and their differences (ΔlogL) compared to the unstrained tree are listed. Constraints with P-values lower than 0.05 are rejected, indicating confidence interval below 95% (marked bold).

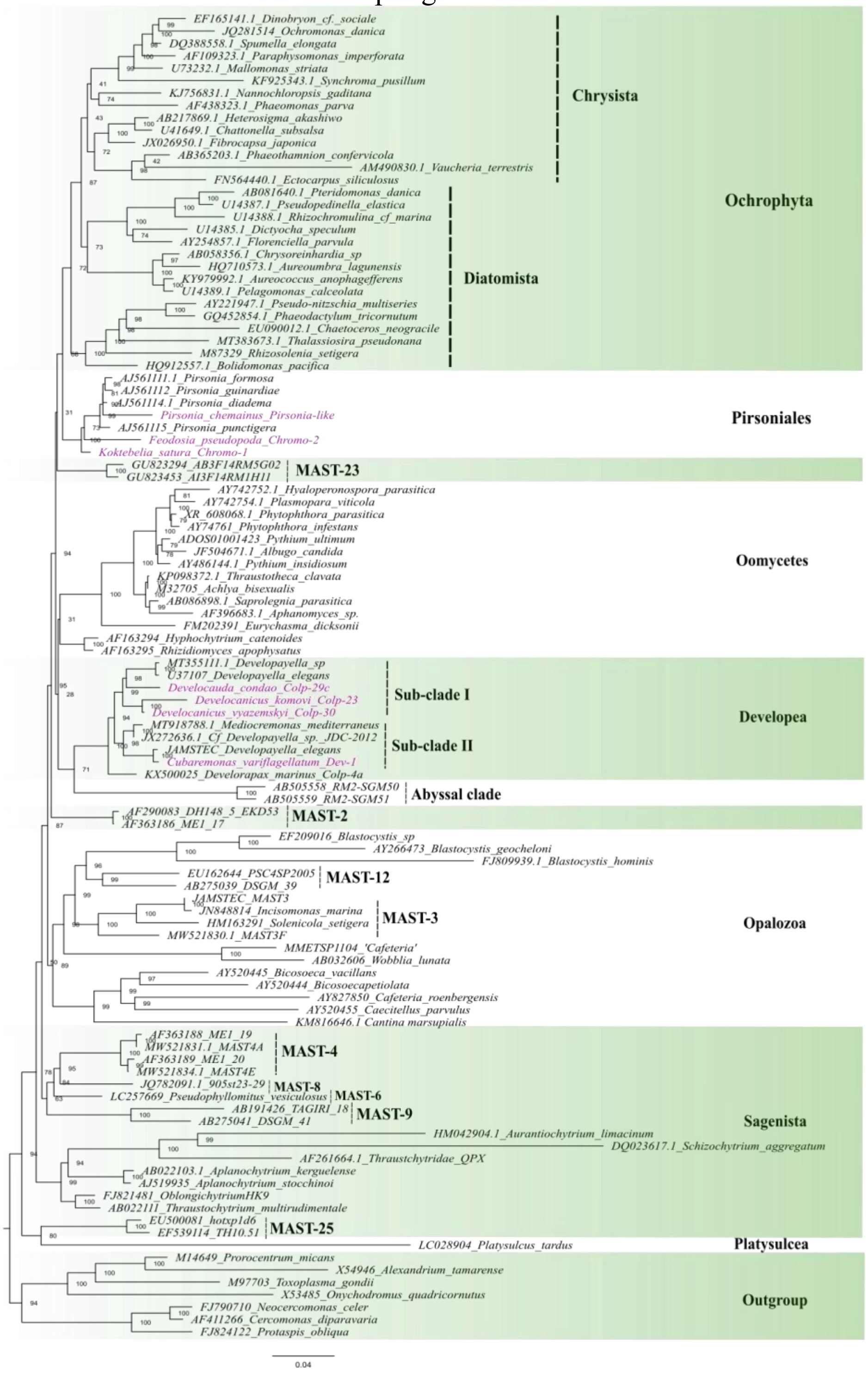
Appendix E: Phylogenomic tree of stramenopiles with the fast-evolving species removed.

Multi-gene phylogenomic tree of stramenopiles with the seven new transcriptomes (pink) added to Gyrista with the fast-evolving species removed (*Cafeteria roenbergensis,* two species of *Blastocystis* sp., and *Cantina marsupialis*). The tree was reconstructed using the Maximum-likelihood (ML) analysis, under the site-heterogenous model (LG+C60+F+G4) implemented in IQ-Tree, comprising 75,798 aa of 247 genes from 72 taxa. Branch support was calculated using 1000 ultrafast bootstrap (UFB). Branches that have different topology from Fig. 2.1 are marked by a star symbol.



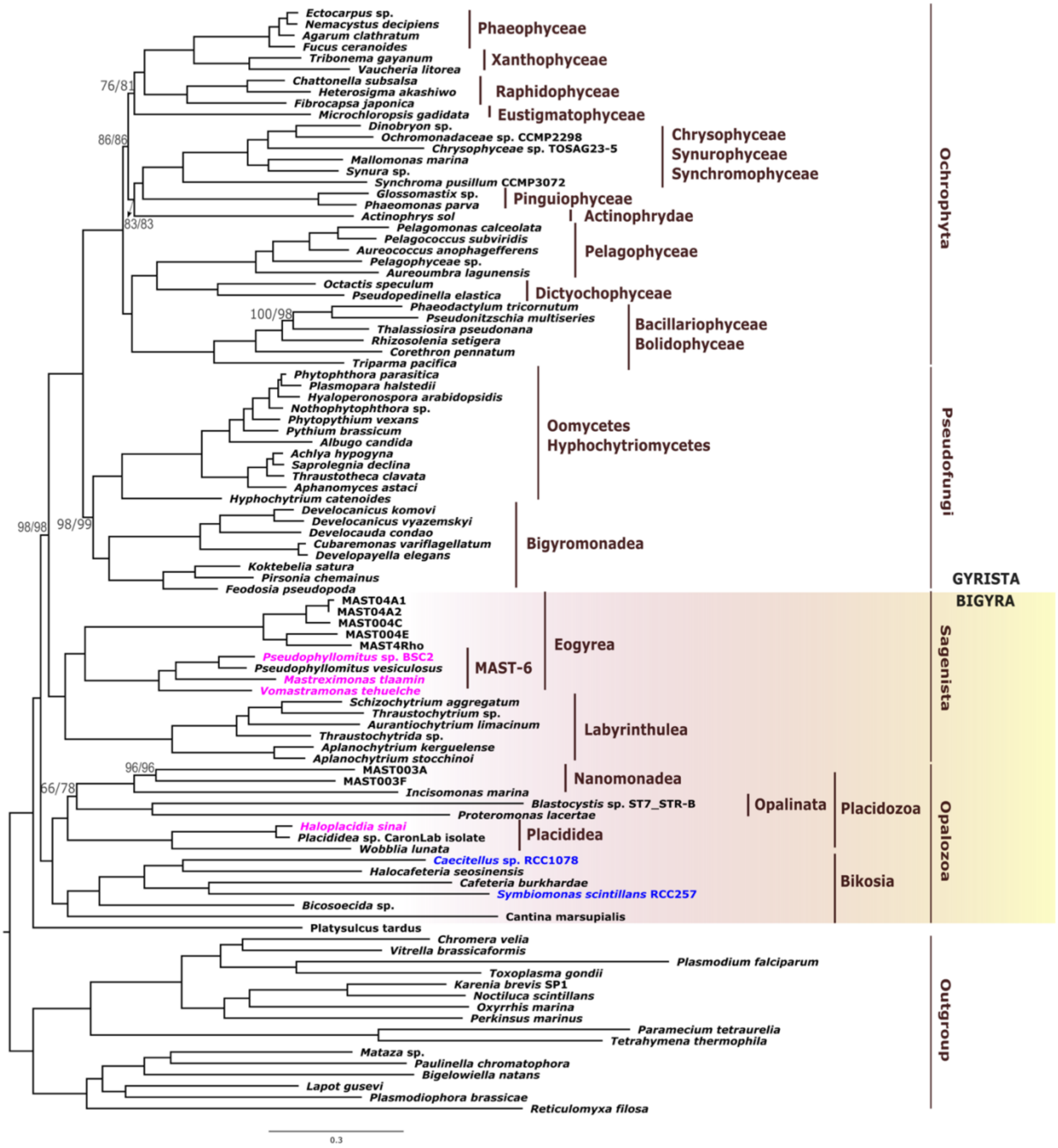
Appendix F: ML tree of a 18S rRNA gene including stramenopile environmental sequences.

ML tree reconstructed under BIC: TIM2+R6 with 1000 UFB from a 18S rRNA gene alignment of 107 taxa (1665 sites) including environmental sequences. The seven new species described in this study are marked as pink: Pirsoniales forming a sisterhood with Ochrophytes and Developea forming a sister clade to 'Abyssal Clade', demonstrating potential expansion of the Bigyromonada clade with further taxon sampling.


EF165141.1_Dinobryon_cf._sociale
JQ281514_Ochromonas_danica
DQ388558.1_Spumella_elongata
AF109323.1_Paraphysomonas_imperforata
U73232.1_Mallomonas_striata
KF925343.1_Synchroma_pusillum
KJ756831.1_Nannochloropsis_gaditana
AF438323.1_Phaeomonas_parva
AB217869.1_Heterosigma_akashiwo
U41649.1_Chattonella_subsalsa
JX026950.1_Fibrocapsa_japonica
AB365203.1_Phaeothamnion_confervicola
AM490830.1_Vaucheria_terrestris
FN564440.1_Ectocarpus_siliculosus
Chrysista
AB081640.1_Pteridomonas_danica
U14387.1_Pseudopedinella_elastica
U14388.1_Rhizochromulina_cf_marina
U14385.1_Dictyocha_speculum
AY254857.1_Florenciella_parvula
AB058356.1_Chrysoreinhardia_sp
HQ710573.1_Aureoumbra_lagunensis
KY979992.1_Aureococcus_anophagefferens
U14389.1_Pelagomonas_calceolata
AY221947.1_Pseudo-nitzschia_multiseries
GQ452854.1_Phaeodactylum_tricornutum
EU090012.1_Chaetoceros_neogracile
MT383673.1_Thalassiosira_pseudonana
M87329_Rhizosolenia_setigera
HQ912557.1_Bolidomonas_pacifica
Diatomista
Ochrophyta
AJ561111.1_Pirsonia_formosa
AJ561112_Pirsonia_guinardiae
AJ561114.1_Pirsonia_diadema
Pirsonia_chemainus_Pirsonia-like
AJ561115_Pirsonia_punctigera
Feodosia_pseudopoda_Chromo-2
Koktebelia_satura_Chromo-1
Pirsoniales
GU823294_AB3F14RM5G02
GU823453_AI3F14RM1H11
MAST-23
AY742752.1_Hyaloperonospora_parasitica
AY742754.1_Plasmopara_viticola
XR_608068.1_Phytophthora_parasitica
AY74761_Phytophthora_infestans
ADOS01001423_Pythium_ultimum
JF504671.1_Albugo_candida
AY486144.1_Pythium_insidiosum
KP098372.1_Thraustotheca_clavata
M32705_Achlya_bisexualis
AB086898.1_Saprolegnia_parasitica
AF396683.1_Aphanomyces_sp.
FM202391_Eurychasma_dicksonii
AF163294_Hyphochytrium_catenoides
AF163295_Rhizidiomyces_apophysatus
Oomycetes
MT355111.1_Developayella_sp
U37107_Developayella_elegans
Develocauda_condao_Colp-29c
Develocanicus_komovi_Colp-23
Develocanicus_vyazemskyi_Colp-30
Sub-clade I
MT918788.1_Mediocremonas_mediterraneus
JX272636.1_Cf_Developayella_sp._JDC-2012
JAMSTEC_Developayella_elegans
Cubaremonas_variflagellatum_Dev-1
Sub-clade II
KX500025_Develorapax_marinus_Colp-4a
Developea
AB505558_RM2-SGM50
AB505559_RM2-SGM51
Abyssal clade
AF290083_DH148_5_EKD53
AF363186_ME1_17
MAST-2
EF209016_Blastocystis_sp
AY266473_Blastocystis_geocheloni
FJ809939.1_Blastocystis_hominis
EU162644_PSC4SP2005
AB275039_DSGM_39
MAST-12
JAMSTEC_MAST3
JN848814_Incisomonas_marina
HM163291_Solenicola_setigera
MW521830.1_MAST3F
MAST-3
MMETSP1104_'Cafeteria'
AB032606_Wobblia_lunata
AY520445_Bicosoeca_vacillans
AY520444_Bicosoecapetiolata
AY827850_Cafeteria_roenbergensis
AY520455_Caecitellus_parvulus
KM816646.1 Cantina marsupialis
Opalozoa
AF363188_ME1_19
MW521831.1_MAST4A
AF363189_ME1_20
MW521834.1_MAST4E
MAST-4
JQ782091.1_905st23-29
MAST-8
LC257669_Pseudophyllomitus_vesiculosus
MAST-6
AB191426_TAGIRI_18
AB275041_DSGM_41
MAST-9
HM042904.1_Aurantiochytrium_limacinum
DQ023617.1_Schizochytrium_aggregatum
AF261664.1_Thraustchytridae_QPX
AB022103.1_Aplanochytrium_kerguelense
AJ519935_Aplanochytrium_stocchinoi
FJ821481_OblongichytriumHK9
AB022111_Thraustochytrium_multirudimentale
EU500081_hotxp1d6
EF539114_TH10.51
MAST-25
Sagenista
LC028904_Platysulcus_tardus
Platysulcea
M14649_Prorocentrum_micans
X54946_Alexandrium_tamarense
M97703_Toxoplasma_gondii
X53485_Onychodromus_quadricornutus
FJ790710_Neocercomonas_celer
AF411266_Cercomonas_diparavaria
FJ824122_Protaspis_obliqua
Outgroup
0.04

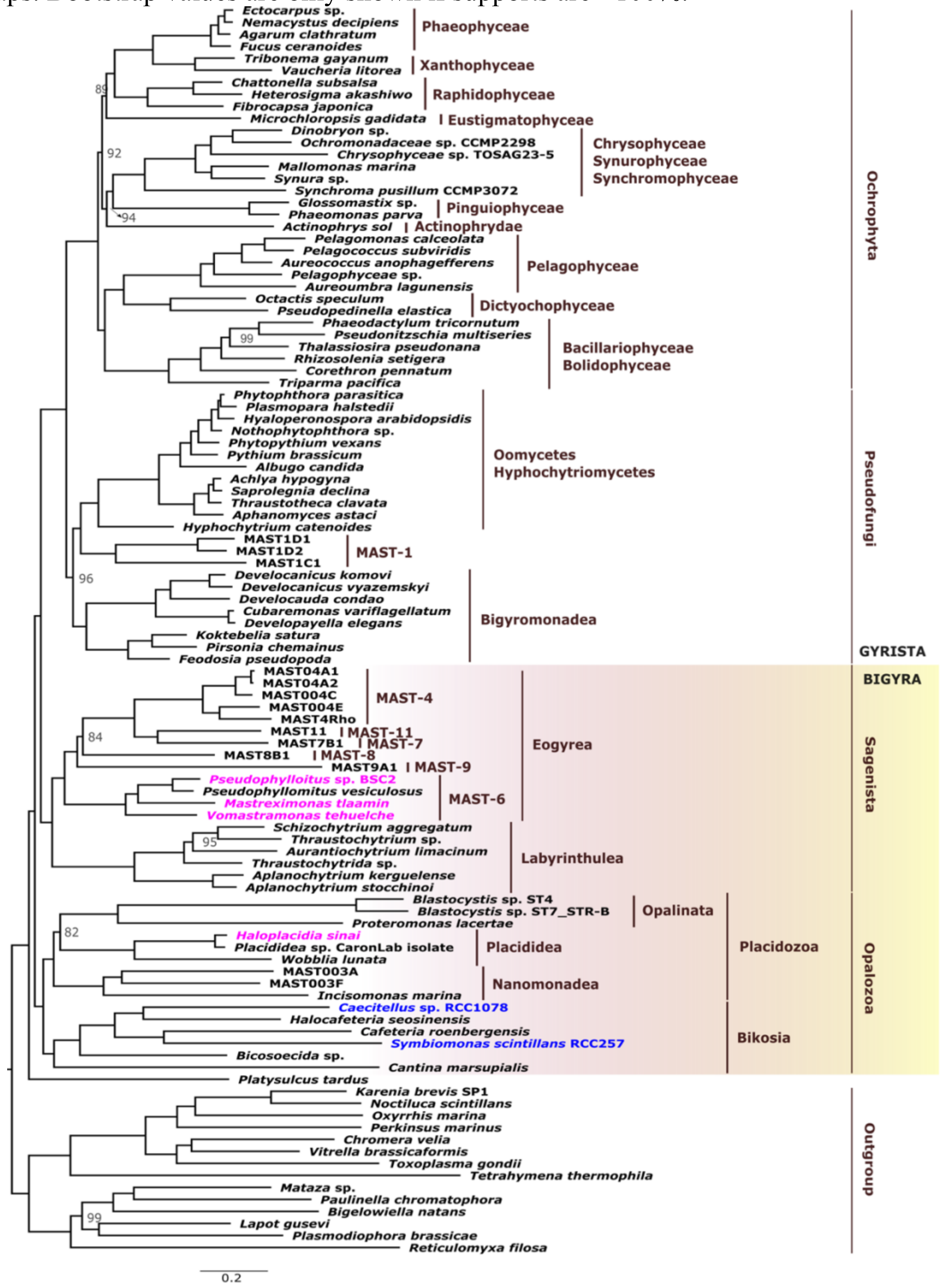
Appendix G: Maximum-likelihood phylogenomic trees inferred from 39per- and 59per-matrix.

Combined maximum-likelihood (ML) multi-gene trees of six new transcriptomes; four from newly described Bigyra (pink) in this study, and two from culture collections (blue). Two trees were constructed from a concatenated alignment of 39per-matrix (233 orthologs of 98 taxa with 74,531 aa), and 59per-matrix (215 orthologs of 98 taxa with 67,630 aa). The tree was estimated under the site-heterogenous model LG+C60+F+G4+PMSF with 100 standard bootstraps. The resulting trees had the same topology and bootstrap values are only shown if supports are <99% or the values between the two trees are different (39per-matrix vs 59per-matrix).

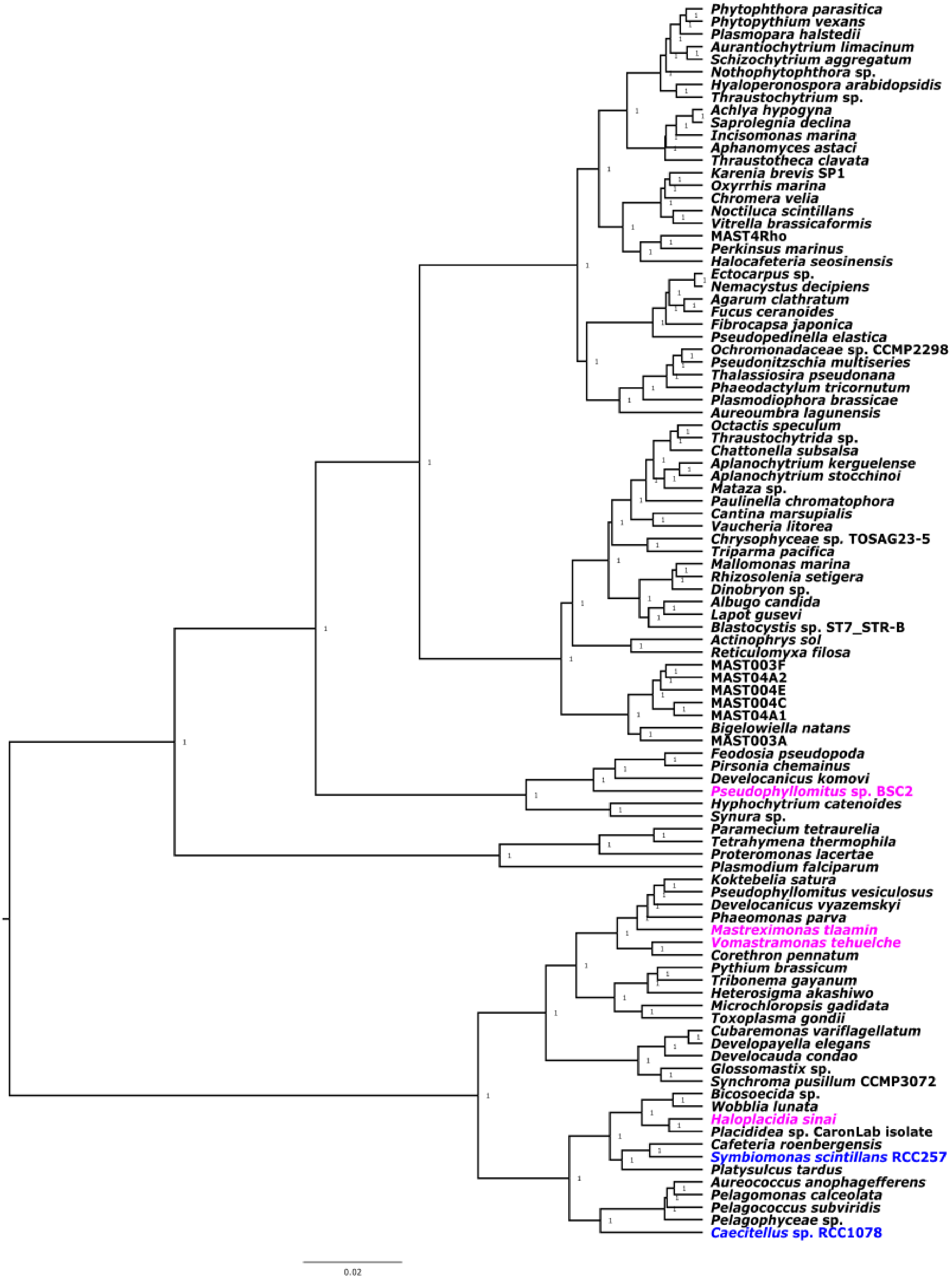


Appendix H: Maximum-likelihood phylogenomic trees inferred from MASTer-matrix.

Maximum-likelihood (ML) multi-gene trees of six new transcriptomes; four from newly described Bigyra (pink) in this study, and two from culture collections (blue). ML tree was constructed from a concatenated alignment of MASTer-matrix (234 orthologs of 104 taxa with 74,898 aa) which includes protein data from MAST-1, -7, -8, -9, and MAST-11. The tree was estimated under the site-heterogenous model LG+C60+F+G4+PMSF with 100 standard bootstraps. Bootstrap values are only shown if supports are <100%.

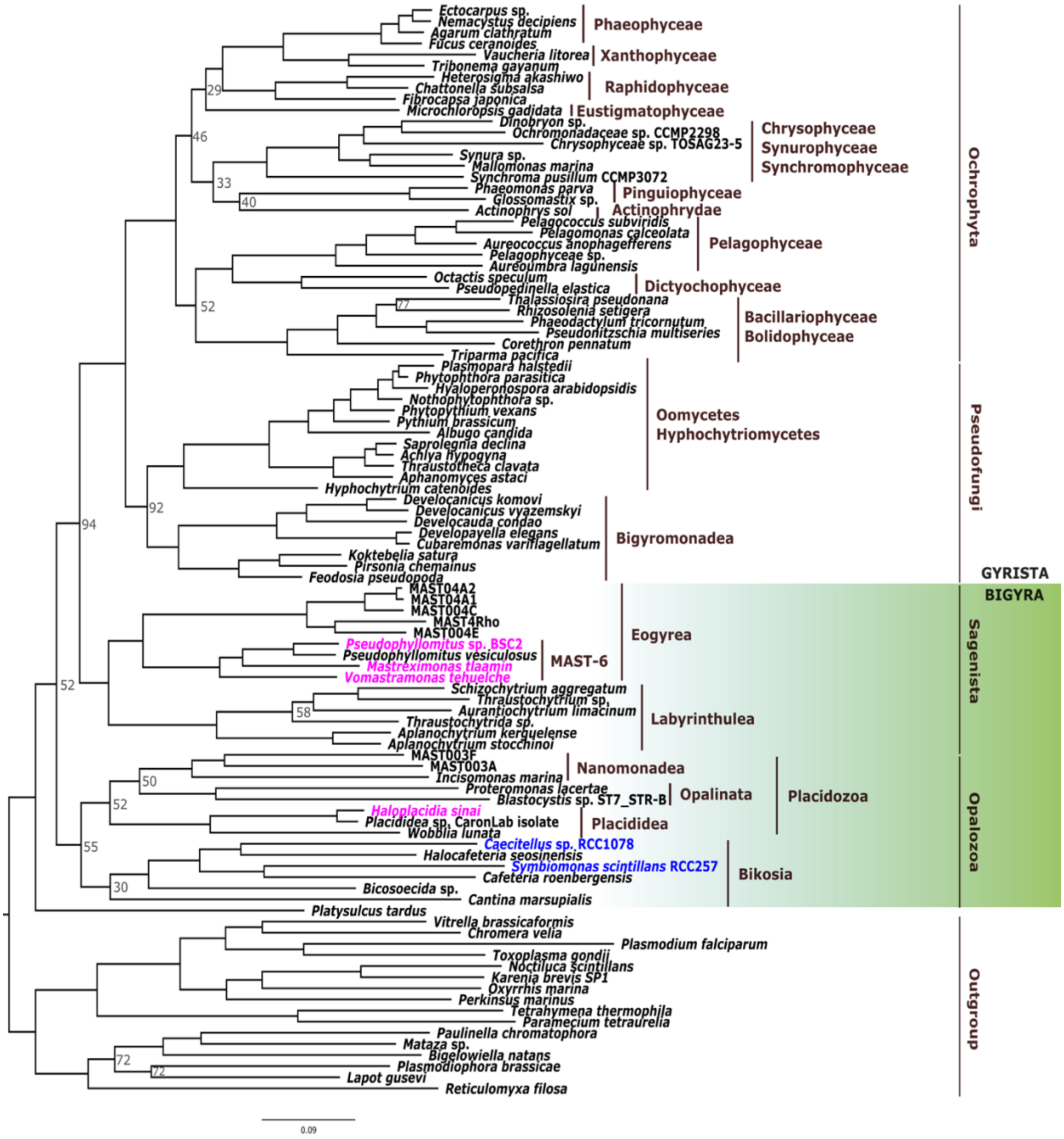


Appendix I: Pairwise-distance tree based on amino acid composition analysis.



