PNAS RESEARCH ARTICLE | EVOLUTION

Phylogenomics reveals the slow-burning fuse of diatom evolution

Andrew J. Alverson[a,1], Wade R. Roberts[a], Elizabeth C. Ruck[a], Teofil Nakov[b], Matthew P. Ashworth[c], Karolina Bryłka[d], Kala M. Downey[a], J. Patrick Kociolek[e], Matthew Parks[f], Eveline Pinseel[g], Edward C. Theriot[h], Simon P. Tye[a], Andrzej Witkowski[i], Jeremy M. Beaulieu[a], and Norman J. Wickett[j]

Affiliations are included on p. 7.

Edited by Fabien Burki, Uppsala Universitet, Uppsala, Sweden; received January 7, 2025; accepted April 17, 2025 by Editorial Board Member David Jablonski

Evolution is often uneven in its pace and outcomes, with long periods of stasis interrupted by abrupt increases in morphological and ecological disparity. With thousands of gene histories, phylogenomics can uncover the genomic signatures of these broad macroevolutionary trends. Diatoms are a species-rich lineage of microeukaryotes that contribute greatly to the global cycling of carbon, oxygen, and silica, which they use to build elaborately structured cell walls. We combined fossil information with newly sequenced transcriptomes from 181 diverse diatom species to reconstruct the pattern, timing, and genomic context of major evolutionary transitions. Diatoms originated 270 Mya, and after >100 My of relative stasis in morphology and ecology, a radiation near the Jurassic–Cretaceous boundary led to the diversity of habitats and cell wall architectures characteristic of modern diatoms. This transition was marked by a genome duplication and high levels of gene tree discordance. However, short generation times increase the probability of coalescence between speciation events, minimizing the impacts of incomplete lineage sorting and implicating sequence saturation and gene tree error as the main sources of discordance. Nevertheless, a rigorous tree-based approach to ortholog selection resulted in strongly supported relationships, including some that were uncertain previously. Three pulses of accelerated speciation were detected, two of which were associated with the evolution of novel traits and ecological transitions. The first 100 My of diatom evolution was a slow-burning fuse that led to a burst of innovations in ecology, morphology, and life history that are hallmarks of contemporary diatom assemblages.

concordance | discordance | diversification | incomplete lineage sorting | microbes

Significance

The tree of life is dotted with inflection points that changed the course of evolution for some lineages, with vast amounts of morphological diversity and species richness generated in short periods of time. We combined fossil information with thousands of genes to reconstruct the evolutionary history of diatoms, a species-rich lineage of photosynthetic algae that produce 20% of Earth's oxygen and form the base of aquatic food webs. Fossil-informed phylogenomics showed that the first 100 My of diatom evolution were highly constrained, but this long and slow start set the stage for a burst of innovations in ecology, morphology, and life history that are hallmarks of modern diatoms.

The evolutionary process conceived by Darwin was one of slow and gradual change, prompting fantastical explanations for examples that countered his theory, such as the sudden rise of angiosperms in the fossil record (1). Although the tree of life contains countless gradually evolving, stalwart lineages of the kind imagined by Darwin, a relatively small number of clades comprise a disproportionate share of Earth's diversity (2), highlighting unevenness in the pace and outcomes of evolutionary change. Characterizing diversification dynamics through time—and correlating them with genomic changes, environmental transitions, or character innovations—are principal goals of modern macroevolutionary biology. We understand now, for example, that the rapid diversification of angiosperms observed by Darwin was not a singular event but involved repeated bursts of speciation linked to intrinsic and extrinsic factors (3–6). Patterns like these have inspired general models that describe, in one form or another, periods of relative stasis separated by abrupt increases in rates of speciation or innovation (7–9).

Diversification histories are commonly inferred from species phylogenies based on a handful of genes that resolve some relationships but fail to resolve others. This has given the impression that hard phylogenetic problems—like rapid diversification events—can be resolved with more data (10–12). The introduction of genome-scale datasets to phylogenetics has shown, however, that some relationships are virtually unresolvable because of conflicting signal (discordance) in the underlying gene histories (11, 13). High levels of discordance often coincide with rapid radiations, morphological innovations, or the emergence of major clades (14), so identifying the causes of discordance can reveal sources of adaptive evolution or population-level processes underway during key points in the history of a lineage (15, 16).

We constructed a large phylogenomic dataset to characterize the pattern, timing, and genomic context of major transitions in diatom evolution. Diatoms are eukaryotic microalgae that play cornerstone roles in marine and freshwater ecosystems. Photosynthesis by

Author contributions: A.J.A., W.R.R., and N.J.W. designed research; A.J.A., E.C.R., T.N., K.B., K.M.D., M.P., E.P., S.P.T., J.M.B., and N.J.W. performed research; A.J.A., M.P.A., J.P.K., E.C.T., and A.W. contributed new reagents/analytic tools; A.J.A. and W.R.R. analyzed data; and A.J.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. F.B. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: aja@uark.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2500153122/-/DCSupplemental.

Published May 29, 2025.

diatoms produces an estimated 20% of Earth's oxygen, driving the "biological pump" that exports fixed atmospheric carbon to the sea floor (17). Their storage products include long-chain fatty acids, making diatoms a preferred food source for invertebrate grazers and placing them at the base of aquatic food webs (18). The elaborate shapes and patterns of their silicon-based cell walls provide the basis for identifying and classifying an estimated 30,000 to 200,000 diatom species (19, 20). This stands in stark contrast to their sister lineage, Parmales, which includes fewer than a dozen species (21). Despite their global ecological importance and outsized species richness, our understanding of the diatom phylogeny is based on a small number of genes (22). Genome-scale sequencing efforts, meanwhile, have focused almost exclusively on marine planktonic species (23), which represent a small fraction of diatom diversity (24). We sequenced transcriptomes for 181 diatom species representing 132 genera. Taxon sampling extended beyond the marine plankton to include species from freshwater, brackish, and terrestrial ecosystems; benthic and epiphytic habitats; and species with novel metabolism or pigmentation. Phylogenomics showed that, after a slow start, a rapid radiation gave rise to the extraordinary diversity in habitats and cell wall architectures characteristic of modern diatoms.

Results

High-Quality Transcriptomes for Diverse Diatoms. We generated transcriptomes for 181 diatom species from diverse habitats and ecosystems (Fig. 1*A*). Representation of nonmarine species increased by >50-fold and benthic species by nearly 10-fold (Fig. 1*A*). These increases largely reflected targeted sampling of pennate diatoms, which were undersampled previously (25). The number of genera available for phylogenomic analysis increased from 43 to 175 (Fig. 1*A*). New transcriptomes (n = 181) were added to previously sequenced genomes (n = 15) and transcriptomes (n = 88) for a total dataset size of 284 taxa (Fig. 1*A*).

