Phycologia

Phylogenetic Reconstruction of the Diatoms Using Seven Genes, Multiple Outgroups and Morphological data for a Total Evidence Approach

--Manuscript Draft--

Manuscript Number:	PH20-10R1
Full Title:	Phylogenetic Reconstruction of the Diatoms Using Seven Genes, Multiple Outgroups and Morphological data for a Total Evidence Approach
Article Type:	Original Article
Keywords:	Diatoms; CMB hypothesis; SG hypothesis, multi-gene phylogeny; multiple outgroups.
Corresponding Author:	Linda Medlin, PhD Marine Biolgoical Association of the UK Plymouth, UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Marine Biolgoical Association of the UK
Corresponding Author's Secondary Institution:	
First Author:	Linda Medlin, PhD
First Author Secondary Information:	
Order of Authors:	Linda Medlin, PhD
	Yves Desdevies, desdevies@obs-banyuls.fr
Order of Authors Secondary Information:	
Abstract:	Version:1.0 StartHTML:0000000311 EndHTML:0000005965 StartFragment:0000003217 EndFragment:0000005929 SourceURL:file://localhost/Users/lindamedlin/MEGA/cur%20work/submitted%20ms%20copy/redo%20theriot%20analysis/submission/phycologia/revised%20Phycologia/submission/revisedfinal%20Phycologia%20submitted.doc Medlin tested multiple outgroups with 18S rRNA dataset and found that haptophytes, ciliates, prasinophytes and chlorophytes recovered monophyletic Coscinodiscophyceae, Mediophyceae, Bacillariophyceae with strong BT support. Theriot et al. added six plastid genes to the diatom dataset but with only one outgroup, Bolidomonas and omitted most of the V4 region of that gene and bases beyond position 1200. They recovered a grade of clades from radial into polar centrics, into araphid pennates into the monophyletic raphid pennates. Their structural gradation hypothesis (SGH) contrasts to the CMB hypothesis of Medlin and Kaczmarska. We selected only those species with all seven genes from their dataset and added the entire 18S RNA gene to make a new dataset to which we sequentially added heterokont, haptophyte, and prasinophyte/chlorophyte outgroups. We analysed it using 1) evolutionary models with parameters relaxed across genes and codon positions for coding sequences (codon partition analysis scheme = CP) and 2) no partitions or evolutionary models as applied to each gene, using only optimised models of evolution for the entire dataset (NCP). CP recovered a monophyletic mediophycean and bacillariophycean clade and three coscinodiscophycean clades. Sequentially adding more outgroups did not change clade topology but dramatically increased BT support. NCP recovered a monophyletic Coscinodiscophyceae and Bacillariophyceae and three Mediophyceae clades, each with strong bootstrap support. Morphological data was added and analyzed similarly. NCP recovered three monophyletic classes and CP recovered the Bacillariophyceae arising from within the Mediophyceae, making the subphylum monophyletic but the class was paraphyletic. Each analysis was tested with SH tests in PAUP and IQTree. Plastid inheritance in the diatoms is not homogenous and thus their phylogenies may not be homologous. If so, then our application of gene models may be overparametrising the data. The application of no partitioning models with morphological data supported the CMB hypothesis.

Cover Letter

Click here to access/download

Cover Letter

cover letter.doc

Response to Reviewers

Click here to access/download

Response to Reviewers

respone to reviewers.doc

Mini Review

Review of the Phylogenetic Reconstruction of the Diatoms Using Molecular Tools with an Analysis of a Seven Gene Data Set Using Multiple Outgroups and Morphological Data for a Total Evidence Approach

Linda K. Medlin [1],* and Yves Desdevises [2]

[1] Marine Biological Association of the UK, Plymouth PL1 2PB UK

[2] Sorbonne Université, CNRS, Biologie Intégrative des Organismes Marins, BIOM, Observatoire Océanologique, F-66550 Banyuls-sur-Mer, France

* Correspondence: lkm@mba.ac.uk

Abstract: Medlin tested multiple outgroups with 18S rRNA dataset and found that haptophytes, ciliates, prasinophytes and chlorophytes recovered monophyletic Coscinodiscophyceae, Mediophyceae, Bacillariophyceae with strong BT support. Theriot *et al.* added six plastid genes to the diatom dataset but with only one outgroup, *Bolidomonas* and omitted most of the V4 region of that gene and bases beyond position 1200. They recovered a grade of clades from radial into polar centrics, into araphid pennates into the monophyletic raphid pennates. Their structural gradation hypothesis (SGH) contrasts to the CMB hypothesis of Medlin and Kaczmarska. We selected only those species with all seven genes from their dataset and added the entire 18S RNA gene to make a new dataset to which we sequentially added heterokont, haptophyte, and prasinophyte/chlorophyte outgroups. We analysed it using 1) evolutionary models with parameters relaxed across genes and codon positions for coding sequences (codon partition analysis scheme = CP) and 2) no partitions or evolutionary models as applied to each gene, using only optimised models of evolution for the entire dataset (NCP). CP recovered a monophyletic mediophycean and bacillariophycean clade and three coscinodiscophycean clades. Sequentially adding more outgroups did not change clade topology but dramatically increased BT support. NCP recovered a monophyletic Coscinodiscophyceae and Bacillariophyceae and three Mediophyceae clades, each with strong bootstrap support. Morphological data was added and analyzed similarly. NCP recovered three monophyletic classes and CP recovered the Bacillariophyceae arising from within the Mediophyceae, making the subphylum monophyletic but the class was paraphyletic. Each analysis was tested with SH tests in PAUP and IQTree. Plastid inheritance in the diatoms is not homogenous and thus their phylogenies may not be homologous. If so, then our

application of gene models may be overparametrising the data. The application of no partitioning models with morphological data supported the CMB hypothesis.

Keywords: diatoms; CMB hypothesis; SG hypothesis; multi-gene phylogeny; multiple outgroups.

Introduction

The diatoms are one of the most diverse groups of unicellular eukaryotic protists. Their origins date from the early Mesozoic as judged by molecular clocks and their fossil records (Kooistra & Medlin 1996; Sims *et al.* 2006, Sorhannus 2007, Medlin 2014). From the Cenozoic, their global diversity has increased (Harwood & Gersonde 1990; Sims *et al.* 2006; Finkel *et al.* 2005). They can be found in all aquatic habitats and in moist terrestrial habitats and are responsible for nearly half of the primary production in the oceans and close to a quarter of the carbon fixed globally (Smetacek 1999). Finkel & Kotrc (2010) report that diatoms export organic carbon into the ocean depths by high sinking rates, relatively large cell sizes and densities and their ability to form large blooms. Relative to other phytoplankton groups, they remove more carbon out of contact with the atmosphere because of their high growth rates (Finkel & Kotrc 2010). Their diversity has increased from their origin to today (Finkel *et al.* 2005).

Diatoms have an absolute requirement for silica in order to initiate DNA replication, thus they have an important impact on silica cycles (see references in Finkel & Kotrc 2010). It is believed that as terrestrial grasslands evolved, they released silica to the global silica pool and the diatoms had an adaptive advantage. Their large storage vacuole enabled them to out-compete other phytoplankton. These hypotheses have been tested by re-analysis of fossil data and have been refuted (Rabosky & Sorhannus 2009). Rabosky & Sorhannus (2009) reported a drop in diatom diversity in the Oligocene, which they believe was correlated with a major drop in CO_2 concentrations as temperatures fell globally. Armbrust (2009) suggested that the divergence dates of the two centric classes as proposed by Medlin and Kaczmarska (2004) were correlated with declining CO_2 levels and their divergence occurred when CO_2 levels rose. She used the molecular clock produced by Sorhannus (2007) to provide divergence dates for her interpretation. Their closest relatives, the Parmales in the Bolidophyceae, do not have an important influence on silica cycles because they do not require silica for cell division (Yamada *et al.* 2014). Finkel & Kotrc (2010) noted that oceanic silicic acid concentrations have declined since diatoms have risen to prominence. Thus, the origin, evolution and diversity of the group is important because they play such an important role in all aquatic ecosystems and they will undoubtedly play an important role in oceanic ecosystems as climate changes.

Despite more than a century of morphological observation and nearly three decades of molecular phylogenetic analyses, the study of diatom phylogeny has progressed slowly, most of which has been controversial (see review in Medlin, 2016b). Medlin *et al.* (1993) produced the first phylogeny of the diatoms using molecular data and suggested that the centric and araphid diatoms were not monophyletic. Based on nearly 20 years of mismatch between molecular and morphological classifications, Medlin & Kaczmarska (2004) revised the classification system of the diatoms, creating two new subphyla, Coscinodiscophytina: with the radial centrics in the amended Coscinodiscophyceae, and Bacillariophytina with two classes: the pennates in the amended Bacillariophyceae and the bipolar centrics in a new class, Mediophyceae. These three classes ((Coscinodiscophyceae = radial centric diatoms) (Mediophyceae = polar centric diatoms + radial Thalassiosirales; Bacillariophyceae = pennate diatoms)) more accurately reflect the evolution and diversity of the diatoms than does the three-class system of centrics, araphid pennates and raphid pennates presented in Round *et al.* (1990). Medlin & Kaczmarska (2004) defined the three classes as follows: (1) the type of sexual reproduction and resultant auxospore formation, (2) the presence/absence of a tube or process (in the case of the centric diatoms) or raphe/sternum (in the pennate diatoms) inside the annulus (the initiation point for silicification in the diatoms), (3) symmetry of the valves and (4) the arrangement of the Golgi bodies in the cells (Medlin & Kaczmarska, 2004). The position of the cribrum in loculate areolae (excluding pseudoloculate areolae, which must have an internal cribrum) was added as another defining character to separate the two centric classes (Medlin 2014). Kaczmarska & Ehrman (2015) added the spore-like structure of the auxospore as another character separating the three classes. A summary of these traits can be found in Table 1. Exceptions to each character have been noted and the placement of the radial Thalassiosirales in the polar centric clade is one of the biggest exceptions to the features defining each class. Medlin (2016a) suggested retention of an ancestral polymorphism (scales) and loss of the ability to make bands to mould a radial centric into a polar one to explain why the radial Thalassiosirales are recovered in the polar diatom lineage, although they possess other valve features that place them in the polar lineage (Table 1). There are other examples in the pennate diatoms where a round morphology is presumed to reflect the loss of bands in the auxospore to squeeze the zygote into a pennate shape (Ashworth *et al.* 2013).

Theriot *et al.* (2009) claimed that one obstacle to obtaining a robust diatom molecular phylogeny has been that the nuclear-encoded small subunit ribosomal (SSU) was the primary gene of choice for phylogenetic analysis (Table S2, refer to most studies by Medlin and co-workers). Analysis of this gene under different taxon sampling schemes and with different optimality criteria has yielded results that differ in detail from one another (Theriot *et al.* 2009, 2010, 2011, 2015) and from that in Medlin and Kaczmarska (2004). In Medlin (2016b), she showed that in Theriot *et al.* (2009)'s re-analysis of Medlin's data, they had misrepresented the 99% tree burn in as the 90% tree burn to determine if her analysis had been run for enough generations. The 90% burn in showed that the analysis had run for a sufficient number of generations so the SSU gene could recover diatom phylogenies when used alone. Thus their analysis was flawed and their conclusion that the SSU gene could not be used to obtain a robust diatom phylogeny was subsequently flawed.

All of the analyses by Theriot and his co-workers (Table S2) have recovered more or less a grade of clades from the so-called radial centrics into polar centrics, which grade into araphid pennates, which themselves grade into the monophyletic raphid pennates, which they have termed the structural gradation hypothesis (SGH) in contrast to the CMB hypothesis (Coscinodiscophyceae, (Mediophyceae, Bacillariophyceae)) of Medlin and Kaczmarska (2004). Some of the later analyses by the Theriot group (Table S1) have recovered one or the other of the two centric classes monophyletic, whereas only those by Medlin and co workers plus the lone analysis by the Theriot group in Li *et al.* (2015), and the analyses done by Vaulot *et al.* (2007), Ehara *et al.* (2000) and Sorhannus (1997) have consistently recovered the two subphyla and the three subclasses using either the SSU alone or multiple genes and mostly with multiple outgroups (see Table S1 for more details on the multiple outgroups used in these papers).