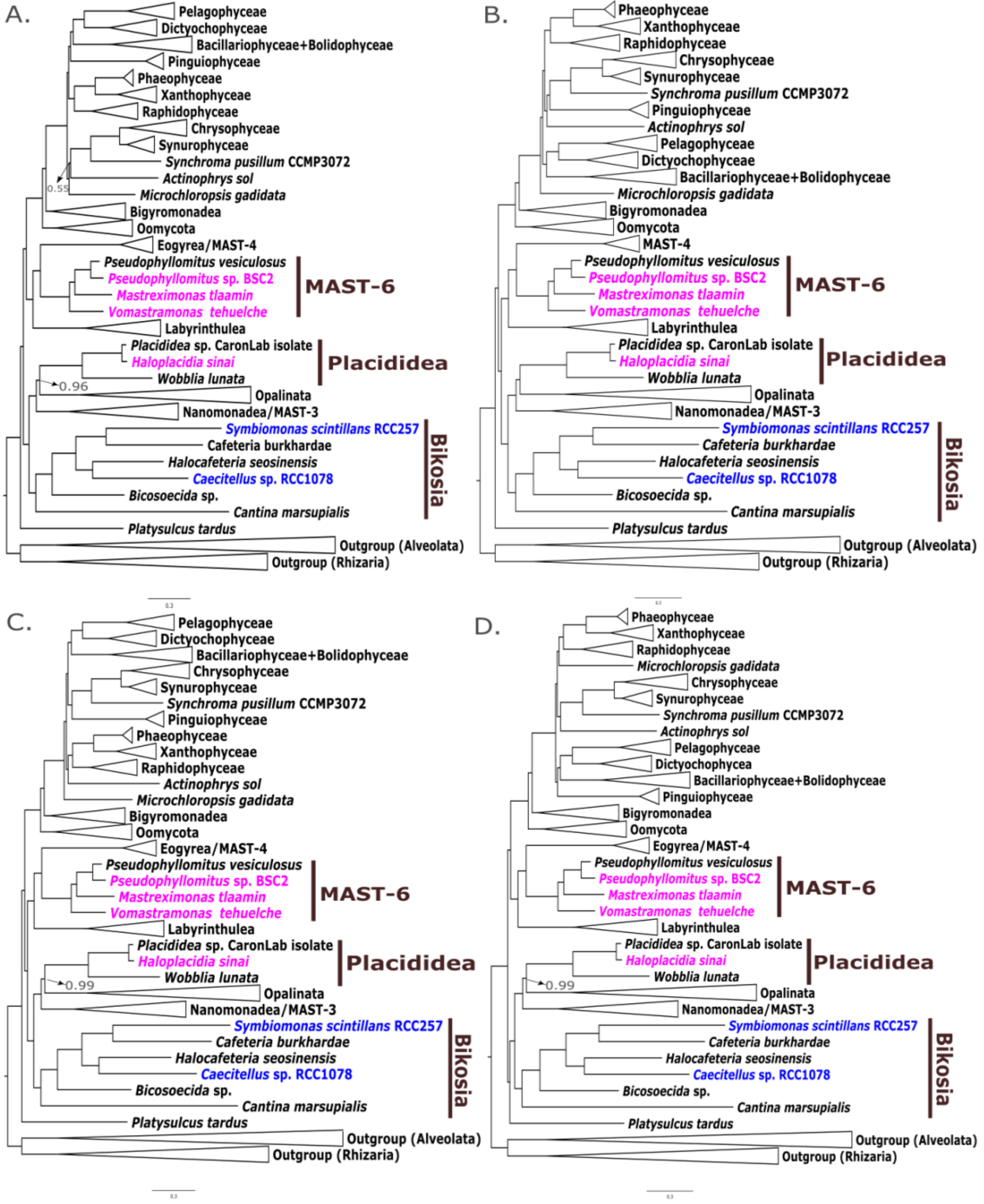
Appendix J: Recoded RAxML-ng multi-gene tree.

Phylogenomic tree of six new transcriptomes; four from newly described Bigyra (pink) in this study, and two from culture collections (blue). The tree was constructed from a recoded alignment of 76,516 sites under the MULTI18_GTR model with 100 standard bootstraps. The bootstrap supports are only labelled if <99%, all other unlabelled nodes indicate 99 or 100% support.



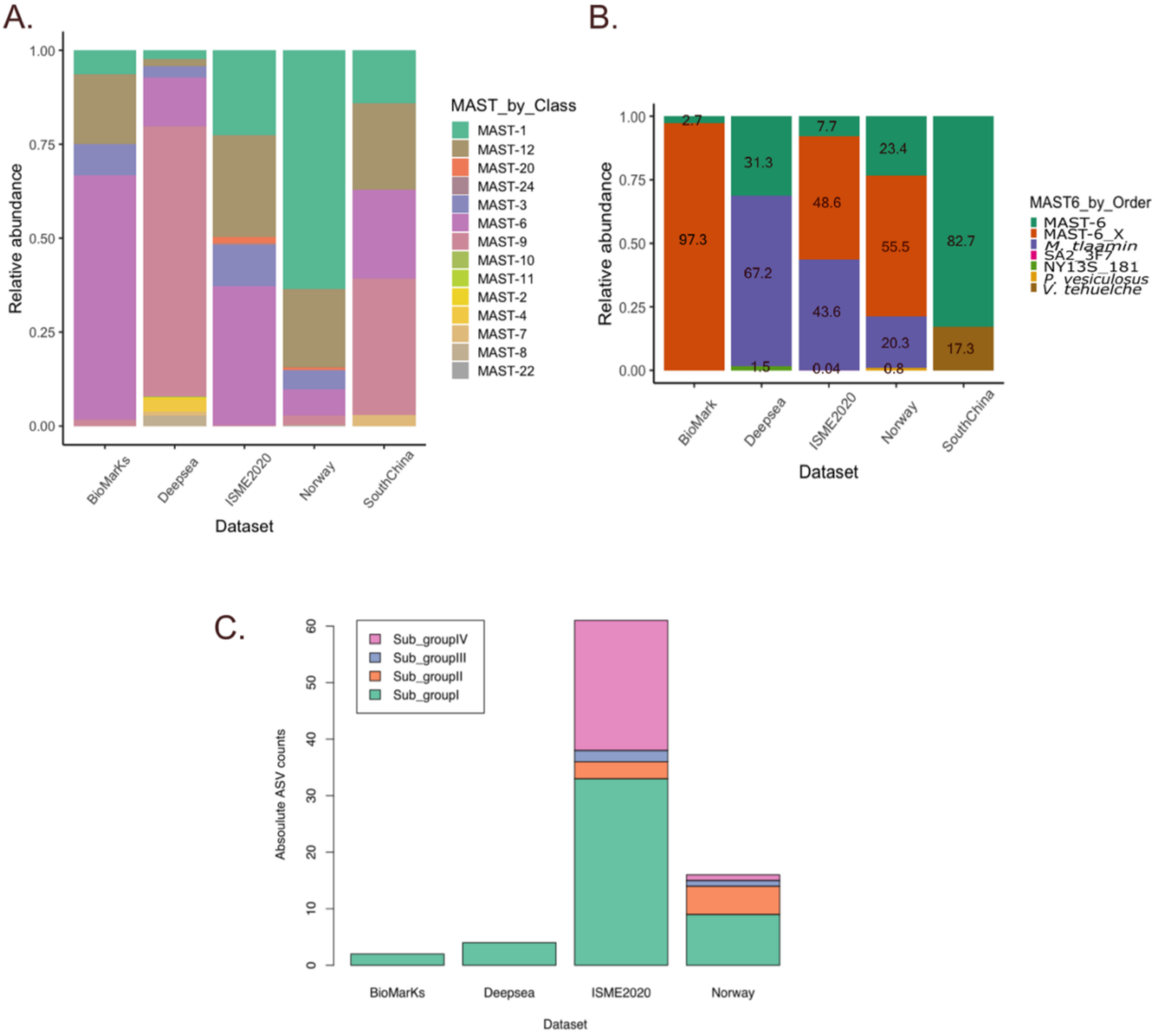
Appendix K: Consensus trees from Bayesian analysis.

Bayesian consensus trees of stramenopiles generated from four independent Markov Chain Monte Carlo (MCMC) chains (maxdiff=1). The tree was reconstructed based on the same matrix (240 genes and 98 taxa) used for the ML-PMSF inference, using PhyloBayes under the CAT+GTR+G4 model. There was no convergence among chains. The four MAST-6 and Placididea transcriptomes generated in this study are marked in pink, two culture collection bikosian species are marked in blue. Only posterior probabilities (PP) lower than 1 are labelled in the figure. All other nodes have PP=1. A=chain 1; B=chain 2; C=chain 3; D=chain 4



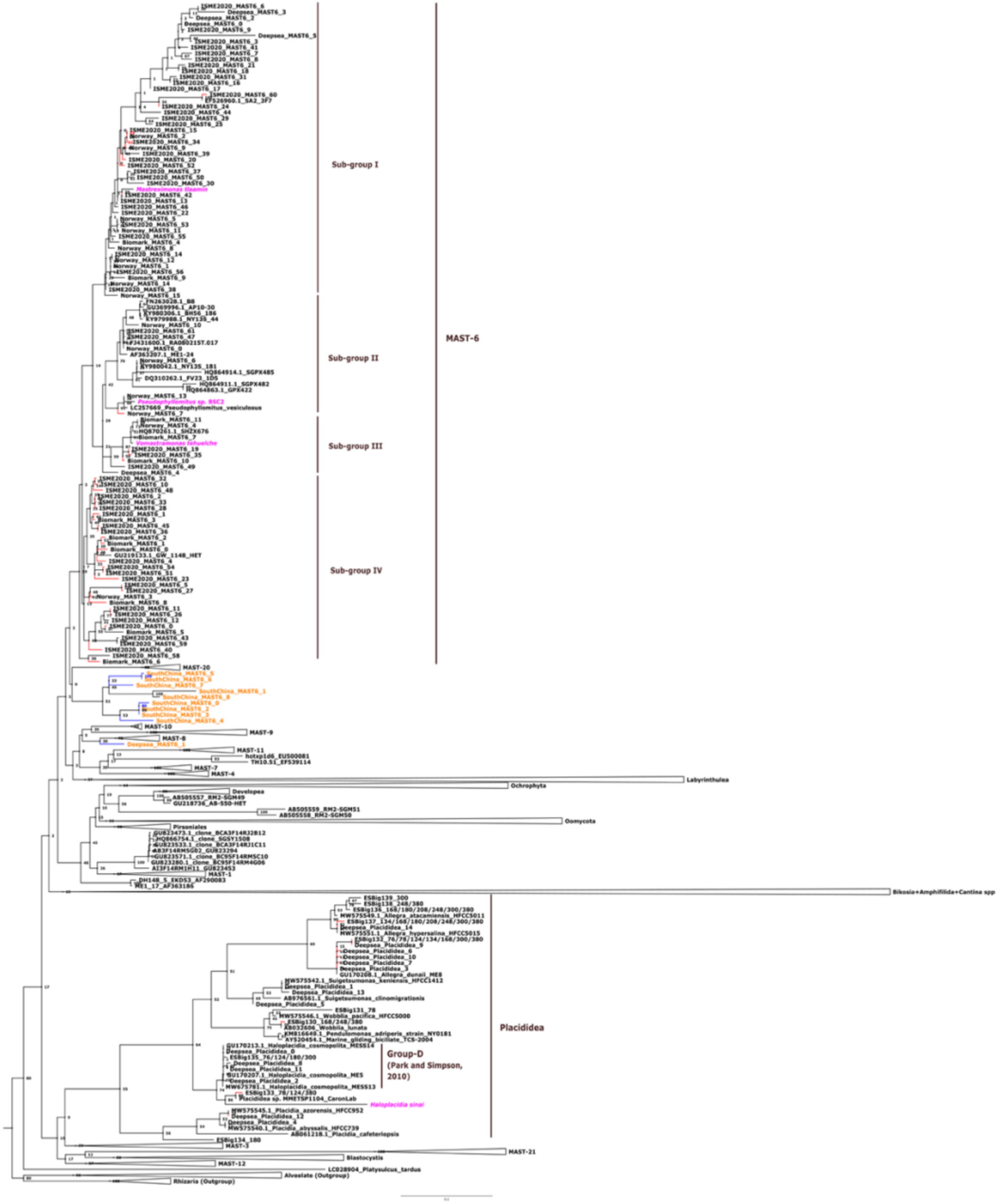
Appendix L: Stacked bar plots showing relative abundance of unique MAST ASVs.

Bar plots showing relative abundance or count of unique ASVs assigned to different MAST groups (A), MAST-6 species (B), and MAST-6 sub-groups in five sediment datasets. SouthChina is the only dataset that used a SSU rRNA gene primer were grouped by class level. **B.** Composition of MAST-6 lineages from each dataset were grouped by order to further show higher taxonomic assignment. “MAST-6_X” represents unknown MAST-6 lineages classified from PR2 database and “MAST-6” represents a potentially new MAST-6 lineage based on the updated taxonomic training database. The new MAST-6 species described in this study are “M. tlaamin” and “V. tehuelche”. “P. vesiculosus” and “NY13S_181” are cell isolates and “SA2_3F7” is an environmental sequence.



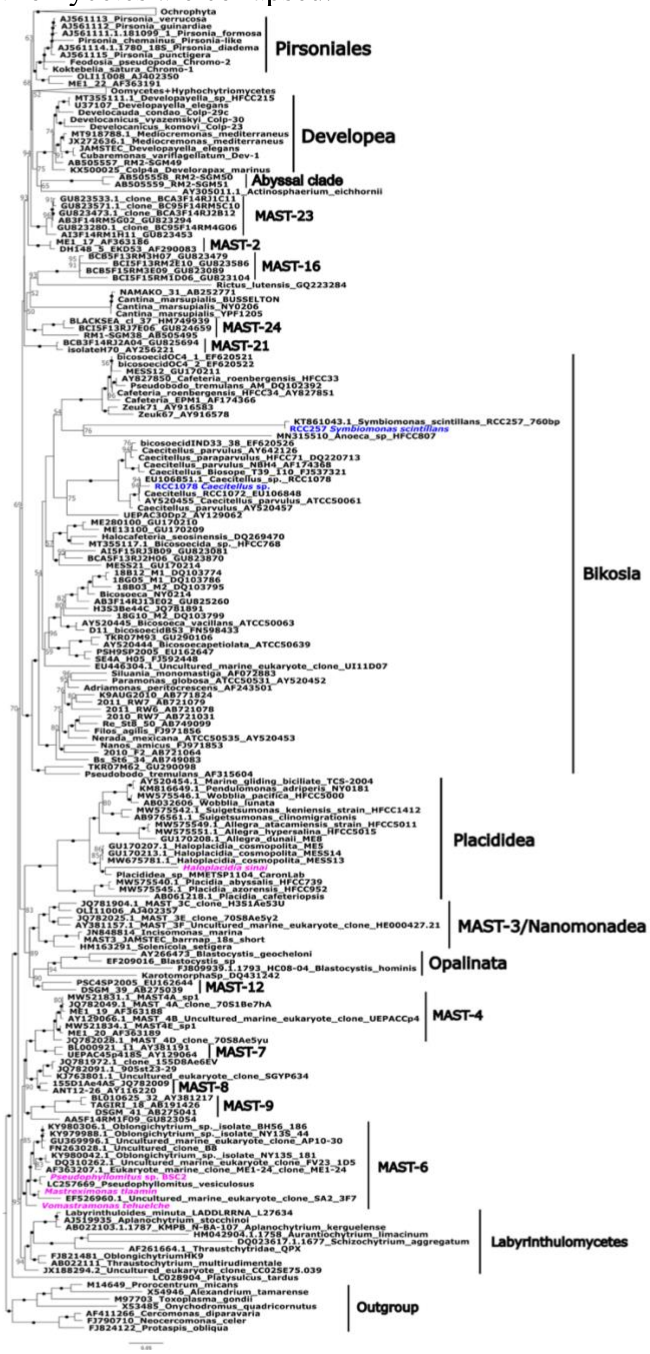
Appendix M: A SSU rRNA phylogenetic tree of stramenopiles with SouthChina dataset.

A RAxML SSU rRNA phylogenetic tree of stramenopiles. The tree was constructed under the GTR+GAMMA model with 1000 rapid bootstrap replicates, using an alignment of 548 stramenopile sequences and seven outgroup sequences, and included 119 extracted ASVs assigned to MAST-6 or Placididea from all the amplicon dataset (including SouthChina dataset) and 10 placididean OTU sequences from the ESBig study. The four new Bigyra species are coloured in pink. High confidence likelihood weight ratio values (LWR ≥95%) are denoted in red. Amplicon sequence variants assigned to MAST-6 in SouthChina and in Deepsea dataset are coloured in orange and excluded from the main figures. Blue nodes in these ASVs indicate low confidence (LWR <95%), indicating an equally likelihood of alternative placements. Clades other than MAST-6 and Placididea are collapsed.



Appendix N: A SSU rRNA phylogenetic tree of stramenopiles without ASVs.

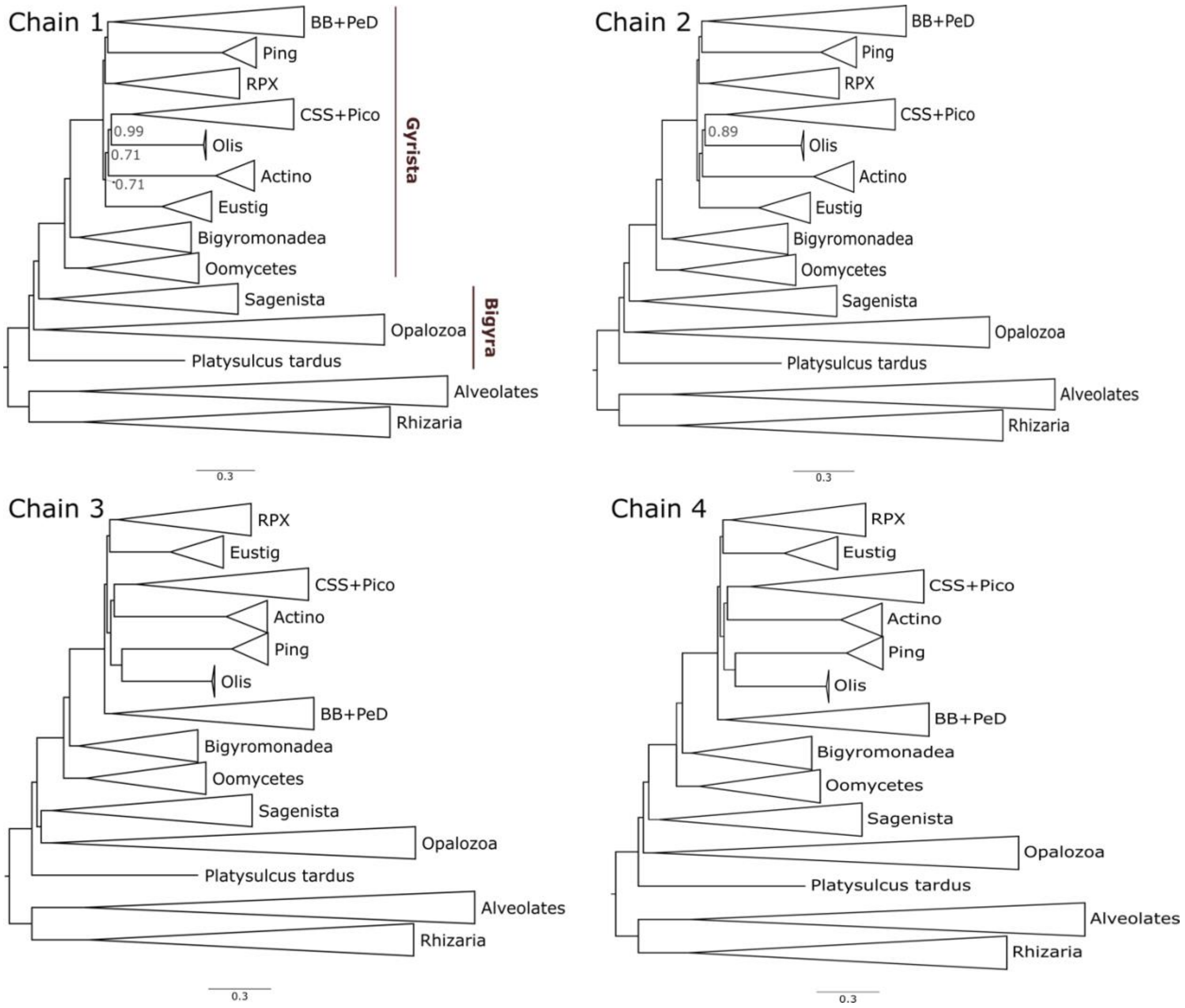
Maximum-likelihood SSU rRNA phylogenetic tree (SSU-tree) constructed using IQ-TREE under TIM2+F+R6 with 1000 ultrafast bootstraps, based on an alignment of 246 taxa and 1649 sites. Bootstrap supports of ≥97% are marked with black dots, while bootstrap supports of <50% are excluded. The tree includes relevant environmental or cell isolate SSU sequences but excludes extracted amplicon sequences from amplicon datasets. Sequences in pink indicate three newly added MAST-6 sequences and *H. sinai* sequence. Blue indicates new sequence data from *S. scintillans* and *Caecitellus* sp. Clades containing Ochrophyta, Oomycetes, and Hyphochytriomycetes are collapsed.



