TITLE: Resolving marine–freshwater transitions by diatoms through a fog of gene tree discordance and hemiplasy

RUNNING HEAD: Phylogenomics of marine and freshwater diatoms

AUTHORS:

Wade R. Roberts¹, Elizabeth C. Ruck¹, Kala M. Downey¹, Andrew J. Alverson¹,¹

¹Department of Biological Sciences, University of Arkansas, 1 University of Arkansas, Fayetteville, AR, 72701, USA

*Corresponding author: Andrew J. Alverson, Department of Biological Sciences, University of Arkansas, 1 University of Arkansas, Fayetteville, AR, 72701, USA; aja@uark.edu
ABSTRACT

Despite the obstacles facing marine colonists, most lineages of aquatic organisms have colonized and diversified in freshwaters repeatedly. These transitions can trigger rapid morphological or physiological change and, on longer timescales, lead to increased rates of speciation. Diatoms are a lineage of ancestrally marine microalgae that have diversified throughout freshwater habitats worldwide. We generated a phylogenomic dataset of genomes and transcriptomes for 59 species to resolve freshwater transitions in one diatom lineage, the Thalassiosirales. Although most parts of the species tree were consistently resolved with strong support, we had difficulties resolving a Paleocene radiation, which affected the placement of one freshwater lineage. This and other parts of the tree were characterized by high levels of gene tree discordance caused by incomplete lineage sorting and low phylogenetic signal. Despite differences in species trees inferred from concatenation versus summary methods and codons versus amino acids, traditional methods of ancestral state reconstruction supported six transitions into freshwaters, two of which led to subsequent species diversification. However, simulations suggested as few as two independent transitions when accounting for hemiplasy, transitions occurring on branches in gene trees not shared with the species tree. This suggested that transitions across the salinity divide were originally facilitated by alleles already present in the ancestral marine populations. Accounting for differences in evolutionary outcomes, in which some taxa became locked into freshwaters while others were able to return to the ocean or become salinity generalists, might help distinguish between the ancestral changes that opened the door to freshwaters versus the subsequent, lineage-specific adaptations that allowed them to stay and thrive.

KEYWORDS: Gene concordance, Phylogenomics, Salinity, Site concordance, Thalassiosirales
INTRODUCTION

From bacteria to animals, the salinity gradient separating marine and freshwater environments poses a significant barrier to the distributions of many organisms (Lozupone and Knight 2007; McCairns and Bernatchez 2010; Kenny et al. 2019). Identifying how different lineages cross the salinity divide will improve our understanding of lineage diversification (Dittami et al. 2017) and the adaptive potential of species to climate change (Dickson et al. 2002; LE E et al. 2022). Diatoms are a diverse lineage of microalgae that occur throughout marine and freshwaters, and despite the numerous obstacles facing marine colonists (Kirst 1990, 1996; Nakov et al. 2020), ancestrally marine diatoms have successfully colonized and diversified in freshwaters repeatedly throughout their history (Nakov et al. 2019). These patterns are based on phylogenetic analyses of a small number of molecular markers, however, so they lack the insights of phylogenomic approaches, which can resolve large-scale macroevolutionary patterns and, at the same time, uncover key processes at play during important evolutionary transitions.

Although phylogenomic datasets have helped resolve historically recalcitrant nodes across the tree of life, they have also revealed how discordance in the evolutionary histories of different genes can confound inferences of species relationships. Gene tree discordance can be caused by biological sources, such as incomplete lineage sorting (ILS), hybridization, and compositional heterogeneity (Maddison 1997; Foster 2004; Degnan and Rosenberg 2006), or methodological sources, such as character sampling and gene tree error (Philippe et al. 2011; Xi et al. 2015; Molloy and Warnow 2018). Each of these challenge our ability to resolve species relationships and impact downstream analyses, such as estimation of divergence times (Smith et al. 2018). Multiple strategies have been proposed to overcome various sources of error, such as excluding third-codon positions from DNA datasets (Sanderson et al. 2000), using site-
heterogeneous models for amino acid data (Wang et al. 2018), and identifying the conditions under which concatenation (Edwards 2009) or gene tree summary approaches (Mirarab et al. 2014; Liu et al. 2015) more accurately resolve species relationships.

Discordance between gene and species trees can confound inferences of trait evolution as well (Hahn and Nakhleh 2016), particularly for complex traits that appear to have evolved convergently. Focusing on the species tree alone, without considering discordant gene trees, can lead to artifactual inferences of molecular convergence (Mendes et al. 2016). This failure occurs when a trait is determined by genes with topologies that do not match the species topology, a condition known as hemiplasy (Avise and Robinson 2008; Hahn and Nakhleh 2016; Storz 2016).