Medlin (2016b) reviewed the evidence as to whether the molecular data have supported or refuted the classification changes made by Medlin & Kaczmarska (2004), i.e. whether scheme 1, CMB model with monophyletic classes, or scheme 2, SGH model of grades of clades, was better supported and to identify where future research areas in diatom phylogeny should be directed. Although the taxonomic changes in the diatoms have not been universally accepted, the general evidence shown in the review by Medlin (2016b) and the detailed analysis by Medlin (2014) and the fact that the trees produced Theriot *et al.* are not significantly different from the CMB hypothesis suggests that the revised classification of scheme 1 as proposed by Medlin and Kaczmarska (2004) should be accepted because of the defining features of each class reflects the morphological and sexual reproductive evolution of the diatoms. However, if the SGH hypothesis is the correct phylogeny, then the acceptance of paraphyletic lineages would have to be invoked to access the classification system proposed by Medlin and Kaczmarska (2004). Paraphyletic lineages are the natural course of evolution (see references in Medlin 2014).

To recover the CMB hypothesis or the three monophyletic classes obtained by Medlin and Kaczmarksa (2004), certain criteria must be met, which have not been followed or met in full by the Theriot group. Medlin and Kaczmarska proposed that the recovery of the two centric clades as monophyletic groups is highly dependent on an alignment based on the secondary structure of the SSU rRNA gene and the use of multiple outgroups. The effect of the secondary structure alignment on the topology of the rRNA tree has been documented in several studies (Medlin et al. , 1993, 2008; Medlin, 2010; Rimet et al. , 2011) and Theriot group only began using a secondary structure analysis in 2009 (Theriot *et al.* 2009), albeit the Gutell model, which does not have a structure for the V4 region of the SSU gene in contrast to the van de Peer model that does (Medlin 2010) so they either do not use it or only use the first helix in their analyses. The use of multiple outgroups has been tested with a single gene (Medlin, 2014) and multiple genes (Sato, 2008; Medlin & Desdevises, 2016), whereas the Theriot group has never tested the multiple outgroup criterion, outside of multiple heterokonts (Theriot *et al.* 2009). The usual number of outgroups the Theriot group use in their multi-gene analyses has been one or two bolidophytes since they began to use a secondary structure alignment (Theriot et al., 2009, 2010, 2013, 2015; Ashworth et al., 2012, 2013, Li *et al.* 2011). Theriot *et al.*

(2009) concluded that the use of the SSU rRNA gene was insufficient to recover the monophyletic classes as proposed by Medlin & Kaczmarska (2004) and directed their subsequent research into multi-gene analysis. However the information contained by the ribosomal RNA genes as compared to the protein-coding genes has been empirically tested by Piganeau *et al.* (2012) who showed that, for protists, the SSU gene contained more information and better resolution as compared to multi-cellular organisms. However, most of this information at the species level is found in the variable V4 region, most of which is omitted in the analyses by Theriot *et al.* (op cit). In the analysis of multiple outgroups with only the SSU rRNA gene, Medlin (2014) showed that the omission of the V4 region reverted the phylogeny recovered to a grade of centric clades, whereas its inclusion recovered monophyletic classes. Further to the Theriot's *et al.* 2009 study, Medlin (2014) provided evidence of an error in their interpretation of the phylogenetic analyses value of the SSU gene, which invalidated their claim that SSU gene was insufficient for resolving the diatom evolutionary history. Medlin (2014) explored the use of the SSU rRNA gene with multiple outgroups for the resolution of the centric classes to determine whether or not they were monophyletic, and if not, how many clades were recovered. She used 34 datasets with different combinations of outgroups, ingroups and numbers of nucleotides to study the effect of multiple outgroups on the ability of analyses of a single gene, the SSU rRNA gene, to recover monophyletic classes. She found that multiple representatives of haptophytes, chlorophytes, ciliates and heterokonts did recover monophyletic classes with high bootstrap support. She also looked at the effects of weighting the frequency of base substitutions per site if maximum parsimony analyses were used for large datasets. In her study, three of the datasets recovered the monophyletic clades. In her analysis, datasets 11 and 25 from Medlin (2014) were examined in more detail, to determine whether the number of nucleotides and the inclusion of short clone library sequences affected the relationships among the diatom taxa in the analyses. In 2016, Medlin and Desdevises expanded the SSU dataset to include 3 plastid genes and tested this with multiple heterokont outgroups and recovered monophyletic classes. In 2015, Theriot *et al.* expanded their data set for diatoms and multiple genes to include 207 taxa and 7 genes SSU plus *atp*B, *psa*A, *psa*B, *psb*A, *psb*C and *rbc*L from the plastid but still used a single outgroup and recovered a grade of clades that they called the structural gradation hypothesis (SGH) relating the four major structural groups (three clades of radial centrics, three clades of bipolar centrics, two clades of araphid pennate diatoms, and the raphid pennate diatoms) but were unable to recover a tree that invalidated those of Medlin & Kacsmarksa (2004).

We explored the addition of multiple outgroups using the Theriot *et al.* (2015) data. We only used their species that had all genes present because we found in Medlin & Desdevises (2016) that the omission of a single gene caused that taxon to have an elongate branch and making it subject to long-branch attraction errors (Figure S1). Using this reduced version of their data set and thirteen outgroups (Table 3), we performed phylogenetic analyses with and without an evolutionary model with parameters relaxed across genes and codon positions for coding sequences (codon partition scheme = CP, no evolutionary models for each gene = NCP). The decision not to use any codon models or partitioning of the data set was based on the evidence in Theriot *et al.* (2015) and Medlin and Desdevises (2016) that the third codon position in the plastid genes was not saturated. All combinations were tested using Shimodeira & Hasegawa tests in IQ-Tree and in PAUP against the monophyletic trees as obtained by Medlin and Kaczmarska (2004) and a reduced version of the Theriot *et al.* (2015) tree, removing all taxa without a complete set of genes. We added morphological data (Table 1) to our dataset and analyzed this in two ways: the morphological data was coded CATG for NCP analysis or numerically for CP analysis and weighted to contribute equally to the molecular data set (Table 2).

Materials and Methods

rRNA sequences from the diatoms in Table S2 were uploaded from Genbank and aligned to the SILVA SSU rRNA sequence alignment in the ARB program Version 5.5 using maximum primary and secondary structural similarity (Ludwig et al., 2004). We found many errors in the Genbank entries for the taxa in Table S1 from the Theriot *et al.* paper. For example, *Syndera hypberborea* was moved to *Synedroposis* in Hasle *et al*. (1995) but all of the sequences for all of its genes in Genbank list the taxon as *Synedra*. In some of the taxa, the same strain is given with a species name for some of the genes and referred to as “sp.” in others. We kept the specific epitat assuming that the specific epitat was the correct and final identification.

The ARB database release (Ref. NR 99, Ludwig *et al.* 2004) used in these analyses contained over 646,151 eukaryotic and prokaryotic sequences. Bases were aligned with one another based on their pairing across a helix. The ARB program generates a most parsimonious (MP) tree from all sequences and all positions in the database as its reference tree. The full SSU gene was used because the accuracy of the SILVA alignment enables the difficult V4 region to be aligned. The plastid protein genes (*rbc*L, *psa*A, *psb*B, *psa*C, *psa*B, *atp*B) were aligned individually using amino acids, then exported to be concatenated into one large file with the SSU gene.

Outgroups were chosen from other closely related algal groups based on the analyses by Medlin (2014). Ciliates could not be included because they are not photosynthetic. Four haptophytes, 2 chlorophytes, 2 prasinophytes, and 4 heterokonts and 2 bolidophytes (Table S2) were used for these analyses. Multiple examples from each group were selected to ensure that long-branch attraction was avoided by breaking up the long branch leading to each outgroup. Most of the outgroup taxa had complete plastid genomes available and their plastid genes were much longer than the amplified partial sequences from the Theriot *et al.* (2015) database. Thus, the plastid genes had to be trimmed so that lengths were almost identical, but we did not trim them as much as was done by Theriot *et al.* (2015), see Table 3. We selected only those species from Theriot *et al.* (2015) who were not missing any of the 7 genes. Our reason for this was that in Medlin and Desdevises (2014) we found that if one gene was missing in the data set, the branch length for that species was elongated relative to the others (Medlin & Desdevises, 2014, Figure S1). Trees were reconstructed from the concatenated alignment of the 7 genes (10565 bp, Table 3) using maximum likelihood (ML) with RaxML (Stamatakis *et al.* 2008), and with IQ-Tree (Nguyen *et al.* 2015), Bayesian Inference (BI) with MrBayes 3.2.6 (Ronquist *et al.* 2012). In ML, branch support was assessed using bootstrap and approximate likelihood-ratio test (Anisimova and Gascuel, 2006). This latter test is a much faster validation method than bootstrapping, and is based on a likelihood ratio test where the null hypothesis is that each tested internal branch has length 0.

BI was performed only on single genes with a mixed amino acid model for the translated coding sequences (except for SSU) and for the total evidence analysis when morphological data were added. Because of the high number of taxa, Bayesian analyses could not be performed on coding DNA sequences, either using a codon model or a codon partition scheme (CP), and on the concatenated dataset. The bootstrap support values from the maximum likelihood analyses are reported as whole numbers. Trees were loaded into FigTree (http://tree.bio.ed.ac.uk) to display them.

The first ML analysis was performed without any partitions for the protein coding genes using a general time reversible model accounting for rate heterogeneity across sites via a Gamma distribution. The best tree obtained was then compared to the taxonomic hypothesis from Medlin & Kaczmarska (2004), which was retrieved in 8% of the trees in the bootstrap analysis, using a SH-Test (Shimodeira & Hasegawa 1999) with PAUP 4b10 (Swofford 2003, Table 4).

For the second analysis, the parameters in the first analysis were also used, with additional parameters relaxed across genes and codon positions for coding sequences (CP) (all except SSU rDNA). Two trees were reconstructed, without and with the topological constraint (Coscinodiscophyceae, (Mediophyceae, Bacillariophyceae)) corresponding to the taxonomic hypothesis tested here (Medlin & Kaczmarska 2004). The outgroups were added sequentially in this order: bolidophytes, heterokonts, haptophytes, chlorophytes/prasinophytes. Each tree with each additional outgroup added was constrained by a similar tree with the CMB hypothesis. These two trees were then compared to each other and to the best tree obtained without CP using SH-Test and Weighted SH-Test (Shimodeira & Hasegawa 1999) using IQ-Tree and PAUP 4b10 (Tables 3 and 4). The WSH test is a less conservative version of the SH test (Shimodaira 2002). SH and WSH tests assess the difference between trees via their likelihoods. The significance of this difference is assessed from a null distribution, and in the WSH, each difference is divided by the estimate of the standard error.

We also took the tree from Theriot *et al.* (2015), pruned the taxa missing one or more of the plastid genes using Mesquite (ver. 3.2) (Maddison and Maddison 2017) and compared that to the tree from NCP analysis and to the final tree obtained with CP, constrained by the tree reflecting the CMB hypothesis with only one bolidomonad outgroup.

The morphological data in Table 1 were treated in two ways. They were first coded as CATG so that they could be used in the ML analysis with NCP (Table 2). Secondly they were coded numerically so that they could be used in a BI analysis with CP.

