Plant Ecology and Evolution FastTrack: 278–296, 2010
doi:10.5091/plecevo.2010.418

REGULAR PAPER

A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research

Edward C. Theriot[1,2,*], Matt Ashworth[3], Elizabeth Ruck[3], Teofil Nakov[3] & Robert K. Jansen[2,4]

[1]Texas Natural Science Center, 2400 Trinity Street, University of Texas, Austin TX 78619, U.S.A.
[2]Section of Integrative Biology, The University of Texas at Austin, 1 University Station (A6700), Austin, TX 78712, U.S.A.
[3]Plant Biology Graduate Program, The University of Texas at Austin, 1 University Station (A6700), Austin, TX 78712, U.S.A.
[4]Institute of Cellular and Molecular Biology, The University of Texas at Austin, 1 University Station (A6700), Austin, TX 78712, U.S.A.
*Author for correspondence: etheriot@austin.utexas.edu

Background and aims – Formal inferences of the diatom phylogeny have largely depended on the nuclear-encoded small subunit of the rDNA gene (SSU). Large parts of the tree remain unresolved, suggesting that new sources of data need to be applied to this question. The next largest dataset consists of the large subunit of the ribulose-bisphosphate carboxylase gene (*rbc*L). The photosystem II gene *psb*C has also been applied to problems at higher levels of the diatom phylogeny. Thus, we sequenced each of these three genes for 136 diatoms in an attempt to determine their applicability to inferring the diatom phylogeny.
Methods – We attempted to obtain a more or less even sampling across the diatom tree. In particular, we increased sampling among the radial and polar centrics and among taxa that morphologically appear to be transitional between polar centrics and araphid pennates. Normal sequencing methods were used. Data were analyzed under maximum likelihood.
Key results – Analysis of SSU and chloroplast data returned many of the same clades and the same general structure of the tree. Combined, the data weakly reject monophyly of the radial centrics. The chloroplast data weakly support monophyly of the polar centrics but SSU and combined data weakly reject polar centric monophyly. There may be an hitherto unrecognized clade of araphid pennates sister to the remaining pennates.
Conclusion – While it is obvious that more genetic data need to be collected, perhaps the greatest obstacle to inferring an accurate, or at least global and robust, diatom phylogeny is the fact that the parts of the diatom tree that appear to be the most intractable to date (relationships among centric groups and between centrics and pennates) are also the most undersampled. This is in part due to major extinctions in the radial and polar centrics. We believe diatomists need to support more effort in both the molecular and morphological studies of these diatoms, and in the search for more information about the first half of the diatom stratigraphic record.

Key words – Diatoms, phylogeny, *rbc*L, *psb*C, SSU, stratigraphy.

INTRODUCTION

Diatoms have traditionally been classified into two major groups, centrics and pennates. For descriptive purposes, major elements of the former are typically around a central point or with no apparent organization, whereas the latter are typically elongate and their structures are organized more or less perpendicular to a longitudinal rib or bar called a sternum. Pennates themselves are often divided into two groups, the raphid pennates (usually with a pair of slits running through the sternum) and araphid pennates (those without such slits).

Traditionally derived diatom phylogenies may present very different hypotheses about relationships of species within these broader groups and also how these groups are related to one another. Steinecke (1931) drew a phylogenetic tree of the diatoms which had the centrics and pennates each as monophyletic sister taxa with raphid pennates as a monophyletic group nested within the paraphyletic araphids. In stark contrast, Simonsen (1979) drew a phylogenetic tree which had centric diatoms as paraphyletic, araphids as monophyletic, and raphid diatoms as paraphyletic (Eunotiaceae, with an unusually foreshortened raphe were placed as the

sister group to araphid diatoms). As an example of a third view, Round & Crawford (1981, 1984) argued that each major lineage (centrics, araphid pennates, raphid pennates) was derived separately from a pool of "Ur-diatom" forms. In short, each of these major groups has been identified as both monophyletic and paraphyletic in one major work or another.

These and other traditional phylogenies are difficult to compare on their own terms because they lacked formal assessment of homology and formal resolution of the inevitable conflict between characters. Homology, as understood by modern systematics (Patterson 1988), was often assessed without applying techniques or concepts that one might associate with modern phylogenetic systematics. Taxonomic groupings might have been made on the basis of some informal assessment of overall similarity, perhaps in conjunction with stratigraphic distribution in order to give direction to presumed evolutionary trends. Different sets of characters were employed in different parts of the hypothesis and by different authors. Evolutionary scenarios, presuming relative ease or difficulty of morphological or ecological transitions, were the explicit or implicit foundation for some hypotheses. In some cases, authors used reasoning that was very similar to modern systematic theory, but even then that reasoning might be applied to one character, while other arguments were applied to other characters. Such was the state of the art in systematics in general until the advent of phylogenetic systematics. Unfortunately, diatomists by and large have continued to eschew phylogenetic thinking about morphology long after the introduction of cladistic methodology.

This history constrains our present understanding of diatom phylogeny. The standard phylogenetic approach is to create a matrix of all available characters and compare them across all taxa for a given problem. The traditional approach often focused on one character system in one taxon in one study and another system in another taxon in another. Mann & Evans (2007) correctly identified a significant resulting problem, that phenotypic data have been gathered irregularly. The morphology of some species has been studied relatively thoroughly, including ultrastructural details of zygotes and gametes as well as details of frustule morphology. In many other species there is no information about even basic frustular elements such as the girdle bands.

Thus, formal phylogenetic analysis of morphology for diatoms as a whole remains lacking. However a number of studies have been performed on small groups of diatoms. Explicit cladistic principles were applied to one or a handful of characters in discussing evidence for certain groups, including *Cyclostephanos* and *Stephanodiscus* (Theriot et al. 1987) and *Mesodictyon* (Theriot & Bradbury 1987). Studies where formal matrices have been produced are limited to smaller groups of diatoms. Examples of such treatments are studies of certain araphid groups (Williams 1990), gomphonemoid and cymbelloid pennates (Kociolek & Stoermer 1993), the surirelloid diatoms (Ruck & Kociolek 2004), fossil and living tangentially undulate *Thalassiosira* species (Julius & Tanimura 2001), and the *Stephanodiscus niagarae* Ehrenb. complex (Theriot 1992). Edgar & Theriot (2004) created a dataset of molecular, and qualitative and quantitative morphological characters for *Aulacoseira*. Jones et al. (2005) created matrices of molecular and qualitative morphological data for raphid pennates.

Formal analyses of the larger diatom phylogeny began with the use of molecular data (Medlin et al. 1993). Alverson & Kolnick (2005), Mann & Evans (2007) and Theriot et al. (2009) summarized most of the available formal phylogenetic analyses done on higher level relationships of diatom molecular data. Analyses of molecular data (mainly nuclear SSU rDNA; henceforth SSU) have generally supported the notion expressed by some traditional phylogenies (Simonsen 1979, Round et al. 1990) that centric diatoms broadly grade into pennates through several nodes (Medlin et al. 1993, Medlin et al. 1996a, Medlin et al. 1996b, 2000, Medlin & Kaczmarska 2004, Sorhannus 2004, 2007). However, the particular relationships among groups vary from study to study. Reasons for this include utilization of different taxa, different optimality criteria, and, in some cases, failure to properly analyze the data under a specific optimality criterion (Theriot et al. 2009).

Centric diatoms have frequently been divided into two groups of convenience. The so-called radial centrics are mainly circular in outline, and the bi-or multipolar diatoms (henceforth simply polar diatoms) consist mainly of diatoms with elongated, triangular, quadrangular, etc. outlines. A notable, although not the only, exception are the Thalassiosirales which are mainly circular in outline but routinely fall in the polar diatoms. In the analyses above, it is generally true that each of these two groups are paraphyletic.

A major reclassification of the diatoms into three Classes has been recently proposed, with radial centrics formally named as the Coscinodophyceae, the polar centrics as the Mediophyceae and the pennates as the Bacillariophyceae (Medlin & Kaczmarska 2004). It has been suggested that the three major groups are each monophyletic. This has been called the CMB hypothesis, the acronym derived from the formal names of the three major clades (Theriot et al. 2009). However, several authors have argued that the classification may not reflect phylogeny, specifically questioning monophyly of the radial and polar centrics (Mann & Evans 2007, Williams & Kociolek 2007, Theriot et al. 2009).

Limited attempts have been made to resolve the diatom phylogeny using other genes. Ehara et al. (2000) recovered centrics as paraphyletic, and araphids and raphids each as monophyletic in a study using the 1.1-kb region of the cytochrome c oxidase subunit I (*cox*I) gene. Sampling was limited to only nine diatoms, however, and this study should be considered a demonstration of the potential of the *cox*I gene for resolving the diatom phylogeny rather than a robust phylogenetic estimate of the diatoms. Fox and Sorhannus (2003) studied eight diatoms using the *rpo*A gene. Tamura et al. (2005) studied seven free-living diatoms and three diatom endosymbionts of dinoflagellates using the large subunit of the ribulose-1,5-bisphosphate carboxylase oxygenase (*rbc*L) gene. Again this study should be understood as a demonstration of the potential of the *rbc*L gene for resolving the diatom phylogeny due to limited taxon sampling. Centric and pennate diatoms were each monophyletic; no araphid diatoms were sampled.

There have been successful applications of *rbc*L data to limited regions of the diatom tree. (Edgar & Theriot 2004, Jones et al. 2005, Alverson et al. 2007). Choi et

al. (2008) expanded *rbc*L sampling across the tree utilizing 36 taxa and included radial and polar centrics, and araphid and raphid diatoms. Their tree resembled the "Ur-diatom" hypothesis in that not only were raphid pennates monophyletic but araphids were as well (with the exception of *Aulacoseira ambigua* (Grunow in Van Heurck) Simonsen being incongruously placed among the araphids). With that exception and *Corethron criophilum* Castracane being sister to all other diatoms, centrics were also monophyletic. It is not clear if this unusual tree was due to still limited taxon sampling and/or properties of the *rbc*L gene. Rampen et al. (2009) included 61 *rbc*L sequences and got similar results. They performed analyses with and without the third codon position. Usually this is done when there is concern about mutational saturation at that position, but it was not stated why it was done here. The studies of Choi et al. (2008) and Fox & Sorhannus (2003) suggest that chloroplast genes may offer additional data useful to inferring the diatom phylogeny.

It is interesting that there have been no efforts, insofar as we are aware, to examine other nuclear genes outside the rDNA family. We have examined several low copy nuclear markers (in press) among heterokont algae and found phylogenetic inference to be highly confounded by paralogy issues. Thus, our lab is focusing on single copy plastid markers in the chloroplast and mitochondrion. This paper reports the results of our initial efforts to add two chloroplast markers that we have applied at the generic and ordinal level in diatoms (Edgar & Theriot 2004, Alverson et al. 2007), *rbc*L and *psb*C.

Here we focus on our three gene results and compare them to the CMB hypothesis in order to specifically examine the effect that adding chloroplast genes has on inferring the diatom phylogeny. Formal comparison of molecular to traditional phylogenies is beyond the scope of this paper, but is being conducted elsewhere (Theriot et al. in press).