Ochrophyta
AJ561113_Pirsonia_verrucosa
AJ561112_Pirsonia_guinardiae
AJ561111.1.181099_1_Pirsonia_formosa
Pirsonia_chemainus_Pirsonia-like
AJ561114.1.1780_18S_Pirsonia_diadema
AJ561115_Pirsonia_punctigera
Feodosia_pseudopoda_Chromo-2
Koktebelia_satura_Chromo-1
Pirsoniales
OLI11008_AJ402350
ME1_22_AF363191
Oomycetes+Hyphochytriomycetes
MT355111.1_Developayella_sp_HFCC215
U37107_Developayella_elegans
Develocauda_condao_Colp-29c
Develocanicus_vyazemskyi_Colp-30
Develocanicus_komovi_Colp-23
MT918788.1_Mediocremonas_mediterraneus
JX272636.1_Mediocremonas_mediterraneus
JAMSTEC_Developayella_elegans
Cubaremonas_variflagellatum_Dev-1
AB505557_RM2-SGM49
KX500025_Colp4a_Develorapax_marinus
Developea
AB505558_RM2-SGM50
AB505559_RM2-SGM51
Abyssal clade
AY305011.1_Actinosphaerium_eichhornii
GU823533.1_clone_BCA3F14RJ1C11
GU823571.1_clone_BC95F14RM5C10
GU823473.1_clone_BCA3F14RJ2B12
AB3F14RM5G02_GU823294
GU823280.1_clone_BC95F14RM4G06
AI3F14RM1H11_GU823453
MAST-23
ME1_17_AF363186
DH148_5_EKD53_AF290083
MAST-2
BCB5F13RM3H07_GU823479
BCI5F13RM2E10_GU823586
BCB5F15RM3E09_GU823089
BCI5F15RM1D06_GU823104
MAST-16
Rictus_lutensis_GQ223284
NAMAKO_31_AB252771
Cantina_marsupialis_BUSSELTON
Cantina_marsupialis_NY0206
Cantina_marsupialis_YPF1205
BLACKSEA_cl_37_HM749939
BCI5F13RJ7E06_GU824659
RM1-SGM38_AB505495
MAST-24
BCB3F14RJ2A04_GU825694
isolateH70_AY256221
MAST-21
bicosoecidOC4_1_EF620521
bicosoecidOC4_2_EF620522
MESS12_GU170211
AY827850_Cafeteria_roenbergensis_HFCC33
Pseudobodo_tremulans_AM_DQ102392
Cafeteria_roenbergensis_HFCC34_AY827851
Cafeteria_EPM1_AF174366
Zeuk71_AY916583
Zeuk67_AY916578
KT861043.1_Symbiomonas_scintillans_RCC257_760bp
RCC257 Symbiomonas scintillans
MN315510_Anoeca_sp_HFCC807
bicosoecidIND33_38_EF620526
Caecitellus_parvulus_AY642126
Caecitellus_paraparvulus_HFCC71_DQ220713
Caecitellus_parvulus_NBH4_AF174368
Caecitellus_Biosope_T39_110_FJ537321
EU106851.1_Caecitellus_sp._RCC1078
RCC1078 Caecitellus sp.
Caecitellus_RCC1072_EU106848
AY520455_Caecitellus_parvulus_ATCC50061
Caecitellus_parvulus_AY520457
UEPAC30Dp2_AY129062
ME280100_GU170210
ME13100_GU170209
Halocafeteria_seosinensis_DQ269470
MT355117.1_Bicosoecida_sp._HFCC768
AI5F15RJ3B09_GU823081
BCA5F13RJ2H06_GU823870
MESS21_GU170214
18B12_M1_DQ103774
18G05_M1_DQ103786
18B03_M2_DQ103795
Bicosoeca_NY0214
AB3F14RJ13E02_GU825260
H3S3Be44C_JQ781891
18G10_M2_DQ103799
AY520445_Bicosoeca_vacillans_ATCC50063
D11_bicosoecidBS3_FN598433
TKR07M93_GU290106
AY520444_Bicosoecapetiolata_ATCC50639
PSH9SP2005_EU162647
SE4A_H05_FJ592448
EU446304.1_Uncultured_marine_eukaryote_clone_UI11D07
Siluania_monomastiga_AF072883
Paramonas_globosa_ATCC50531_AY520452
Adriamonas_peritocrescens_AF243501
K9AUG2010_AB771824
2011_RW7_AB721079
2011_RW6_AB721078
2010_RW7_AB721031
Re_St8_50_AB749099
Filos_agilis_FJ971856
Nerada_mexicana_ATCC50535_AY520453
Nanos_amicus_FJ971853
2010_F2_AB721064
Bs_St6_34_AB749083
TKR07M62_GU290098
Pseudobodo_tremulans_AF315604
Bikosia
AY520454.1_Marine_gliding_biciliate_TCS-2004
KM816649.1_Pendulomonas_adriperis_NY0181
MW575546.1_Wobblia_pacifica_HFCC5000
AB032606_Wobblia_lunata
MW575542.1_Suigetsumonas_keniensis_strain_HFCC1412
AB976561.1_Suigetsumonas_clinomigrationis
MW575549.1_Allegra_atacamiensis_strain_HFCC5011
MW575551.1_Allegra_hypersalina_HFCC5015
GU170208.1_Allegra_dunaii_ME8
GU170207.1_Haloplacidia_cosmopolita_ME5
GU170213.1_Haloplacidia_cosmopolita_MESS14
MW675781.1_Haloplacidia_cosmopolita_MESS13
Haloplacidia sinai
Placididea_sp_MMETSP1104_CaronLab
MW575540.1_Placidia_abyssalis_HFCC739
MW575545.1_Placidia_azorensis_HFCC952
AB061218.1_Placidia_cafeteriopsis
Placididea
JQ781904.1_MAST_3C_clone_H3S1Ae53U
OLI11006_AJ402357
JQ782025.1_MAST_3E_clone_70S8Ae5y2
AY381157.1_MAST_3F_Uncultured_marine_eukaryote_clone_HE000427.21
JN848814_Incisomonas_marina
MAST3_JAMSTEC_barrnap_18s_short
HM163291_Solenicola_setigera
MAST-3/Nanomonadea
AY266473_Blastocystis_geocheloni
EF209016_Blastocystis_sp
FJ809939.1.1793_HC08-04_Blastocystis_hominis
KarotomorphaSp_DQ431242
Opalinata
PSC4SP2005_EU162644
DSGM_39_AB275039
MAST-12
MW521831.1_MAST4A_sp1
JQ782049.1_MAST_4A_clone_70S1Be7hA
ME1_19_AF363188
AY129066.1_MAST_4B_Uncultured_marine_eukaryote_clone_UEPACCp4
MW521834.1_MAST4E_sp1
ME1_20_AF363189
JQ782028.1_MAST_4D_clone_70S8Ae5yu
MAST-4
BL000921_11_AY381191
UEPAC45p418S_AY129064
MAST-7
JQ781972.1_clone_155D8Ae6EV
JQ782091.1_905st23-29
KJ763801.1_Uncultured_eukaryote_clone_SGYP634
155D1Ae4AS_JQ782009
ANT12-26_AY116220
MAST-8
BL010625_32_AY381217
TAGIRI_18_AB191426
DSGM_41_AB275041
MAST-9
AA5F14RM1F09_GU823054
KY980306.1_Oblongichytrium_sp._isolate_BH56_186
KY979988.1_Oblongichytrium_sp._isolate_NY13S_44
GU369996.1_Uncultured_marine_eukaryote_clone_AP10-30
FN263028.1_Uncultured_clone_B8
KY980042.1_Oblongichytrium_sp._isolate_NY13S_181
DQ310262.1_Uncultured_marine_eukaryote_clone_FV23_1D5
AF363207.1_Eukaryote_marine_clone_ME1-24_clone_ME1-24
Pseudophyllomitus sp. BSC2
LC257669_Pseudophyllomitus_vesiculosus
Mastreximonas tlaamin
EF526960.1_Uncultured_marine_eukaryote_clone_SA2_3F7
Vomastramonas tehuelche
MAST-6
Labyrinthuloides_minuta_LADDLRRNA_L27634
AJ519935_Aplanochytrium_stocchinoi
AB022103.1.1787_KMPB_N-BA-107_Aplanochytrium_kerguelense
HM042904.1.1758_Aurantiochytrium_limacinum
DQ023617.1.1677_Schizochytrium_aggregatum
AF261664.1_Thraustchytridae_QPX
FJ821481_OblongichytriumHK9
AB022111_Thraustochytrium_multirudimentale
JX188294.2_Uncultured_eukaryote_clone_CC02SE75.039
Labyrinthulomycetes
LC028904_Platysulcus_tardus
M14649_Prorocentrum_micans
X54946_Alexandrium_tamarense
M97703_Toxoplasma_gondii
X53485_Onychodromus_quadricornutus
AF411266_Cercomonas_diparavaria
FJ790710_Neocercomonas_celer
FJ824122_Protaspis_obliqua
Outgroup

Appendix O Consensus trees of stramenopiles with updated an ochrophyte data

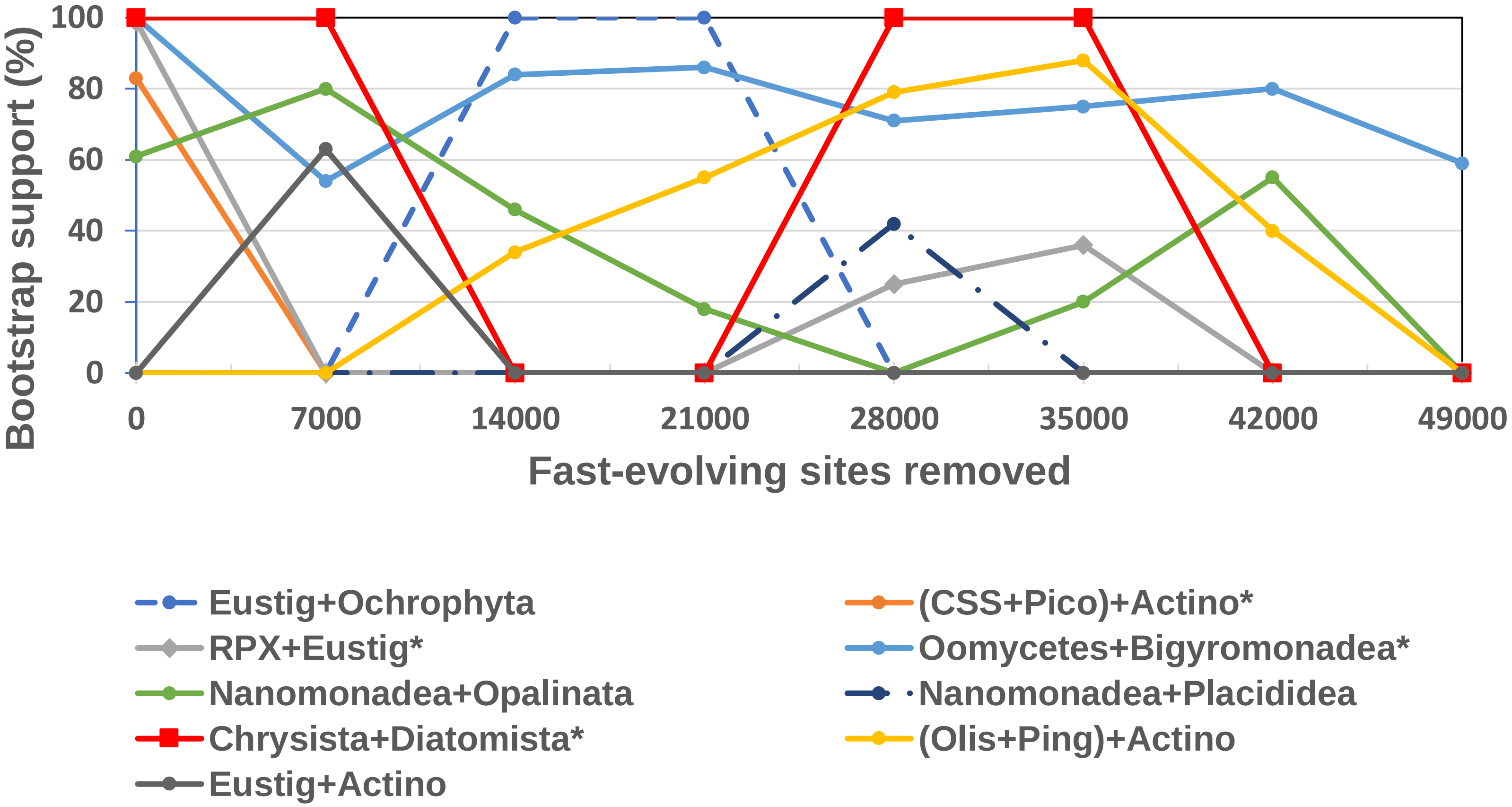
Bayesian consensus trees of stramenopiles generated from four Markov Chain Monte Carlo (MCMC) chains, none of which converged (maxdiff=1). Using PhyloBayes under the CAT+GTR+G4 model, the analysis is based on 231-supermatrix used for the C60+PMSF tree. Only posterior probabilities (PP) lower than 1 are labelled in the figure, otherwise all other unlabelled nodes have PP=1. CSS = Chrysophyceae-Synurophyceae-Synchromophyceae; Pico=Picophagea; Olis=Olisthodiscophyceae; Ping=Pinguiophyceae; BB=Bolidophyceae-Bacillariophyceae; PeD=Pelagophyceae-Dictyochophyceae; RPX=Raphidophyceae-Phaeophyceae-Xanthophyceae; Actino=Actinophryidae; Eustig=Eustigmatophyceae. Oomycetes includes Hyphochytriomycetes. Alveolates and Rhizaria are outgroups.

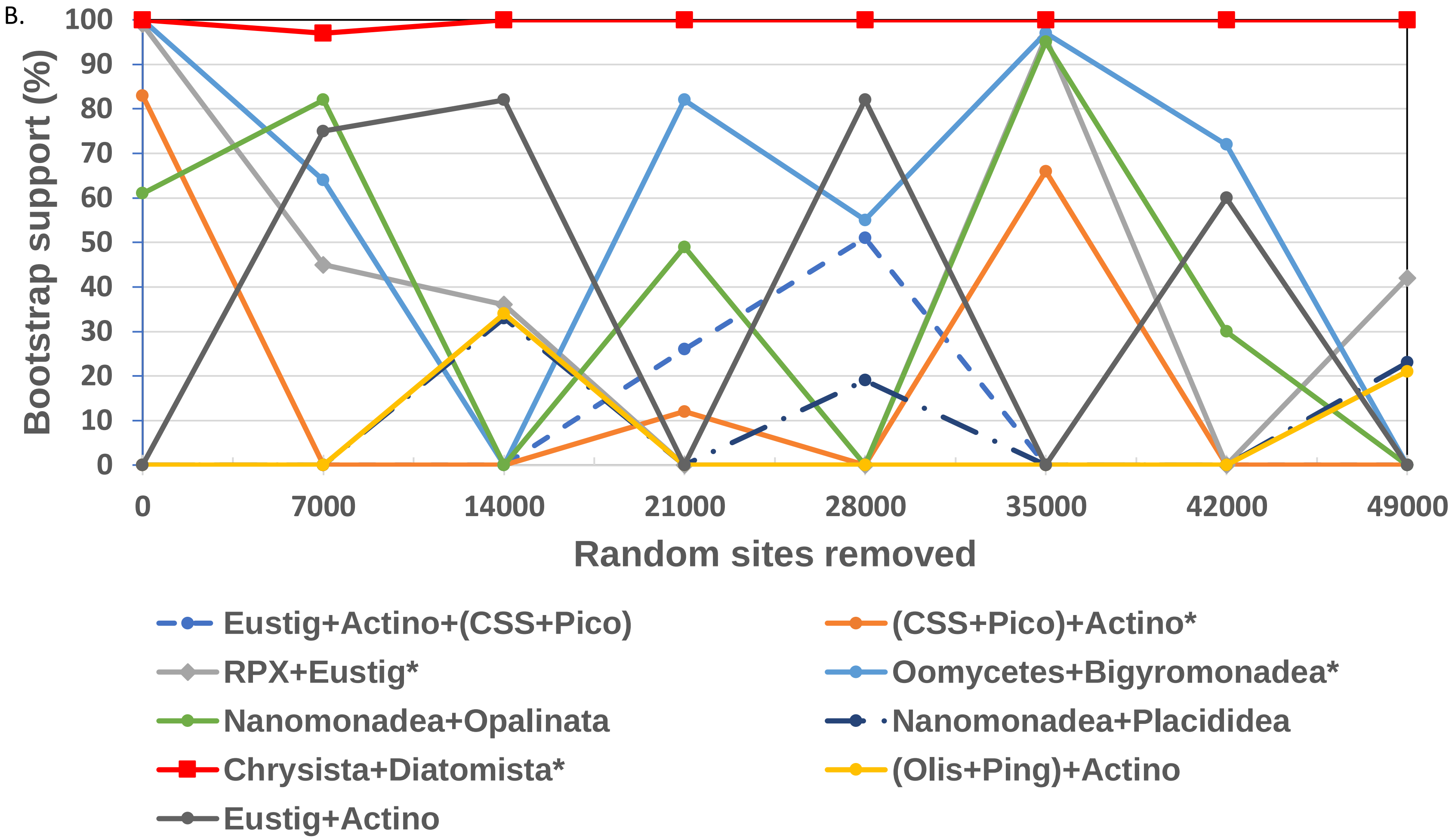


Appendix P: Summary of bootstraps with fast-evolving or random site and genes removed.

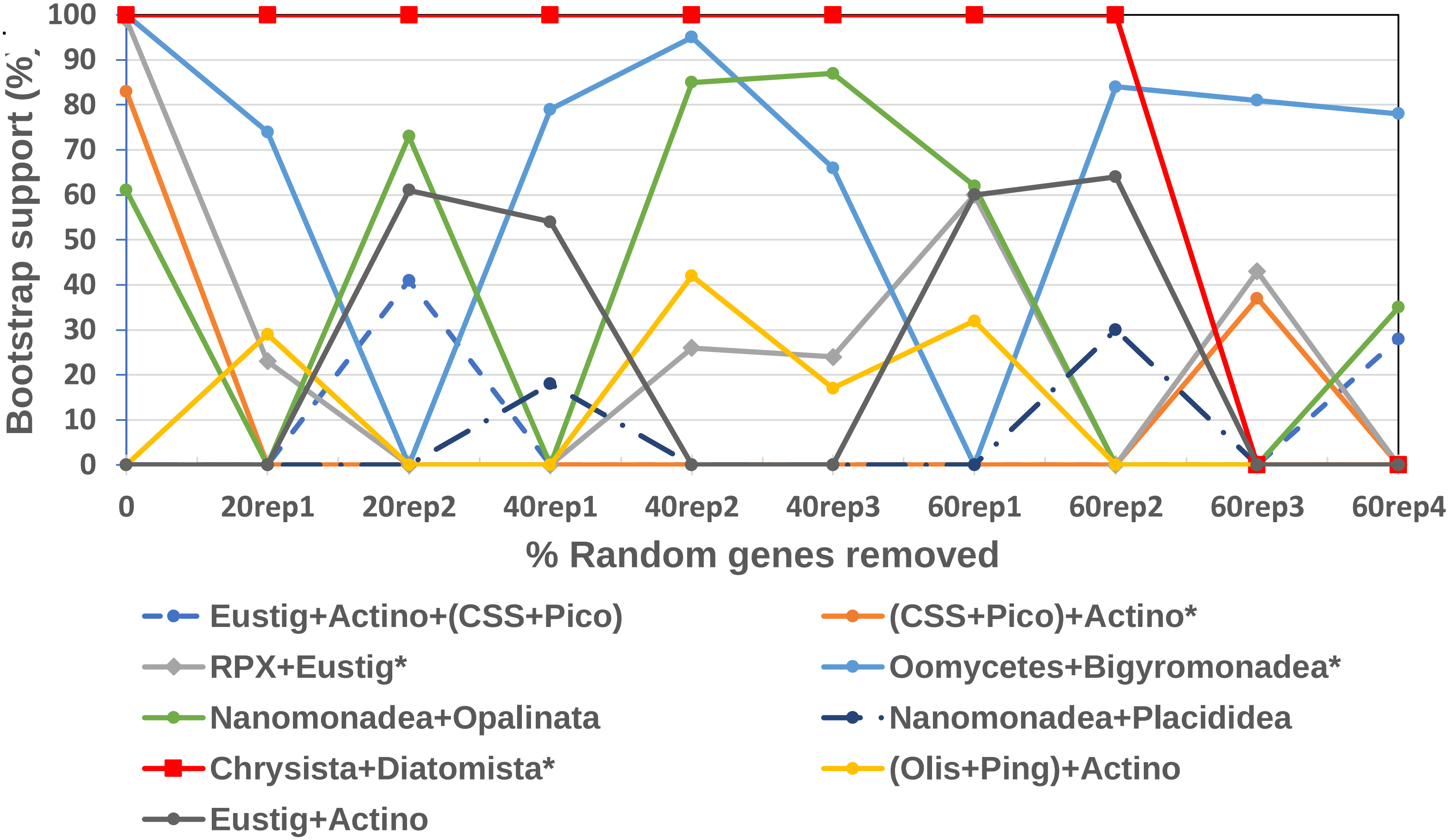
Summary of bootstrap changes of some contentious lineages with incremental removal of (A) fast-evolving sites; (B) random sites; (C) random genes. For (A), (B), increments of 7,000 amino acid sites were removed while for (C) random genes in increments of 20% were removed from '231-supermatrix'. For (C), two to four replicates of each increment were generated and they are denoted by "rep#". (A) Bigyromonadea+Oomycetes (including hyphochytriomycetes) were present in trees generated from removing fast-evolving sites while (B) Chrysista+Diatomista (ignoring placement of Actino, Ping, Olis and Eustig) were the only grouping present in all trees when random sites were removed. Groupings with an asterisk (*) indicate that they are found in '231-supermatrix' C60-PMSF tree. CSS=Chrysophyceae-Synurophyceae-Synchromophyceae; Pico=Picophagea; Olis=Olisthodiscophyceae; Ping=Pinguiophyceae; Platy=Platysulcidae; RPX=Raphidophyceae-Phaeophyceae-Xanthophyceae; Actino=Actinophryidae; Eustig=Eustigmatophyceae.

A.




B.
Bootstrap support (%)
100
90
80
70
60
50
40
30
20
10
0
0
7000
14000
21000
28000
35000
42000
49000
Random sites removed
Eustig+Actino+(CSS+Pico)
(CSS+Pico)+Actino*
RPX+Eustig*
Oomycetes+Bigyromonadea*
Nanomonadea+Opalinata
Nanomonadea+Placididea
Chrysista+Diatomista*
(Olis+Ping)+Actino
Eustig+Actino

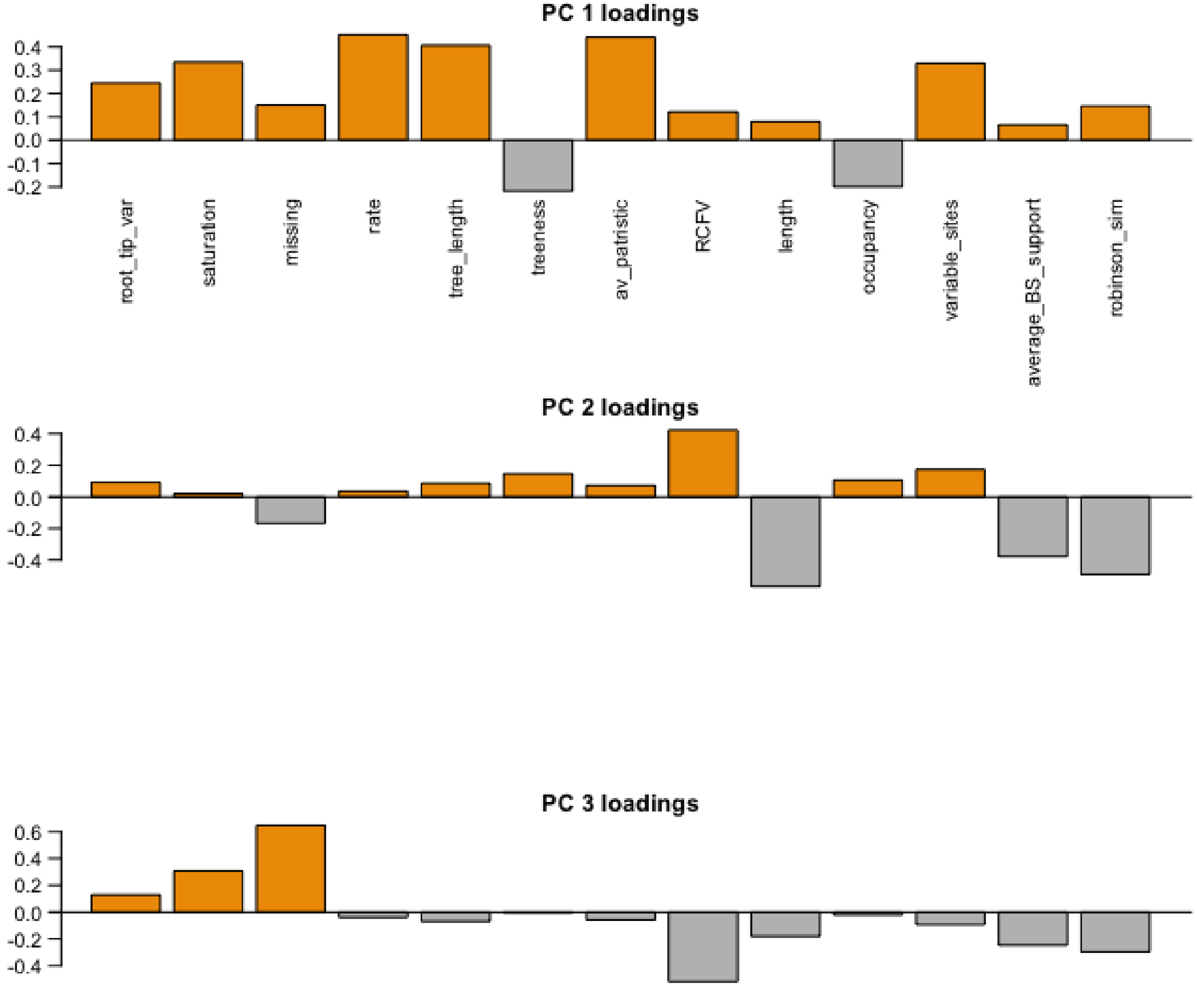
C.



Bootstrap support (%)
100
90
80
70
60
50
40
30
20
10
0
0
20rep1
20rep2
40rep1
40rep2
40rep3
60rep1
60rep2
60rep3
60rep4
% Random genes removed
Eustig+Actino+(CSS+Pico)
(CSS+Pico)+Actino*
RPX+Eustig*
Oomycetes+Bigyromonadea*
Nanomonadea+Opalinata
Nanomonadea+Placididea
Chrysista+Diatomista*
(Olis+Ping)+Actino
Eustig+Actino

Appendix Q: Summary of 13 gene property loadings on three principal components.

Only PC1 and PC2 axes were used for a PCA plot (see Fig. 4.3A).



Appendix R: List of calculated properties for each gene used for building a concatenated matrix.