Characters were treated as unordered in the BI analysis, although initial tests with ordering the auxospore characters produced strange trees and this coding was abandoned. The features in Table 1 represent 7 characters; however it is certain that there are not just seven genes coding for these characters. Thus, the information for the morphology is not equal to the molecular information from the seven genes. Unequal data sets create a bias with regards to one having a greater influence than the other on the results (De Queiroz et al. 1995). Please refer to http://research.amnh.org/ ~siddall/methods/day5.html for a general discussion on weighting of characters. Therefore the morphological data was weighted by repeating the motive for the 7 characters (Table 1) because that essentially multiples each character in the morphological data set, just as one would do in a weighted parsimony analysis using a rescaled consistency index as the weighting tool. We repeated it 230 times making it approximately the same length as the SSU gene, obtaining all three clades, then gradually reduced the repeated motif in large blocks and repeated the analysis until the monophyletic groups disappeared. At that point we decided arbitrarily that one additional morphological motif would make the morphological information approximately equal to that of an additional gene. The final number of repeated motifs was 31 to yield a total of 217 nucleotides (numbers) for the morphological data.

Results

Individual Gene Analysis: Analyses were performed first with each gene individually (Figures S2-7) using both a DNA and an AA based analysis (plastid genes). Of the individual analyses, most of the plastid genes recovered a polytomy of many multiple lineages and only the 18S and the *psa*A (based on AA) and *psa*B (based on DNA) of the plastid genes on their own recovered any phylogenetic reconstruction that could be reconciled with modern diatom systematics in contrast to that recovered by Theriot *et al.* (2015) where *psa*A had the most phylogenetic information and the SSU had the least. In our study the 18S rRNA gene on its own recovered the most meaningful data structure (Figure S2) because it included the V4 region and bases beyond 1200, which were omitted from the Theriot *et al.* (2015) analysis. The dataset used in our analysis is longer than that used in Theriot *et al.* (2015) for two reasons (Table 3). We included the V4 region of the

SSU and bases beyond position 1200 and we did not trim the plastid genes so dramatically as in their study.

CP/NCP Analysis: The first phylogenetic analysis (NCP) on the concatenated dataset (Figure 1) without any codon partitioning or models of evolution applied to each gene displayed a monophyletic Coscinodiscophyceae, three clades of Mediophyceae and a monophyletic Bacillariophyceae. The monophyletic Coscinodicosphyceae (Figure 1) had 100% bootstrap support, which is among the highest support achieved for this clade to date (Table 6, Table S1). The three clades of Mediophyceae recovered in Figure 1 had a range of support from 64 to 96%, and the support for the backbone of three clades was strong (BT = 71-93) except for the sister relationship of the last mediophycean clade to the pennates, which was 43. Taxa in this last mediophyte clade were *Biddulphia* and *Attheya* spp. The pennate clade had 100% BT support. The back bone of our trees also had moderate to high bootstrap support (BT = 57-99, something that is missing from all of the Theriot analyses (BT ranging from 12 to a polytomy).

In Figure 1, *Actinoptychus undulatus* appeared distinct from the rest of the Coscinodiscophyceae and examination of its sequence revealed that its SSU sequence was quite divergent. The fact that this species was pulled out onto its own branch emphases the strong signal in the SSU gene relative to the other genes to the contrary reported by Theriot *et al.* (2010). *Triparma* (= *Bolidomonas*) *pacifica* was also pulled inside the Coscinodiscophyceae. A search of the bootstrap trees reveals about 8% of the trees had a monophyletic Mediophyceae (Figure 2). One of the bootstrap replicates with the three clades (classes) was extracted from the BT analysis (Figure 2) and compared to the tree shown in Figure 1 using a SH-Test in PAUP (Table 4), which suggested that the tree with three clades corresponding to the CMB hypothesis was better but only marginally significantly different from the best tree found by the BT analysis.

The next analyses used evolutionary models determined for each gene partition and codon position for coding genes (CP), with sequentially added outgroups and is presented in Figures 3-6. The first analysis with only Bolidomonads as an outgroup (Figure 3) recovered three clades of Coscinodiscophyceae, monophyletic Mediophyceae and Bacillariophyceae, the latter of which consisted of three monophyletic clades:
basal araphids, core araphids, and raphids. Sequential addition of the other

outgroups: heterokonts, haptophytes, chlorophytes/prasinophytes, (Figures 4, 5, 6 respectively) had the same topology but examination of the BT/aLRT support revealed that with each outgroup added to the analysis, the support for the Mediophyceae grew stronger, reaching a maximum of 90/51 when all outgroups were included (Table 6). The support for the three clades of Coscinosdiscophyceae were more or less the same with increasing outgroups, except for clade 2, which slightly decreased. The addition of the outgroups did not change the topology of the ingroups. The three clades of Coscinodispohyceae always contained the same taxa: Clade 1 had *Corethron* and *Leptocylindrus*; Clade 2 had Melosiraceae and Stephanopyxidaceae; Clade 3 had all remaining radial centrics. The tree with all outgroups built with the CP (Figure 6) had higher bootstrap support for the individual clades (BT = 90-100) than those found in Theriot *et al.* (2015), which ranged from 28 to 81 for the centric clades and 97 for the pennate clade.

Because we wanted to test the monophyly of the three classes, we constrained the CP analyses with the tree shown in Figure 2, but with *Actinoptychus undulatus* inside the Coscinosdiscophyceae and sequentially added of outgroups with the same settings in IQ-Tree, and compared the trees obtained with a several tests within IQ-Tree and within PAUP (Tables 4, 5). The constrained trees with the sequential addition of the outgroups also recovered three clades of Coscinodiscophyceae, a monophyletic Mediophyceae and Bacillariophyceae, as in Figures 3-6 (trees not shown). In these analyses, the topology of the clades did not change with the addition of the increasingly distant outgroup. When these trees were compared to that in Figure 1b using the SH test in PAUP, it was found that they were not significantly different in normal SH tests but were in weighted SH tests (Table 4). As the various outgroups were added to the constrained analysis, the difference in the ln-L decreased from 176 with only bolidomonads to 122 with all heterokonts and haptophytes. When the chlorophytes/prasinophytes were added as outgroups, the ln-L was reduced to 23 and the constrained CMB tree was better. This continued reduction in the difference in the log-likelihood ratio as more outgroups were added, can be interpreted as increased support for the monophyletic classes. In the final analysis with the maximum number of outgroups, the tree with the three monophyletic clades was significantly better than the CP analysis in PAUP.

In IQ-Tree (Table 5), the partitioned analysis selected the best evolutionary model for each gene partition and determined the best codon model for the seven gene dataset. The analysis was constrained by a tree reflecting the CMB hypothesis. In Table 5, the results from the various tests run in IQ-Tree are shown. Of the tests computed by IQ-Tree, the AU test is considered the best replacement for the SH test (Shimodaira, 2002; http://www.iqtree.org/doc/Advanced-Tutorial). In all comparisons, the CP tree was better than the constrained tree and the significance does not seem to have any relationship with the number of outgroups. The log–L difference is the greatest when the green plastid genes (a different primary endosymbiosis than the red algal plastid) and least when only heterokonts were used as outgroups. The most significant difference was obtained when only the bolidomonads were used as outgroups, indicating that the addition of multiple outgroups reduced the significant difference between the constrained CMB tree and the tree based on evolutionary models. From this trend it could be predicted that by adding more outgroups the significance would be reversed, albeit further outgroups should only be added from the red plastid lineage because the codon model analysis is greatly affected by the addition of the green plastid genes.

Morphological Analysis: We coded the morphological data in Table 1 as seven characters. These seven characters were coded in two ways (Table 2). First, each character was coded as a different nucleotide (CATG). This coding was used in the ML analysis with the NCP restrictions. We coded the morphological data as numbers (1234) for the BI analysis in the CP analysis. We repeated the motif 230 times because that placed the morphological sequence just slightly longer than the SSU rRNA gene and gradually reduced the motif until the phylogeny changed, when we assumed that the gene sequence data signal was stronger than morphological data.

In coding the morphological data as nucleotides with the NCP analysis, we recovered the CMB hypothesis (Fig. 7). Coding the nucleotides as numbers with the CP analysis with 230 repetitions of the seven-character motif also produced three clades but they did not correspond to the CMB hypothesis (Figure 8). So strong is the signal for sexual reproduction in the centrics that the radial and the bipolar centrics were sister groups to the pennates in the traditional sense. Reducing the repeats of the motif continued to recover the traditional sense of diatom phylogeny until only 31 repeats of the motif were used. At this point, the bipolar centrics moved their position as sister to radial centrics to be sister to the pennates as has been found in all molecular analysis since Medlin *et al.* (1993), but the pennates arose from within the bipolar centrics (Figure 9). Continued reduction of the character motif removed the monophyly of the radial centrics and they became a grade of clades (data not shown) as seen in Figures 3-.6 Thus, at 31 repeats of the character motif, we reasoned that the weighting of the morphological data balanced the information of the molecular data in the CP analysis. At this point the Coscinodiscophyceae are monophyletic and the Mediophyceae have the pennates arising from within them, making them a grade clades of bipolar centrics and the last clade that diverges before the pennates diverge sister to a clade containing most of the bipolar centrics is the clade containing *Toxarium, Ardissonia* and *Climacophenia* (Figs. 9,10).

We took the nexus file from Theriot *et al.* (2015), pruned the taxa with more than one gene missing and kept those taxa shown in Table S1, reanalyzed it in Mesquite and recovered a tree with a structural grade of taxa with three clades of both Mediophyceae and Coscinodiscophyceae (Figure 11) just as Theriot *et al.* (2015) did. SH tests were made comparing this pruned tree from Theriot *et al.* (2015) to trees in Figures 7-10. The NCP tree that reflected the CMB hypothesis was the better tree (Fig. 7), but it was not significantly different using classical SH but was in weighted SH tests in PAUP (Tables 3). The final CP with the minimum number of repeat motifs (Fig. 8) was also the better tree also but it was not significantly different from the ET tree in PAUP in either test. In IQ-Tree, the ET tree was better than the NCP tree but it was not significantly different. For the CP analysis, the ET tree was significantly different with a very large log-L difference.

Discussion

Modern genomic approaches are now opening the possibility of utilizing a vast number of genes to possibly recover a more robust hypothesis of phylogenetic relationships. The question, however, is which gene compartment(s) might be expected to provide a tractable result. It is the purpose of this paper to bring together these data to update the reviews by Sims *et al.* (2006), Medlin (2016) and Mock & Medlin (2012) and to add analyses based on multiple genes with multiple outgroups and morphological data to examine which trees show concurrent data and which do not.

The diatoms are one of the most successful microalgal groups in both aquatic and terrestrial habitats. Their complex bipartite siliceous cell walls (valves and girdle bands) are unique among the algae. The pattern of cell size reduction in one of the daughter cells following mitosis is also unique and results in a population of cells of smaller sizes that, normally, can only be restored to the cell's maximum size following sexual reproduction (see reviews in Mann & Marchant, 1989; Kaczmarska et al., 2013). Since the 19th century, diatom classification has been based on the intricate designs of their cell walls (for a review of the history of classification see Williams, 2007). The diatoms (Bacillariophyta) have more 10,000 described species and potentially many more cryptic species (Mann, 1999). There are likely at least 30,000 to 100,000 species (Mann & Vanormelingen 2013).

Since the early 1990s, much work has been directed towards understanding diatom classification using molecular tools. In 2006, Sims *et al.* provided a review of the evolution of the group as inferred from molecules, morphology and the fossil record. Mock & Medlin (2012) reviewed the evolution of the group from its origins to its genes. Medlin *et al.* (2007a) commented that where paraphyletic lineages have remained after molecular investigations, investigators are either willing to live with non-monophyletic taxa, not able to find new characters to define the new monophyletic groups, or unwilling to go against conventional wisdom that would lead to the demise of long-standing taxa. Since these two reviews, more molecular data from multiple genes, more information on sexual reproduction and better congruence of molecular clades with morphological features have appeared but paraphyletic lineages continue to appear and authors either describe new taxa or ignore it, e. g., *Hippondonta* arises from within *Navicula* (Ashworth *et al.* 2016, Kulikovsky *et al.* 2019), Mastogloiales is not monophyletic (Ashworth *et al.* 2016), *Pierrecomperia* arises from within *Extubocellulus*, *Campylosira* arises from within *Cymatosira* (Dabek *et al.* 2019), *Epithemia* and *Tetralunata* arising from within *Rhoplaodia*, *Campylodiscus, Cymatopleura, Stenopterobia* and *Petrodictyon* arises from within *Surirella* (Ruck *et al.* 2016).