Hemiplasy has been identified as a likely explanation for patterns of character incongruence in amino acid substitutions in columnar cacti (Copetti et al. 2017), flower and fruit traits in *Jaltomata* (Solanaceae) (Wu et al. 2018), and dietary specialization in *Dysdera* spiders (Vizueta et al. 2019). High levels of gene tree discordance were detected in all of these cases.

Adaptation to low salinity is a complex trait involving many genes and pathways (Jones et al. 2012; Artemov et al. 2017; Hughes et al. 2017; Paver et al. 2018), so the genomic changes associated with successful freshwater colonizations are multifaceted (Cabello-Yeves and Rodriguez-Valera 2019; Rogers et al. 2021) and generally involve mutations in multiple genes (DeFaveri et al. 2011; Terekhanova et al. 2019; Chen et al. 2021). To better understand the pattern, timing, and process of marine–freshwater transitions by diatoms, we assembled a dataset of 45 genomes and 42 transcriptomes—most of them newly sequenced—to resolve species relationships, explore the causes and consequences of gene tree discordance, and determine whether different freshwater transitions were influenced by hemiplasy.
MATERIALS & METHODS

Detailed methods are provided in Supplementary File S1. Briefly, fresh diatom cultures were isolated from natural plankton or acquired from the National Center for Marine Algae and Microbiota (NCMA) or Roscoff Culture Collection (RCC). Collection data, culture conditions, and voucher information are available in Supplementary Table S1. For genome and transcriptome sequencing, we extracted total DNA and RNA from diatom cultures, constructed sequencing libraries, and sequenced them on the Illumina platform. Based on a large multigene phylogeny of diatoms (Nakov et al. 2018), we included Coscinodiscus, Lithodesmiales, and Eunotogramma as outgroups. Accession numbers for reads and assemblies are provided in Supplementary Table S2.

We used OrthoFinder (Emms and Kelly 2019) to cluster amino acid sequences from all genomes and transcriptomes into orthogroups, then aligned orthogroups containing ≥20% of the taxa with MAFFT (Katoh and Standley 2013). For each alignment, we identified the best-fit substitution model using ModelFinder (Kalyaanamoorthy et al. 2017) and estimated gene trees with IQ-TREE (Minh et al. 2020b) or FastTree (Price et al. 2010). We then filtered and trimmed the gene trees using the Rooted Ingroup method to produce final ortholog sets (Yang and Smith 2014). This filtering and trimming procedure was performed twice. We used PAL2NAL (Suyama et al. 2006) to reconcile nucleotide coding sequence alignments against amino acid alignments. We used Degen (Regier et al. 2010; Zwick et al. 2012) to replace synonymous sites in the full coding sequence alignments with degenerate nucleotides. In total, we analyzed datasets consisting of amino acids (AA), the first and second codon positions (CDS12), and degenerate codons (DEGEN). We generated final ortholog alignments and inferred trees as described above, using 1000 ultrafast bootstrap replicates to estimate branch support (Minh et al.)
We generated summary statistics for the alignments and gene trees using AMAS (Borowiec 2016) and PhyKit (Steenwyk et al. 2021) (Supplementary Table S3). Correspondence analysis of amino acid frequencies across all taxa was performed using GCUA (McInerney 1998).

Species trees were inferred using maximum-likelihood analysis of a concatenated supermatrix with IQ-TREE and the summary quartet approach implemented in ASTRAL-III (Zhang et al. 2018). We performed the matched-pairs test of symmetry in IQ-TREE to identify and remove partitions that violated assumptions of stationarity, reversibility, and homogeneity (SRH; Naser-Khdour et al. 2019). For the IQ-TREE analysis, we partitioned supermatrices by gene, used ModelFinder to select the best-fit substitution model for each partition, and estimated branch support with 10,000 ultrafast bootstrap replicates. For ASTRAL, we used the ortholog trees as input, collapsing branches with low bootstrap support (<33) to help mitigate gene tree error (Sayyari and Mirarab 2016; Simmons and Gatesy 2021). We estimated branch support using local posterior probability (Sayyari and Mirarab 2016). In addition to heterogeneity in gene histories, substitutional heterogeneity can also make tree inference difficult at deep time scales (Lartillot and Philippe 2004). To account for this possibility, we also estimated a species tree from the concatenated amino acid matrix using the Posterior Mean Site Frequency (PMSF) model implemented in IQ-TREE (Wang et al. 2018).

We calculated the Robinson-Foulds distance between each pair of species trees and visualized the results with a multidimensional scaling plot made with the R package treespace (Jombart et al. 2017). We characterized discordance using gene and site concordance factors (Minh et al. 2020a) and quartet concordance factors (Pease et al. 2018). We tested the support for competing backbone topologies in our species tree using the approximately unbiased (AU) test.
PHYLOGENOMICS OF MARINE AND FRESHWATER DIATOMS

(Shimodaira 2002) implemented in IQ-TREE. Relative gene tree support for the same set of backbone topologies was further evaluated using gene genealogy interrogation method (Arcila et al. 2017). Finally, we performed the polytomy test on the ASTRAL species tree to test whether any of the unstable backbone branches were better represented as a polytomy (Sayyari and Mirarab 2018).