MATERIALS AND METHODS

Taxon sampling

We attempted to sample broadly and evenly across the diatom tree (table 1). Our laboratory's efforts in the past two years have been particularly focused on obtaining unusual benthic tropical forms from all major structural groups. As a consequence we have greatly expanded sampling among non-Thalassiosirales polar centrics in particular. We have also added several radial centrics to the matrix and pennates such as *Bleakeyela*. Photographic vouchers are available at http://www.protistcentral.org. We have indicated which samples still have preserved DNA for future study, and, where living, those cultures which have been deposited at the Culture Collection for Marine Phytoplankton or the University of Texas Algal Culture Collection.

Molecular methods

DNA extraction and sequencing followed Alverson et al. (2007) for the most part. For species that we were not able to grow in culture in abundance we used the Chelex method for DNA extraction (Richlen & Barber 2005). SSU sequences were aligned against a secondary structure model following Theriot et al. (2009). Chloroplast protein encoding sequences were unambiguously aligned in SEAVIEW (Galtier et al. 1996). Columns that had missing data in more than 25% of the cells were trimmed from either end for each aligned gene. This resulted in matrix of 4237 nucleotide positions: 1705 positions were SSU, 1472 were *rbc*L, and 1060 were *psb*C. The data matrix is available at http://www.treebase.org or from the senior author upon request.

Phylogenetic methods

Analyses were conducted in RAxML 7.0.4 (Stamatakis 2006) on custom built Intel i7 processor based machines running Linux in Ubuntu 9.10. We used the parallel thread version on eight processors. We used the AIC criterion in Modeltest (Posada & Crandall 1998, Posada & Buckley 2004) to identify the most appropriate model for phylogenetic inference on each of seven data partitions: the entire SSU molecule, and the first, second and third codon positions of each of the two protein encoding genes. Modeltest identified GTR+G+I as the most appropriate for all three codon positions of the *psb*C gene, and the third codon position of the *rbc*L gene, TvM+G+I for the remaining *rbc*L codon positions, and TrN+G+I for the SSU gene. These are all special cases of GTR+G+I (all assume unequal base frequencies; GTR assumes six different substitution rate categories, TvM assumes five different rates, and TrN assumes three different rates). RAxML only incorporates the GTR model because the author feels the differences are likely to be trivial with larger datasets, and in order to optimize code (RAxML 7.0.4 manual). Because the three codon positions exhibited very different inferred evolutionary rates in ModelTest even when the same model was selected, we elected to analyze the fully partitioned dataset examined under ModelTest: one for the SSU, and one each for the first, second and third codon positions for each of *rbc*L and *psb*C.

Multiple (10–50) runs each with 100 rapid bootstrap replicates were performed for each analysis. Generally, the SSU and three gene analyses returned two to four identical topologies of nearly identical optimal –lnl scores (identical to 3–4 decimal places) within ten runs (our criterion for stopping analyses). The chloroplast data alone, however, usually returned very different trees within that same span, and typically required 25 or more runs to return multiple instances of the apparent optimal tree.

If all assumptions are met, then the optimal tree produced by a method such as ML will be the true tree if there is an infinite amount of data. With a finite amount of data, however, the optimal tree may or may not reflect the true tree due to stochastic influences. The challenge is then to put confidence intervals on trees, such that one may assume that the true tree lies among them. The standard bootstrap (BS) is one such method. Characters are repeatedly sampled with replacement and a tree is calculated for each dataset so created. Typically, the BS trees are collected and a majority rule consensus is calculated. This tree represents the probability that a node would be supported if one randomly sampled a universe of characters. The accuracy of that estimate assumes character sampling was unbiased. However, sampling is biased. Systematists attempt to sample

Table 1 – List of taxa used in this study grouped by informal name of clades that occurred in all analyses.
The toxariid plus *Lampriscus* clade occurred in all analyses as did the lithodesmiid plus thalassiosiroid clade. We subdivide them here and on the trees only to facilitate discussion. Only those clades which correspond to named genera are given formal names. The informal names are strictly names of convenience and are not meant to suggest a new formal classification nor to endorse any existing classification. Where species determinations are uncertain or are still being investigated we use "cf." to suggest a high degree of similarity. Species with an asterisk (*) lack sequence for the *psb*C gene. All species have sequence for both the SSU and *rbc*L genes. Those sequences obtained using DNA extracted with the Chelex method on single cells are marked with (Chelex) in source. All other sequences were obtained following standard methods. Abbreviations: CA = California, USA; TX = Texas, USA; HI = Hawaii, USA.

informal group **species name**	**source**
bolidophyte	
Bolidomonas pacifica Guillou & Chrét.-Dinet	CCMP 1866
Radial Centrics	
coscinodiscoid	
Coscinodiscus concinniformis Simonsen	Ship channel, Port Aransas, TX
Coscinodiscus granii Gough	Rainbow Harbor, Long Beach, CA
Coscinodiscus radiatus Ehrenb.	CCMP 310
Coscinodiscus wailesii Gran & Angst	Ship channel, Port Aransas, TX
Palmerina hardmaniana (Grev.) Hasle	Ship channel, Port Aransas, TX
Stellarima microtrias (Ehrenb.) Hasle	
melosiroid	
Aulacoseira granulata (Ehrenb.) Simonsen	FD 301 (UTEX)
Hyalodiscus sp.	Achang Reef, Guam
Hyalodiscus stelliger J.W.Bailey	CCMP 454
Melosira nummuloides C.Agardh	CCMP 482
Paralia sulcata (Ehrenb.) Cleve	CCAP 1059
Stephanopyxis turris (Arnott) Grev.	Redfish Bay, Port Aransas, TX
miscellaneous radial centric	
Actinocyclus sp.	Haputo Point, Guam
Actinoptychus sp.	South Africa
Aulacodiscus orientalis Grev.	Talofofo Bay, Guam
Aulacodiscus sp.	Stillwater Cove, Pebble Beach, CA (Chelex)
Corethron hystrix Hensen	CCMP 307
Guinardia delicatula (Cleve) Hasle	Corpus Christi Bay, Corpus Christi, TX
Rhizosolenia imbricata Brightw.	Corpus Christi Bay, Corpus Christi, TX
Rhizosolenia setigera Brightw.	CCMP 1820
Polar Centrics	
biddulphiopsid	
Biddulphiopsis membranacea (Cleve) Stosch & Simonsen	Gab Gab Beach, Guam
Biddulphiopsis titiana (Grunow) Stosch & Simonsen	Haputo Point, Guam
Chrysanthemodiscus sp.	Haputo Point, Guam
*Isthmia enervis** Ehrenb.	Guam (Chelex)
Trigonium formosum	Achang Reef, Guam
cymatosiroid	
Arcocellulus mammifer Hasle, Stosch & Syvertsen	CCMP 132
Brockmanniella brockmannii (Grunow) Hasle, Stosch & Syvertsen	CCMP 151
Campylosira cymbelliformis (A.Schmidt) Grunow ex Van Heurck	Corpus Christi Bay, Corpus Christi, TX
Extubocellulus cribriger Hasle, Stosch & Syvertsen	CCMP 391
Leyanella arenaria Hasle, Stosch & Syvertsen	CCMP 471

Table 1 (continued) – List of taxa used in this study grouped by informal name of clades that occurred in all analyses.

informal group species name	source
cymatosiroid	
Minutocellus polymorphus (Hargraves et Guillard) Hasle, Stosch & Syvertsen	CCMP 497
Papiliocellulus simplex Gardner & Crawford	CS 431 (CSIRO)
Lampriscus	
Lampriscus orbiculatum (Shadbolt) Perag. & H.Perag.*	Pago Bay, Guam
Lampriscus shadboltianum (Grev.) Perag. & H.Perag.*	Gab Gab Beach, Guam
lithodesmiid	
Bellerochea horologicalis Stosch	Redfish Bay, Port Aransas, TX
Lithodesmioides polymorpha Stosch	Taelayag Beach, Guam
Lithodesmium intricatum (West) Perag. & H.Perag.	Rainbow Harbor, Long Beach, CA
Lithodesmium sp.	Kahalu'u, Oahu, HI
Lithodesmium undulatum Ehrenb.	CCMP 1797
miscellaneous polar centric	
Attheya septentrionalis (Østrup) R.M.Crawford	CCMP 2084
Biddulphia tridens Ehrenb.	Rainbow Harbor, Long Beach, CA
Biddulphia alterans (J.W.Bailey) Van Heurck*	Kahana Bay, Oahu, HI
Cerataulina pelagica (Cleve) Hendey	Corpus Christi Bay, Corpus Christi, TX
Chaetoceros muelleri Lemmerman	CCMP 1316
Chaetoceros peruvianus Brightw.	Corpus Christi Bay, Corpus Christi, TX
Hemiaulus sinensis Grev.	Corpus Christi Bay, Corpus Christi, TX
odontelloid	
Amphitetras antediluvianum Ehrenb.	Montana de Oro State Beach, CA
Cerataulus smithii Ralfs*	Corpus Christi Bay, Corpus Christi, TX
Mastodiscus radiatus Prasad & Nienow	Corpus Christi Bay, Corpus Christi, TX
Odontella cf. *aurita*	Talofofo Bay, Guam (Chelex)
Odontella cf. *aurita*	Taelayag Beach, Guam (Chelex)
Odontella cf. *aurita*	Taelayag Beach, Guam (Chelex)
Odontella sinensis (Grev.) Grunow	CCMP 1815
Pleurosira laevis (Ehrenb.) Compère	FD 482 (UTEX)
Triceratium cf. *dubium*	Corpus Christi Bay, Corpus Christi, TX
Triceratium dubium Brightw.	Taelayag Beach, Guam
Triceratium sp.*	CCMP 147
terpsinoid	
Hydrosera sp.*	Austin, TX (Chelex)
Terpsinöe musica Ehrenb.	Brackenridge Field Lab, Austin, TX
thalassiosiroid	
Cyclostephanos dubius (Fricke) Round	Lake Waco, TX
Cyclotella meneghiniana Kütz.	Lake Waco, TX
Cyclotella sp.	Lake Ohrid, Macedonia
Detonula confervacea (Cleve) Gran	CCMP 353
Minidiscus trioculatus	CCMP 495
Planktoniella sol (Wallich) Schütt	CCMP 1608
Porosira glacialis (Grunow) Jørgensen	CCMP 668
Thalassiosira pseudonana Hust. (Hasle & Heimdal)	CCMP 1335
toxariid	
Ardissonea formosa (Hantzsch) Grunow	Gab Gab Beach, Guam
Ardissonea fulgens v. *gigantica* (Lobarzewsky) De Toni	Gab Gab Beach, Guam

Table 1 (continued) – List of taxa used in this study grouped by informal name of clades that occurred in all analyses.