In all of the analyses by Medlin et al., multiple outgroups have been used (Table S2). Where a single outgroup was used (Medlin and Kazcmarska 2004, fig. 3), a grade of clades occurred, which is useful to show the branching order of the taxa to ask specific evolutionary questions, such as what is the last bipolar clade to evolve before pennates. In none of the studies by Theriot *et al.* have they used multiple outgroups outside of one study with multiple heterokonts. When questioned about their reluctance to do this, they have replied that multiple outgroups will only increase long-branch attraction. This is true if only one representative of each outgroup is used but is not the case when multiple representatives of each outgroup are used. In fact, the common advice given to break up long-branch attraction is to add a close relative to break the branch. In our analyses we have used a minimum of four species in each outgroup taxon so that the possibility of long-branch attraction is kept to a minimum. We found in an earlier analysis with multiple outgroups, that the omission of a single gene in the data set produced that taxon on a long branch (Figure S1). Thus, our analysis only included those taxa with a full complement of the seven genes. Also the inclusion of distant outgroups should not disrupt the topology of the ingroup (Ackermann *et al.* 2014). In none of our analysis, did the topology of the ingroup change when more distant outgroups were added. The fact that they did not rearrange the ingroup means that they were not too distant from the ingroup and thus were appropriate for recovering the phylogeny of the diatoms. Future work could be directed to complete the seven gene complement for those taxa in the Theriot *et al.* dataset missing one or more of the plastid genes or to add more outgroups.

Despite this absence of testing of multiple outgroups by the Theriot group, they conclude from their analyses that it is no more or less plausible that there are three clades (Classes) of diatoms (radial centrics, polar centrics plus Thalassiosirales, pennates with the latter two forming a larger monophyletic group) than it is that radial centrics grade into polar centric which then grade into pennates, with Thalassiosirales in the radial grade. They could not determine if the CMB or the SGH was correct.

Theriot *et al.* (2015) found that none of the positions in the codons of the seven genes were saturated so applying codon evolutionary models may not be required. Our NCP analysis is different from the CP analysis in that in the former the Coscinodiscophyceae is monophyletic and in the latter, the Mediophyceae is monophyletic. Clearly applying codon partitioning to the dataset and applying individual models of evolution to each gene, which also consider the base position within each codon is affecting the monophyly of the radial centrics. Our NCP ML analysis (Figure 1) also recovered three classes reflecting the CMB hypothesis (Figure 2) in 8% of the bootstrap trees.

Those trees are not the best tree obtained by the analysis but they are not statistically different from it even though the best trees have a lower log-likelihood ratio. The CP analysis recovers a monophyletic Mediophyceae and a grade of clades in the Coscinodiscophyceae (Figures 3-6).

The difference between the results of the NCP and the CP analysis may be a reflection of the difference in the plastid inheritance in the diatoms, which is certainly not homogenous. This may also likely be the cause of the various resolutions found in the individual plastid trees (Figures S2-6). There are at least three patterns of plastid inheritance in the diatoms: 1) Mereogenous (predominately found in the radial centrics) where all plastids are removed from the sperm during meiosis so inheritance is only maternal: 2) Hologenous (found in the bipolar centrics with one known exception at the genus level) where plastids are retained by the sperm and where the offspring should be a mixture of maternal and paternal plastids assuming no segregative mitoses and in polyphasic plastids, the contribution of the maternal plastid should be greater, and 3) that found in the pennates, with isogamous gametes where there can be a mixture of all maternal, all paternal or both, termed unique, dual or stochastic by Mann (1996). In Table 6 we have reproduced the plastid inheritance table from Jensen *et al.* (2003), correcting some mistakes they made in that paper and adding data from *Corethron* (Crawford 1995). Among the merogenous radial centric diatoms, some species do not loose their plastids during meiosis but do so before the sperm enters the cells. These species are marked with arrows (H➔M). This would make virtually all radial centric plastids maternally inherited with no option of recombination. Notably the two exceptions to this from taxa whose sexual reproduction is noted in from *Corethron* and *Leptocylindrus*, which are the first two divergences in the three clades of radial centrics in Parks *et al.* (2017). Clearly, if the inheritance of the plastid genome is not uniform across the centric diatoms, then this could account for the differences in the NCP and CP trees. The fact that the Coscinodiscophyceae are monophyletic in the NCP analysis suggests that this group is likely the most non-homogeneous plastid gene group (Table 6) and applying different models of evolution for genes that have different modes of inheritance across the radial centrics, likely causes this group to become grade of clades in the CP analysis.

Chepurnov *et al.* (2002) suggested from their studies of *Semiavis* that in biparentally inherited plastids, the plastids are segregated after the initial cell starts to divide so there should be no heterozygous plastids. There is no way to tell morphologically which plastids are maternal or which are paternal. Only different genotypes in plastid genes can be used to trace the genealogy. Ardoor (2017) showed in *Semiavis* there were heterozygous plastids based on *rbc*L genotypes. Ghiron *et al.* (2008) in their study of plastic inheritance in *Pseudo-nitzschia delicatissima* showed that 16 out of 96 strains raised each from single F(1) cells had retained two paternal (PNd(+)) plastids, 20 had two maternal (PNd(-)) plastids and the remaining 60 had one maternal and one paternal plastid. So either two plastids are eliminated stochastically during auxospore development as suggested for *P. delicatissima* by Amato *et al.* (2005), or all survive into the initial cell and then segregate two by two in the first mitotic division. D'Alelio and Ruggerio (2015) also showed that biparental plastids can undergo recombination in *Pseudo-nitzschia*. Crosby and Smith (2012) tested if the mode of plastid inheritance affected genome architecture and found that paternally inherited plastids were more compact.

Thus, the evolutionary pathways of the diatom plastid are not homogeneous. This evolutionary pathway is even more complex in that many of the genes in the diatom plastid can trace their origin to a green endosymbiont rather than a red one. A number of studies have shown that diatoms and other chromalveolates contain nuclear genes of green algal origin that together with those of red algal provenance comprise a chimeric plastid proteome in these taxa (Mustafa *et al.* 2005, Chan *et al.* 2011). In the latter paper, a comparison of membrane transporters in two diatoms showed that 24% of these genes showed non-lineal descent. Either of these facts could account for the differences in the individual plastid phylogenies or the concatenated ones being non congruous and why the NCP tree appears in some tests to be the significant tree. Certainly in the IQ-Tree significance tests in the CP analysis, the addition of the green plastid genes had the largest log–L difference and lowest p–value.

Yu *et al.* (2018) extracted 103 genes from 40 diatom plastid genomes with using only one Bolidomonad as the outgroup, they recovered grades of clades, concluding that two of the three classes of diatoms (Coscinodiscophyceae and Mediophyceae) were not monophyletic. In their study the first two clades of the Coscinodiscophyceae are represented by single taxa and of these *Proboscia* (clade 2) is on a long branch because it has multiple gene losses and and *Leptocylindrus* (clade 1) is also on a long branch likely because it has the largest single copy gene region and the smallest inverted repeats of all of the radial centrics. With a secondary structure analysis of the SSU gene, *Proboscia* falls inside the Mediophyceae (Medlin *et al.* in press). Yu *et al.* recover two clades of Mediophyceae and the last clade before the pennates is that of *Attheya* + *Bidulphia* as in our NCP analysis. The placement of this clade as the last centric one before the pennates has merit in that the male sex cells of *Attheya* may possess the special filament found in other araphid diatoms (Roschin pers. comm.). The majority of bipolar centrics + Thalassiosirales were in one clade and the bipolar taxa had the smallest genome size among the Mediophyceae. Could this be a reflection of paternal plastid inheritance as suggest by Crosby and Smith (2010)? Their analysis also has an araphid taxon (*Plagiogrammopsis vanhuerckii*) in the middle of the bipolar centrics but they do not comment on this irregularity at all. They also discounted the possibility of recombination in the plastid genome, but recombination can only occur if the plastid is biparentially inherited, which is not the case in most of the Coscinodiscophyceae and comparison of the plastid genome should concentrate on those species whose plastid inheritance is well documented. Recombination of the plastid genome is more likely to happen in the pennates because they have fewer plastids. It is unclear how this would occur in the hologeneous radial and even in bipolar centrics whose eggs have multiple plastids with only one sperm fertilizing the egg with more than one plastid.

Parks *et al.* (2017) compared 94 diatom plastid genomes using an amino acid alignment with four heterokont plastids as outgroups and recovered three clades of Coscinodiscophyte, a monophyletic Mediophyceae + *Attheya* and a monophyletic Bacillariophyceae, which is very similar to our CP analysis. They suggested that incomplete lineage sorting disproportionately affects species tree inference at short internodes, such as those separating the nodes of the Coscinodiscophyceae. Incomplete lineage sorting was also invoked as a possible explanation for the radial Thalassioairales being included in the Mediophyceae or bipolar centrics (Medlin 2016a). In Medlin (2014), the addition of only heterokont outgroups recovered almost identical results using only the SSU genes: four clades of Coscinodiscophyceae, a monophyletic Mediophyceae and Bacillariophyceae.

Our total evidence analysis also produced some interesting results. NCP analysis with the morphological data coded as CATG recovered the CMB phylogeny using a 230 times repeat of the morphological motif. CP analysis produced something different. Weighting of the morphological characters 230 times coupled with evolutionary models for each gene created an artefact in that oogamy found in both the radial and bipolar centrics linked them together as sister groups to the exclusion of the pennates in the traditional sense of their relationships: centrics and pennates. Reducing this to a 31 times repeat kept the radial centrics monophyletic and placed the pennates arising from within the Mediophyceae as with most molecular analyses done by the Theriot *et al.* group have recovered.

Lastly, the diatom systematics in the revised version of eukaryotic classification by D.G. Mann in Adl *et al.* (2019), he creates a different classification system by raising every order of radial centrics to its own sub-phylum. This revision is not supported by any of the molecular trees. (Table S2). The revised classification presented by D.G. Mann does, however, recognize the Mediophyceae as a monophyletic class.

Conclusions

Because plastid inheritance in the diatoms is not homologous (Table 6, Mann 1996), the pattern of evolution in each variation is different and therefore the application of codon partition models for the plastid genes could over-parameterize the data. It might be advantageous to investigate more nuclear genes and with the push to add about 100 diatom genomes (T. Mock, pers. comm.), these genes would become available and more heterotrophic organisms could be added as outgroups, which were important in recovering the monophyletic clades in Medlin (2014). Because of the uncertainty regarding linear plastid inheritance for several genes, the inclusion of the SSU gene and possibly the LSU gene would seem to be a pre-requisite for recovering a robust analysis in contrast to the opinion of Theriot et al (2009) that these genes cannot be used.

With additional outgroups in this plastid dataset, the ln-L decreases between the constrained tree and the NCP tree, which suggests that adding even more outgroups could push the significance in favor of the constrained tree. Because the topology of the ingroups does not change with the addition of these distant outgroups in the NCP analysis, more outgroups could be added. However with the CP analysis, only red plastid gene outgroups should be added because this analysis was very sensitive to the addition of the green plastid outgroups to the analysis, pushing the log-L difference to its highest.

The addition of the morphological data supported the CMB phylogeny but only in the NCP analysis. This may come from overparametrization using CP with morphological data. It has also been shown that different partitioning schemes sometimes lead to very different clade supports (Kainer and Lanfear, 2015). De Quieroz et al. (1995) suggested that if the data sets are heterogenous (in our case different plastid inheritance) then the phylogenies obtained would be compromised.