Transcriptomes from nonmodel organisms present challenges for assembly of sequencing reads, transcript redundancy, and ortholog identification, but pipelines that address these issues have encouraged widespread use of de novo transcriptomes in phylogenetics (26, 27). After aggressive filtering, the newly sequenced transcriptomes recovered, on average, 69% of conserved eukaryotic orthologs (BUSCOs; Fig. 1*B*) and 82% of stramenopile BUSCOs (Fig. 1*C*). For comparison, reference genomes from diatoms and other stramenopiles contained, on average, 80% eukaryotic and 95% stramenopile BUSCOs (Fig. 1 *B* and *C*).

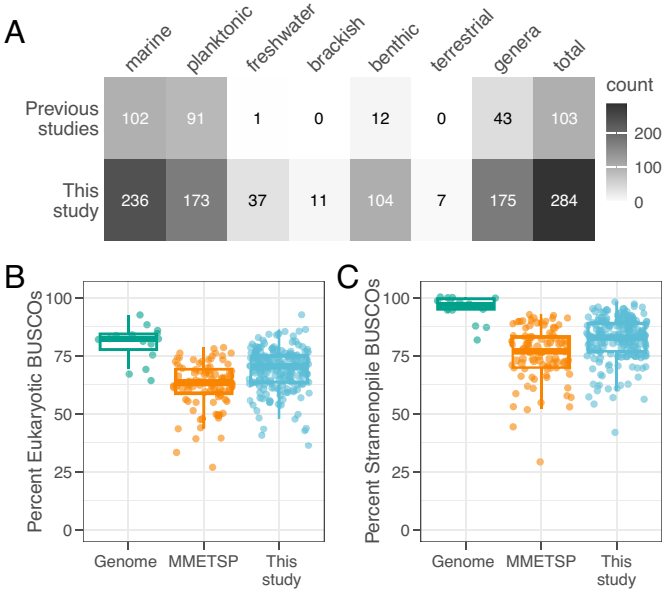

Fig. 1. High-quality transcriptomes from ecologically and phylogenetically diverse diatoms. Numbers represent total diatom strains, most of which are distinct species (*A*). Transcriptomes recovered a large percentage of conserved eukaryotic (*B*) and stramenopile (*C*) orthologs (BUSCOs). The estimated completeness of newly sequenced transcriptomes compared favorably to reassembled diatom transcriptomes from the MMETSP (23).

The Diatom Tree of Life. Twenty-four datasets ranging in size from 127 to 2,890 orthologs were assembled using different cutoffs for species representation (taxon occupancy) and phylogenetic signal within each ortholog (*SI Appendix*, Table S1). For some datasets, an additional quality-filtering step removed orthologs that failed to recover two clades with indisputable evidence for monophyly: Thalassiosirales and raphid pennates (28). Datasets ranged in size from 41,732 to 714,629 aligned amino acids for phylogenetic analysis.

Species trees were reconstructed using gene tree summary methods with ASTRAL or maximum likelihood analysis of concatenated orthologs with IQ-TREE (*SI Appendix*, Table S1). Tree topologies varied based on dataset and optimality criterion (*SI Appendix*, Fig. S1), with most of the variation caused by instability among the major lineages of polar centrics (Mediophyceae) and a subset of raphid pennates (*SI Appendix*, Figs. S2 and S3). Alternative topology tests of the mediophyte arrangements supported species trees inferred by maximum likelihood analysis with profile mixture models and orthologs that passed the monophyly filtering step described above (Fig. 2 and *SI Appendix*, Table S2). Mappings of key traits onto this topology are shown in *SI Appendix*, Fig. S4.

Several key relationships resolved identically in all analyses (*SI Appendix*, Fig. S2). All trees recovered the same paraphyletic arrangement of the oogamous and predominantly marine planktonic radial centric diatoms as the earliest splits in the tree. *Corethron* was sister to all diatoms, followed by a grade of Leptocylindrales and two clades of radial centrics (e.g., *Paralia*, *Rhizosolenia*, and *Coscinodiscus*) (Fig. 2: clades 1 to 4). Although most lineages of polar centrics were consistently recovered as monophyletic, the arrangements of those lineages—particularly the placements of *Attheya* and Toxariales—were among the least stable (*SI Appendix*, Fig. S2). The causes of this uncertainty are discussed below.

The transitions from oogamous to isogamous gametes and radial to bilateral cell symmetry in the common ancestor of araphid pennates, followed by the evolution of gliding motility in raphid pennates, represent some of the most consequential events in diatom evolution (*SI Appendix*, Fig. S4, ref. 29). All analyses recovered monophyletic pennates (Fig. 2: araphid + raphid pennates) and raphid pennates (Fig. 2: clade 10). The species-depauperate Striatellales has been notoriously difficult to place phylogenetically (22, 28). With genome-scale data for *Striatella* and *Pseudostriatella*, this lineage was placed unambiguously as lone sister to the pennate clade (Fig. 2: clade 7). Including Striatellales, araphid pennates were paraphyletic and divided among three clades (Fig. 2: clades 7 to 9).

The raphe (*SI Appendix*, Fig. S4) is a key innovation that enables cellular locomotion, and most raphe-bearing diatoms have a raphe on each half of their bipartite cell wall. Within raphid pennates, the deepest splits were resolved consistently across analyses and comprised a paraphyletic grade of Eunotiales, *Amphora*, Bacillariales, and one of many naviculoid clades (Fig. 2 and *SI Appendix*, Figs. S3–S5). One of the two raphes was lost three times independently (Fig. 2*A*, Achnanthales), and a separate loss of the entire