informal group **species name**	**source**
toxariid	
Ardissonea formosa (Hantzsch) Grunow	Gab Gab Beach, Guam
Ardissonea fulgens v. *gigantica* (Lobarzewsky) De Toni	Gab Gab Beach, Guam
Climacosphenia sp.	Taelayag Beach, Guam (Chelex)
Toxarium hennedyanum (Greg.) Pelletan	Asan Beach, Guam
Toxarium undulatum J.W.Bailey	Gab Gab Beach, Guam
urosoleniid	
Acanthoceros sp.	Lake Okoboji, IA
Urosolenia eriensis (H.L.Smith) Round & R.M.Crawford	Yellowstone Lake
Araphid Pennates	
asterionellopsid	
Asterionellopsis glacialis (Castracane) Round	CCMP 134
Asterionellopsis glacialis (Castracane) Round	CCMP 1717
licmophorid	
Licmophora paradoxa (Lyngb.) C.Agardh	CCMP 2313
Podocystis spathulatum (Shadbolt) Van Heurck	Pago Bay, Guam
miscellaneous araphid	
Bleakeleya notata (Grunow) Round	Pago Bay, Guam
Cyclophora tenuis Castracane	Umatac Bay, Guam
Fragilariforma virescens (Ralfs) D.M.Williams and Round	FD 291 (UTEX)
Grammatophora oceanica Ehrenb.	CCMP 410
Plagiogramma staurophorum (Greg.) Heiberg	Taelayag Beach, Guam
Striatella unipunctata (Lyngb.) C.Agardh	Asan Beach, Guam
rhaphoneid	
Delphineis sp.	CCMP 1095
Rhaphoneis amphiceros (Ehrenb.) Ehrenb.*	Redfish Bay, Port Aransas, TX
staurosiroid	
Nanofrustulum cf. *shiloi*	CCMP 2649
Staurosira construens Ehrenb.	FD 232 (UTEX)
Staurosirella pinnata (Ehrenb.) D.M.Williams & Round	CCMP 330
synedroid	
Centronella reicheltii Voigt	CCAP 1011
Ctenophora pulchella (Kützing) D.M.Williams & Round	FD 150 (UTEX)
Synedra famelica Kütz.	FD 255 (UTEX)
Synedra hyperborea Grunow	CCMP 1423
Synedra ulna (Nitzsch) Ehrenb.	FD 404 (UTEX)
Synedropsis cf. *recta*	CCMP 1620
Tabularia cf. *tabulata*	CCMP 846
tabellarioid	
Asterionella formosa Hassall	UTCC 605
Diatoma elongatum (Lyngb.) C.Agardh	UTCC 62
Diatoma tenue C.Agardh	FD 106 (UTEX)
Tabellaria flocculosa (Roth) Kütz.	FD 133 (UTEX)

characters with a level of variation appropriate to the taxonomic level they are investigating. Other biases are also known (see reviews in Shimodaira 2002, and Verbruggen & Theriot 2008). In this study, we calculated majority rule consensus trees from the pool of all runs that finished within 1 –lnl unit of one another, resulting in populations of

Table 1 (continued) – List of taxa used in this study grouped by informal name of clades that occurred in all analyses.

informal group species name	source
Raphid Pennates	
berkeleyoid	
Berkeleya rutilans (Trentep.) Grunow	Thousand Steps Marine Sanctuary, Laguna Beach, CA
Climaconeis riddleae Prasad	Umatac Bay, Guam
Eunotia	
Eunotia curvata (Kütz.) Lagerstedt	FD 412 (UTEX)
Eunotia glacialis Meister	FD 46 (UTEX)
Eunotia pectinalis (O.F.Müller) Rabenhorst	NIES 461
Fallacia	
Fallacia monoculata (Hust.) D.G.Mann	FD 254 (UTEX)
Fallacia pygmaea (Kütz.) D.G.Mann	FD 294 (UTEX)
gomphonemoid	
Gomphonema affine Kütz.	FD 173 (UTEX)
Gomphonema parvulum Kütz.	FD 241 (UTEX)
Placoneis elginensis (Greg.) E.J.Cox	FD 416 (UTEX)
miscellaneous raphid	
Mastogloia sp.	Mustang Island, TX
Bacillaria paxillifer (O.F.Müller) Hendey	FD 468 (UTEX)
Cocconeis placentula Ehrenb.	FD 23 (UTEX)
Diploneis subovalis Cleve	FD 282 (UTEX)
Gyrosigma acuminatum (Kütz.) Rabenh.	FD 317 (UTEX)
Lemnicola hungarica (Grunow) Round & Basson	FD 456 (UTEX)
Navicula cryptocephala Kütz.	FD 109 (UTEX)
Phaeodactylum tricornutum Bohlin	CCMP 2561
Tryblionella apiculata	FD 465 (UTEX)
Undetermined nitzschiid	FD 185 (UTEX)
neidiid	
Neidium affine	FD 127 (UTEX)
Neidium bisulcatum (Lagerst.) Cleve	FD 417 (UTEX)
Neidium productum (W.Smith) Cleve	FD 116 (UTEX)
Scolioplеura peisonis Grunow	FD 13 (UTEX)
nitzschiid	
Cylindrotheca closterium (Ehrenb.) Reimann & Guillard	CCMP 1855
Denticula kuetzingii Thwaites	FD 135 (UTEX)
Nitzschia filiformis (W.Smith) Hust.	FD 267 (UTEX)
pinnulariid	
Caloneis lewisii Patrick	FD 54 (UTEX)
Pinnularia brebissonii (Kütz.) Rabenh.	FD 274 (UTEX)
Pinnularia termitina (Ehrenb.) Patrick	FD 484 (UTEX)
stauroneid	
Craticula cuspidata (Kütz.) D.G.Mann	FD 35 (UTEX)
Stauroneis acuta W.Smith	FD 51 (UTEX)
surirelloid	
Amphora coffeiformis (C.Agardh) Kütz.	FD 75 (UTEX)
Cymatopleura elliptica (Bréb. & Godey) W.Smith	L1333 (UTEX)
Surirella ovata Kütz.	L1241 (UTEX)
Undetermined surirelloid	CS 782 (CSIRO)

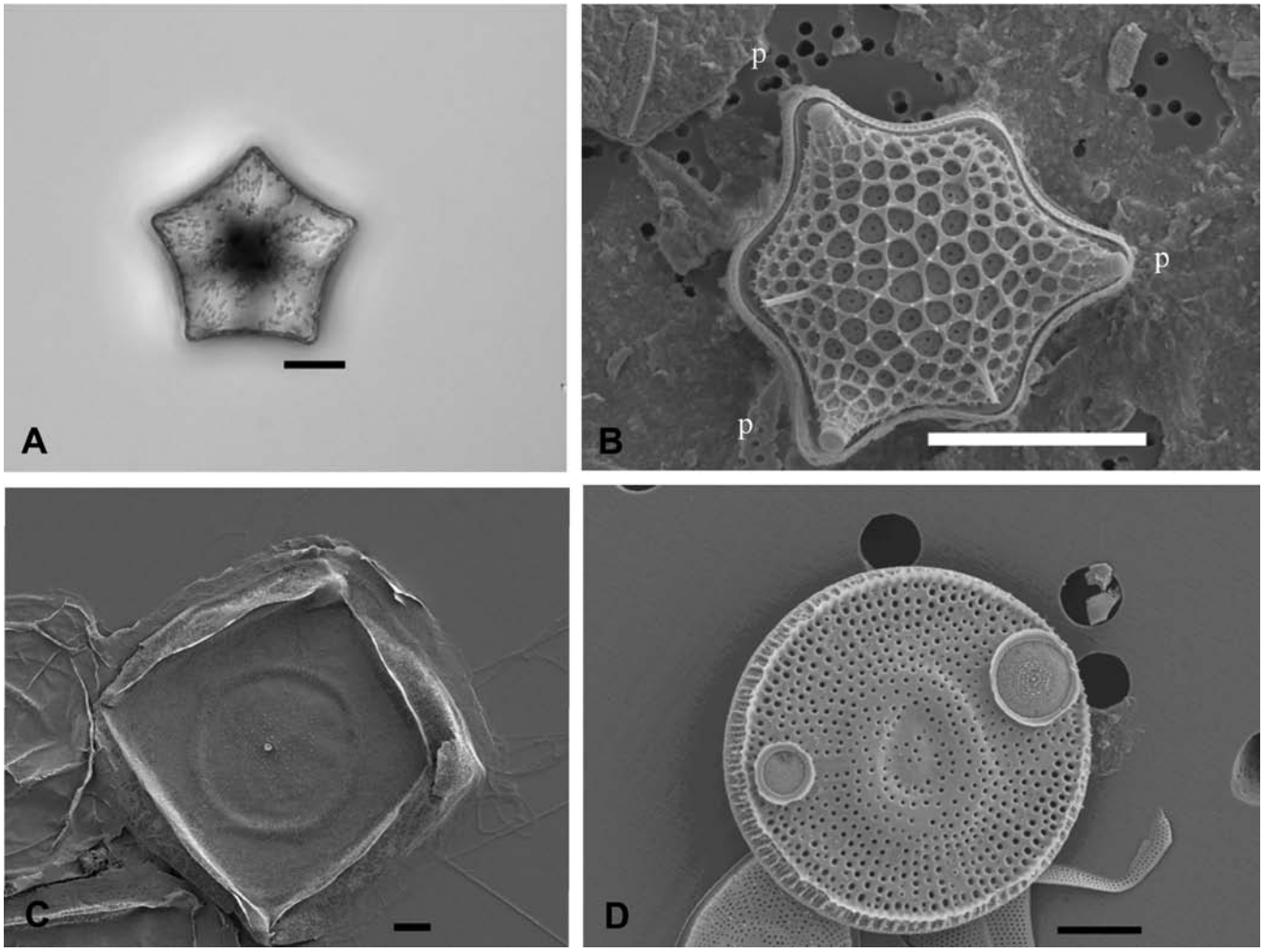

Figure 1 – A, *Trigonium formosum*. LM. Specimen from Achang Reef, Guam; B, *Triceratium dubium*. LM. Specimen from Pago Bay, Guam; C, *Lithodesmioides polymorpha*. SEM. Specimen from Taelayag Beach, Guam; D, *Mastodiscus radiatus*. SEM. Specimen from Corpus Christi Bay, TX. Scale bar = 10 µm. "p" = pore field.

at least 400 BS trees for each analysis.

Because of the many biases in the standard BS approach, other methods for assessing the confidence set of ML trees have been sought. Again, Verbruggen & Theriot (2008) provide a brief review for phycologists. Presently, the best methods seem to be those which attempt to overcome both the biases of standard BS analysis and biases inherent in any statistical test involving multiple comparisons (Shimodaira & Hasegawa 2001, Shimodaira 2002). If assumptions are met, then the odds are good that the true tree is among them. Of course, one cannot be certain that assumptions are met. Nevertheless, this approach can be used to determine if a pre-specified tree is outside the confidence set of trees (and therefore significantly worse than the optimal tree) given the data, assumptions and taxa at hand, regardless of whether or not the true tree is included in the confidence set.