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
BTUB	22	0.02008698	0.54808635	0.05978261	0.03905824	4.49169805	0.43603286	0.21439934	0.06397481	432	0.92	0.61342593	73.25	0.25892857	-6.3509638	0.68434478
RPS20	188	0.00826826	0.6433892	0.18358341	0.06035343	6.21640289	0.44326397	0.34066202	0	99	0.824	0.80808081	72.6	0.39	-4.1113818	0.44308226
ATP6	17	0.02668679	0.62468552	0.02765265	0.06515336	7.23202305	0.47820723	0.336169	0.11342791	144	0.888	0.79861111	79.5277778	0.4537037	-4.499775	1.51172133
VATB	220	0.00614656	0.65638689	0.04155874	0.06665395	6.13216346	0.44989819	0.35387141	0.07434217	470	0.736	0.61489362	87.8426966	0.60674157	-4.3904991	-1.748244
H2A	72	0.01928687	0.72793959	0.03057476	0.07376735	6.56529445	0.3606335	0.40273764	0.14861435	104	0.712	0.71153846	74.9302326	0.3255814	-3.4216898	2.07107281
NSF1-L	111	0.00459541	0.65286736	0.02481524	0.0777782	6.92226013	0.42113917	0.40135104	0.08046821	374	0.712	0.68181818	88.244186	0.61627907	-3.8431137	-1.3340336
IF6	79	0.00304788	0.55673988	0.04687695	0.0780059	8.03460733	0.36513899	0.36593118	0.10617345	233	0.824	0.77682403	84.03	0.58	-3.9546322	-0.3024856
S15P	198	0.00602813	0.66228246	0.02382666	0.07888281	7.25721878	0.43523272	0.38357681	0.12574429	151	0.736	0.72847682	78.3258427	0.5505618	-3.8763	0.70279767
S15A	197	0.00664081	0.62700728	0.0250038	0.08102423	8.26447132	0.38597996	0.38492178	0.13556917	129	0.816	0.72868217	83.2277228	0.52525253	-3.8780457	0.69702386
NSF1-G	106	0.00915086	0.69148636	0.04930746	0.0820762	7.46893395	0.38729602	0.39514606	0.08290885	384	0.728	0.68229167	88.3863636	0.54545455	-3.3874696	-1.1522765
NSF1-M	112	0.00747129	0.71082468	0.04844932	0.08209447	8.12735206	0.39894314	0.39293694	0.08083367	398	0.792	0.69346734	83.90625	0.57291667	-3.362521	-1.0267824
RPS5	192	0.00285813	0.6186278	0.0287226	0.08226134	8.63744053	0.44168342	0.42702084	0.11889633	189	0.84	0.75661376	85.7647059	0.60784314	-3.9093966	0.08107101
ARPC1	14	0.00964489	0.641805	0.05529563	0.08944127	8.31803775	0.36475417	0.41412444	0.09447699	393	0.744	0.72773537	87.0326087	0.63333333	-3.0951512	-1.4691728

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
RPS17	185	0.0224359	0.78203442	0.05363636	0.09095896	9.0958964	0.38991916	0.53079448	0.18077442	110	0.8	0.71818182	73.3402062	0.43298969	-2.5286197	2.03217504
NDF1	99	0.00873459	0.77739384	0.03837342	0.09102972	8.8298826	0.37988616	0.48739415	0.08041657	432	0.776	0.58333333	85.4893617	0.55319149	-3.0501017	-1.2989019
NSF1-K	110	0.01182852	0.68825193	0.05835224	0.09139624	8.86543521	0.42777242	0.42935524	0.09267779	359	0.776	0.71587744	81.4468085	0.56382979	-3.2976639	-0.5406971
NSA2	104	0.01839298	0.73802856	0.03393429	0.09205514	8.83729303	0.36335175	0.45839947	0.10859429	260	0.768	0.73461538	81.2795699	0.49462366	-2.7179496	0.24650304
RPS3	190	0.00812102	0.61563135	0.02343612	0.09238369	9.97743866	0.32620871	0.40274167	0.11372772	209	0.864	0.86124402	82.7428571	0.56190476	-2.9480894	0.25657324
RPS16	184	0.01780399	0.69048765	0.03833992	0.09418748	10.3606229	0.39524623	0.47646219	0.14937963	138	0.88	0.73913043	80.6915888	0.4953271	-3.1343948	1.21570995
RPS11	181	0.0064394	0.74110876	0.0301468	0.09448563	9.92099077	0.45823257	0.51155156	0.15017555	133	0.84	0.7593985	78.245098	0.58823529	-2.9514039	1.09718371
NSF1-J	109	0.00976128	0.73229935	0.06128126	0.09521119	9.71154142	0.44562988	0.45724655	0.08757307	363	0.816	0.70798898	88.7373737	0.56565657	-3.0528752	-0.8131616
AP4S1	9	0.0227263	0.64920428	0.04395604	0.09595532	8.73193425	0.45516573	0.46397238	0.16604647	136	0.728	0.80882353	72.9325843	0.46590909	-3.2579325	1.97802608
APBLC	10	0.00897046	0.66131836	0.0451539	0.09838504	9.54334867	0.37817313	0.4723472	0.08304066	558	0.776	0.80645161	90.2765957	0.72340426	-2.3867012	-2.3761001
IF2G	77	0.01108329	0.75740478	0.04940476	0.10023559	10.0235587	0.32258832	0.42745199	0.08851261	420	0.8	0.68333333	83.8989899	0.59793814	-2.3554488	-1.3573647
RPL44	170	0.02635473	0.79137891	0.01954079	0.10225468	9.10066646	0.3706004	0.49932949	0.15086966	92	0.712	0.67391304	78.0454545	0.25581395	-2.7900278	2.48208368
PSMB-M	139	0.0249813	0.67812516	0.02518233	0.1026702	11.3963925	0.43153743	0.56784198	0.13198825	210	0.888	0.7952381	81.9537037	0.55555556	-2.8099819	0.83145117
SUCA	210	0.00994637	0.8097273	0.04554181	0.10494347	10.07457333	0.34209072	0.49851932	0.09468555	293	0.768	0.62798635	81.2021277	0.48387097	-2.4121358	-0.2161686

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
DHSB3	52	0.00552816	0.72911101	0.03654365	0.10513051	9.56687597	0.3368448	0.44308313	0.11198317	212	0.728	0.6509434	83.2159091	0.45454545	-2.9156828	0.2238067
SAP40	199	0.01021509	0.65565701	0.03820348	0.10689768	11.6518467	0.33050753	0.4591124	0.12810099	201	0.872	0.86069652	79.8018868	0.46226415	-2.5097548	1.0669771
RPS15	183	0.02976633	0.66827776	0.03386426	0.11001225	11.1112372	0.42374691	0.57975508	0.17958499	138	0.808	0.89855072	81.4795918	0.51020408	-2.1585592	1.79554701
AP3S1	7	0.04519769	0.72700448	0.04845469	0.11011109	9.13922079	0.45552818	0.56095877	0.17015986	138	0.664	0.84782609	84.725	0.5	-1.9312889	1.45114004
GTUB	71	0.01587905	0.71212805	0.08297158	0.11054284	9.61722742	0.33063404	0.47726249	0.10342945	421	0.696	0.74584323	83.1547619	0.70238095	-1.7527147	-1.8336508
RAN	146	0.0219883	0.74369907	0.05134425	0.11204555	10.8684181	0.30992923	0.44966163	0.12711072	204	0.776	0.79901961	77.106383	0.54255319	-1.7603047	0.44978767
RPS2	187	0.03244237	0.88857275	0.14845361	0.11246908	10.9095011	0.34289813	0.44260986	0	210	0.776	0.75238095	77.5744681	0.39361702	-1.3253264	-0.2062589
RPL19	158	0.01273668	0.70621984	0.0185937	0.11481444	11.8258868	0.38601839	0.54689838	0.1386279	165	0.824	0.80606061	82.2	0.55	-2.3023852	0.87633105
RPL3	163	0.02442426	0.85044907	0.14138413	0.11602288	12.1824027	0.39394509	0.61498609	0	375	0.84	0.76266667	80.8137255	0.45098039	-1.3776073	-0.8602581
BAT1	21	0.01518605	0.71471529	0.05276972	0.11608206	11.0277961	0.39486128	0.52450942	0.09626955	381	0.76	0.77952756	86.7765957	0.58695652	-2.1296143	-0.8065408
EIF4A3	61	0.02487081	0.70579811	0.0466054	0.11684894	10.0490087	0.32853085	0.4812936	0.09483652	372	0.688	0.71236559	83.1566265	0.55421687	-2.1680811	-0.8700993
DHSA1	51	0.00609136	0.76745099	0.02575536	0.11690628	11.3399091	0.33497675	0.49286025	0.07377695	578	0.776	0.62629758	87.5531915	0.59574468	-2.2887401	-2.093873
NOP5A	103	0.01268329	0.74393959	0.05544425	0.11739598	11.3874098	0.3633125	0.55092407	0.11969183	283	0.776	0.70671378	87.1595745	0.67021277	-1.9037551	-0.9758364
RPL11	151	0.00868092	0.72574593	0.03365385	0.11807674	12.2799809	0.42729952	0.61872314	0.14424543	162	0.832	0.72839506	80.3267327	0.5049505	-2.5764507	1.18660961

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
RPL12	152	0.04786094	0.75813454	0.02683147	0.11845365	12.674541	0.35827248	0.66086244	0.15587349	155	0.856	0.87096774	86	0.36538462	-1.3435243	2.03824448
RPS23	189	0.01057673	0.81920961	0.15328571	0.1188498	11.8849796	0.41734504	0.5972	0	140	0.8	0.63571429	80.5876289	0.50515464	-2.2039789	-0.7017413
RF1	147	0.01006293	0.73180443	0.0386396	0.11894868	11.3001242	0.37180419	0.57625745	0.09974231	407	0.76	0.76412776	88	0.67391304	-1.7814429	-1.3976761
ATSAR2	20	0.03084502	0.73070911	0.03608722	0.1192864	12.0479264	0.36802694	0.58400447	0.13923583	183	0.808	0.80327869	0	0.43877551	-2.5180397	4.8021204
GRC5	69	0.02027505	0.75129002	0.04620986	0.11940668	12.4182951	0.42891872	0.71976159	0.13341076	206	0.832	0.74757282	85.4554455	0.52475248	-1.9869786	0.69158198
MAT	88	0.02406106	0.67352287	0.02598273	0.11974643	11.6154038	0.39960638	0.55910115	0.10379966	369	0.776	0.83197832	89.5851064	0.61702128	-2.0625649	-0.6954274
RPL43	169	0.02106997	0.72406613	0.02854767	0.12061645	9.89054875	0.39744238	0.56704082	0.20032227	88	0.656	0.75	74.6375	0.48101266	-2.1924761	2.03332301
ORF2	119	0.0118293	0.69010599	0.06690501	0.12171674	12.0499575	0.36569601	0.58952305	0.16185913	178	0.792	0.81460674	85.03125	0.57291667	-1.7732866	0.50477961
CCT-A	31	0.01342651	0.68299113	0.04732129	0.12208758	12.6971079	0.36658749	0.57089629	0.09404623	509	0.832	0.82907662	90.0792079	0.69306931	-1.6706214	-1.800047
CCT-E	34	0.00497086	0.67473198	0.03143185	0.12365715	12.6130289	0.32281633	0.53220121	0.08911866	529	0.816	0.79773157	89.1818182	0.70707071	-1.7635001	-2.1290013
GNB2L	67	0.01695068	0.69461218	0.03482774	0.12451301	11.4551966	0.33552808	0.50542416	0.10891533	289	0.736	0.82352941	84.6741573	0.58426966	-1.7738371	-0.3458439
CS	49	0.01434411	0.77625925	0.06311807	0.12708535	13.3439613	0.38117485	0.6214109	0.1090327	365	0.84	0.75342466	87.7403846	0.62745098	-1.4196516	-0.8785209
VATA	219	0.01361688	0.7912	0.06179138	0.1273974	12.4849448	0.35716871	0.63792871	0.08543214	594	0.784	0.64478114	88.6842105	0.73684211	-1.3588996	-2.7798729
RPL15	156	0.00747547	0.75183158	0.0253748	0.12906898	14.0685192	0.35279295	0.62438253	0.12993202	205	0.872	0.73170732	81.1121495	0.53773585	-1.9011091	0.60552265

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
RPS18	186	0.01952431	0.73788019	0.02038014	0.12969133	13.0988241	0.45288998	0.69116144	0.16881807	137	0.808	0.83211679	79.8979592	0.52040816	-1.837319	1.81107406
RPL35	168	0.01188296	0.70976215	0.22163866	0.13022813	14.5855507	0.34641866	0.58865455	0	119	0.896	0.89915966	85.5675676	0.58715596	-0.9519041	-1.1554703
RPL24A	162	0.01842931	0.77486695	0.17272727	0.13025484	14.3280321	0.40086067	0.68026555	0	103	0.88	0.89320388	81.8224299	0.57009346	-0.9103812	-0.4309906
ODO2A	115	0.01324866	0.64913532	0.03913978	0.13065695	12.151096	0.32674308	0.61252482	0.13786427	225	0.744	0.83111111	81.4	0.46666667	-1.8923538	0.84182246
CCT-B	32	0.00672478	0.70693844	0.03510861	0.13169917	12.9065184	0.36527671	0.59653464	0.09874742	497	0.784	0.78269618	89.8247423	0.68421053	-1.5989952	-1.752416
RAD51A	145	0.03265831	0.72600439	0.05968992	0.13216018	11.8944165	0.30553026	0.71578447	0.11307555	301	0.72	0.79734219	80.1264368	0.50574713	-1.0276495	0.08952109
CPN60	46	0.00869956	0.71421696	0.03705901	0.13235055	12.9703538	0.3199411	0.59820138	0.08785079	516	0.784	0.7751938	90.0210526	0.61052632	-1.4591494	-1.6934738
COPG2	43	0.0158542	0.67451532	0.04727075	0.13462811	12.7896703	0.37742252	0.64426937	0.10895304	485	0.76	0.89896907	88.3297872	0.72826087	-1.0781198	-1.5678112
CCT-D	33	0.01344512	0.73814992	0.05848313	0.13481337	13.8857775	0.36177764	0.6592331	0.10104917	505	0.824	0.83960396	89.16	0.69	-0.903266	-1.6570949
RPL31	165	0.01668816	0.79324638	0.14732774	0.13501654	13.9067034	0.40374915	0.67838807	0	103	0.824	0.94174757	80.25	0.51	-0.7419242	0.11124158
ARP3	13	0.05854334	0.74205466	0.05827309	0.13583785	11.1387041	0.37812797	0.58055238	0.10509607	361	0.656	0.79778393	82.1265823	0.51898734	-1.205969	-0.1159641
CCT-G	35	0.00848275	0.71223167	0.04970598	0.13622026	13.4858053	0.34901474	0.62220417	0.1020185	493	0.792	0.82150101	87.7244898	0.65625	-1.2192783	-1.4833186
ODPB	117	0.03422008	0.71356656	0.0333895	0.13724357	13.4498702	0.33123679	0.58240562	0.11003571	327	0.784	0.80428135	82.8736842	0.49473684	-1.4669209	0.1893752
PSMA-B	130	0.02097946	0.72941772	0.03568137	0.13751855	14.4394475	0.4365263	0.73193784	0.14557222	217	0.84	0.83870968	85.745098	0.68627451	-1.2732826	0.21191919

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
MCM-E	93	0.01514015	0.74909456	0.08754554	0.13965754	10.0553426	0.34731865	0.62412687	0.12145828	366	0.576	0.71311475	87.0289855	0.65217391	-1.0801574	-1.4896142
AMP2B	5	0.00990642	0.72171789	0.04101244	0.14006085	14.7063896	0.30658676	0.56370398	0.1058557	333	0.84	0.83783784	86.6470588	0.56862745	-1.1874213	-0.4405505
RPL32	166	0.0147603	0.77315582	0.14323507	0.14197679	14.9075625	0.35654914	0.64563014	0	126	0.84	0.9047619	78.8640777	0.49019608	-0.7939968	-0.000258
CCT-Z	38	0.01810543	0.72330299	0.05837081	0.14218409	14.7871455	0.36665643	0.69068995	0.10517554	496	0.832	0.83870968	85.8613861	0.71287129	-0.8184565	-1.5172463
RPS8	194	0.01618823	0.74864404	0.04922307	0.14273266	15.4151271	0.39011143	0.64697288	0.15688483	174	0.864	0.79885057	83.1238095	0.51428571	-1.4163504	1.168193
C3H4	23	0.05904352	0.76024165	0.04348362	0.14310856	11.7349021	0.4607048	0.94490613	0.22068298	99	0.656	0.81818182	83.6075949	0.5443038	-0.5965723	1.98786508
RPS12	182	0.03893863	0.7179818	0.01832707	0.14313716	16.0313622	0.41081637	0.73806992	0.1896964	114	0.896	0.96491228	84.0091743	0.5412844	-0.7821013	2.11425736
HSP70 MT	74	0.03490921	0.80294509	0.06625963	0.14568816	15.0058805	0.29729943	0.61522933	0.08485776	582	0.824	0.75257732	86.95	0.7	-0.2886688	-2.3339627
RPS10	180	0.04127963	0.71774197	0.05499571	0.14621907	15.4992219	0.40784319	0.69020192	0.23440419	88	0.848	0.88636364	78.1456311	0.45631068	-1.00336	2.86434711
L10A	87	0.00859339	0.70996954	0.03444286	0.14657378	15.5368207	0.32255706	0.67495598	0.13638548	212	0.848	0.86792453	81.5825243	0.49514563	-1.1154957	0.98881314
PSMB-K	137	0.03518657	0.77642665	0.0362254	0.14776723	15.3677924	0.34857626	0.69252459	0.13712769	215	0.832	0.81395349	86	0.45544554	-0.8480087	1.04776885
RPL30	164	0.02156931	0.8558461	0.14146825	0.14790441	15.5299627	0.37229375	0.63260389	0	96	0.84	0.86458333	81.125	0.47058824	-0.4639341	0.13535847
PSMA-A	129	0.01150886	0.69835598	0.03785125	0.14802937	15.2470249	0.37441331	0.67748489	0.13843898	227	0.824	0.85903084	87.46	0.57	-1.2340322	0.4326673
CCT-N	36	0.01314297	0.73613215	0.04199294	0.14818636	14.5222635	0.3373196	0.68770983	0.09646693	503	0.784	0.78131213	89.7578947	0.67368421	-0.8672988	-1.7735738

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
PSMA-C	131	0.03481687	0.74119413	0.0601229	0.14840245	16.0274645	0.46040029	0.81715924	0.15128032	223	0.864	0.85201794	87.8207547	0.63809524	-0.8376647	0.51534068
ODBB	114	0.00782748	0.75308217	0.07012949	0.14933712	12.9923292	0.3173133	0.61309109	0.12818301	316	0.696	0.73734177	81.6511628	0.60714286	-1.0026353	-0.6332066
DRG2	56	0.01615369	0.75341555	0.05422692	0.14996875	13.947094	0.39102555	0.75907418	0.12088485	348	0.744	0.77873563	86.8888889	0.64444444	-0.891818	-0.6943972
RPS6	193	0.02133775	0.77366559	0.04256568	0.15088947	15.9942835	0.36519956	0.7185693	0.14435432	209	0.848	0.81339713	83.7864078	0.60194175	-0.7136515	0.4844978
IF2B	76	0.02665574	0.71652497	0.02748708	0.15128448	15.582301	0.40706168	0.76938854	0.16882345	154	0.824	0.8961039	77.69	0.53	-0.9803782	1.84738228
FTSJ1	62	0.04724632	0.7994088	0.07116105	0.15147637	12.7240154	0.3571001	0.6608813	0.15603792	178	0.672	0.7752809	79.8518519	0.40740741	-0.7069203	1.46263989
GNL2	68	0.06346973	0.82428925	0.04026898	0.15198122	14.8941599	0.38791388	0.83080784	0.14535392	261	0.784	0.77777778	81.7894737	0.54736842	-0.239528	0.77802012
MCM-A	89	0.02101861	0.7554611	0.0569427	0.15244612	10.5187823	0.34519418	0.70287421	0.10047795	451	0.552	0.76053215	87.75	0.68181818	-0.5976103	-1.869209
RPL2	159	0.01581796	0.81928178	0.13449574	0.15247924	15.7053615	0.38795166	0.73373826	0	245	0.824	0.75102041	84.82	0.58	-0.7691884	-1.2354503
CLAT	40	0.02138831	0.71027981	0.05480931	0.15270319	15.4230222	0.36818147	0.76862864	0.07962181	1357	0.808	0.91967576	92.4897959	0.78571429	0.19998416	-4.971663
NDUFV2-MITO	100	0.01557541	0.80562584	0.04564487	0.15338067	15.7982087	0.34534921	0.73376718	0.137053	211	0.824	0.78199052	87.77	0.54	-0.5894677	0.40184558
VPS26B	225	0.02853684	0.72101903	0.06316964	0.15404347	12.3234778	0.33203307	0.63923132	0.16358208	168	0.64	0.86309524	81.961039	0.48051948	-0.7084554	1.10534847
DNAI2	54	0.22103738	0.80324541	0.06460548	0.15720961	10.8474633	0.52568623	0.69603654	0.1040276	450	0.552	0.86222222	88.3636364	0.78787879	0.80108583	-1.1150038
ARPC4	15	0.03383461	0.70410537	0.06336974	0.15726773	12.8959535	0.35959673	0.5979289	0.16573025	163	0.656	0.90797546	75.6375	0.51898734	-0.7471031	1.4132141

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
CRFG	47	0.01906299	0.75050082	0.05675531	0.15796769	12.3214798	0.35641182	0.7517876	0.11846869	431	0.624	0.79118329	91.3866667	0.65333333	-0.4417999	-1.4536616
PSMD6	143	0.04334909	0.87343942	0.12228344	0.15807777	13.9108437	0.41974956	0.73256086	0	297	0.704	0.96296296	83.5058824	0.58823529	0.510207	-0.8497548
MRA1	95	0.02774224	0.78272405	0.05747194	0.15830998	14.4062085	0.37643643	0.81381938	0.16812181	187	0.728	0.88235294	87.0454545	0.61363636	0.0488436	0.57237239
RPL5	172	0.01754576	0.80578427	0.04646994	0.15876865	17.7820888	0.38790732	0.82370574	0.14271575	259	0.896	0.83783784	88.0909091	0.56880734	-0.2841599	0.51439444
DIMT1L	53	0.03571811	0.78359988	0.05573811	0.15900988	14.9469287	0.35281761	0.77911471	0.14491528	259	0.752	0.83397683	83.9010989	0.57142857	-0.1372542	0.37400805
RPL7A	173	0.02023134	0.74138591	0.02778596	0.15945008	16.7422589	0.33432619	0.7484368	0.15781178	194	0.84	0.88659794	85.5294118	0.54901961	-0.3999202	0.96141265
SEC23	202	0.00922164	0.71946682	0.06696259	0.15973766	14.5361267	0.32750419	0.70328906	0.09687037	609	0.728	0.87192118	88.3068182	0.72727273	-0.1644618	-2.3453338
RPF1	150	0.02949295	0.74531153	0.04220119	0.16038032	14.2738484	0.29973376	0.65974495	0.18697811	127	0.712	0.8503937	78.0465116	0.48837209	-0.4241867	1.63823636
RPS4	191	0.01795236	0.78068117	0.04762617	0.16057999	17.1820592	0.35365135	0.75401427	0.13646332	250	0.856	0.848	85.9038462	0.52884615	-0.3846127	0.62165542
NFS1-MITO	101	0.01709761	0.79980826	0.05267152	0.16116819	15.6333143	0.33245925	0.68232198	0.11156924	394	0.776	0.69543147	85.1914894	0.58510638	-0.7432056	-0.8348583
ARP2	12	0.09729413	0.82101956	0.05823632	0.16173743	15.0415806	0.42542649	0.74147558	0.10121841	363	0.744	0.86225895	88.8369565	0.64444444	0.33607365	-0.4722837
RPL9	174	0.01113704	0.69095564	0.04333751	0.16272793	18.2255278	0.38806866	0.75761279	0.15923616	171	0.896	0.9005848	87.1181818	0.52293578	-0.8973562	1.29324902
RPL4B	171	0.01259527	0.77683793	0.0462157	0.16393389	16.5573229	0.34046552	0.73222467	0.13354217	298	0.808	0.82550336	88.6565657	0.62244898	-0.2648668	-0.3194083
MCM-B	90	0.01555689	0.71218543	0.06607904	0.16562267	10.9310965	0.36104505	0.76874124	0.11174523	526	0.528	0.78136882	88.7619048	0.63492063	-0.5854966	-1.7862279

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
RPL20	160	0.02363658	0.8156067	0.1665488	0.16627077	16.7933482	0.38724711	0.90022157	0	168	0.808	0.88095238	85.24	0.47959184	0.18397052	-0.2879884
SRP54	208	0.01767793	0.72867388	0.06624804	0.16637443	15.1400734	0.34978747	0.73537469	0.11733176	385	0.728	0.80519481	87.8522727	0.69318182	-0.3944674	-1.2702629
AP4M	8	0.01438146	0.77201158	0.07250765	0.1671604	16.7160397	0.31485744	0.76113738	0.12813313	327	0.8	0.9204893	83.3298969	0.6185567	0.36951381	-0.2086174
MCM-D	92	0.01545547	0.73030582	0.08314393	0.16894538	12.3330129	0.32657264	0.75084317	0.12026353	446	0.584	0.78251121	84.5714286	0.58571429	-0.3657938	-1.1290766
ATP6V0A1	18	0.06388318	0.73583724	0.0553802	0.16921701	17.2601347	0.3789658	0.81039557	0.10808017	410	0.816	0.80243902	86.1515152	0.65656566	-0.1795657	-0.767519
WD66	228	0.09393167	0.88086836	0.2980963	0.16977069	15.9584452	0.2963834	0.74880472	0	380	0.752	0.93421053	81.0989011	0.54945055	2.2110367	-1.8310927
NSF1-C	105	0.06163959	0.78960052	0.06139818	0.17010976	17.0109765	0.3518838	0.7631094	0.11034914	329	0.8	0.78419453	88.3195876	0.58762887	0.02686153	-0.3918035
RPL17	157	0.03645676	0.74070074	0.04011742	0.17088598	19.1392293	0.40480019	0.86040517	0.18254261	146	0.896	0.9109589	85.027027	0.66972477	0.05446621	1.11230128
MTLPD2	97	0.00732079	0.77425052	0.04349652	0.17163702	18.5367982	0.32768727	0.75098825	0.09981911	463	0.864	0.8012959	82.0754717	0.57142857	-0.3638403	-0.5878303
DNAL1	55	0.04402918	0.72898076	0.0630523	0.17266212	13.6403075	0.40305203	0.70143674	0.16559807	159	0.632	0.87421384	80.3815789	0.42105263	-0.6274131	1.86357418
WBSCR22	226	0.04977581	0.7968154	0.07246901	0.17294505	15.2191643	0.34836094	0.87861019	0.1701529	220	0.704	0.76363636	86.7294118	0.61176471	0.33388503	0.17727068
TRS	215	0.01467618	0.77100786	0.07610189	0.17321869	17.668306	0.382177	0.86739141	0.10560203	585	0.816	0.78974359	90.4747475	0.62626263	-0.0814025	-1.5765604
RPN1B	175	0.0078841	0.72759332	0.08623713	0.17466468	17.8157972	0.34297994	0.77722798	0.10862455	682	0.816	0.91642229	89.5252525	0.72727273	0.41981031	-2.3140251
PSMB-L	138	0.04566227	0.78795563	0.0351574	0.1748157	18.0060166	0.37902911	0.84858434	0.15246563	198	0.824	0.83838384	85.67	0.48	-0.0612092	1.39020567