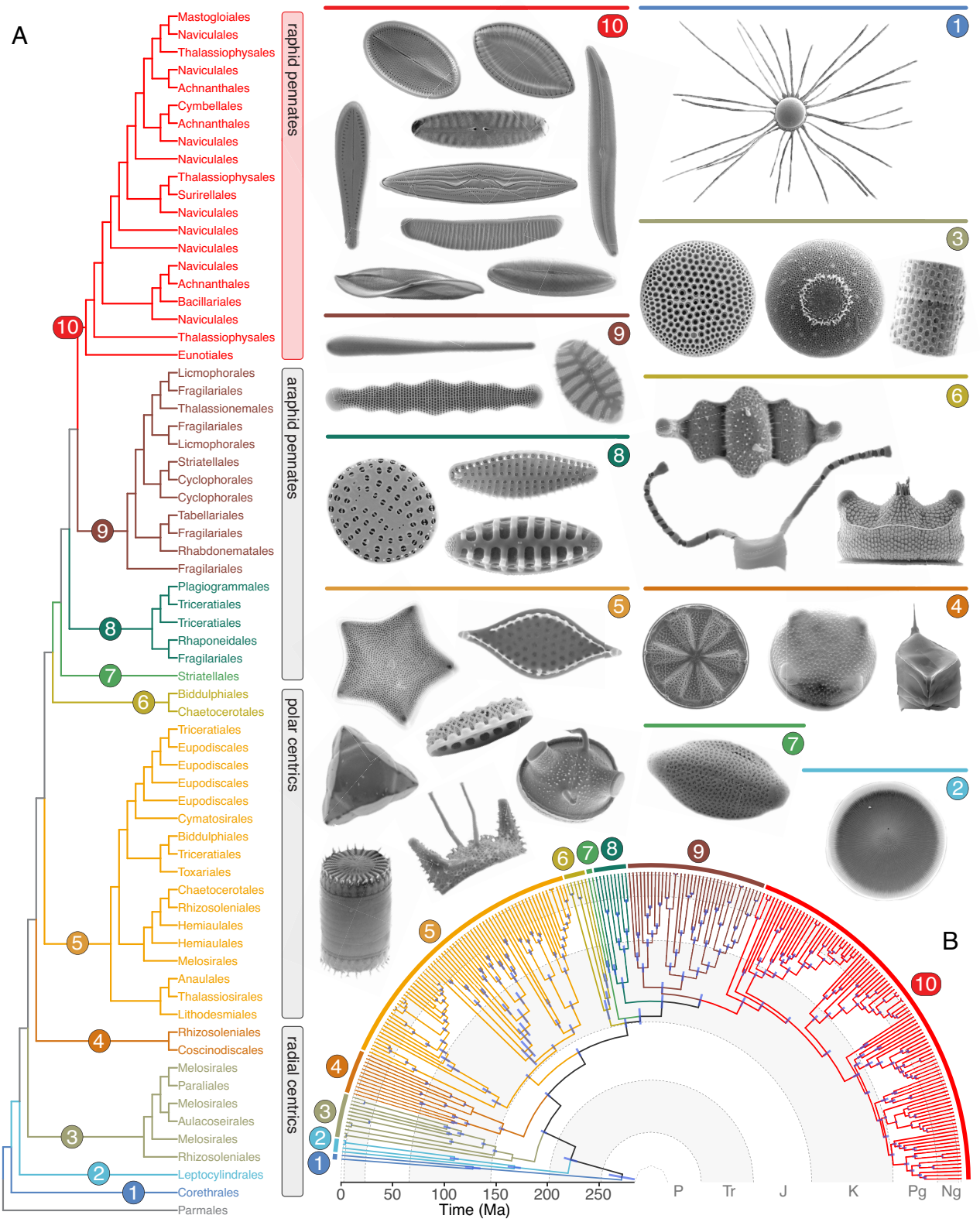

Fig. 2. The diatom tree of life. Relationships are based on maximum likelihood analysis of 240 genes and 171,434 amino acids under the LG+PMSF(C20)+F+G model. Ten numbered clades are highlighted to facilitate discussion. (*A*) Order-level phylogeny with major groups identified by common name. (*B*) Time-calibrated phylogeny. Node bars represent CI of divergence time estimates. Geologic periods are indicated as: P (Permian), Tr (Triassic), J (Jurassic), K (Cretaceous), Pg (Paleogene), and Ng (Neogene). Scanning electron microscope images show representative diatom species. Fully labeled phylogram and chronogram figures are available as *SI Appendix*, Figs. S5 and S6.

raphe system occurred in the cave-dwelling genus, *Diprora* (30). In terms of species richness and biological innovations, Bacillariales represents one of the most remarkable diversifications among all diatoms (31). Despite strong evidence for monophyly from their raphe structure, phylogenetic studies paradoxically have reconstructed nonmonophyletic Bacillariales (31). Our analyses resolved

this anomaly, placing *Craspedostauros*, *Achnanthes*, and *Staurotropis* outside and sister to a monophyletic Bacillariales in all but three species trees (*SI Appendix*, Fig. S3).

The Long Road to Innovate. With 18 validated fossil calibrations (32) and 44 clocklike genes, the diatom crown age was estimated as 260 to 286 Ma ($\bar{x} = 271$) (Fig. 2 and *SI Appendix*, Fig. S6). The first roughly 100 My of diatom evolution was limited to taxa with relatively uniform reproduction (oogamy), cell architecture (rotational symmetry), and ecology (marine plankton, with limited incursions into freshwater) (Fig. 2: clades 1 to 4). The split between these and remaining diatoms, which capture the vast share of their ecological and morphological diversity, occurred roughly 165 Mya (Fig. 2: clades 5 to 10). The diversity of shapes, polarities, and ultrastructural features of their siliceous cell walls are hallmarks of modern diatoms, and most of this diversity traces to the rapid succession of splits leading to the major lineages of polar centric and early pennate diatoms around the Jurassic-Cretaceous boundary roughly 156 to 130 Mya (Fig. 2). This includes the transition from oogamy to isogamy, and rotational to bilateral symmetry, that occurred 140 Mya in the common ancestor of pennate diatoms (Fig. 2: clades 7 to 10). This was followed by the origin of raphe-enabled locomotion 104 Mya (Fig. 2: clade 10).

Key Transitions Are Associated with High Levels of Discordance. The 24 inferred species trees showed important differences in topology (*SI Appendix*, Figs. S1–S3) caused by disagreement among orthologs (gene discordance) and amino acid sites (site discordance) in some parts of the tree. Relationships that resolved identically across all trees had greater gene and site concordance (Fig. 3 *A*–*D* and *SI Appendix*, Fig. S7). The deepest splits had modest concordance levels, and concordance generally increased toward the present (Fig. 3*C* and *SI Appendix*, Fig. S7). Areas of exceptionally low concordance were concentrated around nested branches that split in rapid succession. Two parts of the tree stood out in this regard.

Numerous innovations poured out from the eventual diversification of polar-centric diatoms. As shown by their overlapping age estimates, the major lineages of polar centrics originated within a short timespan (Fig. 3*C*: circles). This rapid diversification was marked by some of the lowest levels of gene and site concordance (Fig. 3*C* and *SI Appendix*, Fig. S7: circles) and, consequently, relationships in this part of the tree were among the most difficult to resolve (*SI Appendix*, Fig. S2). Finally, although the deepest bipartitions in the raphid pennate clade were well resolved, many of the more recent backbone splits were not (*SI Appendix*, Fig. S3). These, too, had among the lowest gene and site concordance factors (Fig. 3*C* and *SI Appendix*, Fig. S7: diamonds).