There are several similar approaches to this problem. We chose the Approximately Unbiased test (Shimodaira 2002), which tends to produce a smaller confidence set and also corrects well for multiple comparisons. It uses a multiscale bootstrap of the site-wise likelihood scores to calculate the likelihood of the BS trees. This saves considerable effort over resampling the actual characters. CONSEL implements the AU test and reports the probability that the optimal tree and each of the suboptimal trees offered for testing belongs to the same confidence set.

Like the standard BS, the AU test can be affected by size of the dataset, number of bootstrap runs, and violation of evolutionary assumptions used to calculate the tree. Thus, we are conservative and do not claim that any tree inside the confidence set is true (and conversely that any tree outside the confidence set is unlikely to be true.) Rather, if the suboptimal tree does not belong to the confidence set, we only conclude that it confers significantly worse likelihood upon the dataset than does the optimal tree.

To test whether the CMB hypothesis confers significantly worse likelihoods upon the dataset(s), we calculated the optimal ML tree for each gene and for all genes concatenated with the taxa constrained to the CMB hypothesis. That is, radial centrics were constrained to monophyly, and polar centrics plus Thalassiosirales were constrained to reciprocal monophyly with the pen-

nates. Sitewise likelihood values were calculated for these constrained trees and for the corresponding unconstrained trees, and analysed with CONSEL.

RESULTS

We obtained SSU and *rbc*L sequences for 136 diatoms plus *Bolidomonas* and *psb*C sequences for 127 diatoms plus *Bolidomonas*. Although this paper is not about morphology and morphological relationships, we illustrate a few crucial diatoms in order to impress upon the reader the tremendous diversity of diatoms. We have added new sequences in all four groups, particularly in the non-Thalassiosirales polar centrics with a diversity of taxa including *Trigonium formosum* (Brightw.) Cleve, *Triceratium dubium* Brightw., *Lithodesmioides polymorpha* Stosch, and *Mastodiscus radiatus* Prasad & Nienow (fig. 1A–D). One important araphid added was *Bleakeleya notata* (Grunow) Round in Round, R.M.Crawford & D.G.Mann (fig. 2A–E). As will be shown, its position in the tree suggests a hitherto unknown clade of araphid pennates, outside of the typical araphid clades found in molecular studies. There is a pore field at that end of the cell which links to the neighboring cells. With the exception of the Thalassiosirales, most polar centrics are marine and only a handful of them are freshwater. In particular, the ocellate/pseudocellate taxa are nearly entirely marine. Whereas, *Pleurosira laevis* (Ehrenb.) Compère has been included in many studies, we have added three species, *Terpsinöe musica* Ehrenb. (fig. 2D), and *Hydrosera* sp. We illustrate another diatom, *Striatella*, which has previously been included in molecular analyses, because of its apparent crucial position in the araphid-raphid part of the tree.

In order to simplify presentation of trees, we calculated the strict consensus of all four trees returned in the analysis. Clades that appeared in each analysis are represented as triangles in the trees reported here. The same six melosiroid taxa are recovered as a clade by SSU, by the combined chloroplast dataset, and by SSU plus chloroplast data, and so are reported as a triangle in

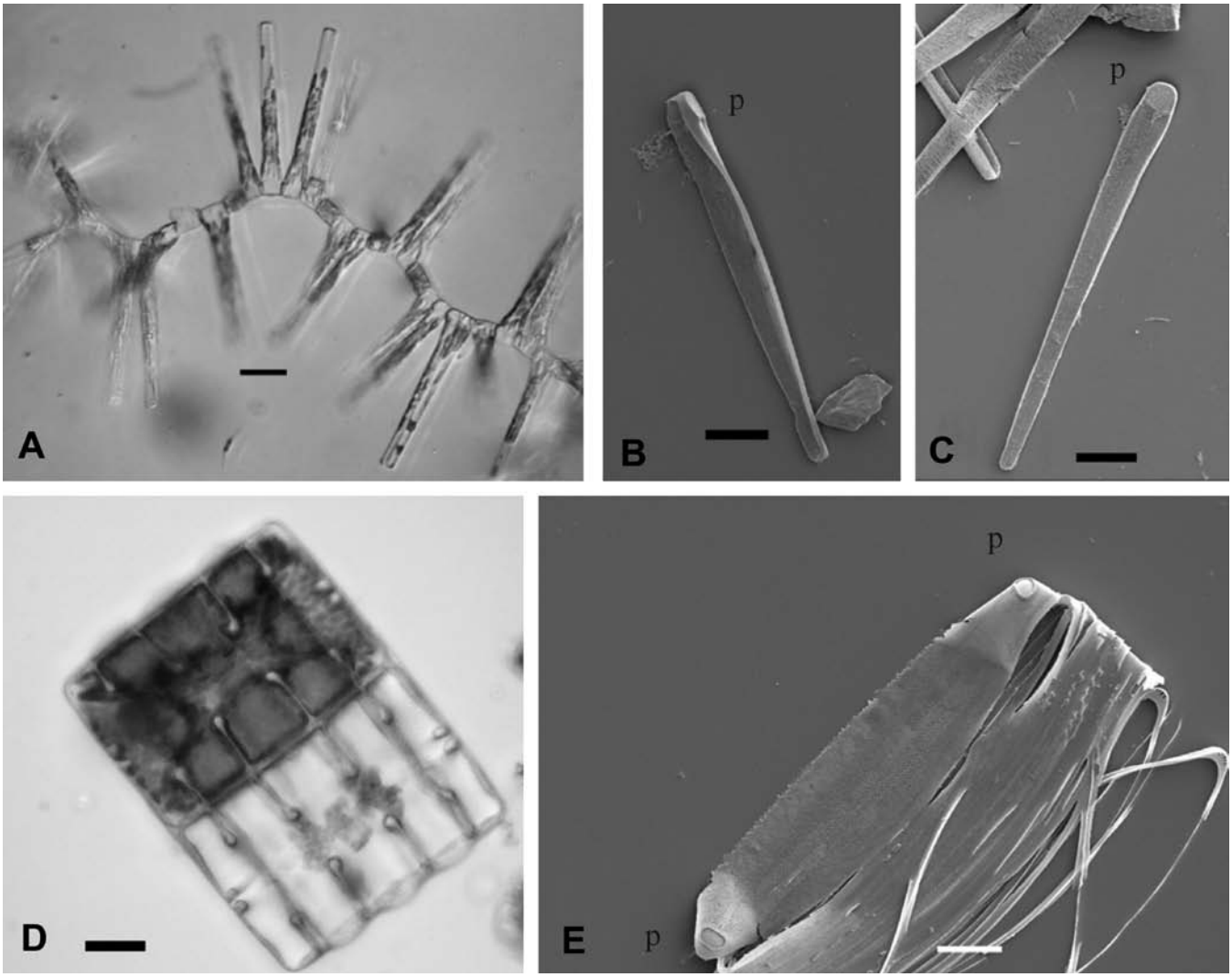

Figure 2 – A–C, *Bleakeyela notata.* Specimens from Pago Bay, Guam; A, LM showing colony formation; B–C, SEM. SEM showing pore field at basal part of cell where cell to cell connection takes place; D, *Terpsinöe musica.* LM from Ylig River, Guam; E, *Striatella unipunctata.* SEM from Agat Beach, Guam. Scale bar = 10 µm. "p" = pore field.

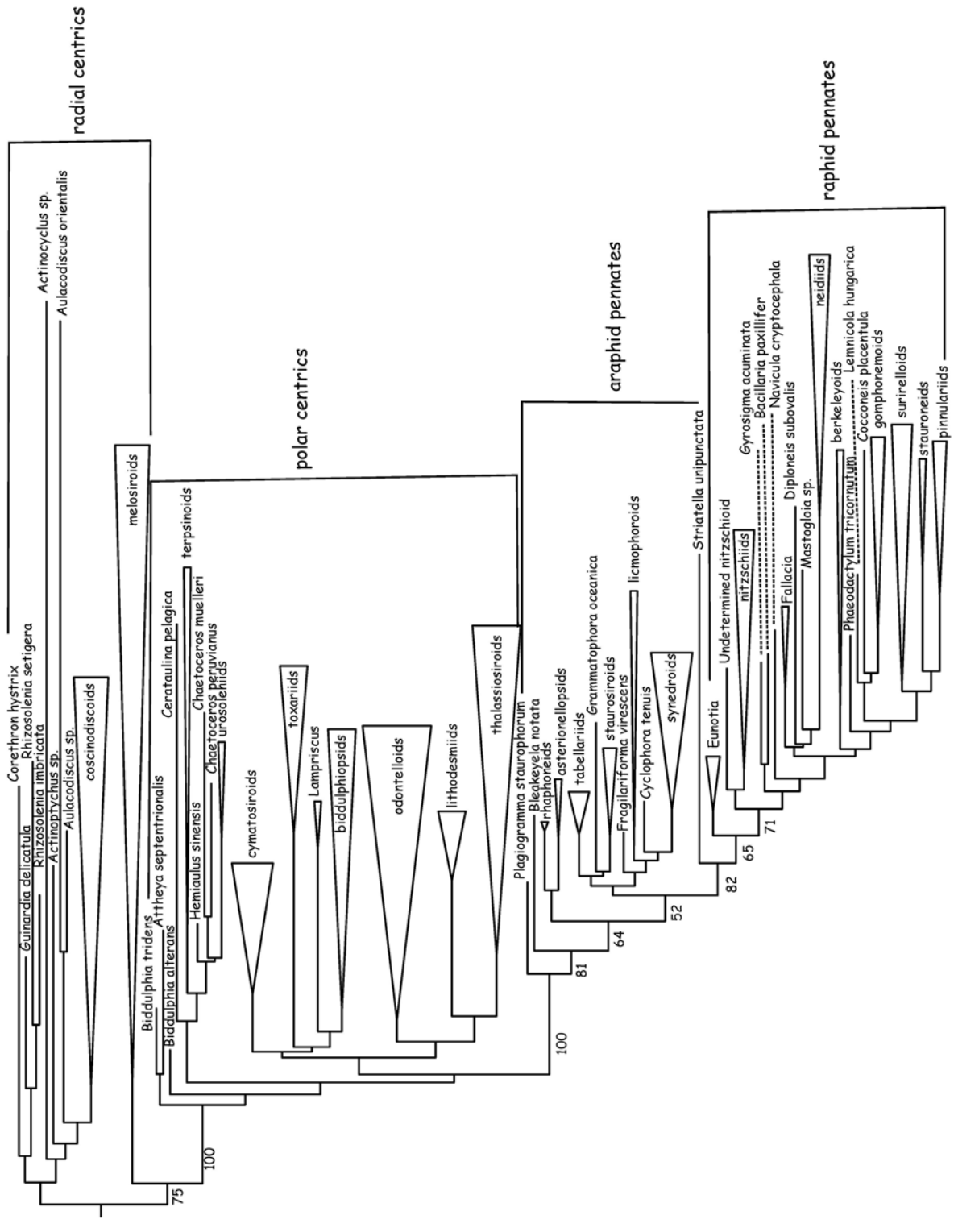

Figure 3 – Maximum likelihood phylogeny inferred from SSU. Height of triangles reflect relative number of taxa sampled. Length of triangles reflects longest distance between a terminal taxon and the most recently shared node for that clade. Names as per table 1. BS values are only shown for nodes on the main "trunk" of the tree.

each tree, for example.