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
IF2P	78	0.01607012	0.74586906	0.03681082	0.17494786	17.1448907	0.36527346	0.88279183	0.10200597	563	0.784	0.8312611	88.2842105	0.72631579	0.045181	-1.7981366
PSD7	128	0.032051	0.7653706	0.04824885	0.17508431	17.6835151	0.37555288	0.79163078	0.14045724	268	0.808	0.87313433	86.622449	0.74489796	0.25091915	-0.42528
RPL13E	154	0.03813884	0.80008371	0.03552972	0.17509863	18.9106519	0.32699172	0.83492167	0.19218432	129	0.864	0.82945736	82.4285714	0.47619048	0.188183	1.94423734
MCM-C	91	0.04325589	0.81116198	0.07784154	0.17513259	12.9598116	0.33084972	0.8299071	0.11275771	438	0.592	0.76027397	88.0704225	0.66197183	0.49173006	-1.5768271
OPLAH	118	0.03716294	0.75986307	0.07390618	0.17672362	10.2499698	0.30426486	0.72074376	0.09190891	677	0.464	0.65288035	86.2909091	0.72727273	-0.1429849	-3.3071086
RPL13A	153	0.01625278	0.76278127	0.02861305	0.17839144	18.5527093	0.37366946	0.84743299	0.17396566	165	0.832	0.86060606	83.8019802	0.4950495	-0.2278427	1.66756906
CC1	26	0.03002746	0.82195773	0.02666667	0.17855102	17.8551016	0.43398798	0.98086241	0.14316118	228	0.8	0.80263158	84.979798	0.58762887	0.01241509	0.81396767
IMP4	84	0.02590962	0.74492248	0.0701488	0.18009503	16.0284577	0.35781322	0.85417212	0.15282026	259	0.712	0.84555985	87.2209302	0.62790698	0.19064573	-0.048012
RPL21	161	0.0774829	0.83148197	0.19244008	0.18207213	19.2996462	0.41349786	0.97198997	0	148	0.848	0.86486486	83.8640777	0.47572816	0.86109947	0.04195193
ATP6V0D1	19	0.03358526	0.74816568	0.06461724	0.18227543	19.1389196	0.38306662	0.86862705	0.13028687	316	0.84	0.94620253	88.372549	0.59803922	0.47992507	0.20168495
PPP2R5C	126	0.05645886	0.77348027	0.05035405	0.18304571	17.0232506	0.26355113	0.75160993	0.11426138	369	0.744	0.88346883	90.2333333	0.57777778	0.87255026	-0.6251423
ODPA2	116	0.01597502	0.78335724	0.04876441	0.18342122	19.62607	0.37994919	0.86353159	0.12113792	326	0.856	0.75766871	88.8018868	0.63461538	-0.1682217	-0.4531672
PACE2-A	120	0.04562225	0.80748677	0.05629078	0.1845387	17.1620994	0.38356181	0.87270253	0.1493866	234	0.744	0.82051282	83.4444444	0.57777778	0.37500966	0.64714677
CTP	50	0.01604509	0.76916233	0.0608854	0.18521328	16.1135555	0.32858344	0.80644273	0.10605819	505	0.696	0.76237624	89.1190476	0.66666667	0.11570093	-1.7799707

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
PSMA-G	134	0.02661923	0.73447281	0.05134011	0.18544634	19.8427586	0.37107084	0.88646732	0.15221751	227	0.856	0.90748899	87.9326923	0.68269231	0.41809927	0.20057211
VAPA	218	0.07195851	0.8041107	0.05396825	0.18593926	11.1563555	0.39968979	0.88274382	0.26010978	63	0.48	0.85714286	66.7457627	0.36842105	0.39084911	3.76544672
WRS	229	0.01376882	0.76342008	0.07169118	0.18615875	12.6587952	0.32490261	0.82993891	0.12921214	368	0.544	0.78804348	79.9384615	0.58461538	0.0920958	-0.5408759
PSMA-E	132	0.00841773	0.71386945	0.05276926	0.18731452	19.6680242	0.37434211	0.86152305	0.1634826	233	0.84	0.9055794	86.9019608	0.52941176	-0.0692242	1.03595924
PSMA-J	136	0.0302628	0.7849033	0.04448717	0.18899137	20.0330849	0.3985665	1.00545246	0.15948042	211	0.848	0.85308057	88.2815534	0.63106796	0.54206162	0.62866701
RPO-C	178	0.02468421	0.7568563	0.08974078	0.19181986	14.3864894	0.31141927	0.83288749	0.08462071	913	0.6	0.74479737	89.5	0.73611111	0.60852548	-3.967325
SF3B2	203	0.020971	0.79295287	0.04419809	0.1926151	17.5279742	0.35040409	0.84663799	0.14879279	227	0.728	0.86784141	84.4382022	0.56818182	0.47919101	0.6343261
IFT88	82	0.0323785	0.70674337	0.07468002	0.19291697	13.6971046	0.29421161	0.75038475	0.11886192	526	0.568	0.91254753	90.057971	0.76470588	0.88201897	-2.3089097
YKT6	231	0.0231544	0.76006705	0.03365009	0.19407696	17.8550799	0.35952139	0.93203994	0.17691981	167	0.736	0.90419162	86.0898876	0.53932584	0.51970379	1.32178937
HMT1	73	0.01856308	0.77949354	0.04117356	0.19573411	13.8971217	0.34030298	0.8368531	0.13166757	313	0.568	0.7827476	83.8571429	0.63235294	0.20874319	-0.5150636
GLCN	66	0.02399735	0.74968315	0.08389738	0.19837065	15.8696518	0.38996881	0.9178111	0.16573507	229	0.64	0.79039301	82.7142857	0.58441558	0.14759844	0.45368855
SRA	207	0.04526008	0.7847934	0.05159743	0.1994104	18.5451668	0.32366539	0.89161184	0.13681743	278	0.744	0.85971223	86.5111111	0.62222222	0.98082253	-0.0293932
RPO-A	176	0.02118258	0.76482458	0.06264547	0.19996867	13.7978383	0.34317421	0.89413212	0.09686072	741	0.552	0.75168691	88.5757576	0.72727273	0.52346993	-3.0289565
SCSB	201	0.02327438	0.77110422	0.08232888	0.2005908	16.2478546	0.32703876	0.84342376	0.12782181	338	0.648	0.80769231	80.7051282	0.46153846	0.30438238	0.34746826

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
COP-BETA	41	0.01282929	0.76170397	0.06671093	0.2018437	18.1659327	0.33249203	0.88713617	0.10183824	753	0.72	0.89110226	89.183908	0.67816092	1.02764242	-2.3527855
PACE2C	121	0.01646748	0.72928846	0.08612415	0.20327145	18.2944309	0.3410786	0.9244037	0.17133777	213	0.72	0.90140845	81.5862069	0.57471264	0.71947783	0.80348957
CCT-T	37	0.02350885	0.77306477	0.04233684	0.20399094	19.3791396	0.35540233	0.91401417	0.12265593	500	0.76	0.908	89.923913	0.70652174	1.08241116	-1.2253513
SCO1-MITO	200	0.06880946	0.7915172	0.05545496	0.20447518	19.2206667	0.34469133	1.01170658	0.19700043	141	0.752	0.82978723	83.1847826	0.42857143	0.99551454	2.16218384
UBA3	216	0.01797886	0.78360854	0.07079591	0.20488952	19.6693941	0.34125034	0.91214405	0.13560744	334	0.768	0.82035928	83.7849462	0.55913978	0.63966852	0.09398159
AGB1	2	0.06884666	0.7877943	0.05873369	0.20541216	21.3628651	0.29872016	0.83646298	0.12752228	277	0.832	0.89891697	81.2079208	0.43564356	1.11423557	1.23761672
RPL33	167	0.06096707	0.85481721	0.14753597	0.20549408	20.3439141	0.38317828	0.94149343	0	99	0.792	0.88888889	81.5208333	0.52083333	1.26993906	0.14584281
IFT46	80	0.02354151	0.76940123	0.06665471	0.20703298	16.9767041	0.32008583	0.87351448	0.18033312	204	0.656	0.89215686	0	0.49367089	0.25061095	4.72590083
RPPO	179	0.02868059	0.78273818	0.0346811	0.2075227	23.8651105	0.36651416	1.00761171	0.14482584	257	0.92	0.90661479	88.6428571	0.58928571	1.05177561	0.74455458
RPO-B	177	0.03269617	0.81049127	0.07144854	0.20817767	17.6951018	0.38765148	0.97203002	0.07682093	1094	0.68	0.82449726	89.8292683	0.73170732	1.33663095	-3.8996306
PSMB-N	140	0.02069089	0.76086859	0.06750228	0.20824304	23.1149771	0.3849668	1.04759382	0.17896993	207	0.888	0.9178744	83.1759259	0.56481481	1.00397136	1.42677885
XPB	230	0.05336875	0.78905534	0.06003877	0.20852064	17.5157339	0.38041622	0.94460375	0.11609124	393	0.672	0.82697201	82.3780488	0.54320988	0.74028377	0.04718059
RRAGD	196	0.098351	0.77125231	0.09143393	0.20925743	20.2977971	0.29537433	0.8859322	0.17308204	275	0.776	0.94181818	87.5520833	0.54255319	1.97555808	0.74313662
RPL14E	155	0.03299574	0.77875862	0.03325243	0.20998521	21.6284764	0.37069869	0.93455879	0.21186804	120	0.824	0.925	75.3861386	0.43	0.76641631	3.05178562

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
PURA	144	0.01283027	0.75076974	0.04550305	0.21031349	13.4600632	0.32563586	0.87940638	0.11295298	410	0.512	0.76097561	87.0655738	0.59016393	0.19842432	-1.0812942
SPTC2	205	0.01377213	0.78823858	0.04973046	0.21068261	21.0682607	0.27935159	0.89652358	0.12484826	371	0.8	0.83018868	81.4747475	0.54639175	0.91134084	0.00272168
CALR	24	0.01601883	0.75735252	0.04779189	0.21088933	18.347372	0.27618612	0.8288464	0.14593119	285	0.696	0.82807018	80.1428571	0.5	0.55447967	0.57876699
NMD3	102	0.02131875	0.73681309	0.06544379	0.21235041	17.4127333	0.38372646	0.98417472	0.14829196	363	0.656	0.92561983	89.85	0.6835443	1.05012803	-0.5487078
EFG-MITO	57	0.06089716	0.89382197	0.08516031	0.21336683	19.4163818	0.31340362	0.88506509	0.10810929	607	0.728	0.80230643	87.4090909	0.67045455	2.02639185	-1.8376419
TM9SF1	212	0.02668938	0.74782881	0.05465884	0.21609639	18.8003856	0.33666908	1.00728453	0.13418681	376	0.696	0.90159574	86.5119048	0.55952381	1.0953606	0.00763905
GCST	63	0.01397668	0.7804613	0.05319981	0.21739908	22.1747057	0.33675643	1.01676384	0.14233271	326	0.816	0.80674847	84.8383838	0.52525253	0.81137592	0.45001314
PPP2R3	125	0.03642804	0.75800117	0.06043637	0.21858163	18.3608569	0.34209881	1.03496513	0.15371712	287	0.672	0.92682927	87.0617284	0.7037037	1.61960697	-0.2628765
PIK3C3	124	0.05529816	0.82539274	0.06146448	0.21959541	19.5439915	0.3804895	1.02868481	0.16070981	236	0.712	0.88983051	82.4767442	0.55813953	1.56342076	1.08323382
ATG2	16	0.0326284	0.75739122	0.05847053	0.22157262	24.151415	0.31177344	0.99580766	0.15480121	292	0.872	0.96917808	88.3679245	0.55660377	1.70742681	0.70255851
AP3M1	6	0.17097612	0.78828168	0.07433016	0.22619654	20.1314922	0.33922644	0.99962136	0.14472694	325	0.712	0.93230769	90.6511628	0.68604651	2.84657064	-0.1664244
RPTOR	195	0.06130577	0.777339	0.09090279	0.22665119	11.5592109	0.25496177	0.7732439	0.14581201	283	0.408	0.83745583	78.3125	0.5	1.31286708	0.14876309
PSMA-H	135	0.01941753	0.73466956	0.06751902	0.22800268	23.7122785	0.37688518	1.00025543	0.18702609	182	0.832	0.96153846	83.8910891	0.46534653	1.11449399	2.06303653
EIF3I	60	0.02378559	0.76055622	0.05171036	0.22843157	22.3862935	0.37910992	1.10151112	0.15579569	281	0.784	0.90747331	90.3157895	0.64210526	1.45603667	0.23914885

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
CAPZ	25	0.06103567	0.81360371	0.07580645	0.22849119	18.2792948	0.3565223	0.96608112	0.17285114	217	0.64	0.88479263	80.2911392	0.45454545	1.52861875	1.6313129
PACE5	122	0.05937038	0.78407015	0.05969332	0.22987575	20.2290664	0.38664892	1.01533299	0.20574323	166	0.704	0.93373494	86.2705882	0.57647059	1.72547173	1.54505474
SYGM1	211	0.01731629	0.79480266	0.06471631	0.23236249	21.8420738	0.33079173	1.03184046	0.11992565	540	0.752	0.82222222	83.8791209	0.64835165	1.49633382	-1.1068038
PSD11	127	0.03503046	0.79345549	0.04717091	0.23571618	24.5144831	0.3344795	1.11489973	0.15631236	329	0.832	0.94832827	87.6633663	0.64356436	2.23338868	0.27912817
VATE	222	0.01512878	0.75746652	0.02304637	0.23728759	23.9660464	0.40593437	1.13979263	0.18189445	177	0.808	0.94915254	87.91	0.67346939	1.53700945	1.07094185
RPAC1	149	0.02724382	0.81932804	0.06624029	0.23733232	20.8852444	0.34296059	1.01835872	0.16251786	199	0.704	0.88442211	82.9195402	0.57647059	1.75459727	0.8953666
METTL1	94	0.02762491	0.79289879	0.06992188	0.23751237	19.0009896	0.30350598	0.87971067	0.17878659	192	0.64	0.81770833	80.1298701	0.48051948	1.17126875	1.24857615
TOPO1	214	0.03271747	0.80077678	0.06356942	0.24185158	22.9759	0.34718982	1.10596576	0.12218707	486	0.76	0.83744856	85.326087	0.61956522	1.8097085	-0.6441309
CRNL1	48	0.02319402	0.7807093	0.059015	0.24257983	20.6192857	0.31677817	1.00886859	0.13869983	553	0.68	0.89692586	86.3902439	0.65853659	1.94807819	-1.1004498
UBE12	217	0.01820689	0.79902613	0.09723517	0.24536617	22.3283214	0.29467094	0.99267311	0.11920822	719	0.728	0.88178025	90.0674157	0.70454545	2.38098695	-2.4004466
CCDC113	27	0.0661906	0.75586273	0.04951981	0.24571458	17.2000204	0.26172197	0.88167559	0.15996148	238	0.56	0.92857143	82.0588235	0.6119403	2.03466531	0.34453009
KDELR2	86	0.03572901	0.83023474	0.07632129	0.24585097	20.8973326	0.28878265	1.00804774	0.16578884	197	0.68	0.89340102	81.4146341	0.5	2.18037373	1.18983185
ADK2	1	0.02145806	0.79141845	0.03915625	0.24855457	24.8554571	0.32564948	1.09461613	0.14758816	320	0.8	0.90625	82.4848485	0.51546392	1.85417738	0.97990505
EFTUD1	58	0.1471788	0.78275882	0.08813362	0.25180723	13.345783	0.37173484	1.17595787	0.14626988	453	0.424	0.89183223	83.4	0.66	2.8343194	-0.5549882

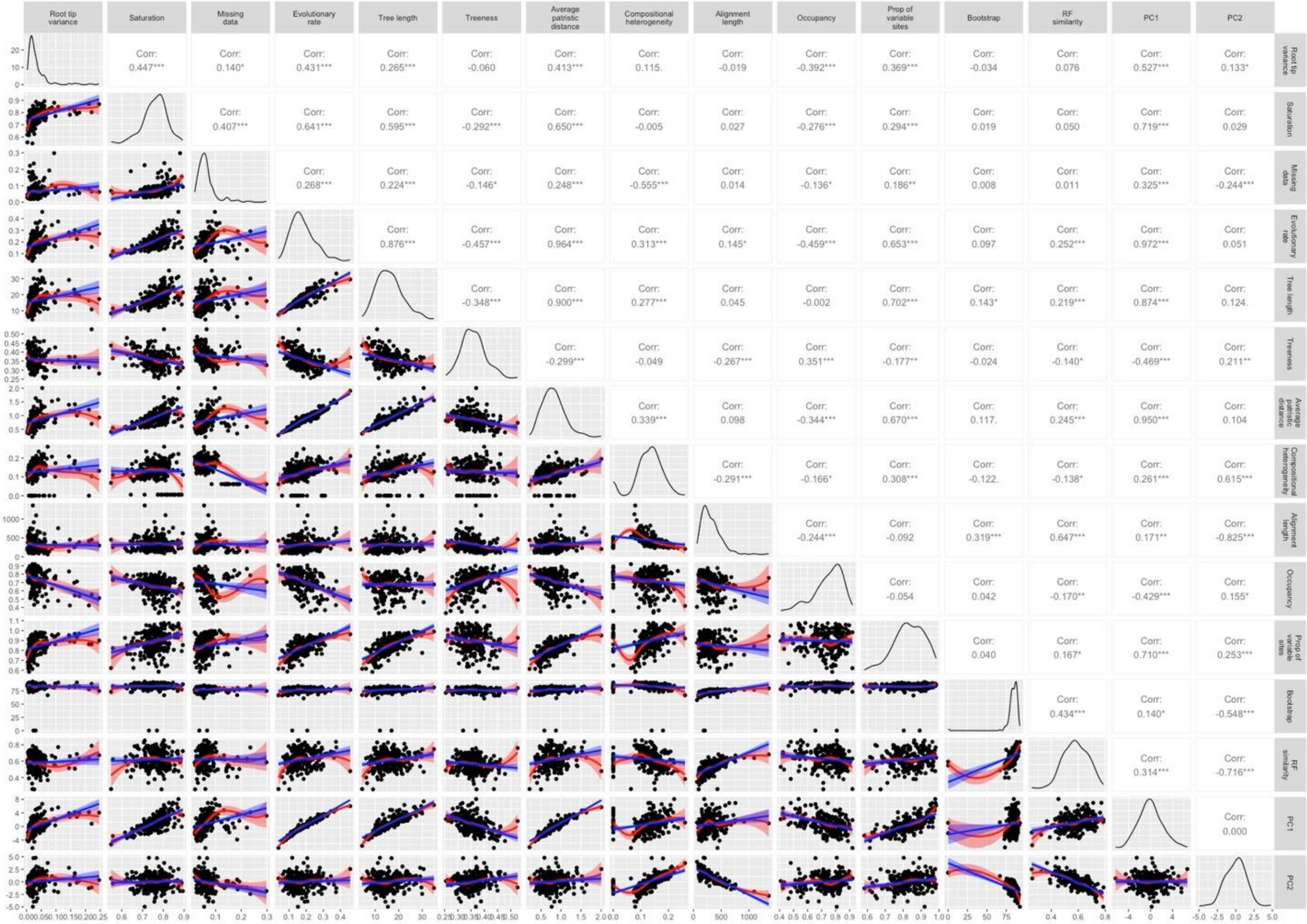
Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
EIF3C	59	0.03191809	0.75257713	0.05829465	0.25414501	23.1271961	0.33584964	1.02390772	0.1477726	374	0.728	0.94652406	89.2272727	0.72727273	2.20553962	-0.6463974
CORO1C	45	0.04481252	0.79332013	0.0571003	0.25648756	27.18768099	0.32553333	1.13387644	0.12863335	380	0.848	0.94210526	87.952381	0.66990291	2.69634354	-0.251476
RHEB	148	0.03365064	0.75143835	0.0678461	0.26066376	18.50712688	0.29372699	1.0140979	0.19885513	164	0.568	0.93902439	86.2571429	0.64705882	2.28275284	0.57789597
CDK5	39	0.02593388	0.80787911	0.10661529	0.26096622	14.09217557	0.27267538	0.88946803	0.16459386	346	0.432	0.7716763	83.2156863	0.47058824	1.59509684	-0.0695554
ODBA	113	0.02244449	0.82777789	0.06007157	0.26211418	22.01759111	0.29720975	1.04392712	0.14590397	326	0.672	0.80981595	83.5121951	0.54320988	2.00079412	0.21415571
GSS	70	0.0235383	0.8290427	0.07035327	0.26232396	24.65845118	0.35064142	1.15019528	0.17824138	212	0.752	0.81132075	81.6521739	0.49450549	1.95854631	1.46624589
SPTLC1	206	0.16200843	0.82548118	0.04796033	0.2659582	23.13836636	0.31545417	1.17557421	0.16983868	255	0.696	0.88627451	84.5714286	0.5952381	3.48206015	1.02859012
HYOU1	75	0.01543057	0.7904898	0.05613649	0.26693795	24.558291	0.33463992	1.15845049	0.13800554	395	0.736	0.9443038	84.988764	0.6741573	2.5732815	-0.2950383
CCDC37	28	0.24630777	0.86905679	0.0613952	0.26978416	17.2661861	0.28120123	0.92668561	0.13200874	297	0.512	0.8989899	74.9180328	0.67213115	4.08249901	0.3393578
MTHFR	96	0.02953958	0.79722973	0.10052508	0.27317551	18.029584	0.29305598	1.09870629	0.1398869	479	0.528	0.86012526	88.7936508	0.76190476	2.80370724	-1.953775
NSF1-H	107	0.04301351	0.8088495	0.0664219	0.27435921	23.046174	0.33239923	1.17974886	0.19810749	214	0.672	0.87850467	79.2771084	0.50617284	2.51070728	1.71281936
AR21	11	0.05038364	0.84836885	0.06980392	0.27473808	23.35273722	0.4097112	1.39009719	0.21698192	150	0.68	0.93333333	79.7195122	0.57317073	3.11258231	2.16206364
PGM2	123	0.02827293	0.78663107	0.0886139	0.27935884	17.32024778	0.27571736	1.06582939	0.1516655	421	0.496	0.81710214	81.6949153	0.59322034	2.28698242	-0.6405565
VBP1	223	0.14413428	0.83399928	0.04362416	0.28013951	29.13450933	0.37301445	1.39365761	0.20821586	149	0.832	0.95302013	81.9405941	0.51485149	3.91529166	2.81265901

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
COPE	42	0.05078291	0.79939772	0.06736501	0.28344141	29.1944656	0.34775455	1.30881633	0.19833086	228	0.824	0.96929825	87.3137255	0.66	3.54053239	1.06426566
VATC	221	0.05175484	0.788575	0.08050388	0.28360378	28.3603777	0.31681777	1.25485729	0.18730601	258	0.8	0.97674419	89.3505155	0.68041237	3.64605648	0.48781045
CCDC65	30	0.17916727	0.80120249	0.05506148	0.28687681	19.2207464	0.31677593	1.11561551	0.13945297	386	0.536	0.96632124	83.25	0.625	3.85786617	0.20463383
ALG11	4	0.02830347	0.82781192	0.10881365	0.28777306	21.0074336	0.28424861	1.14501903	0.19645797	265	0.584	0.81886792	81.2857143	0.47142857	2.71654827	1.01084352
WD	227	0.02055782	0.78972304	0.1856014	0.29103854	24.4472372	0.30325664	1.14553327	0	232	0.672	0.875	87.382716	0.60493827	2.75610518	-1.5092011
PSMD	141	0.02466373	0.80498588	0.23520092	0.29123442	29.9971454	0.33432832	1.32328418	0	288	0.824	0.77083333	90.62	0.71	3.14149767	-2.405255
SND1	204	0.02128991	0.77064894	0.08349045	0.29402466	19.993677	0.3140727	1.19675773	0.15227763	468	0.544	0.94230769	88.8461538	0.69230769	3.07375236	-1.0912231
AGX	3	0.03518448	0.80637923	0.09251769	0.29794636	25.6233873	0.34300465	1.15229815	0.19101946	299	0.688	0.89632107	83.7831325	0.53012048	2.93121071	1.06075966
PSMD12	142	0.05672886	0.84024518	0.1325555	0.29977142	29.6773709	0.36151355	1.32465743	0	374	0.792	0.94385027	86.7916667	0.625	3.67981761	-1.1208252
VPS18	224	0.164006	0.81036858	0.09033613	0.30032075	20.4218111	0.37289834	1.40987724	0.18421853	273	0.544	0.96703297	82.2461538	0.50769231	4.18003918	1.69729919
GDI2	65	0.21192944	0.87027673	0.08415909	0.31013753	16.4372892	0.31452071	1.212268	0	417	0.424	0.86330935	78.1	0.64	4.18678077	-1.2647313
IFT57	81	0.03645936	0.7941574	0.07374332	0.32229247	21.915888	0.32772484	1.24949887	0.18646817	275	0.544	0.95272727	82.2835821	0.64615385	3.54087445	0.54889144
NAA15	98	0.06169604	0.80531646	0.07571688	0.33114655	27.4851639	0.2995256	1.31868631	0.14596348	471	0.664	0.97664544	86.4	0.6375	4.39222615	-0.3064869
COPS6	44	0.08628763	0.8422357	0.10326024	0.34417941	26.8459937	0.34545077	1.52532789	0.20740799	221	0.624	0.95475113	82.2727273	0.66666667	5.06472965	1.10629394