Discordance Is Associated with Low Phylogenetic Signal and High Sequence Saturation. Discordance can have biological causes such as incomplete lineage sorting (ILS) or technical causes such as error in reconstructing gene trees. ILS will be problematic when polymorphism levels are high (e.g., with large populations or high mutation rates) or when polymorphisms have insufficient time between speciation events to become fixed (e.g., when generation times are long). We performed two sets of simulations to test whether the discordance observed here was due to ILS. First, we simulated gene trees under a range of values for theta ($4\ N_e\mu$) and calculated Robinson–Foulds distances between the species tree and simulated gene trees. Across a range of realistic values for theta, levels of discordance in simulated datasets (including ones that incorporated gene tree error) were in all cases substantially lower than observed levels (Fig. 3*D*). Next, we simulated gene trees under a range of plausible mutation rates and generation times, but here, too, the average discordance in simulated datasets was far lower than we observed empirically (Fig. 3*E*). Taken together, models based on relevant population genetic parameters suggest that ILS is not an important source of discordance in diatoms and probably other microbes as well. Partitioning of gene concordance factors at recalcitrant nodes suggested most discordance was due to paraphyly (*SI Appendix*, Fig. S7*A*, high gDFP) rather than bias toward particular alternative topologies (*SI Appendix*, Fig. S7*A*, low gDF1 and gDF2), highlighting low phylogenetic signal as the source of discordance rather than a biological cause.

Next, we explored whether two common causes of error in gene tree estimation—lack of phylogenetic signal and sequence saturation—contributed to gene tree discordance. To determine whether our dataset had sufficient phylogenetic information to resolve recalcitrant nodes, we estimated how much information was necessary to resolve each node in the species tree by simulating 1,290 alignments and subsetting them to produce replicate datasets of size 5 to 1,290 loci (33, 34). We constructed a species tree for each simulated dataset and recorded the minimum number of loci necessary to recover each branch across 100% of the replicates in each dataset size class. Some branches were resolvable with traditionally sized datasets (5 to 10 loci), and most were resolvable with fewer than 200 loci (Fig. 3*A*). Recalcitrant polar centric branches required 500 to 1,000 loci, including the branch placing *Attheya*+Biddulphiales sister to pennates (Fig. 3*A*: circles). Branches surrounding Toxariales were never consistently recovered, even with 1,290 loci. Recalcitrant raphid pennate nodes all required ≥500 loci to resolve, and one branch was unresolvable with 1,290 loci (Fig. 3*A*: diamonds).

We then explored whether sequence saturation, where sites in the genome experience multiple substitutions that overwrite true phylogenetic signal, was a possible source of gene tree error. In saturated datasets, substitution models will not fully compensate for overwritten substitutions, so correlations of pairwise model-corrected versus raw genetic distances between species will be low (11). We calculated the slope of the regression between these two distance measures for each ortholog and found low slopes consistent with strong sequence saturation across nearly all loci (Fig. 3*B*). The joint effects of a rapid radiation, lack of phylogenetic information, and sequence saturation were likely contributors to error in reconstructing gene trees and, consequently, some species relationships.

Net Diversification Rates Are Lowest in Radial Centric and Highest in Pennate Diatoms. A character-independent measure of species diversification showed that net diversification (speciation *minus* extinction) and turnover (speciation *plus* extinction) were lowest in the radial centric lineages (Fig. 4). Pennate diatoms experienced an overall increase in the rate of speciation that, in general, was not balanced by increases in the rate of extinction (Fig. 4). Three clades showed exceptionally high levels of both net diversification and net turnover. One rate shift occurred in Thalassiosirales, a lineage noteworthy for its high abundance in the global ocean (35), repeated colonizations of freshwaters (36), and possession of a unique structure that facilitates colony formation and buoyancy control through the production of β-chitin threads (37, 38). Here, however, the turnover signal was not consistent across all bootstrap trees (Fig. 4*B*, yellow and purple lines in Thalassiosirales), highlighting the importance of accounting for topological uncertainty in diversification analyses. Two other rate increases occurred within the raphid pennate