In the CP analysis, the radial centrics were monophyletic, the bipolar ones a grade of clades with the pennates arising from within them as the last divergence. In PAUP, the addition of morphological data was significantly different from an analysis (ET tree) with no morphological analysis. In IQ-Tree, the ET tree was the better tree and this tree was significantly better when the signal from the morphological data repeat was at a minimum. The task ahead of us is to identify plastid inheritance where possible to determine which are homologous lineages and possibly devise some way to partition paternal, maternal and heterozygous plastid inheritance. Alternatively, with the addition of more whole genome analyses of the diatoms, perhaps more heterotrophic taxa can be added to the outgroup selection. Adding more outgroup plastids outside the heterokont taxa and a total evidence aspect to the data set by coding the morphological features identified in Table 1 has supported the CMB hypothesis in the NCP analyses. Failure to recover the CMB hypothesis in the CP analyses with the morphological data was not significantly different. The evidence presented here suggests that the CMB hypothesis by Medlin and Kaczmarska (2004) is different from an analysis performed with codon partitioning and is different from the trees in Theriot *et al.* (2015), which is likely a result of adding the V4 region, the multiple outgroups and variation in plastid inheritance, which has rendered the grade of clades in the radial centrics.

Literature Cited

Ackerman, M., Brown, D., Loker, D. 2014. Effects of rooting via outgroups on ingroup topology in phylogeny. *International Journal of Bioinformatics and Research Applications* 10:426-46. doi:10.1504/IJBRA.2014.062993.

Adl, S.M., Bass, D., Lane, C.E., Massana, R., Lukeš, J., Schoch, C., Smirnov, A., Agatha, S., Berney, C., Brown, M.W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L, del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., Heiss, A.A., Hoppenrath, M., James, T.Y., Karnkowska, A., Karpov, S.A., Kim, E., Kolisko, M., Kudryavtsev, A., Lahr, Daniel J.G., Lara, E., Le Gall, L. Lynn, D.H., Mann, D.G., Mitchell, E.A.D., Morrow, C., Soo P.J., Pawlowski, J., Powell, M.J., Richter, D.J., Rueckert, S., Shadwick, L., Shimano, S., Spiegel, F.W., Torruella, G., Youssef, N., Zlatogursky, V., Zhang, Q. 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology* 66:4–119.

Amato, A., Orsini, L., D'Alelio, D., Montresor, M. 2005. Life cycle, size reduction patterns, and ultrastructure of the pennate planktonic diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae). *Journal of Phycology* 41:542-556.

Anisimova, M. & Gascuel, O. 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology* 55:539-552.

Ardoor, S. 2017. Characterisation of reproductive behaviour and plastid inheritance in pennate diatoms using *a Seminavis robusta* mapping population. PhD Thesis. University of Ghent. 44 pp.

Armbrust, E.V. 2009. The life of diatoms in the world's oceans. *Nature* 459. doi,10.1033/*Nature*08057.

Ashworth, M. P., Lobban, C. S., Witkowski, A., Theriot, E. C., Sabir, M.J., Baeshen, M.N., Hajarah, N. H., Baeshen, N. A., Sabir, J. S. & Jansen, R. K. 2016. Molecular and morphological investigations of the stauros-bearing, raphid pennate diatoms (Bacillariophyceae): *Craspedostauros* E.J. Cox, and *Staurotropis* T.B.B. Paddock, and their relationship to the rest of the Mastogloiales. *Protist* 168:48–70.

Ashworth, A., Ruck, E., Lobban, C., Romanovicz, R., Theriot, E. C. 2012. Revision of the genus *Cyclophora* and description of *Astrosyne* gen. nov. (Bacillariophyta), two genera with the pyrenoids contained within pseudosepta. *Phycologia* 51:684–699.

Ashworth, M. P., Nako, T., Theriot, E. C. 2013. Revisiting Ross and Sims 1971. Toward a molecular phylogeny of the Biddulphiaceae and Eupodiscaceae (Bacillariophyceae). *Journal of Phycology* 49:1207–1222.

Ashworth, M. P., Ruck, E., Lobban, C. S., Romanovicz, D. K., & Theriot, E. C. 2012. A revision of the genus *Cyclophora* and description of *Astrosyne* gen. nov. (Bacillariophyta), two genera with the pyrenoids contained within pseudosepta. *Phycologia* 51:684–699.

Chan, C. X., Reyes-Prieto, A. & Bhattacharya, D. 2011. Red and green algal origin of diatom membrane transporters, insights into environmental adaptation and cell evolution. *PLoS ONE* 6, e29138. doi,10.1371/journal.pone.0029138

Chepurnov, V. A., Mann, D. G., Vyverman, W., Sabbe, K. & Danielidis, D.B. 2002. Sexual reproduction, mating system, and protoplast dynamics of *Seminavis* (Bacillariophyceae). *Journal of Phycology* 38:1004-1019

Crawford, R.M. 1995. The role of sex in the sedimentation of a marine diatom bloom. *Limnology and Oceanography*. doi.org/10.4319/lo.1995.40.1.0200

Crosby, K. & Smith, D.R. 2012. Does the mode of plastid inheritance influence plastid genome architecture? *PLoS ONE* 7, e46260.

D'Alelio D. & Ruggiero, M.V. 2015. Interspecific plastidial recombination in the diatom genus *Pseudo-nitzschia*. *Journal of Phycology* 51:1024–1028.

Dąbek, P., Ashworth, M.P., Górecka, E., Krzywda, M., Bornman, T.G., Sato, S. & Witkowski, A. 2019. Toward a multigene phylogeny of the Cymatosiraceae (Bacillariophyta, Mediophyceae) II: Morphological and molecular insights into the taxonomy of the forgotten species *Campylosira africana* and of *Extubocellulus*, with a description of two new taxa. *Journal of Phycology* 55:425-441. doi:10.1111/jpy.12831.

De Queiroz, A. Donoghue, M.J., & Kim, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics*. 26:657-681.

Ehara, M., Inagaki, Y., Watanabe, K. I. & Ohama, T. 2000. Phylogenetic analysis of diatom coxI genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. *Current Genetics* 37:29–33.

Finkel Z. V., Katz, M. E., Wright, J. D., Schofield, O. M. E., Falkowski, P. G. 2005. Climatically driven macro-evolutionary patterns in the size of marine diatoms over the Cenozoic. *Proceedings of the National Academy of Science* 102:8927-8932.

Finkel Z.V. & Kotrc B. 2010. Silica use through time, macroevolutionary change in the morphology of the diatom frustule. *Geomicrobiology Journal* 27:596–608.

Ghiron, J. Amato, A., Montresor, M. & Kooistra, W.H.C.F. Plastid inheritance in the planktonic raphid pennate diatom *Pseudo-nitzschia delicatissima* (Bacillariophyceae), *Protist* 2008, 159:91-98.

Harwood D.M. & Gersonde R. 1990. Lower Cretaceous diatoms from ODP Leg 113 Site 693 (Weddell Sea) part 2, resting spores, chrysophycean cysts, and endoskeletal dinoflagellates, and notes on the origins of diatoms. *Proceedings of the Ocean Drilling Program, Scientific Results* 113:403–425.

Hasle, G. R., Medlin, L. K. &Syvertsen, E. E. 1994. *Synedropsis* gen. nov. a genus of araphid diatoms associated with sea ice. *Phycologia* 33:48-270.

Jensen, K. G., Moestrup, O. & Schmid, A. M. 2003. Ultrastructure of the male gametes from two centric diatoms, *Chaetoceros laciniosus* and *Coscinodiscus wailesii* (Bacillariophyceae). *Phycologia* 42:98- 105.

Kaczmarska I. & Ehrman J. M. 2015. Auxosporulation in *Paralia guyana* MacGillivary (Bacillariophyta) and possible new insights into the habit of the earliest diatoms. *PLoS ONE* 10, e0141150. doi, 10.1371/journal. pone.0141150.

Kaczmarska, I., Poulíčková, A., Sato, S., Edlund, M.B., Idei, M., Watanabe, T., & Mann, D.G. 2013. Proposals for a terminology for diatom sexual reproduction, auxospores and resting stages. *Diatom Research* 28:1–32.

Kainer, D. & Lanfear, R. 2015. The effects of partitioning on phylogenetic inference. *Molecular Biology and Evolution* 32:1611-1627.

Kooistra, W.H.C.F. & Medlin, L.K. 1996. Evolution of the diatoms (Bacillariophyta), IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. *Molecular Phylogenetics and Evolution* 6:391–407.

Kulikovskiy, M.S., Maltsev, Ye.I., Andreeva, S.A., Glushchenko, A.M., Gusev, E.S., Podunay, Yu. A., Ludwig, T.V., Tusset, E. & Kociolek, J.P. 2019. Description of a new diatom genus *Dorofeyukea* gen. nov. with remarks on phylogeny of the family Stauroneidaceae. *Journal of Phycology* 55:173–185.

Li, C., Ashworth, M.P., Witkowski, A., Dąbek, P., Medlin, L. K., Kooistra, W.H.C.F., Sato, S., Zgłobicka, I., Kurzydłowski, K.J., Theriot, E.C., Sabir, J.S.M., Khiyami, M.A., Mutwakil, M.H.Z., Sabir, M.H., Alharbi, N.S., Hajara, H.N.H., Qing, S. & Jansen, R.K. 2015. New insights into Plagiogrammaceae (Bacillariophyta) based on multigene phylogenies and morphological characteristics with the description of a new genus and three new species. *PLoS ONE* 10, e0139300.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Kumar, Y., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüßmann, R., May. M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Arndt Bode, A. & Schleifer, K-H. 2004. ARB, a software environment for sequence data. *Nucleic Acids Research* 32:1363–1371.

Maddison, W.P. & Maddison, D.R. 2017. Mesquite, a modular system for evolutionary analysis. Version 3.2. http://mesquiteproject.org.

Mann, D.G. 1996. Chloroplast morphology, movements, and inheritance in diatoms. In: *Cytology, genetics and molecular biology of algae*. (Ed. by B.R. Chaudhary & S.B. Agrawal), pp. 249-274, SPB Academic Publishing, Amsterdam, Netherlands,

Mann, D.G. 1999. The species concept in diatoms. *Phycologia* 38:437–495.

Mann, D.G. & Marchant, H.J. 1989. The origin of the diatom and its life cycle. In: *The Chromophyte Algae, Problems and Perspectives.* (Ed. by J. C. Green, B.S.C. Leadbeater, & W.L Diver), pp. 307–323, Clarendon Press, Oxford.

Mann. D.G. & Vanormelingen, P. 2013. An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology* 60:414–420.

Medlin, L.K. 2010. Pursuit of a natural classification of diatoms, an incorrect comparison of published data. *European Journal of Phycology* 45:155–166.

Medlin, L.K. 2014. Evolution of the diatoms, VIII. Reexamination of the SSU-rRNA gene using multiple outgroups and a cladistic analysis of valve features.

Journal of Biodiversity, Bioprocessing and Development 1:129. doi, 10.4172/2376-0214.1000129.

Medlin, L.K. 2016a. Coalescent models explain deep diatom divergences and argue for acceptance of paraphyletic taxa and for a revised classification for araphid diatoms. *Nova Hedwigia* 102:107–123.

Medlin, L.K. 2016b. Evolution of the diatoms, major steps in their evolution and a review of the supporting molecular and morphological evidence. *Phycologia* 55:79

Medlin, L.K., Boonprakob, A., Lundholm, N. & Moestrup, Ø. On the morphology and phylogeny of the diatom species *Rhizosolenia setigera*: comparison of the type material to modern cultured strains and a taxonomic revision. *Nova Hedwigia***,** Special Volume, Festschrift, in press.