Gene name	Position in dataset	Root-tip var.	Saturation	Missing data	Evolutionary rate	Tree length	Treeness	Av patristic dist.	Comp. heterogeneity	Alignment length	Occupancy	Prop. variable sites	Av. boostrap support	RF similarity	PC1	PC2
STXBP1	209	0.03794078	0.86412304	0.10103496	0.36051759	29.5624424	0.31022605	1.5351851	0.17533783	337	0.656	0.95252226	86.2962963	0.65822785	5.2836845	0.19849101
IPO4	85	0.03107222	0.80310657	0.11508762	0.36260045	26.4698327	0.33917507	1.63379028	0.14520548	573	0.584	0.98429319	86.75	0.6	4.9387556	-0.5651273
CCDC40	29	0.10746314	0.84680552	0.0694224	0.37409951	22.4459706	0.35464048	1.57898888	0.13425542	756	0.48	0.96296296	90.4736842	0.77192982	5.60178857	-2.0546041
IMB1	83	0.04583816	0.82750783	0.0850364	0.37606906	34.9744227	0.31779701	1.52245402	0.12700535	644	0.744	0.98136646	88.8901099	0.71111111	5.58422886	-1.2419882
TMS	213	0.05344802	0.87290306	0.09678412	0.45693679	30.6147651	0.36930828	2.00757123	0.22079557	194	0.536	0.97938144	84.6923077	0.453125	6.65410811	2.55132115

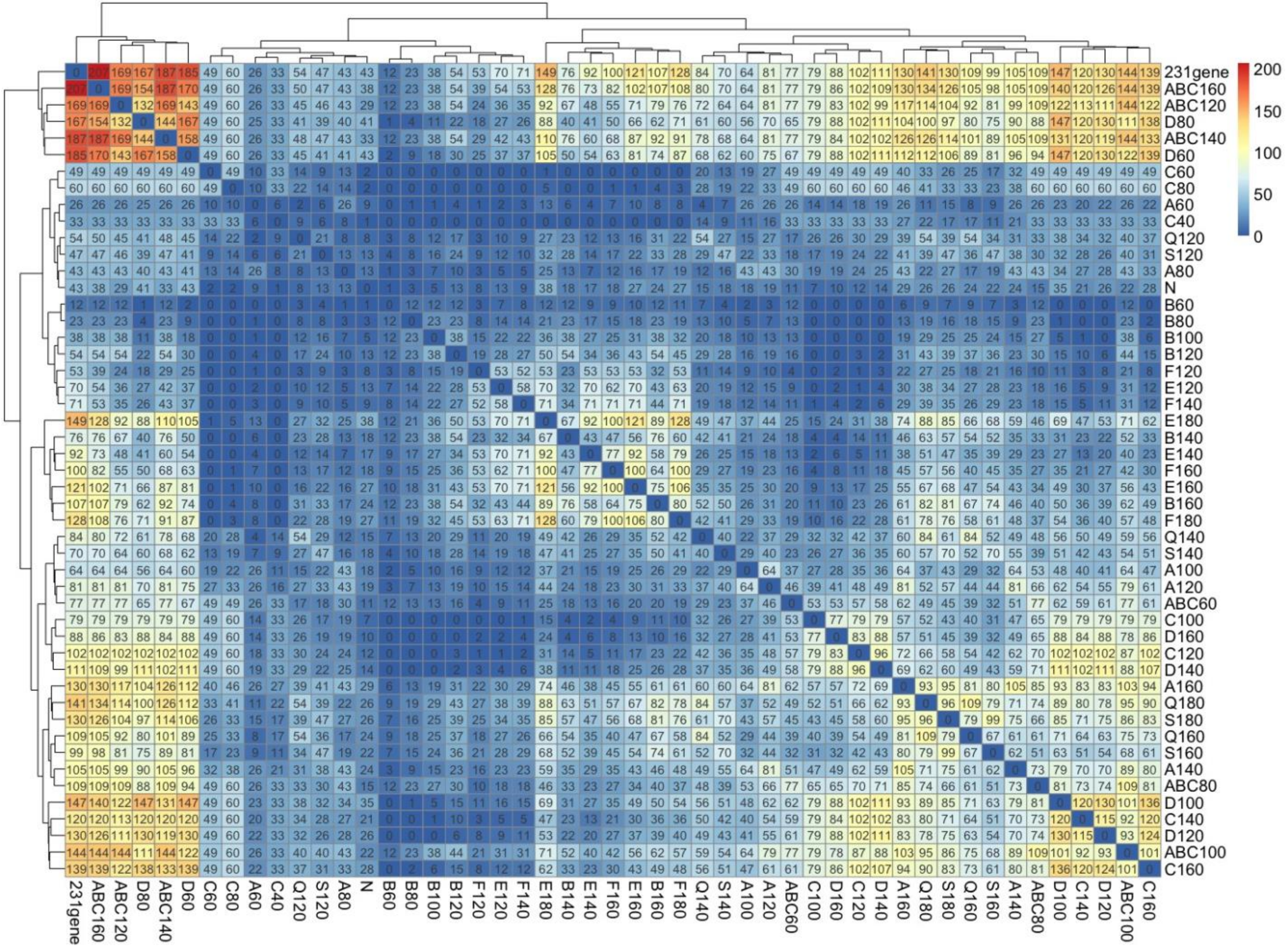
Appendix S: Correlogram of 13 gene properties and PC axes.

Numeric values after “Corr:” indicate the Pearson correlation coefficients followed by the significance of the correlation displayed based on p-values < 0.001 (***); < 0.01 (**); < 0.05 (*); < 0.10 (.). Two regression lines are generated using Linear Model (blue) and Loess (red). Abbreviations: RF = Robinson-Foulds; PC = principal component.



Appendix T: Heatmap of shared genes among different supermatrices.

Genes consisting of each supermatrix are selected by different filtering criteria (i.e., A-F, N, S, Q, and ABC; see Materials and Methods in Chapter 4). The numeric values are the number of shared genes with the red colour indicating high number of shared genes and blue colour indicating low number of shared genes among different supermatrices.



	231gene	ABC160	ABC120	D80	ABC140	D60	C60	C80	A60	C40	Q120	S120	A80	N	B60	B80	B100	B120	F120	E120	F140	E180	B140	E140	F160	E160	B160	F180	Q140	S140	A100	A120	ABC60	C100	D160	C120	D140	A160	Q180	S180	Q160	S160	A140	ABC80	D100	C140	D120	ABC100	C160
231gene	0	207	169	167	187	185	49	60	26	33	54	47	43	43	12	23	38	54	53	70	71	149	76	92	100	121	107	128	84	70	64	81	77	79	88	102	111	130	141	130	109	99	105	109	147	120	130	144	139
ABC160	207	0	169	154	187	170	49	60	26	33	50	47	43	38	12	23	38	54	39	54	53	128	76	73	82	102	107	108	80	70	64	81	77	79	86	102	109	130	134	126	105	98	105	109	140	120	126	144	139
ABC120	169	169	0	132	169	143	49	60	26	33	45	46	43	29	12	23	38	54	24	36	35	92	67	48	55	71	79	76	72	64	64	81	77	79	83	102	99	117	114	104	92	81	99	109	122	113	111	144	122
D80	167	154	132	0	144	167	49	60	25	33	41	39	40	41	1	4	11	22	18	27	26	88	40	41	50	66	62	71	61	60	56	70	65	79	88	102	111	104	100	97	80	75	90	88	147	120	130	111	138
ABC140	187	187	169	144	0	158	49	60	26	33	48	47	43	33	12	23	38	54	29	42	43	110	76	60	68	87	92	91	78	68	64	81	77	79	84	102	102	126	126	114	101	89	105	109	131	120	119	144	133
D60	185	170	143	167	158	0	49	60	26	33	45	41	41	43	2	9	18	30	25	37	37	105	50	54	63	81	74	87	68	62	60	75	67	79	88	102	111	112	112	106	89	81	96	94	147	120	130	122	139
C60	49	49	49	49	49	49	0	49	10	33	14	9	13	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	20	13	19	27	49	49	49	49	49	40	33	26	25	17	32	49	49	49	49	49	49
C80	60	60	60	60	60	60	49	0	10	33	22	14	14	2	0	0	0	0	0	0	0	5	0	0	1	1	4	3	28	19	22	33	49	60	60	60	60	46	41	33	33	23	38	60	60	60	60	60	60
A60	26	26	26	25	26	26	10	10	0	6	2	6	26	9	0	1	1	4	1	2	3	13	6	4	7	10	8	8	4	7	26	26	26	14	14	18	19	26	11	15	8	9	26	26	23	20	22	26	22
C40	33	33	33	33	33	33	33	33	6	0	9	6	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	9	11	16	33	33	33	33	33	27	22	17	17	11	21	33	33	33	33	33	33
Q120	54	50	45	41	48	45	14	22	2	9	0	21	8	8	3	8	12	17	3	10	9	27	23	12	13	16	31	22	54	27	15	27	17	26	26	30	29	39	54	39	54	34	31	33	38	34	32	40	37
S120	47	47	46	39	47	41	9	14	6	6	21	0	13	13	4	8	16	24	9	12	10	32	28	14	17	22	33	28	29	47	22	33	18	17	19	24	22	41	39	47	36	47	38	30	32	28	26	40	31
A80	43	43	43	40	43	41	13	14	26	8	8	13	0	13	1	3	7	10	3	5	5	25	13	7	12	16	17	19	12	16	43	43	30	19	19	24	25	43	22	27	17	19	43	43	34	27	28	43	33
N	43	38	29	41	33	43	2	2	9	1	8	13	13	0	1	3	5	13	8	13	9	38	18	17	18	27	24	27	15	18	18	19	11	7	10	12	14	29	26	26	24	22	24	15	35	21	26	22	28
B60	12	12	12	1	12	2	0	0	0	0	3	4	1	1	0	12	12	12	3	7	8	12	12	9	9	10	12	11	7	4	2	3	12	0	0	0	0	6	9	7	9	7	3	12	0	0	0	12	0
B80	23	23	23	4	23	9	0	0	1	0	8	8	3	3	12	0	23	23	8	14	14	21	23	17	15	18	23	19	13	10	5	7	13	0	0	0	0	13	19	16	18	15	9	23	1	0	0	23	2
B100	38	38	38	11	38	18	0	0	1	0	12	16	7	5	12	23	0	38	15	22	22	36	38	27	25	31	38	32	20	18	10	13	13	0	0	0	0	19	29	25	25	24	15	27	5	1	0	38	6
B120	54	54	54	22	54	30	0	0	4	0	17	24	10	13	12	23	38	0	19	28	27	50	54	34	36	43	54	45	29	28	16	19	16	0	0	3	2	31	43	39	37	36	23	30	15	10	6	44	15
F120	53	39	24	18	29	25	0	0	1	0	3	9	3	8	3	8	15	19	0	53	52	53	23	53	53	53	32	53	11	14	9	10	4	0	2	1	3	22	27	25	18	21	16	10	11	3	8	21	8
E120	70	54	36	27	42	37	0	0	2	0	10	12	5	13	7	14	22	28	53	0	58	70	32	70	62	70	43	63	20	19	12	15	9	0	2	1	4	30	38	34	27	28	23	18	16	5	9	31	12
F140	71	53	35	26	43	37	0	0	3	0	9	10	5	9	8	14	22	27	52	58	0	71	34	71	71	71	44	71	19	18	12	14	11	1	4	2	6	29	39	35	26	29	23	18	15	5	11	31	12
E180	149	128	92	88	110	105	1	5	13	0	27	32	25	38	12	21	36	50	53	70	71	0	67	92	100	121	89	128	49	47	37	44	25	15	24	31	38	74	88	85	66	68	59	46	69	47	53	71	62
B140	76	76	67	40	76	50	0	0	6	0	23	28	13	18	12	23	38	54	23	32	34	67	0	43	47	56	76	60	42	41	21	24	18	4	4	14	11	46	63	57	54	52	35	33	31	23	22	52	33
E140	92	73	48	41	60	54	0	0	4	0	12	14	7	17	9	17	27	34	53	70	71	92	43	0	77	92	58	79	26	25	15	18	13	2	6	5	11	38	51	47	35	39	29	23	27	13	20	40	23
F160	100	82	55	50	68	63	0	1	7	0	13	17	12	18	9	15	25	36	53	62	71	100	47	77	0	100	64	100	29	27	19	23	16	4	8	11	18	45	57	56	40	45	35	27	35	21	27	42	30
E160	121	102	71	66	87	81	0	1	10	0	16	22	16	27	10	18	31	43	53	70	71	121	56	92	100	0	75	106	35	35	25	30	20	9	13	17	25	55	67	68	47	54	43	34	49	30	37	56	43
B160	107	107	79	62	92	74	0	4	8	0	31	33	17	24	12	23	38	54	32	43	44	89	76	58	64	75	0	80	52	50	26	31	20	11	10	23	26	61	82	81	67	74	46	40	50	36	39	62	49
F180	128	108	76	71	91	87	0	3	8	0	22	28	19	27	11	19	32	45	53	63	71	128	60	79	100	106	80	0	42	41	29	33	19	10	16	22	28	61	78	76	58	61	48	37	54	36	40	57	48
Q140	84	80	72	61	78	68	20	28	4	14	54	29	12	15	7	13	20	29	11	20	19	49	42	26	29	35	52	42	0	40	22	37	29	32	32	42	37	60	84	61	84	52	49	48	56	50	49	59	56
S140	70	70	64	60	68	62	13	19	7	9	27	47	16	18	4	10	18	28	14	19	18	47	41	25	27	35	50	41	40	0	29	40	23	26	27	36	35	60	57	70	52	70	55	39	51	42	43	54	51
A100	64	64	64	56	64	60	19	22	26	11	15	22	43	18	2	5	10	16	9	12	12	37	21	15	19	25	26	29	22	29	0	64	37	27	28	35	36	64	37	43	29	32	64	53	48	40	41	64	47
A120	81	81	81	70	81	75	27	33	26	16	27	33	43	19	3	7	13	19	10	15	14	44	24	18	23	30	31	33	37	40	64	0	46	39	41	48	49	81	52	57	44	44	81	66	62	54	55	79	61
ABC60	77	77	77	65	77	67	49	49	26	33	17	18	30	11	12	13	13	16	4	9	11	25	18	13	16	20	20	19	29	23	37	46	0	53	53	57	58	62	49	45	39	32	51	77	62	59	61	77	61
C100	79	79	79	79	79	79	49	60	14	33	26	17	19	7	0	0	0	0	0	0	1	15	4	2	4	9	11	10	32	26	27	39	53	0	77	79	79	57	52	43	40	31	47	65	79	79	79	79	79
D160	88	86	83	88	84	88	49	60	14	33	26	19	19	10	0	0	0	0	2	2	4	24	4	6	8	13	10	16	32	27	28	41	53	77	0	83	88	57	51	45	39	32	49	65	88	84	88	78	86
C120	102	102	102	102	102	102	49	60	18	33	30	24	24	12	0	0	0	3	1	1	2	31	14	5	11	17	23	22	42	36	35	48	57	79	83	0	96	72	66	58	54	42	62	70	102	102	102	87	102
D140	111	109	99	111	102	111	49	60	19	33	29	22	25	14	0	0	0	2	3	4	6	38	11	11	18	25	26	28	37	35	36	49	58	79	88	96	0	69	62	60	49	43	59	71	111	102	111	88	107
A160	130	130	117	104	126	112	40	46	26	27	39	41	43	29	6	13	19	31	22	30	29	74	46	38	45	55	61	61	60	60	64	81	62	57	57	72	69	0	93	95	81	80	105	85	93	83	83	103	94
Q180	141	134	114	100	126	112	33	41	11	22	54	39	22	26	9	19	29	43	27	38	39	88	63	51	57	67	82	78	84	57	37	52	49	52	51	66	62	93	0	96	109	79	71	74	89	80	78	95	90
S180	130	126	104	97	114	106	26	33	15	17	39	47	27	26	7	16	25	39	25	34	35	85	57	47	56	68	81	76	61	70	43	57	45	43	45	58	60	95	96	0	79	99	75	66	85	71	75	86	83
Q160	109	105	92	80	101	89	25	33	8	17	54	36	17	24	9	18	25	37	18	27	26	66	54	35	40	47	67	58	84	52	29	44	39	40	39	54	49	81	109	79	0	67	61	61	71	64	63	75	73
S160	99	98	81	75	89	81	17	23	9	11	34	47	19	22	7	15	24	36	21	28	29	68	52	39	45	54	74	61	52	70	32	44	32	31	32	42	43	80	79	99	67	0	62	51	63	51	54	68	61
A140	105	105	99	90	105	96	32	38	26	21	31	38	43	24	3	9	15	23	16	23	23	59	35	29	35	43	46	48	49	55	64	81	51	47	49	62	59	105	71	75	61	62	0	73	79	70	70	89	80
ABC80	109	109	109	88	109	94	49	60	26	33	33	30	43	15	12	23	27	30	10	18	18	46	33	23	27	34	40	37	48	39	53	66	77	65	65	70	71	85	74	66	61	51	73	0	81	73	74	109	81
D100	147	140	122	147	131	147	49	60	23	33	38	32	34	35	0	1	5	15	11	16	15	69	31	27	35	49	50	54	56	51	48	62	62	79	88	102	111	93	89	85	71	63	79	81	0	120	130	101	136
C140	120	120	113	120	120	120	49	60	20	33	34	28	27	21	0	0	1	10	3	5	5	47	23	13	21	30	36	36	50	42	40	54	59	79	84	102	102	83	80	71	64	51	70	73	120	0	115	92	120
D120	130	126	111	130	119	130	49	60	22	33	32	26	28	26	0	0	0	6	8	9	11	53	22	20	27	37	39	40	49	43	41	55	61	79	88	102	111	83	78	75	63	54	70	74	130	115	0	93	124
ABC100	144	144	144	111	144	122	49	60	26	33	40	40	43	22	12	23	38	44	21	31	31	71	52	40	42	56	62	57	59	54	64	79	77	79	78	87	88	103	95	86	75	68	89	109	101	92	93	0	101
C160	139	139	122	138	133	139	49	60	22	33	37	31	33	28	0	2	6	15	8	12	12	62	33	23	30	43	49	48	56	51	47	61	61	79	86	102	107	94	90	83	73	61	80	81	136	120	124	101	0

Appendix U: Summary of different strains and sequencing data of *S. scintillans* examined

RCC=Roscoff Culture Collection, France; NIES=Microbial Culture Collection at the National Institute of Environmental Studies, Japan; WGA=Whole genome amplification; AF-SMG=Amplification-free shotgun metagenome.

	Culture collection	Isolation location	FISH location	Sequencing method	Sequencing location	Library Prep Date (M/Y)	Reported strain in Guillou et al., 1999
RCC24	RCC	Pacific Ocean	Canada	WGA	Canada	04-2022	Yes
RCC257	RCC	Atlantic Ocean	Canada	WGA	Canada	04-2022	No
RCC257-late	RCC	Atlantic Ocean	Canada	WGA	Canada	06-2022	No
RCC257-jp	RCC	Atlantic Ocean	Japan	——	——	——	No
RCC24-jp (NIES-2589)	NIES	Pacific Ocean	Japan	AF-SMG	Japan	03,05,06, 10-2022	Yes
RCC25	RCC	Mediterranean Sea	——	One of the two strains reported by Guillou et al., in 1999. Lost at RCC in 2008			

Appendix V: Summary of genomic characteristics and references of prasinoviruses

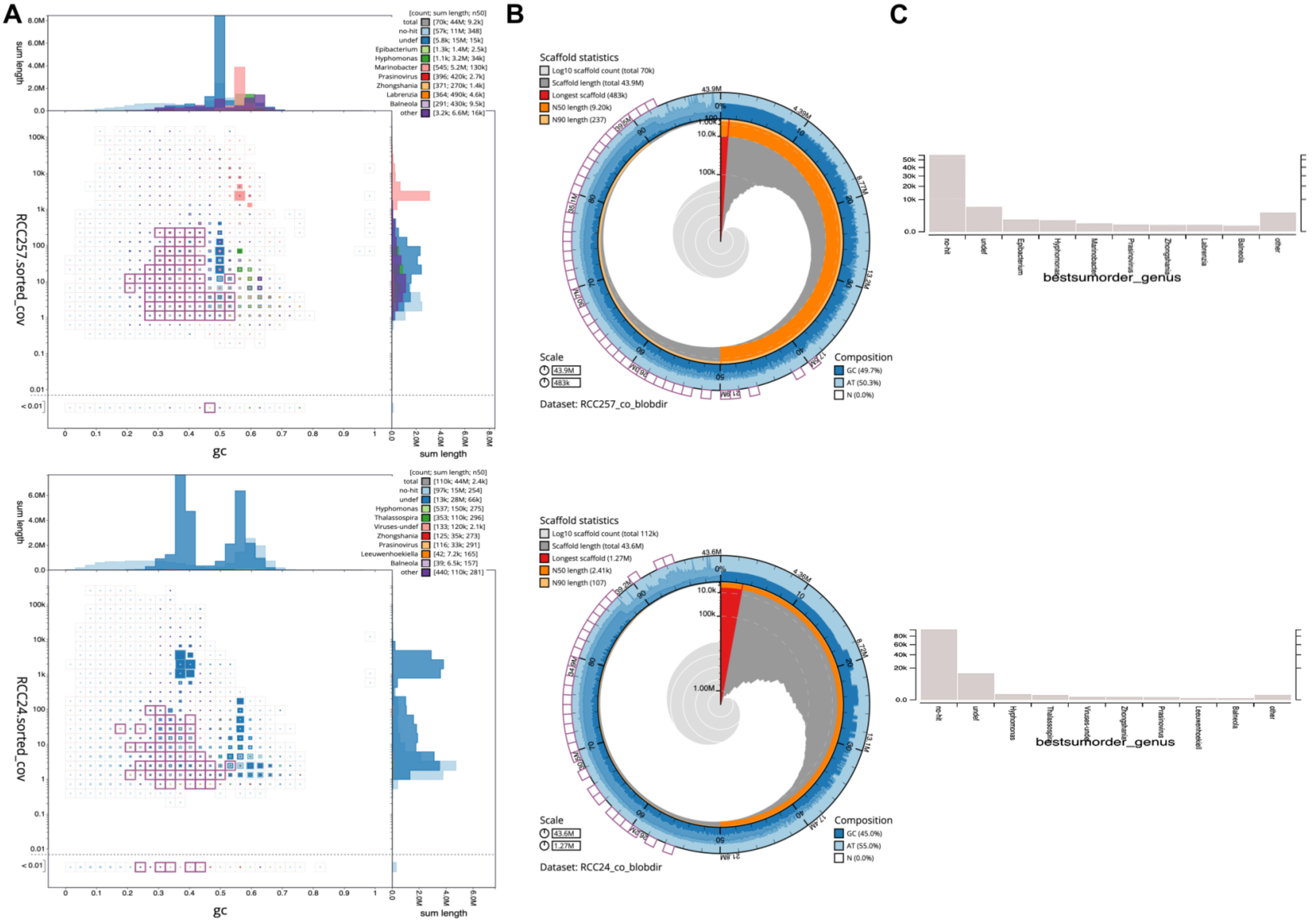
List of genomes used to guide assemblies and the subsequent vMAGs from RCC24 and RCC257. CheckV% indicates completeness for each vMAG assemblies.