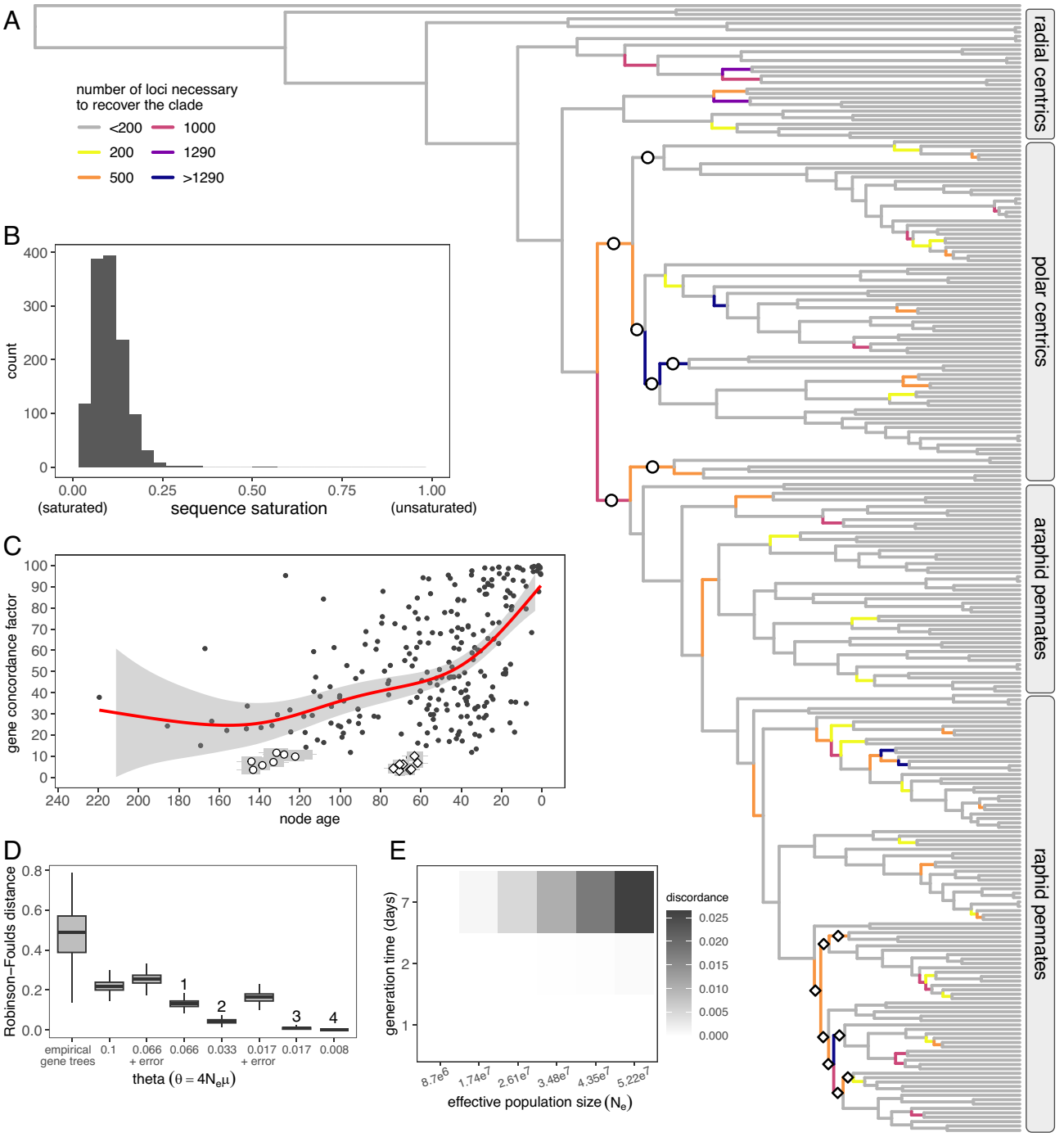

Fig. 3. The rapid radiations of polar centric and raphid pennate diatoms were marked by low gene concordance and low phylogenetic signal. (*A*) Species phylogeny with branches colored by the amount of data necessary to resolve each branch across simulated datasets. Most branches were consistently resolved with <200 loci, but some were unresolvable with 1,290 loci. (*B*) Sequence saturation across 1,290 loci, calculated as the slope of a regression between pairwise patristic distances and uncorrected genetic distances. (*C*) Gene concordance factor for each node in the species tree. Nodes with open circles or diamonds correspond to identically marked nodes in panel (*A*); gray bars represent 95% CI around age estimates; the red line indicates the fit of a generalized additive model. (*D*) Average distance (discordance) between the species tree and empirical gene trees ($\bar{x} = 0.48$) far exceed expected levels based on coalescent simulations across a range of plausible values for theta, parameterized with values from diatoms (*Fragilariopsis* [1] and *Phaeodactylum* [3]), green algae (*Chlamydomonas* [2]), and coccolithophores (*Gephyrocapsa* [4]). (*E*) Average discordance between the species tree and empirical gene trees ($\bar{x} = 0.48$) far exceeded discordance levels with gene trees simulated across a range of realistic values for effective population size and generation time.

clade, a lineage of predominantly benthic species already known to have an increased diversification rate associated with the raphe (*SI Appendix*, Fig. S4), a structure that enables attachment and active locomotion along substrates (29). One upward shift occurred in a subclade of Bacillariales marked by adaptations to two recent ecological transitions—one involving the evolution of ice-binding proteins coincident with the colonization of polar ice habitats in *Fragilariopsis* (39) and a second involving the evolution of toxin production associated with secondary recolonization of the plankton in *Pseudo-nitzschia* (40). A third rate shift occurred in a clade of Naviculales that includes the model species, *Seminavis robusta* (41).

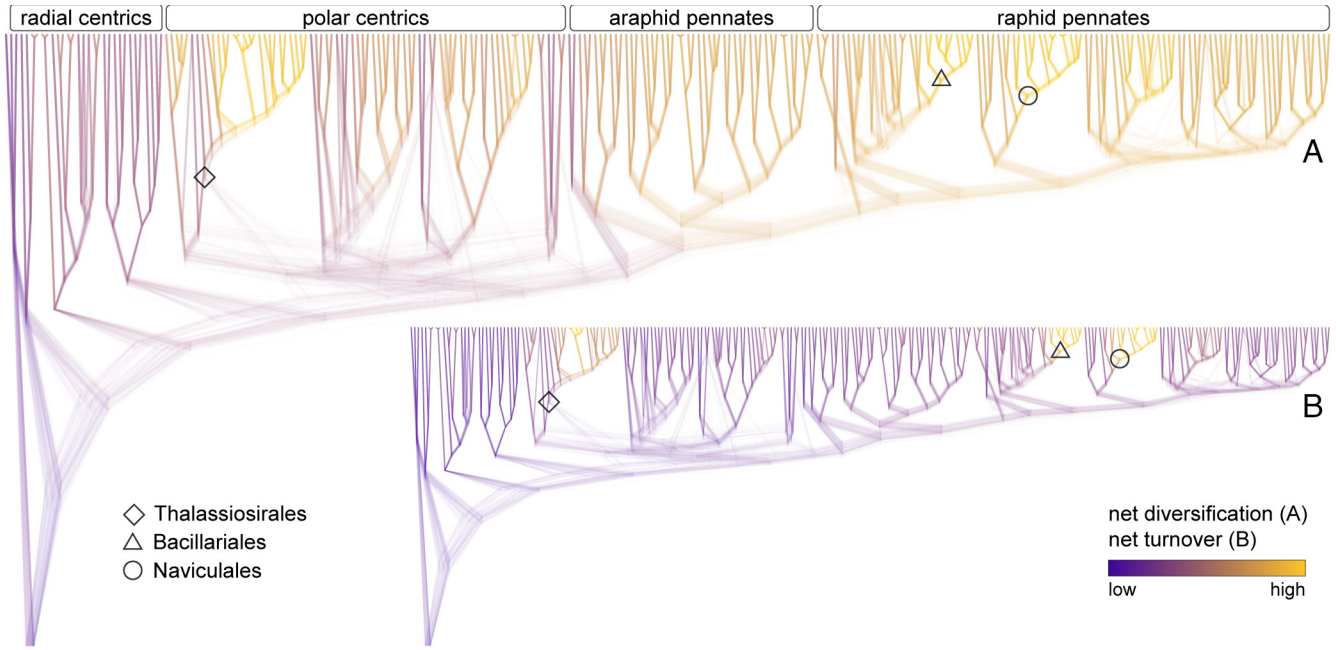

Fig. 4. Species diversification and turnover are lowest in radial centric diatoms and highest in pennate diatoms. (*A*) Three pulses of net diversification (speciation *minus* extinction) occurred in marked lineages. (*B*) Net turnover (speciation *plus* extinction) was also highest in Bacillariales and Naviculales but was sensitive to tree topology in Thalassiosirales. The analysis was run on 100 bootstrap trees that captured topological uncertainty within polar centrics and raphid pennates.

The Diatom Classification System Is Broken. The current class- and order-level taxonomic classifications of diatoms are poor surrogates for phylogenetic diversity. Our analyses decisively divided Coscinodiscophyceae (radial centrics) into a grade of four separate lineages (Fig. 2, clades 1 to 4). Although four analyses recovered a monophyletic Mediophyceae (polar centrics), the remaining 20 analyses placed *Attheya* and relatives outside mediophytes (Fig. 2 and *SI Appendix*, Fig. S2). The currently recognized pennate diatom class (Bacillariophyceae) is monophyletic. We adopted the National Center for Biotechnology Information's widely used Taxonomy database to illustrate the mismatch between the phylogeny and order-level classification, and among the 34 taxonomic orders of diatoms in the dataset, 14 of them were nonmonophyletic and spread across two to as many as nine distinct clades. Some orders were divided between different taxonomic classes (Fig. 2*A*).