Medlin, L.K. & Desdevises, Y. Phylogeny of 'araphid' diatoms inferred from SSU and LSU rDNA, *rbc*L and *psb*A sequences. *Vie et Millieu* 65:129–154.

Medlin, L.K. & Kaczmarska, I. 2004. Evolution of the diatoms, V. Morphological and cytological support for the major clades and a taxonomic revision. *Phycologia* 43:245–270.

Medlin, L.K., Metfies, K., John, U. & Olsen, J. 2007. Algal molecular systematics, a review of the past and prospects for the future. In: *Unravelling the algae, the past, present and future of algal systematics*. (Ed. by J. Broadie, & J. Lewis) *Systematics Association Special Volume* 75, pp. 234-253.

Medlin, L.K., Sato, S., Mann, D.G. & Kooistra, W.C.H.F. 2008. Molecular evidence confirms sister relationship of *Ardissonea, Climacosphenia*, and *Toxarium* within the bipolar centric diatoms (Bacillariophyta, Mediophyceae), and cladistic analyses confirm that extremely elongated shape has arisen twice in the diatoms. *Journal of Phycology* 44:1340-1348.

Medlin, L.K., Williams, D.M. & Sims, P.A. 1993. The evolution of the diatoms (Bacillariophyta. I. Origin of the group and assessment of the monophyly of its major divisions. *European Journal of Phycology* 28:261–275.

Mock, T. & Medlin, L. K. 2012. Genomics and Genetics of Diatoms. In: *Genomic Insights into the Biology of Algae*. (Ed. by G. Piganeau), *Advances in Botanical Research Volume* 64, pp. 245–284, Academic Press, London.

Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K.U. & Bhattacharya, D. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324:1724–1726.

Nguyen, L-T., Schmidt, H.A., von Haeseler, A., & Minh, B.Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Molecular Biology and Evolution* 32:268-274. https://doi.org/ 10.1093/molbev/msu300

Parks, M.B., Wickett, N.J. & Alverson, A.J. 2017. Signal, uncertainty, and conflict in phylogenomic data fora diverse lineage of microbial eukaryotes (diatoms, Bacillariophyta., *Molecular Biology and Evolution* doi,10.1093/molbev/msx268.

Piganeau, G., Eyre-Walker, A., Grimsley, N. & Moreau, H. 2012. How and why DNA barcodes underestimate the diversity of microbial eukaryotes. *PLoS ONE* 7:10. 1371/annotation /c12aac06-71d2-4749-91de- 46c458e7a4eb.

Rabosky D.L. & Sorhannus U. 2009. Diversity dynamics of marine phytoplankton diatoms across the Cenozoic. *Nature* 457:183–186.

Rimet, F., Kermarrec, L., Bouchez, A., Hoffmann, L., Ector, L. & Medlin, L.K. 2011. Molecular phylogeny of the family Bacillariaceae based on 18S rDNA sequences, focus on freshwater *Nitzschia* of the *Lanceolatae* section. *Diatom Research* 26:1–20.

Ronquist, F., Teslenko, M., van der Mark, P., L. Ayres, D.L., Darling, A., Höhna, S., Large, B., Liu, L., Suchard, M.A. & Huelsenbeck, J.P. 2012. MrBayes 3.2, efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biol*ogy 61:539-542.

Round F.E., Crawford R.M. & Mann D.G. 1990. The Diatoms, Biology and Morphology of the Genera. Cambridge University Press, Cambridge, UK. 747 pp.

Ruck, E.C., Nakov, T., Alverson, A. J. & Theriot, E. C. 2016. Phylogeny, ecology, morphological evolution, and reclassification of the diatom orders Surirellales and Rhopalodiales., *Molecular Phylogenetics and Evolution* 103:155-171.

Sato, S. 2008. Phylogeny of araphid diatoms inferred from morphological and molecular data. PhD Dissertation. University of Bremen. http://elib. suub. uni-bremen. de/diss/docs /00011057.pdf.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology* 51:492-508.

Shimodaira, H. & Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* 16:1114-1116.

Sims, P.A., Mann, D.G. & Medlin, L.K. 2006. Evolution of the diatoms, insights from fossil biological and molecular data. *Phycologia* 45:361–402.

Smetacek V. 1999. Diatoms and the ocean carbon cycle. *Protist* 150:25–32.

Sorhannus, U. 1997. The origination time of diatoms, an analysis based on ribosomal RNA data. *Micropaleontology* 43:215–218.

Sorhannus, U. 2007. A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Marine Micropaleontology* 65:1–12.

Stamatakis, A., Hoover, P. & Rougemont, J. A 2008. Rapid Bootstrap Algorithm for the RAxML Web-Servers. *Systematic Biology* 75:758-771.

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods. Version 4. Sinauer Associates, Sunderland, Massachusetts.

Theriot, E., Alverson, A. & Gutell, R. 2009. The limits of nuclear-encoded SSU rDNA for resolving the diatom phylogeny. *European Journal of Phycology* 44:277–290.

Theriot, E.C. Ruck, E. Ashworth, M. Nakov, T. & Jansen, R.K. 2011. Status of the pursuit of the diatom phylogeny, are traditional views and new molecular paradigms really that different? *In*: *The Diatom World*, (Ed. by J. Seckbach & P. Kociolek), pp. 119-144, CRC Publications, Boca Raton, Fl.

Theriot, E.C., Ashworth, M., Nakov, T., Ruck, E. & Jansen, R.K. 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., Ashworth, M., Ruck, E., Nakov, T. & Jansen, R.K. 2010. A preliminary multigene phylogeny of the diatoms. *Plant Ecology and Evolution* 143:278–296.

Vaulot, D., Eikrem, W., Viprey, M. & Moreau, H. 2007. The diversity of small eukaryotic phytoplankton in marine ecosystems. *FEMS Microbiology Re*view 32:795–820.

Williams D.M. 2007. Classification and diatom systematics, the past, the present and the future. In: *Unravelling the algae, the past, present and future of algal systematics*. (Ed. by *J.* Brodie & J. Lewis) CRC Press, Boca Raton, Florida, pp. 57–91.

Yamada, K., Yoshikawa, S., Ichinomiya, M., Kuwata, A., Kamiya, M. & Ohki, K. 2014. Effects of silicon-limitation on growth and morphology of *Triparma* *laevis* Nies-2565 (Parmales, Heterokontophyta). *PLoS ONE* 9, e103289. doi,10.1371/journal.pone. 0103289

Yu, M., Ashworth, M.P., Hajrah, N.H., Khiyami, M.A., Sabir, M.J., Alhebshi, A.M., Al-Malki, A.L., Sabir, J.S.M., Theriot, E.C. & Jansen, R.K., 2018. Evolution of the plastid genomes in diatoms. *Advances in Botanical Research* https://doi.org/10.1016 /bs.abr.2017.11.009.

Figure Legends

Figures 1-2. Phylogenetic reconstruction of the diatoms without coding for any codon positions or applying any models. 1. Best tree found in the bootstrap analysis, 2. Tree reflecting CMB hypothesis found in 8% of the bootstrap replicates.

Figures 3-6. Phylogenetic reconstructions using a ML analysis coding for each codon position and applying models of evolution for each gene. 3. only two Bolidomonads as outgroups. 4. Heterokonts and bolidomonads as outgroups. 5. Haptophytes, heterokonts and bolidomonads as outgroups. 6. Prasinophytes/chlorophytes, haptophytes, heterokonts and bolidomonads as outgroups. See Table 6 for bootstrap support for each of the major clades.

Figures 7-10. Phylogenetic reconstruction with morphological data added to the gene sequence data set. 7. NCP analysis with morphological data coded as nucleotides, 230 repeats, ML analysis. 8. CP data with morphological data coded as unordered numbers, BI analysis, 230 repeats. 9. CP data with morphological data coded as unordered numbers, BI analysis, 31 repeats. 10. Detail of the pennate divergence within the polar centrics

Figure 11. Phylogenetic reconstruction of the Theriot data set pruning those taxa missing one or more of the genes.

Table 1. Summary of the morphological features used in the total evidence analysis supporting the classification of the diatoms in Medlin & Kaczmarska (2004). NCP = the coding of the morphological data in this analysis and CP = the coding of the morphological data in that analysis. These data are extracted below for ease of interpretation.

Taxon	**Name**	**1. Sexual Reproduction**	**ncp**	**cp**	**2. Male sex cell**	**ncp**	**cp**	**3. Auxospore structure**	**ncp**	**cp**	**4. Structure in Annulus**	**ncp**	**cp**	**5. Position of cribrum in locualte areolae pseudolocuate excluded**	**ncp**	**cp**	**6. Golgi Postion**	**ncp**	**cp**	**7. Spore like nature of auxospore, i.e. heterovalvate and large dissimilarity between the vegetative and initial cell valve**	**ncp**	**cp**	**Exceptions to listed characters**
Class	Coscinodiscophyceae	oogamy	c	1	sperm	c	1	scales	c	1	none			extern	c	1	GERM[b]	c	1	Yes, where known	c	1	Golgi
Class	Mediophyceae	oogamy	c	1	sperm	c	1	Scales + properizonium bands	a	2	Yes, strutted or labiate process	a	2	intern	a	2	Peri-nuclear	a	2	partially	a	2	Auxospore and Golgi
Class	Bacillariophyceae	anisogamy or isogamy	a	2	Sperm with threads or no sperm	g	4	Scales + properizonium c perizonium banc or both	t	3	Yes, sternum	t	3	None found	t	3	Peri-nuclear	a	2	no	t	3	none
Sub class	Uneidiophycidae	anisogamy	t	3	Sperm wi filaments	a	2	Scales + properizonium AND perizonium bands	t	3	Yes, sternum	t	3	None found	t	3	Peri-nuclear	a	2	no	t	3	None Where known
Sub class	Fragilariophycidae	isogamy	g	4	No sperm	t	3	Scales + perizonium bands	g	4	Yes, sternum	t	3	None found	t	3	Peri-nuclear	a	2	no	t	3	None where known
Sub class	Bacillariophycidae	isogamy[a]	g	4	No sperm	t	3	Scales + perizonium bands	g	4	Yes, sternum + raphe	g	4	None found	t	3	Peri-nuclear	a	2	no	t	3	None where known

[a] can be physiological anisogamic

[b] Golgi/ Endoplasmic Reticulum/ Mitochondria Association

Table 2 Coding of the morphological data from table 1 to be used in the CP and NCP analyses

Taxon	**NCP coding**	**CP coding**
Coscinodiscophyceae	CCCCCCC	1111111
Mediophyceae	CCAAAAA	1122222
Uneidiophycidae	AATTTAT	2233323
Fragilariophycidae	TTGTTAT	3343323
Bacillariophycidae	TTGGTAT	3344323

Table 3. Comparison of the Theriot et al. (2015) data set with the current study in terms of nucleotides/gene and taxa.

	Theriot et al.	**This study**
Number of taxa	208	161
Number of outgroups	1	14
Number of nucleotides	9349	10575
SSU	1450	2068
*atp*B	1185	1297
*psa*A	1517	1627
*psa*B	1937	1933
*psb*A	853	920
*psb*C	1058	1484
*rbc*L	1352	1240

Table 4. Shimodaira-Hasegawa test results using RELL bootstrap (one-tailed test) and 10000 bootstrap replicates in PAUP.