Virus Genomes	Length (bp)	GC %	ORFs	tRNAs	Genes	CheckV %	Accession	Publication
BIIV1	174,426	35.2	220	3		NA	MK522034- MK522037	Bachy et al., 2021
BIIV2	207,870	36.5	235	2	249	NA	MK522038	
BIIV3	211,597	36.3	230	3	241	NA	MK522039	
BpV1	198,519	37.2	203	3	203	NA	NC014765	Moreau et al., 2010
BpV2	187,069	37	210	4	225	NA	HM004430	
OlV1	194,022	40.9	254	5	255	NA	NC014766	Zimmerman et al., 2019
OlV2	196,300	41.2	269	5	274	NA	NC028091	Derelle et al., 2015
OlV4	216,925	40.3	256	5	319	NA	JF974316	
OlV5	186,468	41.6	254	4	263	NA	NC020852	Derelle et al., 2015
OlV6	184,949	41.7	251	5	257	NA	HQ633059	
OlV7	182,309	41	243	5	248	NA	NC028093	Zimmerman et al., 2019
OmV1	193,301	44.6	252	5	257	NA	NC028092	Derelle et al., 2015
OtV1	189,567	44.5	240	4	233	NA	JN225873	
MpV1	184,095	39	244	6	244	NA	NC014767	Moreau et al., 2010
MpV_12T	205,622	39.8	253	7	265	NA	NC020864	
MpV_Pl1	197,060	43.3	259	5	270	NA	HQ633072	Finke et al., 2017
RCC257_vMAG_BIIV1	104,406	36.17	153		137	54.61		This study
RCC257_vMAG_BIIV2	98,732	35.35	138		130	51.51		This study
RCC257_vMAG_BIIV3	83,238	36.04	116		109	43.54		This study
RCC257_vMAG_BpV1	193,823	36.13	295	2	254	100		This study
RCC257_vMAG_BpV2	195,514	36.27	297	2	259	100		This study

Virus Genomes	Length (bp)	GC %	ORFs	tRNAs	Genes	CheckV %	Accession	Publication
RCC257_vMAG_MpV1	19,211	39.79	23		21	10.03		This study
RCC257_vMAG_MpV12T	15,031	40.58	15		15	7.85		This study
RCC257_vMAG_MpVPl1	34,707	41.52	47		40	18.11		This study
RCC257_vMAG_OlV1	102,789	39.97	149		146	53.93		This study
RCC257_vMAG_OlV2	87,373	39.91	119		113	45.83		This study
RCC257_vMAG_OlV4	56,877	39.55	88		82	29.77		This study
RCC257_vMAG_OlV5	89,948	39.99	116		114	47.18		This study
RCC257_vMAG_OlV6	89,926	39.98	119		117	47.17		This study
RCC257_vMAG_OlV7	102,138	40.01	152		143	53.59		This study
RCC257_vMAG_OmV1	67,431	39.84	811		71	35.37		This study
RCC257_vMAG_OtV1	68	39.7	85	1	73	35.89		This study
RCC24_vMAG_BIIV1	125,432	37.12	183	4	165	65.61		This study
RCC24_vMAG_BIIV2	97,551	37.27	113	3	108	51.03		This study
RCC24_vMAG_BIIV3	90,981	37.68	112	3	105	47.6		This study
RCC24_vMAG_BpV1	223,996	36.51	341	4	291	100		This study
RCC24_vMAG_BpV2	221,235	36.53	338	4	286	100		This study
RCC24_vMAG_MpV1	40,992	40.41	64	1	60	21.4		This study
RCC24_vMAG_MpVPl1	80,383	40.7	137	1	132	41.98		This study
RCC24_vMAG_OlV1	123,654	40.18	221	2	190	64.54		This study
RCC24_vMAG_OlV2	149,739	39.93	244	2	220	78.52		This study
RCC24_vMAG_OlV4	72,632	40.29	125		119	38.1		This study
RCC24_vMAG_OlV5	121,783	40.15	196	2	188	63.87		This study
RCC24_vMAG_OlV6	119,913	40.14	196	2	182	62.59		This study
RCC24_vMAG_OlV7	120,063	40.46	203	2	176	62.69		This study
RCC24_vMAG_OmV1	125,261	39.89	196	2	186	65.39		This study
RCC24_vMAG_OtV1	92,373	39.38	137		121	48.43		This study

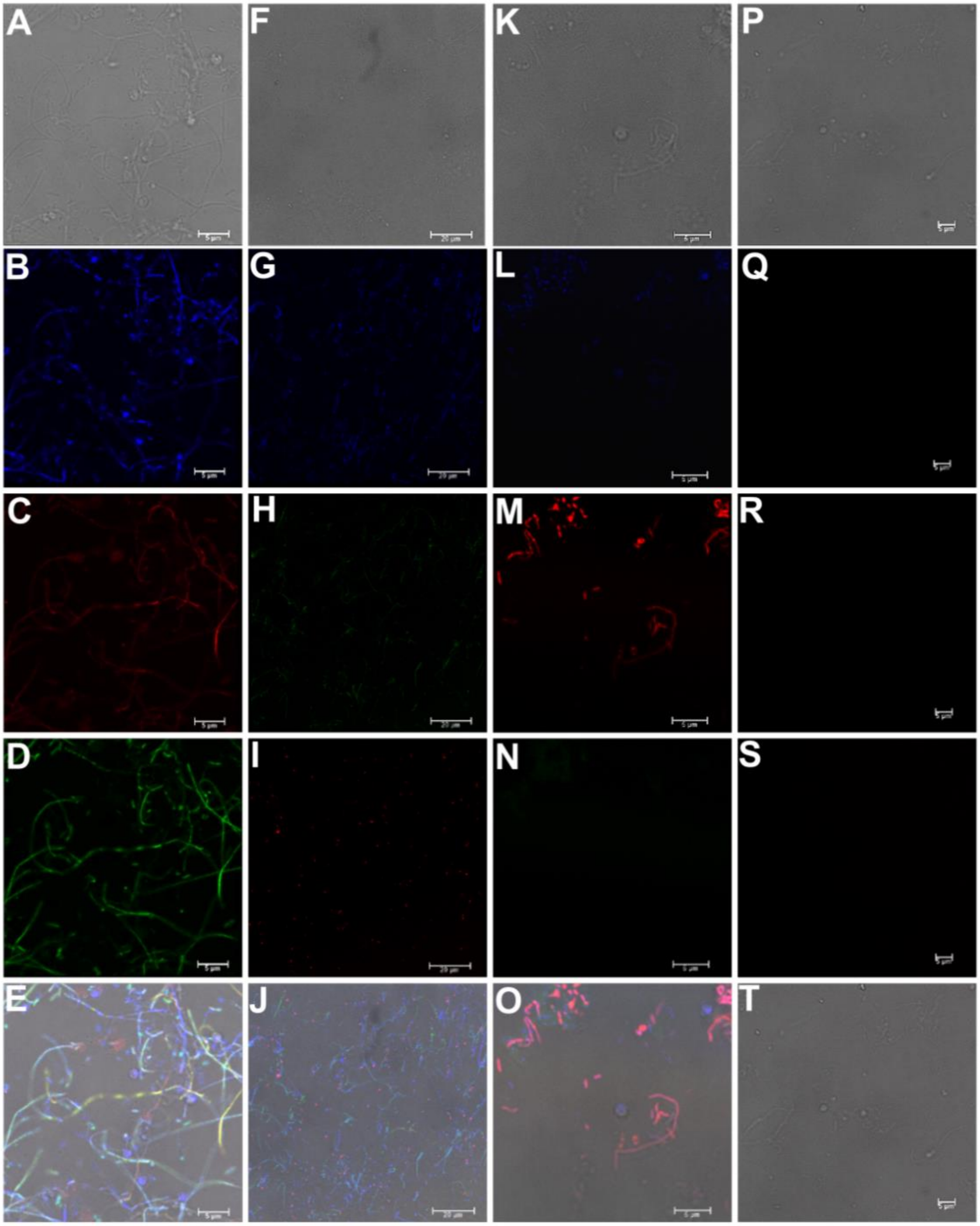
Appendix W: Summary of BlobToolKit analyses of the initial WGA assembles

Visualization of sequencing results of RCC257 (top row) and RCC24 (bottom row) without filtering or sub-setting. (A) Blob plots based on mean coverage (per-base) in y-axis and mean GC contents in x-axis. Each “blob” represents a square-root scaled size (showing max size) of a scaffold with its size representing the length or span. The blobs are coloured according to the top ten taxonomic assignment at the genus level (‘bestsum’ taxrule), based on coverage. Sum lengths along each axis are plotted on histograms. All reads assigned to prasinoviruses are highlighted with purple squares around each blob. (B) Snail plots visualizing quality of the initial assembly represented by N50 and N90. The purple squares in the blob plots and ones positioned at the outermost part of the plots are scaffolds assigned to prasinoviruses. (C) Histograms showing coverage (y-axis) for top ten genus (including “no-hit”, “undefined” and “others”).


A
B
C
sum length
8.0M
6.0M
4.0M
2.0M
0.0
[count; sum length; n50]
total [70k; 44M; 9.2k]
no-hit [57k; 11M; 348]
undef [5.8k; 15M; 15k]
Epibacterium [1.3k; 1.4M; 2.5k]
Hyphomonas [1.1k; 3.2M; 34k]
Marinobacter [545; 5.2M; 130k]
Prasinovirus [396; 420k; 2.7k]
Zhongshania [371; 270k; 1.4k]
Labrenzia [364; 490k; 4.6k]
Balneola [291; 430k; 9.5k]
other [3.2k; 6.6M; 16k]
RCC257.sorted_cov
100k
10k
1k
100
10
1
0.1
0.01
< 0.01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
gc
0.0 2.0M 4.0M 6.0M 8.0M
sum length
Scaffold statistics
Log10 scaffold count (total 70k)
Scaffold length (total 43.9M)
Longest scaffold (483k)
N50 length (9.20k)
N90 length (237)
43.9M
4.39M
8.77M
13.2M
17.5M
21.9M
26.3M
30.7M
35.1M
39.5M
Scale
43.9M
483k
Composition
GC (49.7%)
AT (50.3%)
N (0.0%)
Dataset: RCC257_co_blobdir
50k
40k
30k
20k
10k
0.0
no-hit
undef
Epibacterium
Hyphomonas
Marinobacter
Prasinovirus
Zhongshania
Labrenzia
Balneola
other
bestsumorder_genus
sum length
6.0M
4.0M
2.0M
0.0
[count; sum length; n50]
total [110k; 44M; 2.4k]
no-hit [97k; 15M; 254]
undef [13k; 28M; 66k]
Hyphomonas [537; 150k; 275]
Thalassospira [353; 110k; 296]
Viruses-undef [133; 120k; 2.1k]
Zhongshania [125; 35k; 273]
Prasinovirus [116; 33k; 291]
Leeuwenhoekiella [42; 7.2k; 165]
Balneola [39; 6.5k; 157]
other [440; 110k; 281]
RCC24.sorted_cov
100k
10k
1k
100
10
1
0.1
0.01
< 0.01
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
gc
0.0 2.0M 4.0M 6.0M
sum length
Scaffold statistics
Log10 scaffold count (total 112k)
Scaffold length (total 43.6M)
Longest scaffold (1.27M)
N50 length (2.41k)
N90 length (107)
43.6M
4.36M
8.72M
13.1M
17.4M
21.8M
26.2M
30.5M
34.9M
39.2M
Scale
43.6M
1.27M
Composition
GC (45.0%)
AT (55.0%)
N (0.0%)
Dataset: RCC24_co_blobdir
80k
60k
40k
20k
0.0
no-hit
undef
Hyphomonas
Thalassospira
Viruses-undef
Zhongshania
Prasinovirus
Leeuwenhoekiell
Balneola
other
bestsumorder_genus

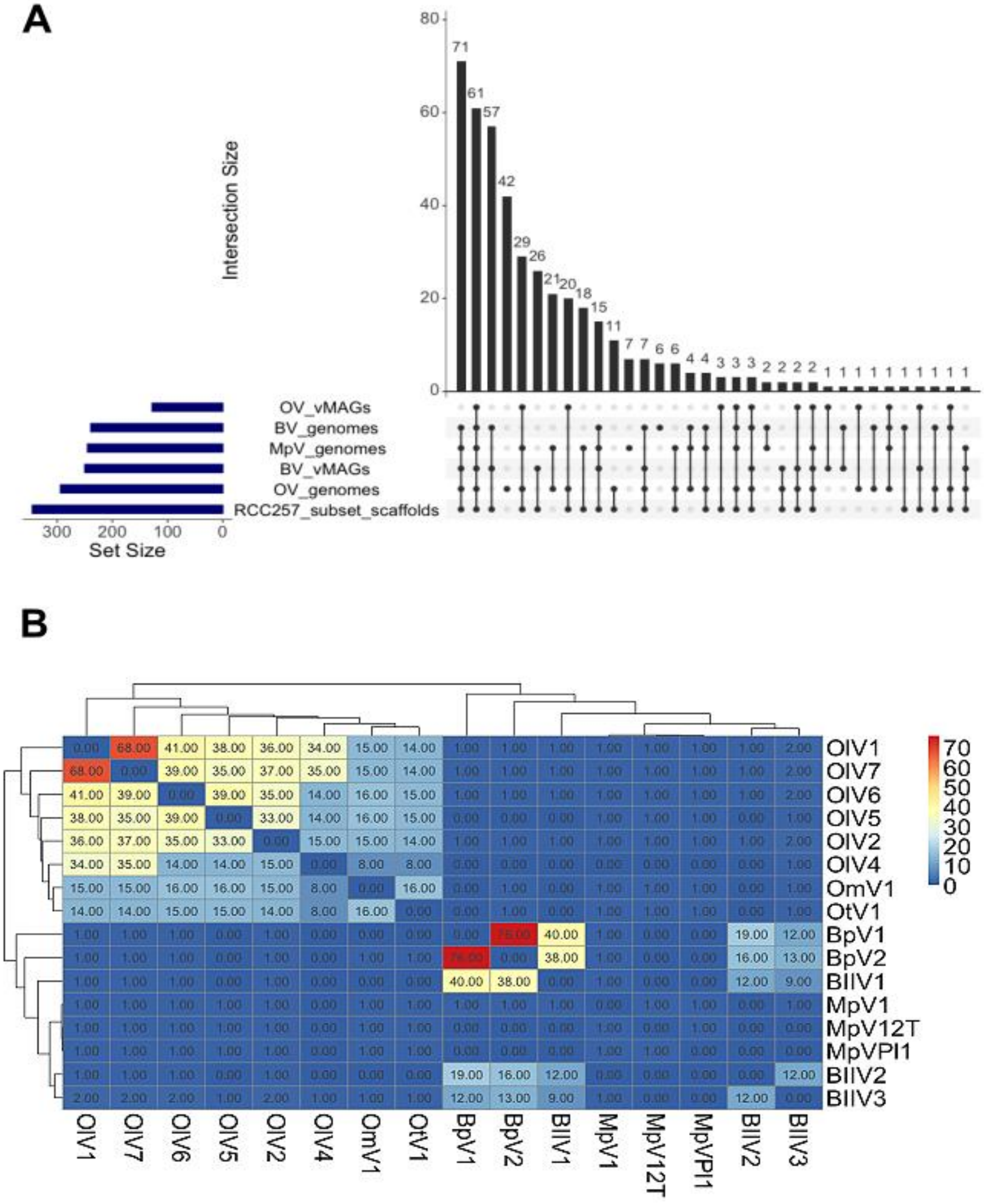
Appendix X: Fluorescence *in situ* hybridzation on RCC24

FISH analysis on *S. scintillans* RCC24-jp showing no endobacterial signals. (A), (F), (K), and (P) Brightfield; (B), (G), and (L) DAPI; (C) CF319 probe under 647 nm; (D) and (M) EUB388 probe under 488 and 647 nm; (H) γ-proteobacteria probe; (I) α-proteobacteria probe; (N) Planctomycete probe; (E), (J), (O), and (T) merged image of (A-D), (F-I), (K-N), and (P-S); (R-S) unstained controls under three different channels for DAPI, 488 and 647 nm. Scale bars = 5 µm for A-E and K-T; 20 µm for F-J.



Appendix Y: Number of shared orthologs and scaffolds among vMAGs

(A) Upset plot showing shared number of ortholog clusters among vMAGs, reference genomes and RCC257 viral-subset-scaffolds. (B) Heatmap showing shared number of recruited scaffolds from RCC257 viral-subset-scaffolds for each genome. Red colour indicates more shared numbers of scaffolds to assemble vMAGs. OV_vMAG= combined orthologs predicted from OlVs-, OtV1-, OmV1-vMAGs; BV_vMAGs=combined orthologs predicted from BpVs-, BIIVs-vMAGs; BV-genomes=combines orthologs predicted from reference genomes of BpVs and BIIVs; OV_genomes=combined orthologs predicted from reference genomes of OlVs, OtV1 and OmV1; RCC257_subset_scaffolds=RCC257 viral-subset-scaffolds



	OlV1	OlV7	OlV6	OlV5	OlV2	OlV4	OmV1	OtV1	BpV1	BpV2	BIIV1	MpV1	MpV12T	MpVPI1	BIIV2	BIIV3
OlV1	0.00	68.00	41.00	38.00	36.00	34.00	15.00	14.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00
OlV7	68.00	0.00	39.00	35.00	37.00	35.00	15.00	14.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00
OlV6	41.00	39.00	0.00	39.00	35.00	14.00	16.00	15.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00
OlV5	38.00	35.00	39.00	0.00	33.00	14.00	16.00	15.00	0.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00
OlV2	36.00	37.00	35.00	33.00	0.00	15.00	15.00	14.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00
OlV4	34.00	35.00	14.00	14.00	15.00	0.00	8.00	8.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
OmV1	15.00	15.00	16.00	16.00	15.00	8.00	0.00	16.00	0.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00
OtV1	14.00	14.00	15.00	15.00	14.00	8.00	16.00	0.00	0.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00
BpV1	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	76.00	40.00	1.00	0.00	0.00	19.00	12.00
BpV2	1.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00	76.00	0.00	38.00	1.00	0.00	0.00	16.00	13.00
BIIV1	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	40.00	38.00	0.00	1.00	0.00	0.00	12.00	9.00
MpV1	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00
MpV12T	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
MpVPI1	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00
BIIV2	1.00	1.00	1.00	0.00	1.00	0.00	0.00	0.00	19.00	16.00	12.00	0.00	0.00	0.00	0.00	12.00
BIIV3	2.00	2.00	2.00	1.00	2.00	1.00	1.00	1.00	12.00	13.00	9.00	1.00	0.00	0.00	12.00	0.00

Appendix Z: Summary BlobToolKit table for green algae and prasinoviruses

Sequencing results of RCC257 summarized using BlobToolkit showing hits assigned to green algal lineage. Scaffold IDs are omitted. “Uni” refers to Uniprot database.

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
46	Eukaryota	Viridiplantae	Chlorophyta	Trebouxiophyceae	Trebouxiales	Trebouxiaceae	Lobosphaera	Lobosphaera incisa	331.4455	89.9	99	9.69E-26	no hit	no hit	no hit	no hit
31	Eukaryota	Viridiplantae	Chlorophyta	Mamiellophyceae	Mamiellales	Bathycoccaceae	Ostreococcus	Ostreococcus tauri	185.0727	91.5	94	9.69E-26	no hit	no hit	no hit	no hit
30	Eukaryota	Viridiplantae	Chlorophyta	Mamiellophyceae	Mamiellales	Bathycoccaceae	Ostreococcus	Ostreococcus tauri	99.633	92.5	93	2.08E-27	no hit	no hit	no hit	no hit
32	Eukaryota	Viridiplantae	Chlorophyta	Mamiellophyceae	Mamiellales	Bathycoccaceae	Ostreococcus	Ostreococcus tauri	81.7545	90.625	96	9.69E-26	no hit	no hit	no hit	no hit
54	Eukaryota	Viridiplantae	Chlorophyta	Chlorophyceae	Chlamydomonadales	Chlamydomonadaceae	Chlamydomonas	Chlamydomonas reinhardtii	22.7182	81.2	232	1.08E-41	A0A6G0XEB7_9STRA	35	1064	3.33E-200
51	Eukaryota	Viridiplantae	Chlorophyta	Chlorophyceae	Chlamydomonadales	Dunaliellaceae	Dunaliella	Dunaliella primolecta	6.5282	84.4	141	7.86E-26	A0A5A8CRY8_CAFRO	64	292	5.67E-125
56	Eukaryota	Viridiplantae	Chlorophyta	Chlorophyceae	Chlamydomonadales	Chlamydomonadaceae	Chlamydomonas	Chlamydomonas reinhardtii	2.9457	83.1	207	7.67E-43	A8JAX1_CHLRE	78	76	2.39E-38
59	Eukaryota	Viridiplantae	Chlorophyta	Chlorophyceae	Sphaeropleales	Selenastraceae	Monoraphidium	Monoraphidium neglectum	2.9033	73.2	508	1.92E-33	A0A812N4R2_9DINO	65	266	1.14E-109
60	Eukaryota	Viridiplantae	Chlorophyta	Mamiellophyceae	Mamiellales	Mamiellaceae	Micromonas	Micromonas pusilla	2.7721	83.2	179	2.65E-34	C1MTF8_MICPC	49	162	8.52E-45
48	Eukaryota	Viridiplantae	Chlorophyta	Trebouxiophyceae	Chlorellales	Chlorellaceae	Auxenochlorella	Auxenochlorella protothecoides	2.374	82.2	365	1.88E-78	no hit	no hit	no hit	no hit
44	Eukaryota	Viridiplantae	Chlorophyta	Trebouxiophyceae	Trebouxiophyceae-undef	Trebouxiophyceae-undef	Picochlorum	Picochlorum sp. 'soloecismus'	2.0939	88	175	1.71E-46	no hit	no hit	no hit	no hit
49	Eukaryota	Viridiplantae	Chlorophyta	Trebouxiophyceae	Chlorellales	Chlorellaceae	Micractinium	Picochlorum sp. 'soloecismus'	1.6617	95.2	269	9.45E-115	no hit	no hit	no hit	no hit
59	Eukaryota	Viridiplantae	Chlorophyta	Trebouxiophyceae	Chlorellales	Chlorellaceae	Auxenochlorella	Auxenochlorella protothecoides	1.6427	84.5	162	4.10E-35	A0A835ZE30_9STRA	75	1290	4.63E-53

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
30	Eukaryota	Viridiplantae	Chlorophyta	Trebouxiophyceae	Chlorellales	Chlorellaceae	Chlorella	Chlorella variabilis	1.4009	87.3	212	7.77E-60	A0A4P9ZP24_9FUNG	89	70	5.55E-37
31	Eukaryota	Viridiplantae	Chlorophyta	Trebouxiophyceae	Chlorellales	Chlorellaceae	Helicosporidium	Helicosporidium sp. ex Simulium jonesi	1.2377	78.8	226	9.51E-30	A0A4Q5PZF8_9CYAN	79	75	4.62E-31
35	3499	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
34	862	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	5819	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
32	2996	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
35	973	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
35	825	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
43	588	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
40	1430	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	4562	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	1007	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC716 virus 3							
39	1190	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC716 virus 1							
38	770	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
33	710	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
45	764	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
40	654	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
40	694	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
34	5830	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
35	373	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
38	399	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
39	268	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
32	333	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	6382	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	1079	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	2571	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	1223	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	9870	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	801	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
41	153	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	6208	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	2348	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	399	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	2544	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	281	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	534	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	3047	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
30	274	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	177	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	1641	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	977	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	819	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	853	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	472	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	387	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
41	445	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	333	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	1001	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	2064	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	2468	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	9879	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 2						
36	11754	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	1313	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	1009	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	447	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
41	1393	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
31	3472	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	343	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	3408	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	16987	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
36	1628	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	834	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	478	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	442	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	775	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	1365	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	510	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	5882	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
32	695	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	3953	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	1545	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	1149	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
30	378	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	454	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	4004	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
30	844	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	9690	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	1008	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	5153	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	321	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	367	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	4114	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	374	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
31	6174	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	414	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	706	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
28	307	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	1087	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	2409	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	352	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	996	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
37	367	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
31	622	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	343	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
41	14865	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
39	397	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Phycodnaviridae-undef		Bathycoccus virus BpV178						
37	474	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	157	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	359	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	213	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	858	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	2698	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	247	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	502	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	2127	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	864	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	714	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	340	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	487	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	294	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
21	263	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	4074	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
29	461	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	233	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
34	314	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	418	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
27	264	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	486	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
29	204	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	319	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	688	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	570	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	447	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
52	1848	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Dishui Lake phycodnavirus 3						
37	474	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	1076	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	7309	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
24	306	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	4276	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	679	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	291	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
30	374	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
30	605	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	397	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	345	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	503	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	846	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
31	1688	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 2						
36	2292	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	5322	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus 1						
38	10264	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
35	4802	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
32	466	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	305	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
26	363	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	388	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	585	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	177	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	435	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
21	173	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	549	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	1528	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
36	1131	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	2017	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	1574	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus OlV5						
40	8102	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus OlV5						
34	409	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	3570	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	536	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	467	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
28	277	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
35	258	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	3968	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
33	158	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC716 virus 1							
36	1425	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC716 virus 3							
34	1091	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
38	388	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
33	246	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
39	2272	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
33	173	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
43	902	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
30	291	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
40	371	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	2189	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
31	220	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC716 virus 1							
29	206	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	526	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
41	1770	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
45	257	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
35	1555	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
38	745	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
32	338	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	285	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
32	594	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
30	763	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	290	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
45	1733	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
41	5875	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
33	716	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	165	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
29	615	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
31	1084	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	2224	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	213	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	525	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	1179	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
35	737	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 2						
38	251	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	270	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	669	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
34	280	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	1043	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
43	742	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
33	288	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
50	844	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
34	579	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
33	254	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
53	213	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Micromonas sp. RCC1109 virus MpV1						
48	650	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus 1						
42	543	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	786	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
32	443	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
48	1186	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
38	1331	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
39	1138	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
48	639	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
32	305	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
46	596	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
35	555	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
49	647	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	354	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
47	593	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	185	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	730	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
46	338	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
39	422	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
45	1864	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
32	644	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
44	499	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 3						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
41	1108	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
43	1280	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
43	640	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
38	570	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
45	252	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	207	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
45	266	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
35	307	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	271	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
45	654	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	381	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	388	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
44	668	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Micromonas sp. RCC1109 virus MpV1						
46	281	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	521	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
27	341	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	627	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
39	570	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
40	520	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
45	521	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Micromonas sp. RCC1109 virus MpV1						
40	230	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	161	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
40	293	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
26	223	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	303	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
40	535	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
34	332	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 1							
44	659	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
40	213	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
43	1281	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
43	785	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
44	364	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus tauri virus 2							
28	734	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC716 virus 2							
50	1062	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
51	251	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
39	254	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
40	497	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
44	320	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 1							
44	838	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 1							
36	390	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus mediterraneus virus 2							
46	933	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
37	333	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 7							
36	464	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							
49	477	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 1							
45	274	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Ostreococcus lucimarinus virus 1							
37	347	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus	Bathycoccus sp. RCC1105 virus BpV							

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
46	623	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
42	835	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
47	251	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
44	211	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	895	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
45	213	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
40	717	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
49	265	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
40	288	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	295	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
47	443	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
40	296	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
53	369	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Micromonas sp. RCC1109 virus MpV1						
42	290	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	435	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
41	266	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
33	297	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
40	252	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
29	224	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
37	374	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	524	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 1						
49	750	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 2						
34	225	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
32	373	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	458	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 1						
41	310	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
41	235	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	1091	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
41	469	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	398	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 1						
44	316	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
37	392	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	323	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 1						
42	322	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
27	244	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	404	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 2						
47	211	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
46	245	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
33	245	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
30	372	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	328	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
43	396	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
39	576	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
32	247	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
52	371	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
36	337	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
33	516	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
29	519	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 1						
41	259	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
45	346	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
26	163	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	347	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
40	263	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
49	350	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
39	444	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
40	441	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
43	249	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
40	267	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 2						
36	178	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 1						
41	324	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
40	210	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
39	811	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
40	546	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
34	275	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus mediterraneus virus 1						
35	370	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
37	561	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
43	281	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
47	285	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
43	286	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						