Discussion

Species trees inferred from genome-scale datasets provide near certainty in some relationships and highlight evolutionary inflection points that make other relationships difficult, if not impossible, to resolve (16). The influx of phylogenomic data presented here advanced our understanding of the diatom phylogeny beyond what was possible with a small number of genes. Resolving some relationships, such as the placement of Striatellales next to pennates or the recovery of Bacillariales as monophyletic, simply required more data. Other historically recalcitrant relationships, such as the placement of Toxariales (42, 43), are less certain, but the thousands of genes analyzed here revealed why. Toxariales is one of several lineages descended from the near-simultaneous diversification of polar centric and pennate diatoms, which together account for the vast share of diatom diversity in life history, morphology, ecology, reproductive biology, and species richness (19). Genome duplications have preceded the origins or diversifications of major plant and animal groups (4, 44), and similarly, a genome duplication was predicted in the common ancestor of polar centric and pennate diatoms (45). Identifying the duplicated genes and pathways that trace back to this event may shed light on one of the most pivotal periods of diatom evolution.

The diversification of polar centrics was marked by high levels of gene tree discordance, which has been linked to morphological innovations and the emergence of major plant and animal clades (5, 14, 46). Discordance was also rife in the rapid diversification of a subclade of raphid pennates whose members include the morphologically derived Surirellales; asymmetrically shaped, tube-forming Cymbellales; and species that lost one or both raphes. Although genome duplication may have contributed to conflict in deep polar centric branches (16), simulations parameterized with realistic values for effective population size, mutation rate, and generation time suggest that another biological source of conflict, ILS, is unlikely for diatoms and other microbes, setting them apart from plants and animals with longer generation times or smaller effective population sizes (13). Our simulations assumed uniform population genetic properties across the tree and through time, which is probably unrealistic for a group as large and diverse as diatoms. As a result, some parts of the tree may be more susceptible to ILS, including branches with discordance factors showing substantial, near-equal support for specific alternative relationships (e.g., *Surirella*+*Campylodiscus*) (47). Introgression is another potential source of conflict in diatoms (48) but at deep phylogenetic scales, most conflict appears to be due to error in gene tree reconstruction caused by low phylogenetic signal and high sequence saturation (25, 36).

The phylogeny and genomic data presented here provide a foundation for identifying the developmental enablers, key characters, and environmental drivers of diatom diversification (49). This is challenged, however, by our limited understanding of the functional morphology and adaptive significance of most cell wall structures. To the extent that shape itself is important, the transition from scaled auxospores to ones with perizonial bands and incunabular strips allowed for reshaping of ancestral rotationally symmetric cells into complex multipolar (Fig. 2: clades 5 and 6)

and longitudinal forms (Fig. 2: clades 7 to 10) (19, 50). This was followed by the evolution of other structures whose arrangements track cell polarity, such as the ocelli of *Pleurosira*, pseudocelli of *Biddulphia*, and setae of *Chaetoceros* (Fig. 2 and *SI Appendix*, Fig. S4). Finally, taxonomic classifications are often used as surrogate measures of biodiversity (51). The results presented here highlight the need to establish a phylogenetic classification of diatoms that better reflects their evolutionary history.

Diatoms have numerous features that distinguish them from their closest relatives (52, 53) and may be associated with their vast species richness and global biomass (54). At least two of these features—a bipartite cell wall and diplontic life cycle—were already present in the common ancestor of extant diatoms. Nevertheless, the genomic resources and time-calibrated phylogeny presented here suggest the first 100 My of diatom evolution was relatively constrained, limited mainly to rotationally symmetric taxa in the marine environment. Although the rich Cenozoic fossil record of diatoms has provided insights into broad macroevolutionary trends (29, 55), the earlier record is either sparse (Cretaceous) or nonexistent (Jurassic and below) (56, 57), so we cannot rule out that extinction erased a greater degree of morphological and ecological diversity than suggested by analyses of extant species only (56, 57). The oldest fossil diatom deposits from the Lower Cretaceous contain dozens of morphologically diverse radial and polar centric species which, together with the results presented here, point to a much deeper history than captured by the existing fossil record alone (57). Fossil-informed phylogenomics of diatoms suggest an extended early phase of limited innovation—consistent with a long but not invisible fuse (8)—which set the stage for a series of diversifications that led to their extraordinary species richness, morphological and genomic complexity, and rise to prominence in aquatic ecosystems worldwide.

Materials and Methods

A detailed overview of the methods is available as *SI Appendix* and through a permanent online Zenodo repository (58). Culture metadata and database accession numbers for sequencing reads are available in Dataset S1. Alignments, ortholog trees, and species trees necessary to replicate these analyses are available on Zenodo.

Laboratory Methods. Individual cells were isolated from environmental samples or procured from culture collections and grown in conditions matching their native environment. Total RNA was extracted from cells in exponential growth, and stranded mRNA libraries were sequenced on the Illumina DNA sequencing platform.

Transcriptome Processing and Ortholog Selection. Transcriptome assembly and filtering steps were performed for newly generated transcriptomes and, to improve and standardize assemblies, Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) transcriptomes (23). Processing and assembly of RNA-seq reads followed the Oyster River Protocol (59). Transcriptome filtering and ortholog selection followed a modified version of the phylogenomic pipeline of Yang and Smith (26), which was designed to overcome the challenges of de novo transcriptomes from nonmodel organisms. Transcriptome quality and completeness was estimated with BUSCO (60).

Orthologs were selected for phylogenetic analyses based on cutoffs for taxon occupancy and proportion of parsimony-informative sites (*SI Appendix*, Table S1). Two lineages, Thalassiosirales and raphid pennates, have unequivocal evidence for monophyly from morphological synapomorphies and decades of phylogenetic work (28). An additional topology-based filtering step removed orthologs that failed to recover monophyly of these groups due to weak or misleading signal.

Species Trees and Coalescent Simulations. Species trees were inferred using gene tree summary methods with ASTRAL (61) or ASTRAL-Pro (62) and maximum likelihood analysis of concatenated ortholog matrices with IQ-TREE (63). Maximum likelihood analyses used standard models and profile mixture models, some with Posterior Mean Site Frequency (PMSF) approximation (64), to account for among-site heterogeneity in amino acid composition (64, 65). These models do not account for rate or compositional heterogeneity across lineages, however. Coalescent simulations were performed with DendroPy (66) and IQ-TREE.