As the bulk of formal diatom phylogenetic inference has been conducted using SSU rDNA, our first analysis is of the SSU gene alone.

SSU tree

The best tree recovered a grade of clades, through paraphyletic radial centrics, paraphyletic polar centrics, paraphyletic araphid diatoms, with raphid diatoms monophyletic (fig. 3). The radial centrics are composed of three clades, two with long pervalvar axes (one composed of *Corethron*, *Guinardia* and *Rhizosolenia*, and the other of melosiroids). The third clade is composed of diatoms with a relatively short pervalvar axis (coscinodiscoids). Melosiroids are sister to polar centrics plus pennates with a BS value of 75%.

Polar centrics plus pennates are resolved with 100% BS support. There are four clades in the polar centrics. *Biddulphia tridens* (Ehrenb.) Ehrenb. and *Attheya septentrionalis* (Oestrup) R.M.Crawford in R.M.Crawford, Gardner & Medlin form a small clade sister to remaining diatoms. *Biddulphia alterans* (J.W.Bailey) Van Heurck forms a single point clade. None of the internodes in the polar centric backbone had BS support values of 50% of greater.

Pennate diatoms have BS support of 100%. Nodes along the backbone among araphids and basal raphids received low to modest support (52%–82%). *Plagiogramma*, then *Bleakeyela* are sister to remaining pennates. *Striatella unipunctata* (Lyngb.) C.Agardh is sister to raphid pennates, and *Eunotia* is sister to all other raphids.

Chloroplast tree

The SSU and chloroplast trees had 25 clades in common and the two resembled each other in general appearance. The chloroplast tree again returned a radial centric grade, in this case with modest BS support separating *Corethron hystrix* Hensen from the remaining diatoms (fig. 4). No other backbone nodes received BS support above 50%. The polar centrics were arranged in a clade corresponding to the Mediophyceae. *Biddulphia* and *Attheya* did not form a grade at the base of the polar centrics as they did in the SSU tree, but formed a clade nested well inside the polar centric clade. The araphid pennates were arranged in a grade. Raphids are paraphyletic because *Cyclophora* and *Eunotia* were sister taxa.

Pairwise versus k2p corrected distances were plotted for each codon position for each gene. A curvilinear relationship was especially pronounced for the third codon position (not shown). Such a plot is sometimes interpreted to indicate that the gene or codon position is saturated with mutations and is not useful for phylogenetic analysis (but see the Discussion section). The correspondence between the results of the chloroplast and SSU genes, however, suggests that they share a similar signal and we combined them for a third analysis.

Three gene tree

The combined SSU plus chloroplast dataset returned a tree in which again only the raphid pennates were monophyletic (fig. 5). Nodes separating the various radial centric subgroups into a ladder-like structure had modest to high BS support, supporting the concept that radial centrics are a grade.

Polar centrics were arranged into two clades. One small clade, consisting of *Biddulphia* spp. and *Attheya septentrionalis*, was sister to a clade containing the remaining polar centrics clade plus the pennates clade. BS support was below 50% for this node.

Pennate diatoms were well supported as monophyletic (100% BS support). *Plagiogramma staurophorum* (Greg.) Heiberg and *Bleakeyela notata* were again separated from the remaining pennates (as they were in the SSU tree) but are a clade in the three gene analysis. *Striatella unipunctata* is again the sister group to raphid pennates which received low support as monophyletic (58%).

Tests of hypotheses

The CMB hypothesis was suboptimal but not significantly so in both the SSU and three gene analysis ($p = 0.243$ and $p = 0.113$, respectively). However, the chloroplast dataset, in spite of recovering a monophyletic polar centric clade, did very strongly reject the CMB hypothesis ($p = 0.003$), suggesting that the chloroplast data strongly reject a monophyletic radial centric clade (given that the polar centrics and the raphids were each monophyletic and sister to one another in the unconstrained chloroplast dataset). This possibly was related to the non-monophyly of the raphid pennates in the optimal chloroplast tree, so we compared a tree in which raphids were constrained to monophyly but polar and radial centrics were unconstrained to the CMB hypothesis. Even with raphids constrained to monophyly, the chloroplast data still rejected the CMB hypothesis ($p = 0.004$).

These results plus the fact that no tree returned high BS values among backbone nodes for the polar centrics (or high BS support for the monophyly of polar centrics in the case of chloroplast data), suggest that SSU and chloroplast data both support a paraphyletic radial centric hypothesis, but are more or less indecisive about higher level polar centric relationships.

DISCUSSION

It is fashionable to infer molecular phylogenies and then to speculate (or make firm conclusions) about diatom evolution on that basis. However, it is arguably difficult to find much about the diatom tree that is, in fact, robust to addition of new taxa or new data. Our own study is not immune to this. When we mention below that a study lacks a certain species whereas ours includes it or vice versa, it is not a criticism of that study (or ours) but we only mean to illustrate the effect of inclusion or not of that species by comparison and thus point the way for future research. There are somewhere between 100,000 and 1,000,000 or more diatom species that are alive or have lived. The entire genetic database only includes 1,000–2,000 entities. There is much to be done.

Our main interest in this study is whether or not the addition of new genes can robustly resolve the backbone of the diatom tree (assuming that it is a grade), or conversely resolve the diatoms into three (or at least a

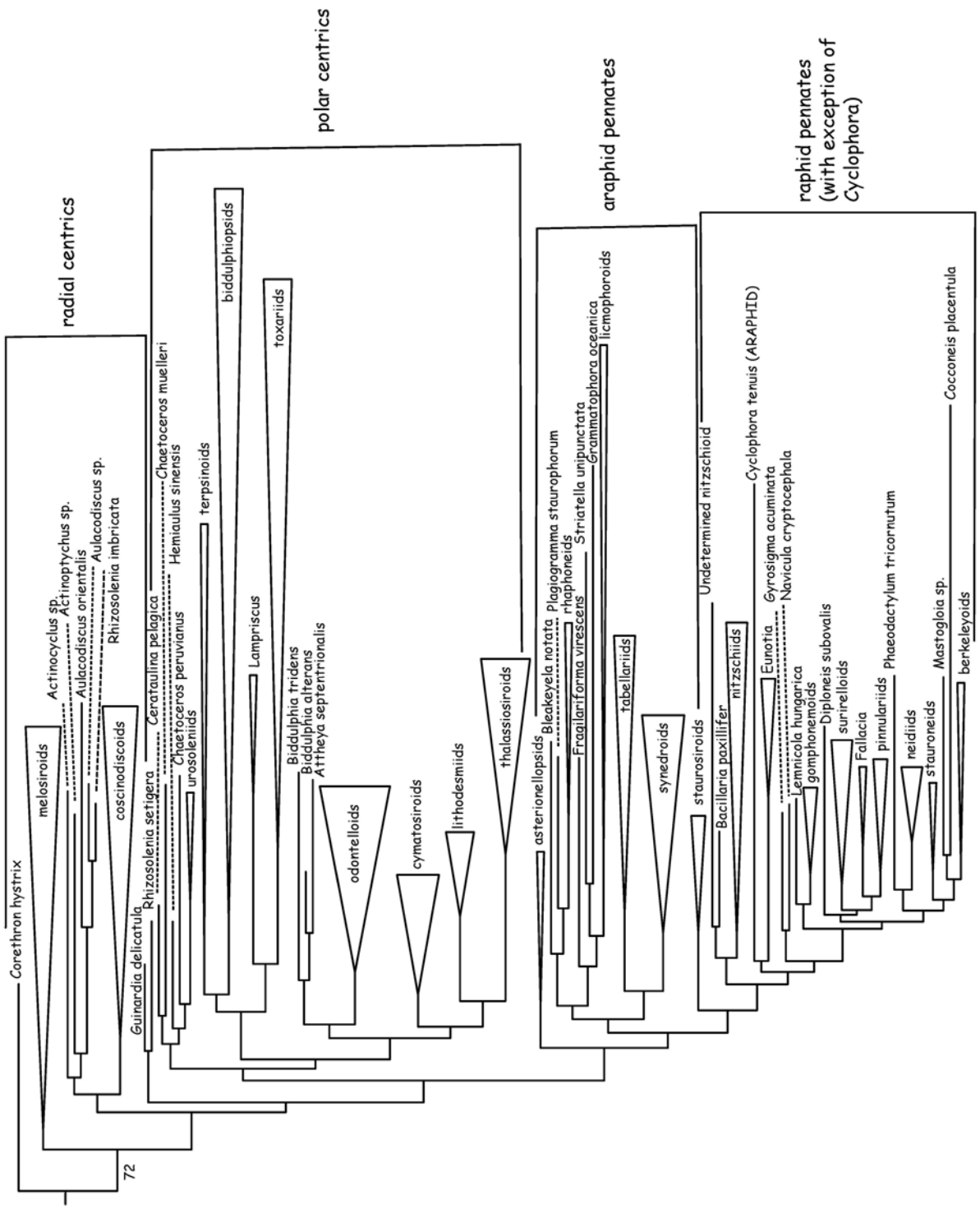

Figure 4 – Maximum likelihood phylogeny inferred from combined chloroplast data (*rbc*L plus *psb*C). Height of triangles reflect relative number of taxa sampled. Length of triangles reflects longest distance between a terminal taxon and the most recently shared node for that clade. Names as per table 1. BS values are only shown for nodes on the main "trunk" of the tree.