Divergence times were estimated using MCMCtree (Yang 2007; Reis and Yang 2011), using five fossil-based calibrations (Supplemental File S1), autocorrelated rates, and the approximate likelihood approach. We performed marginal ancestral state reconstruction for marine and freshwater habitat across the species tree using hidden state speciation and extinction (SSE) models in the R package hisse (Beaulieu and O’Meara 2016). Lastly, we explored the probability of hemiplasy in our habitat reconstructions using the models and approaches implemented in the R package pepo and the program HeIST (Guerrero and Hahn 2018; Hibbins et al. 2020). Due to the computational demands of gene tree simulations, HeIST simulations were based on a reduced 16-taxon tree.

Assembled genomes and transcriptomes have been deposited at NCBI under BioProject PRJNA825288. Predicted proteomes, phylogenomic datasets, and scripts have been deposited in the Dryad Digital Repository: ____________.

RESULTS

We combined 42 newly sequenced draft genomes and 50 newly sequenced transcriptomes with publicly available genome or transcriptome data for a final dataset of 87 taxa representing 59 distinct species. A total of 17 transcriptomes were used directly in phylogenomic analyses, and the other 33 were used for genome annotation. From the combined dataset of
genomes and transcriptomes, we generated alignments and gene trees for 6262 orthologs, with each taxon represented in an average of 3275 (52.3%) orthologs.

Compositional heterogeneity and dataset construction

Relative Composition Variability (Phillips and Penny 2003) indicated greater compositional heterogeneity in third codon positions compared to amino acids and first and second codon positions (Supplementary Fig. S1). We also found substantial variation in GC content of third codon positions across taxa and genes (Supplementary Fig. S2), ranging from an average proportion of 0.40 ± 0.07 in *Cyclotella kingstonii* to 0.76 ± 0.14 in *Shionodiscus oestrupii*. To minimize potentially misleading signal in the nucleotide data, we removed third codon positions and, to examine the effects of saturation and GC heterogeneity in coding sequences (CDS), we replaced synonymous sites by recoding CDS sequences with degenerate nucleotides (Zwick et al. 2012). Based on these results, we created three datasets to estimate phylogenetic relationships: amino acids (AA), first and second codon positions (CDS12), and degenerate codons (DEGEN).

Dataset and alignment characteristics are summarized in Table 1 and detailed in Supplementary Table S3. Each dataset initially contained 6262 orthologs (Table 1). Gene trees constructed from all datasets were well supported (80 ± 7% average bootstrap; Supplementary Table S3). To reduce systematic errors due to model misspecification, we removed orthologs that failed assumptions of stationarity, reversibility, and homogeneity (SRH; *P* < 0.05), which retained 5522 (AA), 3259 (CDS12), and 3788 (DEGEN) loci in the datasets (hereafter referred to as the “complete” datasets; Table 1). We additionally subset the complete datasets to maximize signal or minimize missing data. To maximize signal and reduce stochastic error, we sorted
orthologs by the percentage of parsimony informative (PI) sites and retained the top 25% ranked orthologs (“top-PI”; Table 1). To minimize the amount of missing data and maximize taxon occupancy, we sorted complete datasets by the number of taxa and subset these to include the top 25% ranked orthologs with highest taxon occupancy (“top-Taxa”; Table 1). Orthologs in the top-Taxa datasets contained an average of 76 ± 8% of the total taxa.