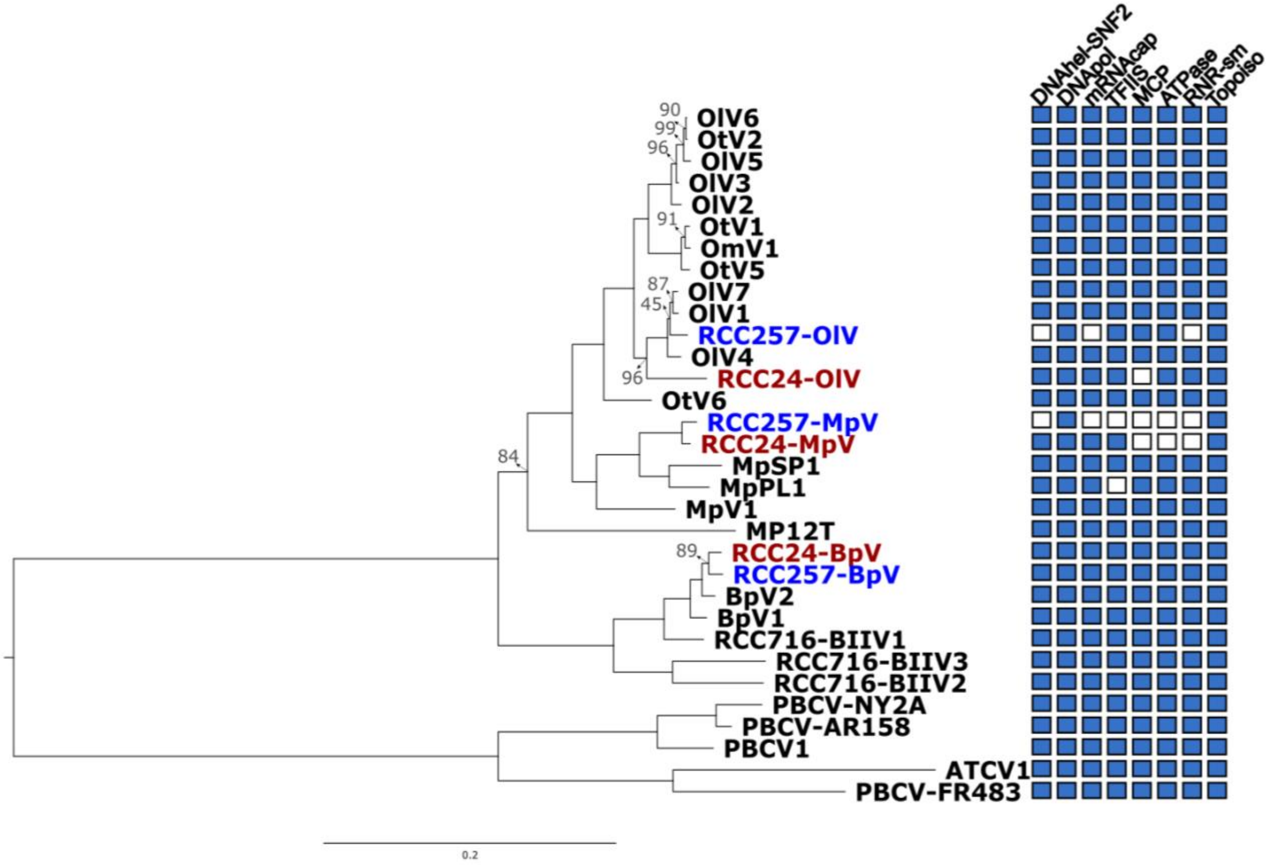
GC %	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
38	547	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
53	298	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus 2						
44	305	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	199	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
37	309	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	409	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
31	307	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
29	208	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
41	213	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
36	316	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
43	216	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
41	218	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
46	222	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus 1						
47	223	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
45	224	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	226	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
47	228	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
48	341	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
46	229	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
40	230	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
39	344	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
34	225	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	697	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
46	232	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
40	233	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
41	217	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
39	537	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
42	247	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
41	248	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus OtV5						
38	369	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus RT-2011						
41	247	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
41	249	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
42	254	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC716 virus 3						
48	254	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
50	259	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus 2						
46	518	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
32	259	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
29	268	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
45	268	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
38	266	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
34	266	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
46	265	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
43	267	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
33	272	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
32	272	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
31	270	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

GC%	superkingdom	kingdom	phylum	class	order	family	genus	species	coverage	%ID	length	evalue	Uniprot ID	Uni_%ID	Uni_length	Uni_evalue
53	271	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
42	246	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Micromonas sp. RCC1109 virus MpV1						
37	294	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 7						
36	278	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
38	305	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
43	326	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus lucimarinus virus 1						
41	335	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Ostreococcus tauri virus OtV5						
39	168	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						
45	161	Viruses	Bamfordvirae	Nucleocytoviricota	Megaviricetes	Algavirales	Phycodnaviridae	Prasinovirus		Bathycoccus sp. RCC1105 virus BpV						

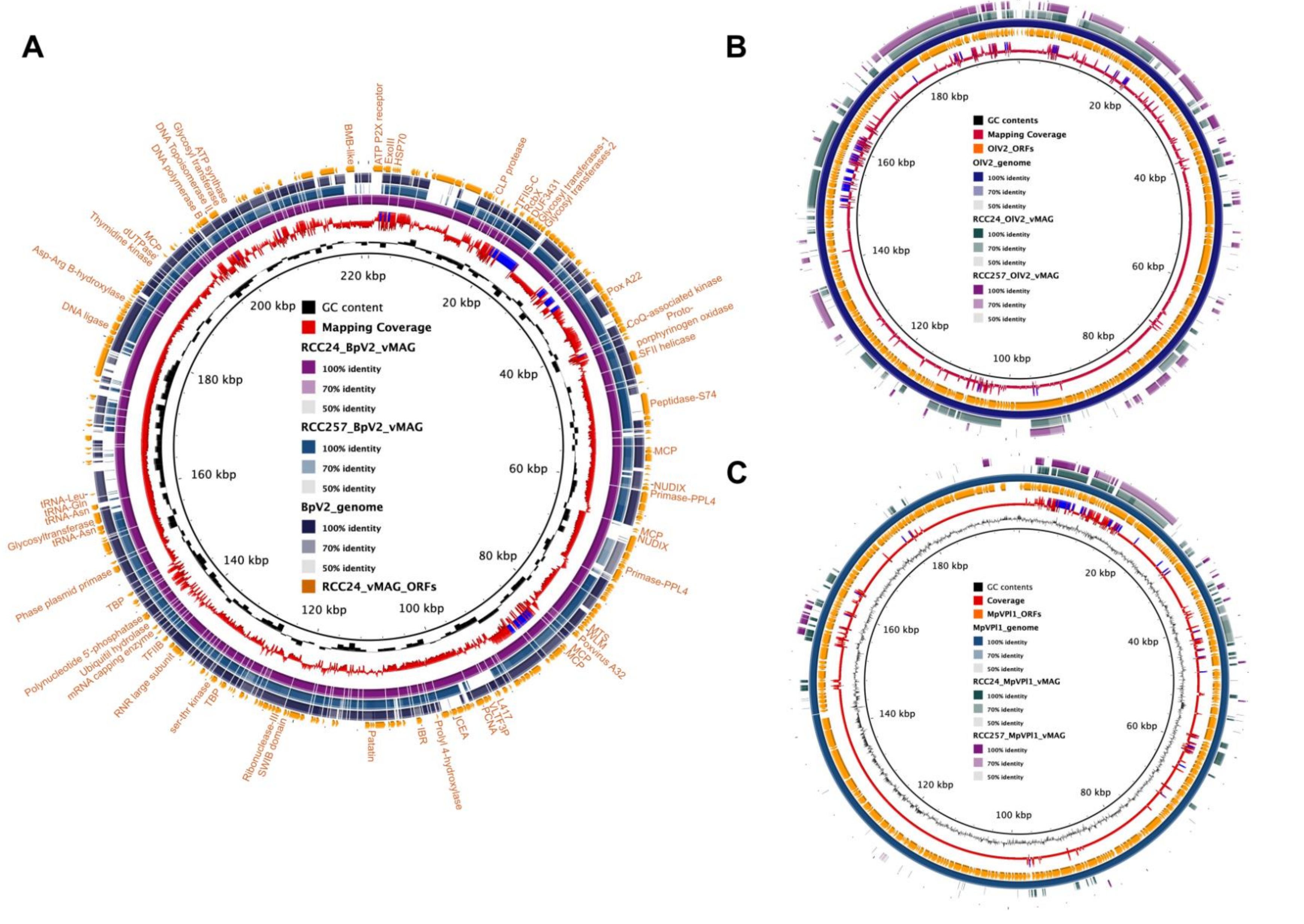
Appendix AA: Prasinovirus phylogeny including both RCC24 and RCC257 data

A multi-gene prasinovirus phylogeny reconstructed from 22 prasinovirus core genes (5,355 sites) using IQ-TREE2 LG+F+G4 model, including genes searched from WGA data of two different *S. scintillans* strains, RCC24 and RCC257. The right panel shows presence-absence of select core genes. Single-copy genes are DNApol (DNA polymerase B), DNAhel-SNF2 (SNF2 helicase), mRNAcap (mRNA capping enzyme), ATPase, and RNR-sm (RNR small subunit). The tree is rooted with Chlorovirus (PBCVs and ATCV) for visualization. Only nodes <100% ultrafast bootstrap supports are labelled. OlV=*Ostreococcus lucimarinus* virus; OtV=*Ostreococcus tauri* virus; OmV=*Ostreococcus mediterraneus* virus; MpV=*Micromonas pusilla* virus; BpV=*Bathycoccus prasino* virus; BIIV=*Bathycoccus* sp. virus clade BII. PBCV=*Paramecium bursaria chlorella* virus; ATCV=*Acanthocystis turfaceae chlorella* virus.



Appendix AB: vMAG genome overview from RCC24 and RCC257

Genome overview and comparison of the most complete BpV-, OlV-, and MpV-vMAGs to corresponding reference genomes. (A) Circularized representation of (A) RCC24 BpV-vMAG compared to BpV2 genome and RCC257 BpV-vMAG; (B) OlV2 genome compared to RCC24 and RCC257 OlV2-vMAGs, (C) MpV-Pl1 genome compared to RCC24 and RCC257 MpVPl1-vMAGS, in an ordered set of coding sequences, represented by blocks shaded by similarity. (A) Mapping coverage is based on RCC24 BpV-vMAG mapped to RCC24 WGA viral-subset-scaffolds and regions with the coverage more than one standard deviation [59.9] from the mean coverage [44.5] are shown in blue spikes. The outermost ring represents predicted ORFs of the vMAG with manually annotated protein from Prodigal-gv and Viralrecall. (B) Mapping coverage is based on OlV2 genome mapped to RCC24 WGA viral-subset-scaffolds and regions with the coverage more than one standard deviation [5.1] from the mean coverage [1.6] shown in blue spikes. Only ORFs from the reference OlV2 genome is shown and the partial RCC24 and RCC257 OlV2-vMAG CDS are shown in the outer rings. (C) Mapping coverage is based on MpV-Pl1 mapped to RCC24 WGA viral-subset-scaffolds and regions with the coverage more than one standard deviation [2.1] from the mean coverage [0.6] are shown in blue spikes. Only ORFs from the reference MpV-Pl1 genome is shown and the partial RCC24 and RCC257 MpVPl1-vMAGs CDS are shown in the outer rings. See Table S2 for annotation in a tabular format.


A
BMB-like
ATP P2X receptor
ExoIII
HSP70
CLP protease
TFIIS-C
RcbX
DUF3431
Glycosyl transferases-1
Glycosyl transferases-2
Pox A22
CoQ-associated kinase
Proto-
porphyrinogen oxidase
SFII helicase
Peptidase-S74
MCP
NUDIX
Primase-PPL4
MCP
NUDIX
Primase-PPL4
MTS
WLM
Poxvirus A32
MCP
MCP
L417
VLTF3P
PCNA
ICEA
Prolyl 4-hydroxylase
IBR
Patatin
Ribonuclease-III
SWIB domain
TBP
ser-thr kinase
RNR large subunit
TFIIB
mRNA capping enzyme
Ubiquitil hydrolase
Polynucleotide 5'-phosphatase
TBP
Phase plasmid primase
tRNA-Asn
Glycosyltransferase
tRNA-Asn
tRNA-Gln
tRNA-Leu
DNA ligase
Asp-Arg B-hydroxylase
Thymidine kinase
dUTPase
MCP
DNA polymerase B
DNA Topoisomerase II
Glycosyl transferase
ATP synthase
220 kbp
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
180 kbp
200 kbp
GC content
Mapping Coverage
RCC24_BpV2_vMAG
100% identity
70% identity
50% identity
RCC257_BpV2_vMAG
100% identity
70% identity
50% identity
BpV2_genome
100% identity
70% identity
50% identity
RCC24_vMAG_ORFs
B
180 kbp
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
GC contents
Mapping Coverage
OlV2_ORFs
OlV2_genome
100% identity
70% identity
50% identity
RCC24_OlV2_vMAG
100% identity
70% identity
50% identity
RCC257_OlV2_vMAG
100% identity
70% identity
50% identity
C
180 kbp
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
GC contents
Coverage
MpVPl1_ORFs
MpVPl1_genome
100% identity
70% identity
50% identity
RCC24_MpVPl1_vMAG
100% identity
70% identity
50% identity
RCC257_MpVPl1_vMAG
100% identity
70% identity
50% identity

Appendix AC: Select annotation of RCC257 BpV2-vMAG

genome	vog	virbit	vdesc
BpV2_HM004430_RagTag	GVOG13327	26.2354722	unknown | DNA polymerase III is a complex, multichain enzyme responsible for most of the replicative synthesis in bacteria. The epsilon subunit contain the editing function and is a proofreading 3'-5' exonuclease | PFAM Exonuclease, RNase T and DNA polymerase
BpV2_HM004430_RagTag	GVOGm0548	31.7694822	Cell shape determining protein MreB Mrl | ribosomal large subunit binding | Heat shock 70 kDa protein | unfolded protein binding | ATP binding | nuclear pore complex assembly | Heat shock 70 kDa protein | ATP binding | unknown | ATP binding | ethanolamine
BpV2_HM004430_RagTag	GVOG11896	16.1245155	ATP-dependent specificity component of the Clp protease. It directs the protease to specific substrates. Can perform chaperone functions in the absence of ClpP | ATPase which is responsible for recognizing, binding, unfolding and translocation of substrate
BpV2_HM004430_RagTag	GVOGm0173	12.9073622	termination of RNA polymerase III transcription | termination of RNA polymerase I transcription | Belongs to the archaeal rpoM eukaryotic RPA12 RPB9 RPC11 RNA polymerase family | transcription, DNA-templated
BpV2_HM004430_RagTag	GVOGm1326	17.6720118	RbcX protein | PFAM RbcX protein
BpV2_HM004430_RagTag	GVOG06656	19.6468827	Catalyzes the transfer of a two-carbon ketol group from a ketose donor to an aldose acceptor, via a covalent intermediate with the cofactor thiamine pyrophosphate | Transketolase, thiamine diphosphate binding domain | PFAM Transketolase domain protein | t
BpV2_HM004430_RagTag	GVOG03757	16.6072273	dehydrogenase e1 component | PFAM Transketolase central region | Transketolase | Catalyzes the acyloin condensation reaction between C atoms 2 and 3 of pyruvate and glyceraldehyde 3-phosphate to yield 1-deoxy-D-xylulose-5-phosphate (DXP)
BpV2_HM004430_RagTag	GVOGm0477	10.0697567	dTDP-4-dehydrorhamnose reductase | NAD-dependent epimerase dehydratase | racemase and epimerase activity, acting on carbohydrates and derivatives | Catalyzes the two-step NADP-dependent conversion of GDP- 4-dehydro-6-deoxy-D-mannose to GDP-fucose, involvi
BpV2_HM004430_RagTag	GVOGm0363	14.0641388	unknown | negative regulation of septation initiation signaling | unknown | negative regulation of septation initiation signaling | Glycosyl transferases group 1
BpV2_HM004430_RagTag	GVOGm1048	18.6064505	COG0463 Glycosyltransferases involved in cell wall biogenesis | Glycosyl transferase family 2
BpV2_HM004430_RagTag	GVOGm1447	14.8290256	no_annot
BpV2_HM004430_RagTag	GVOGm0873	11.2915898	unusual protein kinase | regulation of tocopherol cyclase activity | kinase activity | Is probably a protein kinase regulator of UbiI activity which is involved in aerobic coenzyme Q (ubiquinone) biosynthesis | ubiquinone biosynthetic process

genome	vog	virbit	vdesc
BpV2_HM004430_RagTag	GVOGm0071	18.568791	Catalyzes the 6-electron oxidation of protoporphyrinogen-IX to form protoporphyrin-IX | oxidoreductase activity | Protoporphyrinogen oxidase | oxidoreductase activity | COG1233 Phytoene dehydrogenase and related proteins | tRNA (5-methylaminomethyl-2-thio
BpV2_HM004430_RagTag	GVOGm0013	23.2163735	PFAM helicase domain protein | Type III restriction enzyme res subunit | DEAD DEAH box helicase | Type I site-specific restriction-modification system, R (Restriction) subunit and related | DEAD DEAH box helicase domain protein | ATP-dependent DNA helicase
BpV2_HM004430_RagTag	GVOG03539	18.1741575	unknown
BpV2_HM004430_RagTag	GVOGm0003	13.6601611	Large eukaryotic DNA virus major capsid protein
BpV2_HM004430_RagTag	GVOGm0003	15.5627761	Large eukaryotic DNA virus major capsid protein
BpV2_HM004430_RagTag	GVOG10311	18.1135309	no_annot
BpV2_HM004430_RagTag	GVOGm0085	19.1598539	DNA primase activity
BpV2_HM004430_RagTag	GVOGm0760	19.7737199	Poxvirus A32 protein | unknown
BpV2_HM004430_RagTag	GVOGm0003	18.5418446	Large eukaryotic DNA virus major capsid protein
BpV2_HM004430_RagTag	GVOGm0003	16.8937858	Large eukaryotic DNA virus major capsid protein
BpV2_HM004430_RagTag	GVOGm0303	10.397115	ICEA Protein | ICEA Protein
BpV2_HM004430_RagTag	GVOG10021	17.4985714	Possesses two activities a DNA synthesis (polymerase) and an exonucleolytic activity that degrades single-stranded DNA in the 3'- to 5'-direction. Has a template-primer preference which is characteristic of a replicative DNA polymerase | unknown | DNA pac
BpV2_HM004430_RagTag	GVOGm0018	12.1696343	triglyceride mobilization | esterase of the alpha-beta hydrolase superfamily | Esterase of the alpha-beta hydrolase superfamily | phosphatidylethanolamine catabolic process | esterase of the alpha-beta hydrolase superfamily | Patatin-like phospholipase |

genome	vog	virbit	vdesc
BpV2_HM004430_RagTag	GVOGm0160	11.9205705	regulation of transcription by RNA polymerase I | SWIB/MDM2 domain
BpV2_HM004430_RagTag	GVOGm0031	16.0592652	YqaJ-like viral recombinase domain | YqaJ-like viral recombinase domain | YqaJ-like viral recombinase domain | unknown | YqaJ-like viral recombinase domain | YqaJ viral recombinase family | YqaJ-like viral recombinase domain | YqaJ-like viral recombinase
BpV2_HM004430_RagTag	GVOGm0088	20.2089089	Catalyzes the reduction of ribonucleotides to deoxyribonucleotides. May function to provide a pool of deoxyribonucleotide precursors for DNA repair during oxygen limitation and or for immediate growth after restoration of oxygen
BpV2_HM004430_RagTag	GVOGm0088	28.4587421	Catalyzes the reduction of ribonucleotides to deoxyribonucleotides. May function to provide a pool of deoxyribonucleotide precursors for DNA repair during oxygen limitation and or for immediate growth after restoration of oxygen
BpV2_HM004430_RagTag	GVOG12766	24.6048776	defense response to oomycetes | Serine threonine protein kinase | histone kinase activity (H3-T3 specific) | 3-phosphoinositide-dependent protein kinase activity | protein serine/threonine kinase activity | protein serine/threonine kinase activity
BpV2_HM004430_RagTag	GVOGm0172	17.5840837	Stabilizes TBP binding to an archaeal box-A promoter. Also responsible for recruiting RNA polymerase II to the pre-initiation complex (DNA-TBP-TFIIB) | TBP-class protein binding | RNA polymerase III type 3 promoter DNA binding
BpV2_HM004430_RagTag	GVOGm0868	21.881499	mRNA guanylyltransferase activity | unknown
BpV2_HM004430_RagTag	GVOGm0214	13.1377319	thiol-dependent ubiquitin-specific protease activity | Opioid growth factor receptor (OGFr) conserved region | catalytic activity, acting on a protein | mitochondrion organization | metalloendopeptidase inhibitor activity | ubiquitinyl hydrolase activity
BpV2_HM004430_RagTag	GVOGm1594	17.0205758	polynucleotide 5'-phosphatase activity
BpV2_HM004430_RagTag	GVOGm1694	30.8074666	unknown | Phage plasmid primase, P4 | Phage plasmid primase P4 family
BpV2_HM004430_RagTag	GVOG12168	13.6821051	glycosyltransferase involved in LPS biosynthesis
BpV2_HM004430_RagTag	GVOGm0003	11.827933	Large eukaryotic DNA virus major capsid protein
BpV2_HM004430_RagTag	GVOGm0054	16.6973052	SMART DNA-directed DNA polymerase B | DNA polymerase | DNA replication proofreading | chloroplast mRNA modification | DNA replication proofreading | leading strand elongation | DNA polymerase activity

genome	vog	virbit	vdesc
BpV2_HM004430_RagTag	GVOG03001	12.5499004	DNA replication proofreading | leading strand elongation | DNA polymerase | DNA polymerase activity
BpV2_HM004430_RagTag	GVOGm0003	13.0613935	Large eukaryotic DNA virus major capsid protein
BpV2_HM004430_RagTag	GVOG05686	12.973049	DNA topoisomerase type II (ATP-hydrolyzing) activity | A type II topoisomerase that negatively supercoils closed circular double-stranded (ds) DNA in an ATP-dependent manner to modulate DNA topology and maintain chromosomes in an underwound state.
BpV2_HM004430_RagTag	GVOG10851	23.0195569	PFAM Glycosyl transferases group 1 | glycogen (starch) synthase activity | COG0438 Glycosyltransferase | PFAM Glycosyl transferase, group 1 | PFAM Glycosyl transferase, group 1 | Glycosyl transferases group 1 | PFAM Glycosyl transferase, group 1
BpV2_HM004430_RagTag	GVOGm0872	13.685759	unknown | unknown | unknown | atp synthase

Appendix AD: Annotated HMM clusters found only in BV vMAGs from RCC257 viral-subset-scaffolds

HMM_cluster	Uniprot or VOG	Description	E-value	score	bias
I2320	YHDJ_ECOLI	DNA adenine methyltransferase YhdJ OS=Escherichia coli (strain K12	1.60E-25	94.8	0.1
I2321	YP_009052178.1	putative methyltransferase [Aureococcus anophagefferens virus]	3.60E-41	145.7	4.9
I2340	YP_004061542.1	DUFF5855 similar to neurofilament protein [BpV1]	5.50E-166	557.5	67
I2341	YP_004061557	DUF5756 [BpV1]	2.70E-20	77.5	1.3
I2346	N/A	N/A			
I2347	N/A	N/A			
I2348	YP_001648133.2	hypothetical [OtV5]	8.50E-07	34	1
I2349	YP_004061690.1	DUF5773 [OlV1]	7.50E-09	40.8	5.9
I2350	YP_009465930.1	DUF5773 [Dishui lake phycodnavirus 1]	1.10E-12	53.1	3.3
I2351	YP_004063626.1	hypothetical [OtV2]	2.60E-40	143.6	15.5
I2369	YP_004061518.1	DUF5762 [BpV1]	3.00E-26	95.8	3.1
I2370	YP_004061552.1	similar to Ribonuclease III [BpV2]	9.62E-12	50.3	0.5
I2371	YP_004061535.1	hypothetical [BpV1]	5.50E-12	51	9.1
I2372	ARF08749.1	hypothetical [Catovirus 1]	4.30E-13	54.1	10.5
I2374	YP_004061915.1	4-hydroxy-2-oxopentanoic acid aldolase [MpV1]	1.80E-10	45.5	2.4
I2375	YP_009665084.1	Pyrimidine (PYR) binding domain of DXS, and transketolase protein [MpV_SP1]	2.50E-78	266.5	0.5
I2376	YP_009665083.1	Transketolase domain-containing protein [MpV_SP1]	1.30E-107	363.1	7.1

HMM_cluster	Uniprot or VOG	Description	E-value	score	bias
I2377	YP_009665082.1	adenosylmethionine-dependent methyltransferases (SAM or AdoMet-MTase) [MpV_SP1]	8.60E-88	298.9	8
I2380	YP_009173577.1	ICEA 1 virulence factor [Chrysochromulina ericina virus]	3.90E-04	24.9	2.5
I2381	YP_004061450.1	hypothetical [BpV1]	4.00E-18	69.5	3.6
I2405	YP_009465878.1	hypothetical [Dishui lake phycodnavirus 1]	3.40E-09	41.4	0.2
I2407	N/A	N/A			
I2411	N/A	N/A			
I2412	AYV75274.1	hypothetical [Terrestrivirus sp.]	1.20E-25	95.5	7.8
I2416	YP_004061591.1	nuclease family [BpV1]	6.10E-07	34.3	1.2
I2417	YP_004061616.1	hypothetical [BpV1]	1.70E-22	84	3.4