Tree	**-ln L**	**Diff -ln L**	**SH**	**WT SH**	**Significance**
Fig. 1a vs Fig. 1b					
1a	479976.45099	179.57072	0.094		
1b	479796.88027	(best)			
Only Bolidomonads (CP vs Constrained)					
1	354588.14536	(best)			
2	354763.78655	175.64119	0.23	0.0000	P < 0.05
Heterokonts (CP vs Constrained)					
1	372307.35541	(best)			
2	372455.28637	147.93096	0.2337	0.0000	P < 0.05
Haptophytes (CP vs Constrained)					
1	391874.36905	(best)			
2	391996.97037	122.60132	0.2640	0.0000	P < 0.05
Chlorophytes/Prasinophytes (CP vs Constrained)					
2	416777.05492	(best)			
1	416804.68386	27.62894	0.2857	0.0000	P < 0.05
ET tree vs. Fig. 3a					
2	349082.13795	(best)			
1	349115.05346	32.91551	0.1315	0.0000*	P < 0.05
Fig. 3c vs. ET tree					
1	356926.10172	(best)			
2	359209.10474	2283.00302	0.7224	0.7224	

Table 5. IQ-tree test results of comparing trees under different analyses using 10000 RELL replicates. Those values with a (+) indicate no significance, whereas those with a (–) indicate significance at the 0.05 level and the tree is rejected.

Tree	**ln L**	**Diff –ln L**	**p-SH**	**p-WSH**	**p-AU**
all outgroups (Constrained vs. CP)					
1	-384716.358		1.0000+	1.0000+	1.0000+
2	-385214.014	497.656	0.0000-	0.0000-	0.0000-
Haptophytes (Constrained vs. CP)					
1	-342881.466		1.0000+	0.9483+	0.9518+
2	-342916.306	34.840	0.0517+	0.0517+	0.0482-
Heterokonts (Constrained vs. CP)					
1	-324859.448		1.0000+	0.9582+	0.9622+
2	-324890.836	31.388	0.0418-	0.0394-	0.0378-
only bolidomonads (Constrained vs. CP)					
1	-308673.146		1.0000+	0.9984+	0.9993+
2	-308728.365	55.219	0.0016-	0.0016-	0.0007–
ET vs. Fig. 3a					
1	-320657.9565	26.362	0.293+	0.293+	0.307+
2	-320631.5949		1.0000+	0.707+	0.693+
Fig. 3c vs ET					
1	- 310748.0897		1.0000+	1.0000+	0.998+
2	- 314468.9609	3720.9	0.000-	0.000-	0.00164-

Diff-L : log -L difference from the maximum log -L in the set.
p-SH : p-value of Shimodaira-Hasegawa test.
p-WSH : p-value of weighted SH test.

p-AU : p-value of approximately unbiased (AU) test

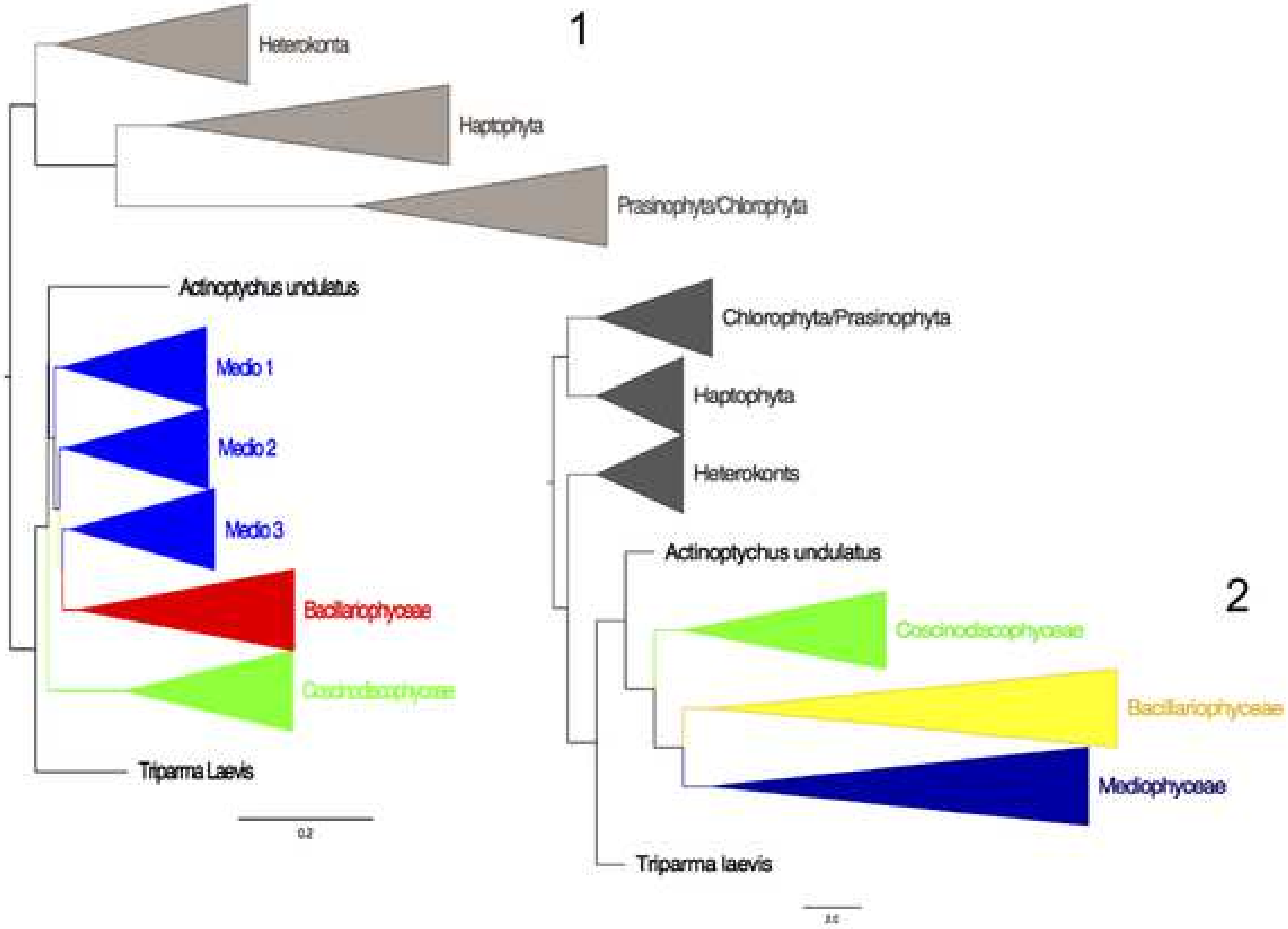
Table 6. Comparison of BT/aLRT in the ML CP analysis after sequentially adding outgroups and with all outgroups in the ML NCP analysis.

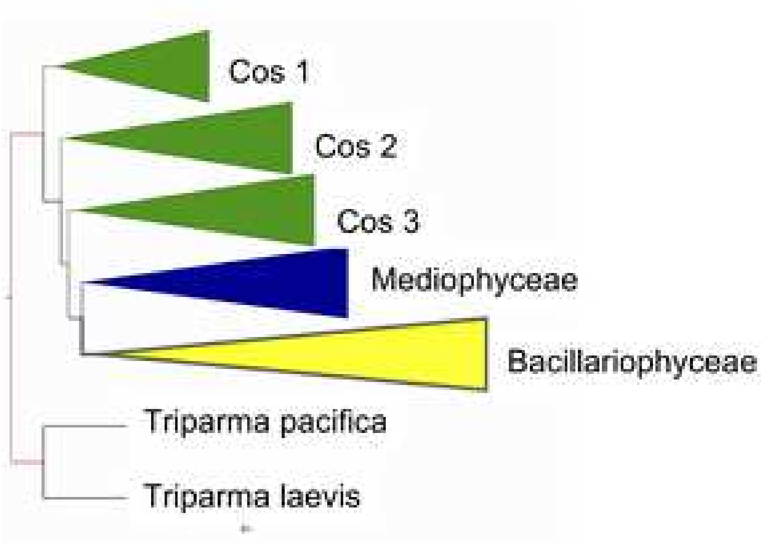
Clades as found in the CP analysis in Figure 2 and in the NCP anlaysis in Figure 1	**Only Bolidos**	**Only Heterokonts**	**Heterokonts + Haptophytes**	**Heterokonts + Haptophytes + Chlorophyceae /Prasinophyceae**	**No models No partitions**
Cos 1	94/99	95/98	94/99	92/98	
Cos 2	59/95	43/86	17/83	21/67	
Cos 3	98/100	99/100	99/99	98/99	
Mediophyceae	86/28	86/30	90/42	90/51	
Bacillariophyceae	100/100	100/100	100/100	100/100	
Coscinodiscophyceae					100
Medio 1					84
Medio 2					65
Medio 3					96
Bacillariophyceae					100

Table 6. Overview of the type of gametogenesis (hologenous (H) or merogenous (M)) of diatoms reported in the literature shown in Jensen et al. (2007) with errors corrected for the correct class (*). See Jensen et al. (2007) and Crawford (1995) for the original references for each species in the table. H→M* refers to taxa with hologenous gametogenesis but whose plastids degrade before fertilization making the plastid inheritance only maternally inherited or merogenous. The two taxa marked in a box are the early divergences in Parkes et al. (2016).

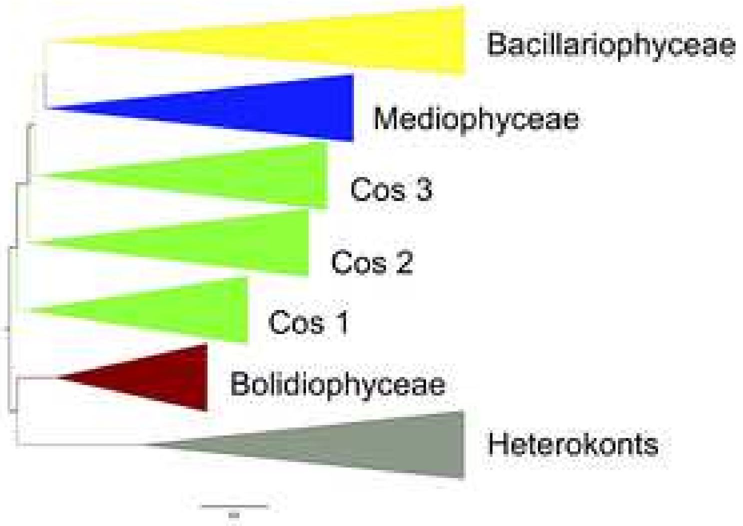
Taxon	**Type**
Coscinodiscophyceae	
Actinocyclus sp.	M
Coscinodiscus granii Gough	H→M*
Guinardia delicatula (Cleve) Hasle	M
Leptocylindrus danicus Cleve	H
Melosira moniliformis (O.F. Mull.) C. Ag.	M
Melosira moniliformis var. *octagol1a* (Grun.) Hust.	H→M*
Melosira varians C. Ag.	M
Rhizosolenia sp.	H
Stephanpyxis turris (Arnott in Gre) Ralfs in Prich.	M
Stephanopyxis palmeriana (Grev.) Grun.	M
Actinoptychus undulatus (Bailey) Ralfs in Pritchard *	M
Corethron pennatum (Grun.) Ost	H→M*
Mediophyceae	
Attheya decora T. West	H
Bacteriastrum hyalinum Laud.	H
Bellerochea malleus (Brightwell) V. H.	H
Chaetoceros spp.	H
Cyclotella meneghiniana Kütz.	H
Helicotheca tamensis (Shrub.) Ric.	H
Lithodesmium undulatum Ehr.	H
Odontella granulata (Rop.) R. Ross	M
Odontella mobiliensis (J.W. Bail.) Grun.	M
Odontella regia (Schultze) Sim.	H→M*

Odontella rhombus (Ehr) Kütz	M
Odontella sinensis (Grev,) Grun.	H→M*
Pleurosira laevis (Ehr.) Comp.	M
Skeletonema costatum (Grev.) Cleve	M
Thalassiosira lacustris (Grun.) Hasle in Hasle & Fryx.	H
Thalassiosira eccentrica (Ehr.) Cleve	M