Table 1. Summary of datasets used in the study.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Dataset name</th>
<th>Num loci</th>
<th>Total sites</th>
<th>Parsimony informative sites (%)</th>
<th>Num loci passing SRH</th>
<th>Thalassiosira grade topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acids (AA)</td>
<td>complete</td>
<td>6262</td>
<td>3,777,062</td>
<td>60</td>
<td>5522</td>
<td>Topology 4 Topology 1</td>
</tr>
<tr>
<td></td>
<td>top-PI</td>
<td>1588</td>
<td>996,027</td>
<td>76</td>
<td>1314</td>
<td>Topology 4 Topology 1</td>
</tr>
<tr>
<td></td>
<td>top-Taxa</td>
<td>1574</td>
<td>808,477</td>
<td>64</td>
<td>1374</td>
<td>Topology 5 Topology 1</td>
</tr>
<tr>
<td>top-PI top-Taxa + PMSF model</td>
<td>488</td>
<td>246,300</td>
<td>77</td>
<td>488</td>
<td>Topology 5 Topology 1</td>
<td></td>
</tr>
<tr>
<td>1st + 2nd codon positions (CDS12)</td>
<td>complete</td>
<td>6262</td>
<td>7,554,124</td>
<td>55</td>
<td>3259</td>
<td>Topology 4 Topology 5</td>
</tr>
<tr>
<td></td>
<td>top-PI</td>
<td>1570</td>
<td>1,950,228</td>
<td>71</td>
<td>661</td>
<td>Topology 4 Topology 3</td>
</tr>
<tr>
<td></td>
<td>top-Taxa</td>
<td>1574</td>
<td>1,616,954</td>
<td>59</td>
<td>782</td>
<td>Topology 5 Topology 3</td>
</tr>
<tr>
<td>Degenerate codons (DEGEN)</td>
<td>complete</td>
<td>6262</td>
<td>11,311,186</td>
<td>33</td>
<td>3788</td>
<td>Topology 4 Topology 3</td>
</tr>
<tr>
<td></td>
<td>top-PI</td>
<td>1569</td>
<td>2,884,728</td>
<td>44</td>
<td>771</td>
<td>Topology 4 Topology 2</td>
</tr>
<tr>
<td></td>
<td>top-Taxa</td>
<td>1574</td>
<td>2,425,431</td>
<td>36</td>
<td>903</td>
<td>Topology 5 Topology 3</td>
</tr>
</tbody>
</table>

1 Matched-pairs test of Symmetry, Reversibility, and Homogeneity.
2 Posterior Mean Site Frequency model.
We initially estimated 18 species trees using amino acids, codon positions 1 and 2, and degenerate codon sequences with different cutoffs for taxon occupancy or proportion of parsimony informative sites, using both concatenation and summary quartet approaches (Table 1). Correspondence analysis of amino acid frequencies separated taxa principally by habitat (marine vs. freshwater) rather than phylogeny (Supplementary Fig. S3), which led us to explore whether substitutional heterogeneity affected phylogenetic reconstructions. To do so, we estimated an additional species tree using the PMSF mixture model, which can accommodate heterogeneity in the amino acid substitution process between sites (Wang et al. 2018). Due to the computational demands of implementing this model, we applied it only to a reduced amino acid dataset with orthologs that met both the top-PI and top-Taxa filtering criteria (“AA-top-PI-top-Taxa”; Table 1). Gene trees from this dataset were also used as input to ASTRAL.

Previous phylogenetic analyses resolved freshwater taxa into two main clades: the genus Cyclotella, which also includes several marine and brackish species, and the “cyclostephanoids”, comprised of several stenohaline genera confined exclusively in freshwaters (Alverson et al. 2011). Given the potential implications for uncovering the mechanisms of freshwater adaptation, we were primarily interested in the placements of these two clades. Gross differences among data types and methods were evident in an ordination of species trees based on pairwise Robinson-Foulds distances, which showed a clear separation between IQ-TREE and ASTRAL topologies, with further separation of IQ-TREE trees estimated from datasets that maximized signal (top-PI) or minimized missing data (top-Taxa) (Supplementary Fig. S4). The phylogenetic position of Cyclotella was strongly supported and robust to differences in data type (codons vs. amino acids) and analysis (IQ-TREE vs. ASTRAL) (Fig. 1). The cyclostephanoids were placed
consistently within a large clade of marine *Thalassiosira* and relatives (Fig. 1), but the
arrangements of five main subclades—*Thalassiosira* I–IV and the freshwater
cyclostephanoids—varied depending on data type and analysis (Table 1; Fig. 2a). We refer to
this part of the tree as the *Thalassiosira* grade.

One resolution of the *Thalassiosira* grade (topology 1), recovered only by ASTRAL
analysis of amino acid gene trees, placed the freshwater cyclostephanoids as sister to a clade of
*Thalassiosira* I–IV (Table 1; Fig. 2a). All other species trees placed cyclostephanoids as sister to
*Thalassiosira* III (Table 1; Fig. 2a). ASTRAL analyses of codon-based gene trees (CDS12 and
DEGEN) alone recovered topology 3, which placed cyclostephanoids and *Thalassiosira* III as
sister to the remaining *Thalassiosira* (Table 1; Fig. 2a). Topologies 4 and 5 were recovered by
both data types, but only topology 5 was robust to both data type and analysis, having been
recovered by IQ-TREE analysis of both amino acid and codon alignments, and ASTRAL
analysis of codon-based gene trees (Table 1; Fig. 2a). Moreover, topology 5 was also recovered
by IQ-TREE analysis with the PMSF model, with almost all branches in the *Thalassiosira* grade
receiving maximum support (Fig. 1, branches A–E). Notably, the PMSF analysis also recovered
a monophyletic *Stephanodiscus*, which matches expectations based on morphology (Theriot et
al. 1987). *Stephanodiscus* was paraphyletic, with strong support, in relation to *Cyclostephanos* in
18 of the 20 species trees. Based on all of these results, we chose topology 5 from the PMSF
analysis as the reference species tree (Fig. 1).
Figure 1. Phylogram based on maximum likelihood analysis of amino acids using the posterior mean site frequency (PMSF) model and a dataset of 488 loci with the highest proportions of taxa and informative sites (“AA-top-PI-top-Taxa” dataset; Table 1). Backbone nodes of the *Thalassiosira* grade are indicated by the letters A–E. All branches had bootstrap support (BS).
values of 100 except for those with black circles which had BS = 90. Pie charts on backbone nodes show the proportion of gene trees that support the clade (gCF), the proportion that support both discordant topologies (gDF1, gDF2), and the proportion that are discordant due to polyphyly (gDFP). The size of the pie charts is only for aesthetics.