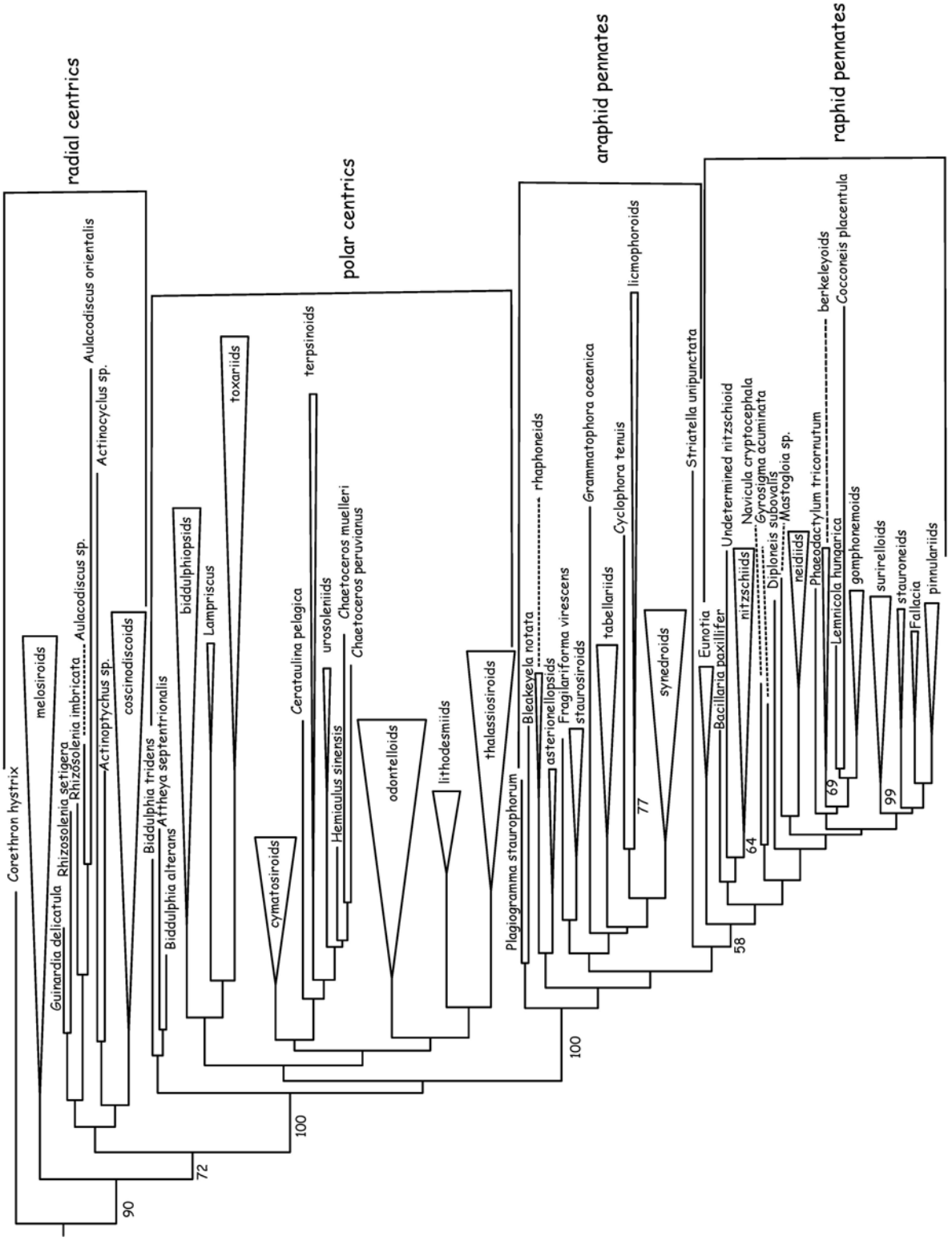

Figure 5 – Maximum likelihood phylogeny inferred from combined SSU, *rbcL* and *psbC* data. Height of triangles reflect relative number of taxa sampled. Length of triangles reflects longest distance between a terminal taxon and the most recently shared node for that clade. Names as per table 1. BS values are only shown for nodes on the main "trunk" of the tree.

few) major clades. Theriot et al. (2009) have previously remarked on the lack of strong support for or against the monophyly of each of the radial centrics and the polar centrics. Our study suggests that addition of chloroplast genes may lead to rejection of this hypothesis. Certainly these data reinforce the grade-like nature of the radial centrics. We strongly agree with many that the current diatom classification is flawed, largely because it does not reflect phylogeny. We also agree with Williams & Kociolek (2007) that we should not reject one paraphyletic system and replace it with another. Right now, the prevailing molecular evidence is that the three-taxon classification of diatoms is no more well supported (and maybe less well supported) than many alternatives (Theriot et al. 2009).

Mann & Evans (2007) provided a good summary of the issues. We expand on a few of them which we think are the most critical. Conclusions from our study and others are tempered by two related facts, data and taxon sampling remain inadequate. Mann & Evans (2007), citing Wortley et al. (2005), suggested at least 10,000 aligned nucleotides might be necessary to resolve the diatom phylogeny. With this study, we are nearly halfway there. Yet the phylogeny does not appear to be more resolved, if resolution is measured by nodes supported by high BS values along the backbone of the tree.

First we discuss the chloroplast data themselves. The chloroplast data added appear to have a relatively low signal to noise ratio. While the chloroplast tree generally resembles the SSU tree and those traditional phylogenies that have centrics related through a long grade, there is only one node with a BS value greater than 50% along the backbone. Related to this, we are obviously missing many taxa from this analysis (as any diatom study is destined to be for some time), including several which appear to strongly influence inferences about the diatom phylogeny.

Early analysis of *rbc*L resulted in phylogenies with topologies highly incongruent with those produced by SSU-based analyses (Choi et al. 2008, Rampen et al. 2009). There were fewer than forty diatom sequences in the former study, and fewer than seventy in the latter. This is important because there are theoretical reasons to believe that the results of these studies reflected taxon sampling issues, and not appropriateness of the *rbc*L gene *per se*. A general criticism of protein encoding genes is that the third codon position is saturated in mutations to the point where any historical signal is destroyed, and therefore homoplasy is so common that it cannot be untangled with standard evolutionary models. This idea could explain why Rampen et al. (2009) thought the third codon position might be problematic and elected to perform analyses with and without the third codon position.

The problem is that there is no way to *a priori* test that saturation will actually mislead an analysis. Indeed in this study, the third codon position is saturated according to the usual standard (plots of pairwise differences versus adjusted distances – not shown). However, this approach is a poor guide to actual utility of a molecule in phylogenetic inference (Verbruggen & Theriot 2008). For example, although saturated according to this test, the *rbc*L molecule was still shown to be useful in resolving deep branches in the heterokont algae as a whole, including providing strong corroboration for the sister group status of *Bolidomonas* and diatoms (Daugbjerg & Andersen 1997a, Daugbjerg & Andersen 1997b, Daugbjerg & Guillou 2001, Goertzen & Theriot 2003). Thus, it is hard to justify excluding *rbc*L (and *psb*C by extension) from our study. Note that the one backbone node supported by chloroplast data at a BS value greater than 50% is the deepest branch in the diatom tree.

The reason that the standard plot of pairwise versus corrected distance does not adequately assess saturation is that it does not take into account the number and distribution of taxa analysed. The patristic distance between any two taxa will, for example, always be greater than the pairwise distance if any character along the path between the taxa demonstrates homoplasy. Sampling taxa more densely will result in more homoplasy that might be correctly identified. This is the core reason that increased taxon sampling within the group of interest improves accuracy in phylogenetic inference (Hillis 1998, Poe & Swofford 1999, Pollock et al. 2002, Zwickl & Hillis 2002, Goertzen & Theriot 2003, Hillis et al. 2003, Hedtke et al. 2006, Dunn et al. 2008).

One might also believe that our results differ because we have both *rbc*L and *psb*c data (i.e. more data). Certainly, this has some effect, but we note that both genes have similar properties in terms of variability in this and in Alverson et al. (2007). Also, as we added *rbc*L and *psb*c sequences to our dataset, we also recovered trees with unusual topologies initially (not shown). However, as we approached 100 taxa, the tree topologies began to resemble SSU topologies. Thus, theory and empirical results both support the idea that the chloroplast genes we selected require relatively dense taxon sampling in order to recover phylogenetic signal from them for the deeper branches of the diatom phylogeny.

This point is significant. We are on the verge of being able to effectively sequence dozens of genes over the course of a single study. Phylogenomic studies are not far off. Yet, as this study shows, more data alone is not sufficient. Indeed, even in studies utilizing a huge part of the entire genome, taxon sampling remains a paramount issue (Dunn et al. 2008).

Radial and polar centrics remain undersampled, in spite of the fact that we have added many more species than previously available, particularly among ocellate and pseudocellate taxa. Centrics have many extinct species that immediately signals potential taxon sampling problems. When one considers that these lineages are probably older than pennates and relatively also have many more extinctions, the potential for long-branch problems becomes immediately apparent. This problem of "… the extinction of evidence …" was elegantly addressed by Williams (2007).

Sampling of radial centrics has been greatest among species of *Aulacoseira* (Edgar & Theriot 2004) and *Rhizosolenia*. Otherwise, the diatom SSU datasets analyzed have generally had but a few discoid centrics (*Stellarima*, *Coscinodiscus* and *Actinocyclus*), and a few species each of *Corethron* and *Leptocylindrus*. Our study added species of *Aulacodiscus*, *Actinoptychus*, and the hemidiscoid-shaped *Palmerina* (Garcia & Odebrecht 2008) but these seemed to have no dramatic effect on topology. Nevertheless, there are still many living genera and species still not sampled in this region of the tree. More problematic is the fact that there are more extinct genera alone that probably belong to the

radial centrics than there are genera in the most broadly sampled tree. Harwood & Nikolaev (1995) diagram the stratigraphic distribution of numerous probably radial centrics from the Upper and Lower Cretaceous.

The melosiroid lineage needs further sampling. What is now known as *Ellerbeckia sol* is sister to the remaining diatoms in several studies (e.g. Medlin & Kaczmarska 2004), even though SSU places it rather distantly from other melosiroids (Theriot et al 2009). This needs to be re-examined because the base of the tree is very much in question, and this in turn affects conclusions about the monophyly of the radial centrics. Rampen et al. (2009), for example, return a monophyletic radial centric group, but do not include *Corethron* in their analysis.

Similar comments can be made about missing taxa in the polar centrics. Both Harwood & Nikolaev (1995) and Sims et al. (2006) discuss many extinct genera that are polar in morphology. Among living taxa, our study demonstrates that *Biddulphia* may be a particularly undersampled genus. The diatom *Attheya* has been placed close to the pennates by both SSU and *rbc*L analyses (Rampen et al. 2009). However, that study did not include *Biddulphia*. Our two species of *Biddulphia* group with *Attheya septentrionalis* in the SSU and three gene analysis, separate from all other polar centrics (making the polar centrics non-monophyletic).

But important extinctions are not limited to the centrics. One example is the diatom *Adoneis* (Kociolek et al. 2007) which has the areolar pattern of a pennate diatom but is characterized by a circle of rimoportulae. This is similar to *Pseudostriatella oceanica* Sato, Mann & Medlin (Sato et al. 2008b). As Kociolek et al. (2007) wrote, this sort of arrangement further challenges present definitions of the pennate diatoms.

But this problem is not new. Several diatoms now considered polar centrics have historically been relevant to the discussion of exactly where one draws the line between centrics and pennates (e.g. cymatosiroids, toxariids), and others that have historically been considered centrics (e.g. *Plagiogramma*) are now considered pennates. It is clear that more attention needs to be paid to sampling these groups and taxa that appear to have many derived features of the pennates and plesiomorphic features of the centrics. Again the work of Sato and colleagues (Sato et al. 2008a, 2008b, 2008c) and Kociolek (Kociolek & Rhode 1998, Kociolek et al. 2007) illustrates that there is a diversity of species that may fall at the transition of the centrics to pennates.