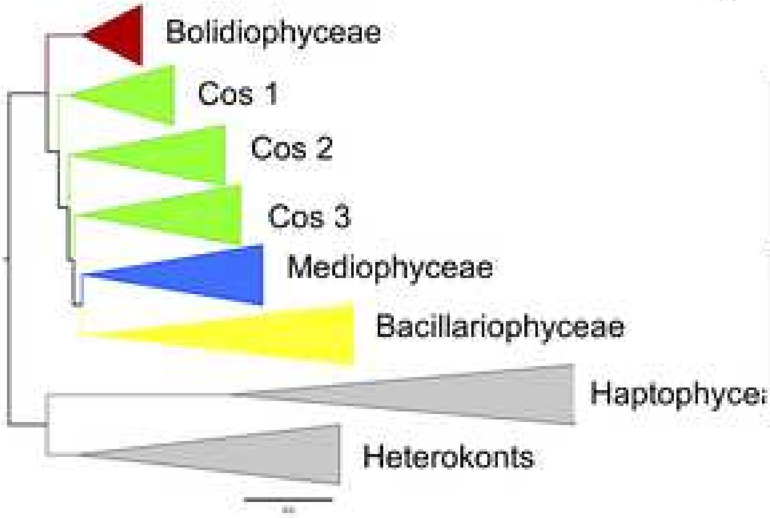

1
Heterokonta
Haptophyta
Prasinophyta/Chlorophyta
Actinoptychus undulatus
Medio 1
Medio 2
Medio 3
Bacillariophyceae
Coscinodiscophyceae
Triparma Laevis
0.2
Chlorophyta/Prasinophyta
Haptophyta
Heterokonts
Actinoptychus undulatus
2
Coscinodiscophyceae
Bacillariophyceae
Mediophyceae
Triparma laevis


Cos 1
Cos 2
Cos 3
Mediophyceae
Bacillariophyceae
Triparma pacifica
Triparma laevis

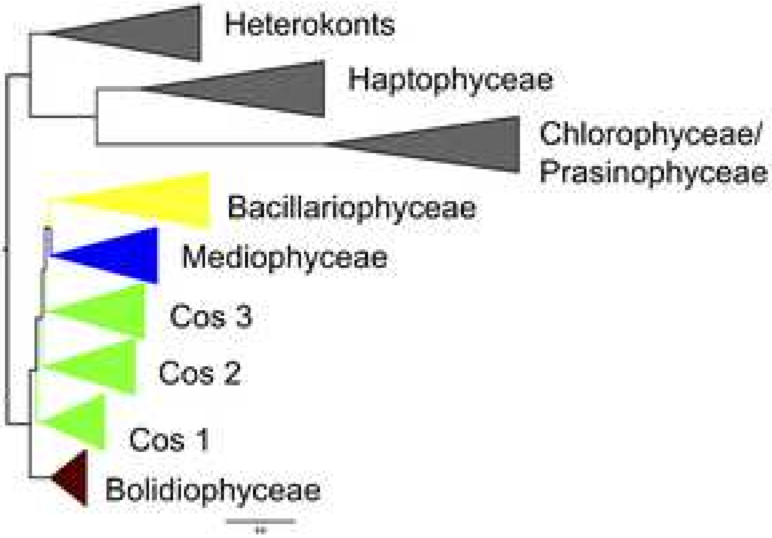
3


Bacillariophyceae
Mediophyceae
Cos 3
Cos 2
Cos 1
Bolidiophyceae
Heterokonts

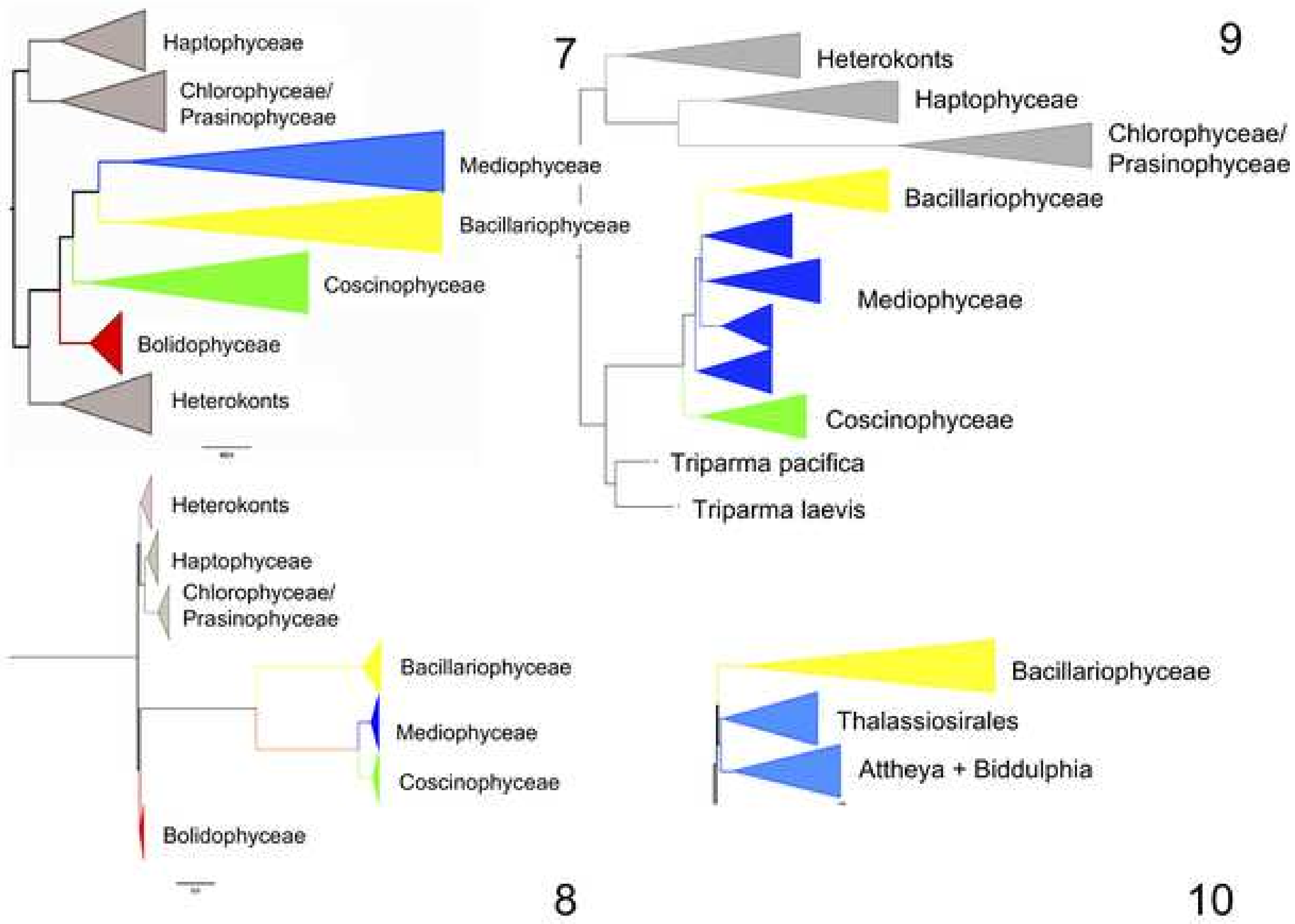
4

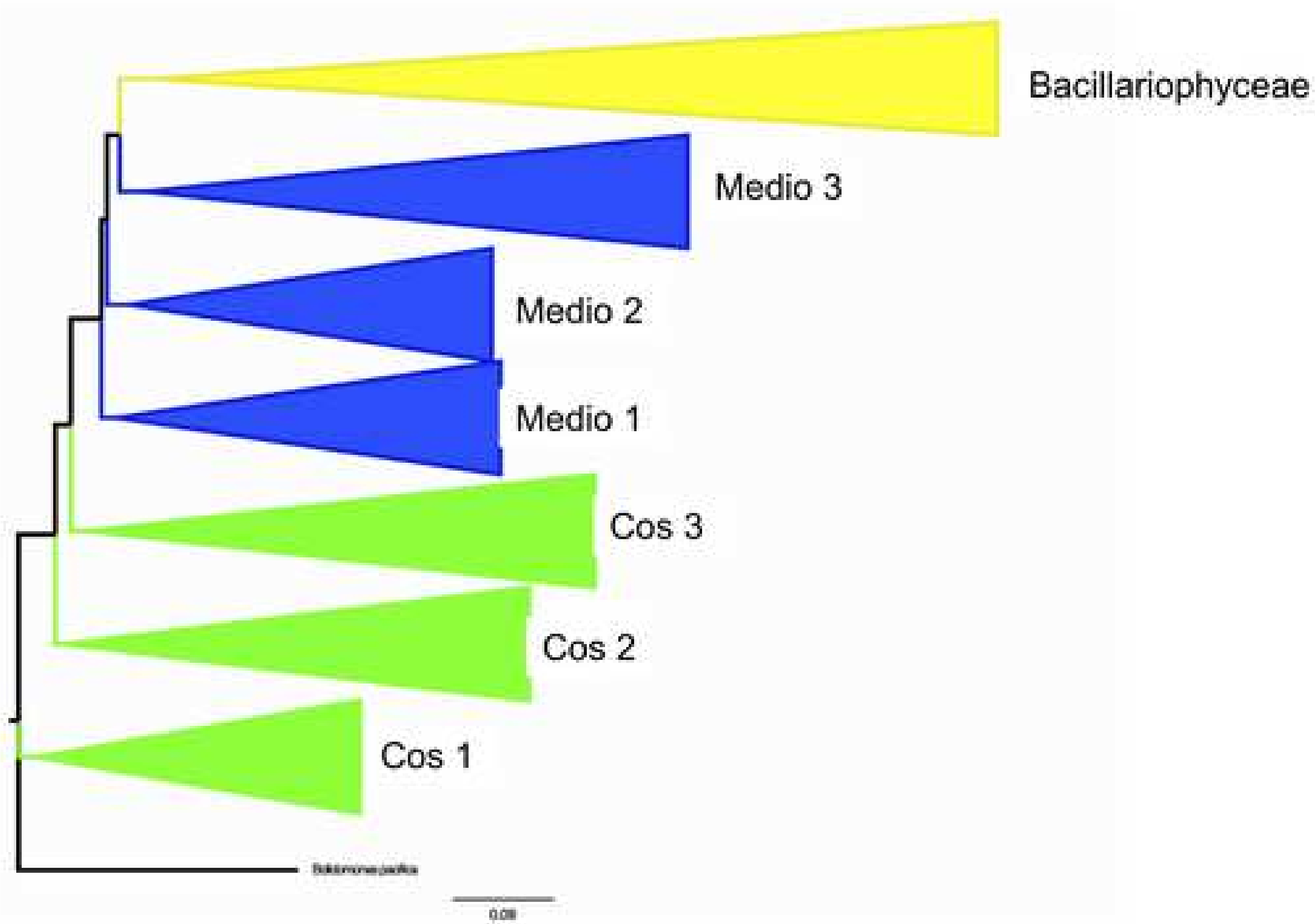

Bolidiophyceae
Cos 1
Cos 2
Cos 3
Mediophyceae
Bacillariophyceae
Haptophyce
Heterokonts

5


Heterokonts
Haptophyceae
Chlorophyceae/
Prasinophyceae
Bacillariophyceae
Mediophyceae
Cos 3
Cos 2
Cos 1
Bolidiophyceae

6


7
Haptophyceae
Chlorophyceae/
Prasinophyceae
Mediophyceae
Bacillariophyceae
Coscinophyceae
Bolidophyceae
Heterokonts
9
Heterokonts
Haptophyceae
Chlorophyceae/
Prasinophyceae
Bacillariophyceae
Mediophyceae
Coscinophyceae
Triparma pacifica
Triparma laevis
Heterokonts
Haptophyceae
Chlorophyceae/
Prasinophyceae
Bacillariophyceae
Mediophyceae
Coscinophyceae
Bolidophyceae
8
Bacillariophyceae
Thalassiosirales
Attheya + Biddulphia
10



Bacillariophyceae
Medio 3
Medio 2
Medio 1
Cos 3
Cos 2
Cos 1
Bolidomonas pacifica
0.08

Supplemental Material