Discordance underlies topological uncertainty

Discordance among genes and sites is an important factor impacting phylogenetic reconstruction (Degnan and Rosenberg 2006; Mallet et al. 2016) and multiple tools now exist for identifying and characterizing discordance (Salichos et al. 2014; Smith et al. 2015; Pease et al. 2018). We characterized discordance by calculating gene, site, and quartet concordance factors for each branch (Figs. 1 and 2a; Supplementary Figs. S5 and S6). Gene (gCF) and site concordance factors (sCF) represent the proportion of genes or sites that are in agreement with a particular branch in the species tree (Minh et al. 2020a). Gene concordance factors range from 0 to 100, and site concordance factors typically range from one-third (33) to 100, with values near 33 indicative of no signal for that branch (Minh et al. 2020a). Quartet concordance factors (QC) provide an alternative, likelihood-based estimate of the relative support at each branch for the three possible resolutions of four taxa (Pease et al. 2018). Quartet concordance factors range from -1 to 1, with positive values showing support for the focal branch, negative values supportive for an alternate quartet, and values of zero indicating equal support among the three possible quartets (Pease et al. 2018).

All three concordance factors were high for most branches in the tree, indicating that a majority of genes, sites, and quartets supported those relationships (Supplementary Figs. S5 and S6). For example, the branch subtending the Skeletonema clade had high concordance
(gCF/sCF/QC = 93/75/0.87; Supplementary Figs. S5 and S6). Despite having maximum
bootstrap and local posterior probability support, however, concordance factors were low for
many of the backbone branches (Fig. 1; Supplementary Figs. S5 and S6). These included
backbone branches within the *Thalassiosira* grade (Fig. 1, nodes A–E) that affected placement of
the freshwater cyclostephanoids (Fig. 2). In this part of the tree, concordance was generally low
for the backbone branches (gCF = 4–26; sCF = 32–42; QC = −0.04–0.36) (Figs. 1 and 2;
Supplementary Figs. S5 and S6). Gene concordance factors were lowest for branches C and D,
the only two branches in the species tree with <100% bootstrap support (Fig. 1). Gene
concordance factors were only slightly higher for branches A, B, and E (Fig. 1). These had low
site concordance (Fig. 2a) and near-zero quartet concordance factors (Supplementary Fig. S6),
consistent with expectations for little to no signal. Site concordance factors did not change
appreciably when we used the CDS12 data. However, repeating quartet concordance factor
calculation with the CDS12 data revealed a minor switch in support for branches C (QC_C = 0.004
→ −0.006) and E (QC_E = −0.04 → 0.07) (Supplementary Fig. S6). These patterns of shifting
support combined with the lowest site concordance factor for branch E (sCF_E = 32) suggest that
very few sites support the sister relationship of *Thalassiosira* III and cyclostephanoids, despite its
recovery in 16 of the 20 inferred species trees (Figs. 1 and 2a).

Although bootstrap and posterior probability values were uniformly high across our trees,
traditional measures of branch support often fail to capture the underlying agreement or
disagreement among genes or sites for a given tree topology (Minh et al. 2020a). For many of the
branches on our tree, there were more discordant than concordant genes and sites (Fig. 1;
Supplementary Figs. S5 and S6), suggesting that more genes and sites supported an alternative
relationship. This was the case with *Stephanodiscus*, for example, where more genes supported
paraphyly than monophyly, though more sites supported monophyly (Supplementary Figs. S5 and S6). Across the tree, support for alternative relationships was neither strong nor consistent among genes or sites. Illustrative of this, discordance in more than one-third of the tree (29 of 83 branches) was due to polyphyly (Supplementary Fig. S5), indicative of a widespread lack of signal in our gene trees. Taken together, these analyses highlighted extensive gene and site discordance, some well-supported and much of it not, across Thalassiosirales.