The relationship of *Striatella unipunctata* to the raphid pennates is of significance given the general result that raphid pennates are monophyletic. Sato et al. (2008b) recovered *Striatella* and the new genus *Pseudostriatella* as sister to *Eunotia*, and questioned whether or not this relationship was an artifact of long-branch problems. They also reviewed the molecular and morphological relationships of *Striatella* to the raphid pennates. They noted that various studies reported *Striatella* as embedded within the araphids (Sims et al. 2006, Sorhannus 2007), although it was only one node removed from the raphid pennates in Sorhannus (2007). Interestingly, given the results of Alverson et al. (2006), Sims et al. (2006) recovered *Rhabdonema* as sister to raphid pennates, with *Striatella unipunctata* only one node removed from *Rhabdonema* plus raphids. We included neither *Rhabdonema* nor *Pseudostriatella* in our study. Clearly both taxa need to be considered in future studies as well as a number of important taxa recently added to the SSU dataset (Sato et al. 2008a, Sato et al. 2008b, Sato et al. 2008c) in the family Plagiogrammaceae and relatives of *Bleakeyela*. Incidentally, the instability in the position of *Striatella*, and the chloroplast data placing *Cyclophora* with *Eunotia*, illustrate that the sister group to raphid pennates remains as uncertain as the relationships at the transition between centrics and pennates.

There are other indications of regions where the molecular tree is undersampled. Several regions indicate the possibility of massive morphological convergence that may suggest phylogenetic error. The fultoportula morphology of *Cyclotella* s. str. appears to be convergent on that of *Discotella*, *Cyclostephanos*, *Puncticulata* and *Stephanodiscus* (Alverson et al. 2007). Whereas elongation may have arose once early in diatom history as implied by trees of the topology of the three gene tree (e.g. Alverson et al. 2006), it appears that extreme elongation has arisen multiple times, and arose through different developmental causes as well (Medlin et al. 2008). The complicated canal raphe of nitzschiid and surirelloid taxa may have also arisen at least twice. These two taxa are separated by several nodes in various molecular studies cited here as well as in our analyses. In these and other cases, sister group relationships are surprising not because there is conflicting morphological evidence but rather because there is simply a lack of evidence in terms of synapomorphic characters. The sister group of the *Discotella-Cyclostephanos-Stephanodiscus* clade, for example, includes *Bacterosira* which bears no obvious derived similarity to any of these diatoms (Alverson et al. 2007).

Assuming, for the moment, that the genes we are using are appropriate to the goal, and that we have at least achieved some modestly effective level of taxon sampling, one might entertain the idea that the lack of resolution in parts of the diatom tree is, in fact, real. That is they represent massive and rapid radiations over a very brief period of time such that any gene that might have evolved rapidly enough to record the branching pattern will have had its historical signal completely obscured. One might ask if the fossil record demonstrates a gradual increase in the number of diatom taxa or are there sudden bursts of diversity?

There remains an alternative explanation. The lack of morphological synapomorphy (as opposed to conflict of putative morphological synapomorphies) for many relationships implied by molecular data strongly suggests that there may be entire lineages missing from our analysis. Indeed it suggests that there might be entire lineages missing from scientific observation. That is, can we assume that we, as a community, have achieved some modestly effective level of taxon sampling in our pursuit of the diatom phylogeny?

The problem with addressing this question is that the fossil record itself has quite large gaps (Harwood & Nikolaev 1995, Sims et al. 2006) and molecular clocks imply the existence of several "ghost lineages" or discrepancies between clock estimates and stratigraphy (Sorhannus 2007). Other summaries of the quality of the fossil record of diatoms are available elsewhere (Harwood & Nikolaev 1995, Sims et al. 2006). The fol-

lowing account is taken from those papers and papers cited therein.

The main features are that, according to these works, the accepted marine record starts with a single early Jurassic deposit (c. 190 My), and then there is a gap of nearly 70 million years in which there are no or few well preserved and well studied deposits. There are several deposits from the lower Cretaceous (Aptian and Albanian epochs) then few or no studied deposits for another 20 million years. More regular recovery of diatom bearing sediments in deep sea cores and from terrestrial deposits begins in the Santonian epoch through the Maastrichtian (end of the Cretaceous). Yet, the number of well studied deposits still only number in the few dozens (more deposits are available, but the diatoms are heavily pyritized and so provide little detailed information). The early freshwater record is at least equally poorly known. The earliest known freshwater deposit is from South Korea (175 My). The next oldest reported deposit is 105 million years younger (Chacón-Baca et al. 2002). While there are many archived Eocene and Oligocene deposits from the Ocean Drilling Project, there are still but a handful deposits studied, with increasing numbers of deposits (and taxa) occurring in the Miocene and later.

This again is not meant as a criticism of any particular work or body or work. Rather it is simply a caution that it is quite likely that our understanding of diatom diversity through time is quite likely biased by the quality of the fossil record. What this means to us is that search for and study of older lacustrine and marine deposits should be encouraged by diatomists, even in this age of molecular systematics, if we are ever to truly understand the diversification of diatoms. The gaps in the fossil record and the gaps in our molecular datasets appear to be highly correlated.

However, this cannot be tested without formal analysis of morphological data. At best, diatomists are somewhat ambiguous about the role and importance of morphology. For example, Mann & Evans (2007) did not dismiss morphology. In fact, they wrote it was necessary to create a matched dataset of both morphology and molecules, and made a good case for further investigating morphology. Nevertheless, in several areas where traditional hypotheses conflict with molecular results they concluded that morphology was incorrect and always favoured the molecular result.

Medlin & Kaczmarska (2004) are more explicit and less ambiguous. The introduction to that paper argued that morphology can only be used after molecular phylogenies clarify which morphological characters are homologous. On page 246, they stated that the purpose of the paper is to place all other types of evidence in the context of the molecular tree, clearly putting into practice the philosophy of the introduction. Morphological similarity is, for the most part, expressedly treated as convergence or parallelisms, or explained non-canonically in terms of ancient polymorphisms in which it is claimed that diatom ancestors retained conditions that have never been observed (a cell with two types of Golgi arrangements) and that these polymorphisms degraded through time so that descendants just had one or the other condition (page 265).

However, Medlin & Kaczmarska (2004, page 252) also demonstrated an ambiguous attitude towards the role of morphology when they placed *Paralia* in a clade with other radial centrics despite molecular evidence to the contrary. There is only one other case we could find in which a molecular result was rejected on the basis of morphology (e.g. Sato et al. 2008b – see discussion below).

Strictly from the point of recovering the diatom phylogeny, why do we need a morphological analysis? Are the fossil taxa so important? Are missing taxa so important? Will we not overcome all this with data from thousands and thousands of nucleotides? Not necessarily.

First of all, in practice at least today, even molecular datasets are finite in size. With finite datasets, including morphological data for extinct taxa has been shown to overturn hypotheses based on other datasets (Gauthier et al. 1988). Even with the advent of sequence data for multiple genes, morphological data from extant and fossil taxa are important. In fact, it is not unusual to find that the incongruence between morphological and molecular trees vanishes or is greatly reduced when morphological data from extinct taxa are formally included, or that neither dataset alone provides a robust solution (Eernisse & Kluge 1993, O'Leary 1999, Gatesy & O'Leary 2001, Springer et al. 2001, Gatesy et al. 2003, Mallatt & Chen 2003, Strait & Grine 2004).

The entire literature on long-branch attraction began with a concern about what happens as more and more data are added (Felsenstein 1978). All methods available to us now, and likely to be available to us in the future are statistically inconsistent when assumptions are not met, this includes likelihood methods. In simple terms, if your assumptions are wrong then the more molecular data you get, the more certain you are to be wrong. The fewer number of character states a character system has, the worse the problem. A fundamental difference between molecular and morphological data is that while the former has a practically infinite number of characters, it has but four character states (five if one codes gaps as a character). That is, even phylogenomic approaches are liable to get statistically supported results, regardless of whether or not they are correct. Morphology appears to have a more limited number of characters, but many characters have a great number of character states judging from the morphological diversity of life.

While there might be technical difficulties with comparing or weighing thousands upon thousands of nucleotides against morphological characters, this is only compounded when morphology is just informally analyzed. Thus, we agree with Mann & Evans (2007) that a morphological dataset comparable to that of the molecular dataset must ultimately be built. We applaud efforts made to generate such data where it is done. Without formal analysis of morphology, we might have little to suggest that a molecular tree might be wrong as we gather more and more molecular data. Without morphology, for example, Sato et al. (2008b) would not have been able to write: "Some features of our tree, such as the sister relationship between the *P. oceanic* – *S. unipunctata* clade and the raphid genus *Eunotia*, have high support but are frankly implausible because of morphological and reproductive evidence."

In summary, while considerable progress has been made towards understanding diatom phylogeny in the last several decades, much remains to be done. We be-

lieve that several critical areas have now been identified as priority research for reconstructing the diatom tree: 1) enhanced sampling of taxa, particularly among radial and polar centrics, 2) broadly increased sampling of morphological characters, particularly developmental features as advocated by Mann & Evans (2007), 3) support for research on the early diatom fossil record, 4) formal integration of morphological data for living and extinct taxa, and 5) finally increased sampling of the diatom genome, particularly markers from the mitochondrion and chloroplast.

ACKNOWLEDGMENTS

This research was supported by NSF EF 0629410. The senior author was also supported by the Jane and Roland Blumberg Centennial Professorship in Molecular Evolution. This work could not have been accomplished without the friendship and support of Chris Lobban and Maria Schefter who introduced us to Guam diatoms and provided excellent scientific support and great companionship on field trips there, in California and to the Texas Gulf Coast.

REFERENCES

Alverson A.J., Kolnick L. (2005) Intragenomic nucleotide polymorphism among small subunit (18S) rDNA paralogs in the diatom genus Skeletonema (Bacillariophyta). Journal of Phycology 41: 1248–1257.

Alverson A.J., Cannone J.J., Gutell R.R., Theriot E.C. (2006) The evolution of elongate shape in diatoms. Journal of Phycology 42: 655–668.

Alverson A.J., Jansen R.K., Theriot E.C. (2007) Bridging the Rubicon: Phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. Molecular Phylogenetics and Evolution 45: 193–210.

Chacón-Baca E., Beraldi-Campesi H., Cevallos-Ferriz S.R.S., Knoll A.H., Golubic S. (2002) 70 Ma nonmarine diatoms from northern Mexico. Geology 30: 279–281.

Choi H.-G., Joo H.M., Jung W., Hong S.S., Kang J.-S., Kang S.-H. (2008) Morphology and phylogenetic relationships of some psychrophilic polar diatoms (Bacillariophyta). Nova Hedwigia Beihefte 133: 7–30.

Daugbjerg N., Andersen R.A. (1997a) A molecular phylogeny of the heterokont algae based on analyses of chloroplast-encoded rbcL sequence data. Journal of Phycology 33: 1031–1041.

Daugbjerg N., Andersen R.A. (1997b) Phylogenetic analyses of the rbcL sequences from haptophytes and heterokont algae suggest their chloroplasts are unrelated. Molecular Biology and Evolution 14: 1242–1251.