Divergence Times. An extensive survey identified strongly supported diatom fossils for calibration of molecular clocks (32). From these, we identified 18 clades with ≥2 fossils suitable for estimating minimum and maximum bounds of clade ages (*SI Appendix*, Fig. S8 and Table S3). Divergence times were estimated with treePL (67) on 100 bootstrap trees that captured diverse resolutions of mediophyte and raphid pennate relationships, accounting for uncertainty in these and other parts of the tree (Fig. 4).

Diversification Analysis. Rates of net diversification were estimated using the trait-independent model implemented in MiSSE (68). The results were summarized across 100 time-calibrated bootstrap trees to account for uncertainty in the tree topology.

Data, Materials, and Software Availability. Sequencing reads were deposited in NCBI's SRA database under BioProject PRJNA1086848 (69). Data, code, and GitHub links to specific analyses are available from Zenodo (58).

ACKNOWLEDGMENTS. Jan Rines and Cory Gargas provided diatom strains. This work was supported by the NSF (Grants DEB-1353131 and DEB-1651087 to A.J.A.; DEB-1353152 to N.J.W.). This research used computational resources available through the Arkansas High Performance Computing Cluster, which was funded through multiple NSF grants and the Arkansas Economic Development Commission.

Author affiliations: [a]Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701; [b]Escuela de Biología, Centro de Investigación en Ciencias del Mar y Limología, Universidad de Costa Rica, San José 11501-2060, Costa Rica; [c]Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712; [d]Department of Geology, Lund University, Lund 223 62, Sweden; [e]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309; [f]Department of Biology, University of Central Oklahoma, Edmond, OK 73034; [g]Laboratory of Protistology en Aquatic Ecology, Department of Biology, Ghent University, Ghent 9000, Belgium ; [h]Department of Integrative Biology, University of Texas at Austin, Austin, TX 78712; [i]Faculty of Physical, Mathematical and Natural Sciences, Institute of Marine and Environmental Sciences, University of Szczecin, Szczecin 70-383, Poland; and [j]Department of Botany and Biodiversity Research, University of Vienna, Vienna 1030, Austria

1. W. E. Friedman, The meaning of Darwin's "abominable mystery". *Am. J. Bot.* **96**, 5–21 (2009).
2. B. B. Larsen, E. C. Miller, M. K. Rhodes, J. J. Wiens, Inordinate fondness multiplied and redistributed: The number of species on Earth and the new pie of life. *Q. Rev. Biol.* **92**, 229–265 (2017).
3. B. C. O'Meara *et al.*, Non-equilibrium dynamics and floral trait interactions shape extant angiosperm diversity. *Proc. Biol. Sci.* **283**, 20152304 (2016).
4. J. B. Landis *et al.*, Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
5. A. R. Zuntini *et al.*, Phylogenomics and the rise of the angiosperms. *Nature* **629**, 843–850 (2024).
6. S. Magallón, L. L. Sánchez-Reyes, S. L. Gómez-Acevedo, Thirty clues to the exceptional diversification of flowering plants. *Ann. Bot.* **123**, 491–503 (2019).
7. J. C. Uyeda, T. F. Hansen, S. J. Arnold, J. Pienaar, The million-year wait for macroevolutionary bursts. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15908–15913 (2011).
8. A. Cooper, R. Fortey, Evolutionary explosions and the phylogenetic fuse. *Trends Ecol. Evol.* **13**, 151–156 (1998).
9. S. J. Gould, N. Eldredge, Punctuated equilibrium comes of age. *Nature* **366**, 223–227 (1993).
10. A. H. Wortley, P. J. Rudall, D. J. Harris, R. W. Scotland, How much data are needed to resolve a difficult phylogeny?: Case study in Lamiales. *Syst. Biol.* **54**, 697–709 (2005).
11. H. Philippe *et al.*, Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol.* **9**, e1000602 (2011).
12. A. Rokas, S. B. Carroll, More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol. Biol. Evol.* **22**, 1337–1344 (2005).
13. J. C. Oliver, Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* **67**, 1823–1830 (2013).
14. C. Parins-Fukuchi, G. W. Stull, S. A. Smith, Phylogenomic conflict coincides with rapid morphological innovation. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2023058118 (2021).
15. J. B. Pease, D. C. Haak, M. W. Hahn, L. C. Moyle, Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* **14**, e1002379 (2016).
16. G. A. Bravo *et al.*, Embracing heterogeneity: Coalescing the tree of life and the future of phylogenomics. *PeerJ* **7**, e6399 (2019).

17. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
18. D. C. Müller-Navarra, M. T. Brett, A. M. Liston, C. R. Goldman, A highly unsaturated fatty acid predicts carbon transfer between primary producers and consumers. *Nature* **403**, 74–77 (2000).
19. F. E. Round, R. M. Crawford, D. G. Mann, *Diatoms: Biology and Morphology of the Genera* (Cambridge University Press, 1990).
20. D. G. Mann, P. Vanormelingen, An inordinate fondness? The number, distributions, and origins of diatom species *J. Eukaryot. Microbiol.* **60**, 414–420 (2013).
21. M. Ichinomiya *et al.*, Diversity and oceanic distribution of the Parmales (Bolidophyceae), a picoplanktonic group closely related to diatoms. *ISME J.* **10**, 2419–2434 (2016).
22. E. C. Theriot, M. Ashworth, E. Ruck, T. Nakov, R. K. Jansen, A preliminary multigene phylogeny of the diatoms (Bacillariophyta): Challenges for future research. *Plant Ecol. Evol.* **143**, 278–296 (2010).
23. P. J. Keeling *et al.*, The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
24. T. Nakov, J. M. Beaulieu, A. J. Alverson, Diatoms diversify and turn over faster in freshwater than marine environments. *Evolution* **73**, 2497–2511 (2019).
25. M. B. Parks, N. J. Wickett, A. J. Alverson, Signal, uncertainty, and conflict in phylogenomic data for a diverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *Mol. Biol. Evol.* **35**, 80–93 (2018).
26. Y. Yang, S. A. Smith, Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* **31**, 3081–3092 (2014).
27. One Thousand Plant Transcriptomes Initiative, One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685 (2019).
28. P. A. Sims, D. G. Mann, L. K. Medlin, Evolution of the diatoms: Insights from fossil, biological and molecular data. *Phycologia* **45**, 361–402 (2006).
29. T. Nakov, J. M. Beaulieu, A. J. Alverson, Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). *New Phytol.* **219**, 462–473 (2018).
30. J. P. Kociolek, J. G. Stepanek, R. L. Lowe, J. R. Johansen, A. R. Sherwood, Molecular data show the enigmatic cave-dwelling diatom *Diprora* (Bacillariophyceae) to be a raphid diatom. *Eur. J. Phycol.* **48**, 474–484 (2013).
31. D. G. Mann *et al.*, Ripe for reassessment: A synthesis of available molecular data for the speciose diatom family Bacillariaceae. *Mol. Phylogenet. Evol.* **158**, 106985 (2021).
32. K. Bryłka, M. P. Ashworth, A. J. Alverson, D. J. Conley, The Cretaceous Diatom Database: A tool for investigating early diatom evolution. *J. Phycol.* **60**, 1090–1104 (2024).
33. J. Stiller *et al.*, Complexity of avian evolution revealed by family-level genomes. *Nature* **629**, 851–860 (2024).
34. M. L. Borowiec *et al.*, Evaluating UCE data adequacy and integrating uncertainty in a comprehensive phylogeny of ants. *Syst. Biol.*, 10.1093/sysbio/syaf001 (2025).
35. S. Malviya *et al.*, Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1516–E1525 (2016).
36. W. R. Roberts, E. C. Ruck, K. M. Downey, E. Pinseel, A. J. Alverson, Resolving marine–freshwater transitions by diatoms through a fog of gene tree discordance. *Syst. Biol.* **72**, 984–997 (2023).
37. A. E. Walsby, A. Xypolyta, The form resistance of chitan fibres attached to the cells of *Thalassiosira fluviatilis* Hustedt. *Br. Phycol. J.* **12**, 215–223 (1977).
38. W. Herth, A special chitin-fibril-synthesizing apparatus in the centric diatom *Cyclotella*. *Sci. Nat.* **65**, 260–261 (1978).
39. M. Bayer-Giraldi, C. Uhlig, U. John, T. Mock, K. Valentin, Antifreeze proteins in polar sea ice diatoms: Diversity and gene expression in the genus *Fragilariopsis*. *Environ. Microbiol.* **12**, 1041–1052 (2010).
40. J. K. Brunson *et al.*, Biosynthesis of the neurotoxin domoic acid in a bloom-forming diatom. *Science* **361**, 1356–1358 (2018).
41. C. M. Osuna-Cruz *et al.*, The *Seminavis robusta* genome provides insights into the evolutionary adaptations of benthic diatoms. *Nat. Commun.* **11**, 3320 (2020).
42. L. K. Medlin, S. Sato, D. G. Mann, W. H. C. F. Kooistra, 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. *J. Phycol.* **44**, 1340–1348 (2008).
43. A. J. Alverson, J. J. Cannone, R. R. Gutell, E. C. Theriot, The evolution of elongate shape in diatoms. *J. Phycol.* **42**, 655–668 (2006).
44. S. Laurent, N. Salamin, M. Robinson-Rechavi, No evidence for the radiation time lag model after whole genome duplications in Teleostei. *PLoS One* **12**, e0176384 (2017).
45. M. B. Parks, T. Nakov, E. C. Ruck, N. J. Wickett, A. J. Alverson, Phylogenomics reveals an extensive history of genome duplication in diatoms (Bacillariophyta). *Am. J. Bot.* **105**, 330–347 (2018).
46. T. Nosenko *et al.*, Deep metazoan phylogeny: When different genes tell different stories. *Mol. Phylogenet. Evol.* **67**, 223–233 (2013).
47. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
48. O. Çiftçi *et al.*, Phylotranscriptomics reveals the reticulate evolutionary history of a widespread diatom species complex. *J. Phycol.* **58**, 643–656 (2022).
49. M. J. Donoghue, Key innovations, convergence, and success: Macroevolutionary lessons from plant phylogeny. *Paleobiology* **31**, 77–93 (2005).
50. R. Trobajo, D. G. Mann, V. A. Chepurnov, E. Clavero, E. J. Cox, Taxonomy, life cycle, and auxosporulation of *Nitzschia fonticola* (Bacillariophyta). *J. Phycol.* **42**, 1353–1372 (2006).
51. T. Nakov, J. M. Beaulieu, A. J. Alverson, Insights into global planktonic diatom diversity: The importance of comparisons between phylogenetically equivalent units that account for time. *ISME J.* **12**, 2807–2810 (2018).
52. C. R. Kessenich, E. C. Ruck, A. M. Schurko, N. J. Wickett, A. J. Alverson, Transcriptomic insights into the life history of bolidophytes, the sister lineage to diatoms. *J. Phycol.* **50**, 977–983 (2014).
53. H. Ban *et al.*, Genome analysis of Parmales, the sister group of diatoms, reveals the evolutionary specialization of diatoms from phago-mixotrophs to photoautotrophs. *Commun. Biol.* **6**, 697 (2023).
54. M. J. Behrenfeld *et al.*, Thoughts on the evolution and ecological niche of diatoms. *Ecol. Monogr.* **91**, e01457 (2021).
55. Z. V. Finkel, M. E. Katz, J. D. Wright, O. M. E. Schofield, P. G. Falkowski, Climatically driven macroevolutionary patterns in the size of marine diatoms over the Cenozoic. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 8927–8932 (2005).
56. K. Bryłka, S. Richoz, A. J. Alverson, D. J. Conley, Looking for the oldest diatoms. *Mar. Micropaleontol.* **190**, 102371 (2024).
57. K. Bryłka, A. J. Alverson, R. A. Pickering, S. Richoz, D. J. Conley, Uncertainties surrounding the oldest fossil record of diatoms. *Sci. Rep.* **13**, 8047 (2023).
58. A. J. Alverson, Diatom phylogenomic vouchers, data, and analyses. Zenodo. https://doi.org/10.5281/zenodo.14595827. Deposited 25 April 2025.
59. M. D. MacManes, The Oyster River Protocol: A multi-assembler and kmer approach for de novo transcriptome assembly. *PeerJ* **6**, e5428 (2018).
60. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
61. C. Zhang, S. Mirarab, Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Mol. Biol. Evol.* **39**, msac215 (2022).
62. C. Zhang, C. Scornavacca, E. K. Molloy, S. Mirarab, ASTRAL-Pro: Quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* **37**, 3292–3307 (2020).
63. B. Q. Minh *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
64. H.-C. Wang, B. Q. Minh, E. Susko, A. J. Roger, Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018).
65. L. S. Quang, O. Gascuel, N. Lartillot, Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* **24**, 2317–2323 (2008).
66. J. Sukumaran, M. T. Holder, DendroPy: A Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).
67. S. A. Smith, B. C. O'Meara, treePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).
68. T. Vasconcelos, B. C. O'Meara, J. M. Beaulieu, A flexible method for estimating tip diversification rates across a range of speciation and extinction scenarios. *Evolution* **76**, 1420–1433 (2022).
69. A. J. Alverson, Diatom transcriptomes. NCBI. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1086848. Deposited 12 March 2024.