Figure 2. a) Phylogenetic hypotheses of the *Thalassiosira* grade inferred using concatenation and summary methods on the amino acid (AA), codon positions 1 and 2 (CDS12), and recoded codon (DEGEN) datasets. Nodes are labeled with the percentage of amino acid sites concordant with the branch (site concordance factor). The principal clade of interest, the freshwater cyclostephanoids, is colored orange and labeled ‘FW’. The four focal clades of marine
Thalassiosira and allies are labeled I–IV. Below, results of gene genealogy interrogation tests of alternative hypotheses of relationships within the Thalassiosira grade. These tests used datasets filtered to include only the top 25% of orthologs based on the percentage of parsimony informative sites (top-PI) for b) amino acids, c) codon positions 1 and 2, and d) recoded codons. Lines correspond to the cumulative number of genes (x-axis) supporting topology hypotheses with the highest probability and their P values (y-axis) from the Approximately Unbiased (AU) topology tests. Values above the dashed line indicate topological hypotheses that are significantly better than the alternatives (P < 0.05). For example, the green line in (b) shows that there were a total of 88 genes that best supported topology 3, while only four of those genes were above the dotted line and were significantly better supported than the other four alternative topologies.

There are both biological (e.g., ILS) and technical causes (e.g., gene tree error) of gene discordance, but the proportion attributable to each factor can be difficult to discern (Morales-Briones et al. 2020; Cai et al. 2021). To better assess the importance of gene tree error in our dataset—whether genes with low phylogenetic signal produced inaccurate gene trees—we recalculated gene and site concordance using just the 1588 amino acid orthologs with the highest percentage of parsimony informative sites (top-PI dataset; Table 1). Average gene concordance increased modestly, from 56.6% to 62.1%, but site concordance was unchanged (Supplemental Fig. S7). Gene concordance factors for branches A, B, and E in the Thalassiosira grade increased by 5–7% but were largely unchanged for branches C and D (Supplemental Fig. S7). These increases in gene concordance when using the most signal-rich genes suggest that errors in gene tree estimation contributed to the lack of resolution in several critical branches (Chan et al. 2020;
Vanderpool et al. 2020). Deeper nodes in the tree may be more prone to technical errors caused by long-branch attraction, poor alignments, or model misspecification, despite our attempts to minimize these during dataset construction (Supplementary File S1). We found slight negative correlations between node age and both gene concordance ($R^2 = 0.10$) and site concordance ($R^2 = 0.26$) (Supplementary Fig. S8), which suggests that older branches more likely suffered from saturation due to recurrent substitutions.

Placement of the freshwater cyclostephanoids

We used two additional tree-based methods to test the relative support for competing topologies within the *Thalassiosira* grade. Approximately unbiased (AU) tests on the concatenated alignment for each dataset in Table 1 were used to assess the relative statistical support for topologies 1–5 (Fig 2a). Like the original species tree inferences, results of AU tests largely reflected data type (Table 1), with amino acid characters supporting topology 1 ($P = 0.01$, AU test) and the codon and degenerate codon datasets supporting topology 4 ($P = 0$, AU test; Supplementary Table S4). Although concordance factors suggested that most gene trees were uninformative for relationships in the *Thalassiosira* grade, we looked for secondary signal for one or more of the five competing topologies using gene genealogy interrogation (GGI; Arcila et al. 2017). To do this, we performed constrained gene tree searches on the most information-rich (top-PI) orthologs and compared their likelihoods using the AU test. The GGI test assumes monophyly of the tested clades, so using our time-calibrated tree, we converted branch lengths (in millions of years) to coalescent time units using a range of plausible effective population sizes and generation times for diatoms (Supplementary File S1). The estimated stem branch lengths for the five clades in the *Thalassiosira* grade were all >5 coalescent units, suggesting
sufficient time to reach monophyly (Rosenberg 2003). After ranking likelihood scores from the AU tests and selecting constraint topologies with the best score (rank 1 trees), no single topology was strongly favored in a majority of constrained gene trees across the three datasets, implying similar levels of support (Fig. 2b–d; Supplementary Table S5). Support from the amino acid dataset was split between topologies 1 and 4, which were recovered by both summary and concatenation methods (Table 1; Fig. 2b; Supplementary Table S5). The most frequent best-fit topology for the codon and degenerate codon datasets corresponded to topology 4 (Fig. 2c,d), one originally recovered by concatenation only (Table 1).