Daugbjerg N., Guillou L. (2001) Phylogenetic analyses of Bolidophyceae (Heterokontophyta) using rbcL gene sequences support their sister group relationship to diatoms. Phycologia 40: 153–161.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452: 745–749.

Edgar S.M., Theriot E.C. (2004) Phylogeny of Aulacoseira (Bacillariophyta) based on morphology and molecules. Journal of Phycology 40: 772–788.

Eernisse D.J., Kluge A.G. (1993) Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution 10: 1170–1195.

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.

Felsenstein J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27: 401–410.

Fox M.G., Sorhannus U.M. (2003) RpoA: A useful gene for phylogenetic analysis in diatoms. Journal of Eukaryotic Microbiology 50: 471–475.

Galtier N., Gouy M., Gautier C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Computer Applications in the Biosciences 12: 543–548.

Garcia M., Odebrecht C. (2008) Morphology and ecology of the planktonic diatom Palmerina hardmaniana (Greville) Hasle in southern Brazil. Biota Neotropica 8: 85–90.

Gatesy J., Amato G., Norell M., Desalle R., Hayashi C. (2003) Combined support for wholesale taxic atavism in gavialine crocodylians. Systematic Biology 52: 403–422.

Gatesy J., O'leary M.A. (2001) Deciphering whale origins with molecules and fossils. Trends in Ecology & Evolution 16: 562–570.

Gauthier J., Kluge A.G., Rowe T. (1988) Amniote phylogeny and the importance of fossils. Cladistics 4: 105–209.

Goertzen L.R., Theriot E.C. (2003) Effect of taxon sampling, character weighting, and combined data on the interpretation of relationships among the heterokont algae. Journal of Phycology 39: 423–439.

Harwood D.M., Nikolaev V.A. (1995) Cretaceous diatoms: morphology, taxonomy, biostratigraphy. In, Siliceous Microfossils: 18th Annual Short Course of the Paleontological Society: 81–106. New Orleans, Louisiana, The Paleontological Society.

Hedtke S., Townsend T., Hillis D. (2006) Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Systematic Biology 55: 522–529.

Hillis D.M. (1998) Taxonomic sampling, phylogenetic accuracy and investigator bias. Systematic Biology 47: 3–8.

Hillis D.M., Pollock D.D., Mcguire J.A., Zwickl D.J. (2003) Is sparse taxon sampling a problem for phylogenetic inference? Systematic Biology 52: 124–126.

Jones H.M., Simpson G.E., Stickle A.J., Mann D.G. (2005) Life history and systematics of Petroneis (Bacillariophyta), with special reference to British waters. European Journal of Phycology 40: 61–87.

Julius M.L., Tanimura Y. (2001) Cladistic analysis of plicated Thalassiosira (Bacillariophyceae). Phycologia 40: 111–122.

Kociolek J.P., Fourtanier E., Rubinstein J., Encinas A. (2007) Adoneis miocenica, a new species from Chile, with comments on the morphological separation of centric and pennate diatoms. Diatom Research 22: 309–316.

Kociolek J.P., Rhode K. (1998) Raphe vestiges in "Asterionella" species from Madagascar: Evidence for a polyphyletic origin of the araphid diatoms? Cryptogamie Algologie 19: 57–74.

Kociolek J.P., Stoermer E.F. (1993) Freshwater gomphonemoid diatom phylogeny: Preliminary results. Hydrobiologia 269/270: 31–38.

Mallatt J., Chen J.-Y. (2003) Fossil sister group of craniates: Predicted and found. Journal of Morphology 258: 1–31.

Mann D.G., Evans K.M. (2007) Molecular genetics and the neglected art of diatomics. In: Brodie J., Lewis J. (eds) Unravelling the algae – the past, present and future of algal systematics: 231–265. Boca Raton, Florida, CRC Press.

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., 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.

Medlin L.K., Kooistra W.H.C.F., Gersonde R., Wellbrock U. (1996a) Evolution of the diatoms (Bacillariophyta): II. Nuclear-encoded small-subunit rRNA sequence comparisons confirm a paraphyletic origin for the centric diatoms. Molecular Biology and Evolution 13: 67–75.

Medlin L.K., Kooistra W.H.C.F., Gersonde R., Wellbrock U. (1996b) Evolution of the diatoms (Bacillariophyta): III. Molecular evidence for the origin of the Thalassiosirales. Nova Hedwigia Beiheft 112: 221–234.

Medlin L.K., Kooistra W.H.C.F., Schmid A.-M.M. (2000) A review of the evolution of the diatoms – a total approach using molecules, morphology and geology. In: Witkowski A., Sieminska J. (eds) The Origin and Early Evolution of the Diatoms: Fossil, Molecular and Biogeographical Approaches: 13–36. Kraków, Szafer Institute of Botany, Polish Academy of Sciences.

Medlin L.K., Sato S., Mann D.G., Kooistra W.H.C.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 elongate shape has arisen twice in the diatoms. Journal of Phycology 44: 1340–1348.

O'leary M.A. (1999) Parsimony analysis of total evidence from extinct and extant taxa and the cetacean-artiodactyl question (Mammalia, Ungulata). Cladistics 15: 315–330.

Patterson C. (1988) Homology in classical and molecular biology. Molecular Biology and Evolution 5: 603–625.

Poe S., Swofford D.L. (1999) Taxon sampling revisited. Nature 398: 300–301.

Pollock D.D., Zwickl D.J., Mcguire J.A., Hillis D.M. (2002) Increased taxon sampling is advantageous for phylogenetic inference. Systematic Biology 51: 664–671.

Posada D., Buckley T.R. (2004) Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology 53: 793–808.

Posada D., Crandall K.A. (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818.

Rampen S.W., Schouten S., Panoto F.E., Brink M., Andersen R.A., Muyzer G., Abbas B., Damsté J.S.S. (2009) Phylogenetic position of Attheya longicornis and Attheya septentrionals (Bacillariophyta). Journal of Phycology 45: 444–453.

Richlen M.L., Barber P.H. (2005) A technique for the rapid extraction of microalgal DNA from single live and preserved cells. Molecular Ecology Notes 5: 688–691.

Round F.E., Crawford R.M. (1981) The lines of evolution of the Bacillariophyta .1. Origin. Proceedings of the Royal Society of London Series B-Biological Sciences 211: 237–260.

Round F.E., Crawford R.M. (1984) The lines of evolution of the Bacillariophyta. 2. The Centric Series. Proceedings of the Royal Society of London Series B-Biological Sciences 221: 169–188.

Round F.E., Crawford R.M., Mann D.G. (1990) The diatoms: Biology and morphology of the genera. Cambridge, Cambridge University Press.

Ruck E.C., Kociolek P. (2004) Preliminary phylogeny of the family Surirellaceae. Bibliotheca Diatomologica 50: 1–236.

Sato S., Kooistra W.H.C.F., Watanabe T., Matsumoto S., Medlin L.K. (2008a) A new araphid diatom genus Psammoneis gen. nov. (Plagiogrammaceae, Bacillariophyta) with three new species based on SSU and LSU rDNA sequence data and morphology. Phycologia 47: 510–528.

Sato S., Mann D.G., Matsumoto S., Medlin L.K. (2008b) Pseudostriatella (Bacillariophyta): a description of a new araphid diatom genus based on observations of frustule and auxospore structure and 18S rDNA phylogeny. Phycologia 47: 371–391.

Sato S., Watanabe T., Crawford R.M., Kooistra W.H.C.F., Medlin L.K. (2008c) Morphology of four plagiogrammacean diatoms; Dimeregramma minor var. nana, Neofragilaria nicobarica, Plagiogramma atomus and Psammogramma vigoensis gen. et sp. nov., and their phylogenetic relationship inferred from partial large subunit rDNA. Phycological Research 56: 255–268.

Shimodaira H. (2002) An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51: 492–508.

Shimodaira H., Hasegawa M. (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17: 1246–1247.

Simonsen R. (1979) The diatom system: Ideas on phylogeny. Bacillaria 2: 9–71.

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.

Sorhannus U. (2004) Diatom phylogenetics inferred based on direct optimization of nuclear-encoded SSU rRNA sequences. Cladistics 20: 487–497.

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

Springer M.S., Teeling E.C., Madsen O., Stanhope M.J., De Jong W.W. (2001) Integrated fossil and molecular data reconstruct bat echolocation. Proceedings of the National Academy of Sciences of the United States of America 98: 6241–6246.

Stamatakis A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.

Steinecke F. (1931) Die Phylogenie der Algophyten. Schriften der Königsberger Geleherten Gesellschaft 8: 127–298.

Strait D.S., Grine F.E. (2004) Inferring hominoid and early hominid phylogeny using craniodental characters: the role of fossil taxa. Journal of Human Evolution 47: 399–452.

Tamura M., Shimada S., Horiguchi T. (2005) Galeidinium rugatum gen. et sp. nov. (Dinophyceae), a new coccoid dinoflagellate with a diatom endosymbiont. Journal of Phycology 41: 658–671.

Theriot E., Bradbury J.P. (1987) Mesodictyon, a new fossil genus of the centric diatom family Thalassiosiraceae from the Miocene Chalk Hills Formation, western Snake River Plain, Idaho. Micropaleontology (New York) 33: 356–367.

Theriot E., Stoermer E., Håkansson H. (1987) Taxonomic interpretation of the rimoportula of freshwater genera in the centric diatom family Thalassiosiraceae. Diatom Research 2: 251–265.

Theriot E.C. (1992) Clusters, species concepts, and morphological evolution of diatoms. Systematic Biology 41: 141–157.

Theriot E.C., Cannone J.J., Gutell R.R., Alverson A.J. (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. (in press) Status of the pursuit of the diatom phylogeny: Are traditional views and new molecular paradigms really that different? In: Seckbach J., Kociolek, J.P. (eds) Cellular Origins, Life in Extreme Environments. CRC Press, Boca Raton.

Verbruggen H., Theriot E.C. (2008) Building trees of algae: Some advances in phylogenetic and evolutionary analysis. European Journal of Phycology 43: 229–252.

Williams D.M. (1990) Cladistic analysis of some freshwater araphid diatoms (Bacillariophyta) with particular reference to Diatoma and Meridion. Plant Systematics and Evolution 171: 89–97.

Williams D.M. (2007) Diatom phylogeny: Fossils, molecules and the extinction of evidence. Comptes Rendus Palevol 6: 505–514.

Williams D.M., Kociolek J.P. (2007) Pursuit of a natural classification of diatoms: History, monophyly and the rejection of paraphyletic taxa. European Journal of Phycology 42: 313–319.

Wortley A.H., Rudall P.J., Harris D.J., Scotland R.W. (2005) How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. Systematic Biology 54: 697–709.

Zwickl D.J., Hillis D.M. (2002) Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology 51: 588–598.

Paper based on a keynote presented during the Symposium "Diatom Taxonomy in the 21st Century" (Meise 2009). Manuscript received 8 Jan. 2010; accepted in revised version 15 Jun. 2010.

Communicating Editor: Bart Van de Vijver.