Gene genealogy interrogation can also be used to explore the effects of gene tree error on summary quartet methods by filtering the input to ASTRAL to include only the highest ranking constrained genes (Arcila et al. 2017; Mirarab 2017). For each of the three character types, we performed two ASTRAL analyses using as input either all the top scoring (rank 1) constrained gene trees ($n_{AA} = 1588$, $n_{CDS12} = 1570$, $n_{DEGEN} = 1569$) or just the subset that had statistical support ($P < 0.05$) above the AU-based rank 2 topology ($n_{AA} = 142$, $n_{CDS12} = 56$, $n_{DEGEN} = 71$). In all six cases, the inferred trees were consistent with topology 5, despite it being best supported (rank 1) in just 13–16% of the constrained gene trees (Fig. 2b–d; Supplementary Table S5). We originally chose topology 5 as the reference species tree (Fig. 1) because it was recovered by both amino acids and codons, ASTRAL and IQ-TREE analysis with the PMSF model, and because it recovered monophyly of *Stephanodiscus*. Coalescent theory predicts that in severe cases of ILS, short internal branches can produce gene trees that conflict with the species tree more often than they agree, creating the so-called “anomaly zone” (Degnan and Rosenberg 2006). Our recovery of a species tree topology here that is not the most frequent among the GGI gene trees could indicate that the backbone of the *Thalassiosira* grade lies in the anomaly zone.
Despite this, polytomy tests in ASTRAL using each dataset rejected the null hypothesis that any of these branches is a polytomy ($P < 0.05$).

The temporal sequence of marine–freshwater transitions

Divergence time estimates dated the crown Thalassiosirales to the upper Cretaceous, around 113 Ma (96–120, 95% CI; Fig. 3a; Supplementary Fig. S9). One of the two main freshwater lineages, *Cyclotella*, originated in the late Cretaceous (79 Ma [66–86, 95% CI]) and the other freshwater lineage, the cyclostephanoids, originated later in the Eocene (43.5 Ma [36–48, 95% CI]) (Fig. 3a; Supplementary Fig. S9). Radiation of the *Thalassiosira* grade lineages occurred during the Paleocene, from 57 Ma [48–63, 95% CI] to 73 Ma [61–80, 95% CI] (Fig. 3a; Supplementary Fig. S9). The overlapping confidence intervals allow for the possibility that these lineages diverged in much more rapid succession than suggested by the mean ages.

Hemiplasy impacts marine–freshwater transitions

We estimated the number of marine–freshwater transitions in two ways. A traditional trait reconstruction method, HiSSE, was used to estimate the maximum number of independent marine–freshwater transitions under a “homoplasy-only” model. A second model allowed us to estimate how many of the transitions were non-independent due to hemiplasy, i.e., transitions to freshwater occurring on branches of discordant gene trees that are not found in the species tree (Hibbins et al. 2020). The best-fit HiSSE model was a character independent model (CID-4, $AIC = 535.3$; Supplementary Table S6), which suggests that shifts in diversification rate occurred independently of marine–freshwater transitions. Using parameter estimates from the CID-4 model, we inferred six transitions from marine to freshwater habitats (Fig. 2a). With this dataset,
the most recent common ancestor of *Cyclotella* was marine, with the transition to freshwater occurring shortly thereafter (Fig. 2a). To assess the possible role of hemiplasy in our trait reconstructions, we calculated hemiplasy risk factors (HRF), which are the ratio of the probabilities of hemiplasy to homoplasy in different parts of the species tree (Guerrero and Hahn 2018). This analysis revealed an increased chance of character state transitions due to hemiplasy along most of the backbone branches of the species tree (Fig. 3b). Gene trees simulated from a reduced 16-taxon species tree using HeIST returned 1133 loci with a simulated site pattern that matched the marine–freshwater character states (i.e. freshwater taxa shared the same mutation). With this approach, the most likely scenario is the one supported by the largest number of gene trees. In our dataset, the scenarios could range from strict homoplasy (n = 6 marine–freshwater transitions) to strict hemiplasy (n = 1 transition). A total of 732 of the 1133 simulated gene trees (64.6%) corresponded to a scenario of combined hemiplasy and homoplasy, with two independent transitions into freshwaters (Supplementary Table S7). Under this model, freshwater taxa grouped into one of two clades in a simulated gene tree—each clade representing a freshwater transition—where at least one transition was discordant with the species tree. Just 23% of the simulated trees supported a hemiplasy-only model with a single freshwater transition, and there was no support for a homoplasy-only history of six independent marine–freshwater transitions (Supplementary Table S7).
Figure 3. **a)** Divergence times and ancestral state reconstruction of marine and freshwater habitat in the Thalassiosirales. Conspecific taxa were removed prior to ancestral state reconstruction, leaving one tip per species. Tip labels are colored according to their habitat (blue=freshwater, orange=euryhaline, grey=marine). Pie charts denote the probability of each node reconstructed as either marine (grey) or freshwater (blue) using parameters estimated from the HiSSE CID–4 model. Euryhaline taxa were coded as marine for the purposes of ancestral state reconstruction. Divergence times for the full set of taxa can be found in Supplementary Fig. S9. **b)** Hemiplasy Risk Factors (HRF) on internal branches show increased values on most short internal branches.
demonstrating an increased potential for hemiplasy to influence trait reconstruction. Branch
lengths in coalescent units were inferred using ASTRAL.

DISCUSSION

Transitions to freshwaters

Marine–freshwater transitions have been key events in the diversification of lineages
across the tree of life (Logares et al. 2009; e.g., Tedesco et al. 2017), including diatoms where
freshwater taxa have experienced significantly increased rates of both speciation and extinction
compared to their marine ancestors (Nakov et al. 2019). Our study focused on one model clade,
the Thalassiosirales, which is among the most abundant and diverse lineages in the marine and
freshwater plankton and where genetic and genomic resources are readily available (Armbrust et
al. 2004; Nawaly et al. 2020; Roberts et al. 2020). The 42 draft genomes presented here greatly
expand the phylogentic and ecological diversity of sequenced genomes for Thalassiosirales, and
diatoms as a whole, and will greatly facilitate efforts to identify the genomic basis of freshwater
adaptation in diatoms.

We identified six different marine–freshwater transitions in Thalassiosirales, more than
previous studies (Alverson et al. 2007). We tested whether these transitions were fully
independent, owing to separate mutations (homoplasy) in each of the six freshwater lineages, or
whether some of the transitions were attributable to hemiplasy, i.e., shared ancestral
polymorphisms in discordant gene trees (Avise and Robinson 2008). Hemiplasy simulations
supported a scenario of as few as two independent transitions at the genic level, a result that
followed naturally from the extensive history of gene tree discordance across the backbone of the
tree. Long considered an ecological “Rubicon” that is rarely crossed (Mann 1999), phylogenetic
studies have shown that diatoms move much more frequently across the salinity divide than was assumed historically (Alverson et al. 2007; Nakov et al. 2019). The discovery of hemiplasy, which reduced the number of independent freshwater transitions in Thalassiosirales by a factor of three, suggests that transitions across steep environmental gradients might occur more commonly than expected when colonists can leverage ancestral polymorphisms rather than relying exclusively on new mutations.

With hundreds of species, a crown age of roughly 100 Mya, and freshwater transitions dating as far back as 80 Mya, the breadth and depth of the Thalassiosirales phylogeny far exceeds other, much younger lineages with traits that have been impacted by hemiplasy (Copetti et al. 2017; Wu et al. 2018). Relatively deep, and short, backbone branches that predated freshwater transitions were the ones identified as most susceptible to hemiplasy. The coalescent branch lengths were short in these parts of the tree (Fig. 3b), and the substitution- (Fig. 1) and clock-based branch lengths (Fig. 3a) are also shorter than suggested by our trees due to incomplete taxon sampling (Alverson et al. 2007). Although many of the conditions for hemiplasy were clear (high levels of gene tree discordance coinciding with a rapid radiation, particularly along the Thalassiosira grade), other properties of the habitat transitions were indicative of homoplasy: ancestral state reconstructions were unambiguous, the two main freshwater transitions were not paraphyletic but were aligned with phylogenetic relationships, and gene concordance was high on the long internal branches subtending the two main freshwater lineages (Hahn and Nakhleh 2016; Wu et al. 2018).

Given the sheer length of time that has elapsed since the transitions occurred, the shared ancestral polymorphisms that opened the door to freshwaters for Thalassiosirales will be difficult to distinguish from the augmenting, lineage-specific adaptive mutations that followed (Zou and
Zhang 2015; Mendes et al. 2016). A large number of genes and pathways have been implicated in the response to low salinity (Nakov et al. 2020; Downey et al. 2022; Pinseel et al. 2022), so an adaptive allele in one of the many possible target genes might have made it possible simply to survive in freshwaters initially. In the tens of millions of years since then, the hemiplasious alleles could have been overwritten in one or more freshwater lineages, leaving the shared mechanism itself (e.g., enhanced transport of sodium ions) as the only remaining direct evidence of the hemiplasy. The divergent evolutionary outcomes might also help distinguish hemiplasy from homoplasy. The genus *Cyclotella* includes freshwater, secondarily marine, and generalist euryhaline species that can tolerate a wide range of salinities (Guillard and Ryther 1962; Nakov et al. 2020; Downey et al. 2022). Similarly, most freshwater transitions at the tips of the tree (Fig. 1) involved species with populations that also grow in marine habitats (*Conticribra weissflogii* and *Cyclotella nana*) or can tolerate slightly brackish water (*Thalassiosira lacustris* and *Skeletonema potamos*). The cyclostephanoids are, by contrast, stenohaline specialists found exclusively in freshwaters, suggestive of a different genetic trajectory into freshwaters.

Although we have treated salinity as a categorical variable, salinity varies along a continuum from freshwater to marine and even hypersaline habitats. Moreover, salinity fluctuations are common in brackish and marine systems, such as coastlines influenced by precipitation and river discharge. Adaptation to variable environmental conditions may be more likely linked to modifications of gene expression (Wray 2007). In diatoms, exposure to freshwater induces profound changes gene expression related to cellular metabolism, ion transport, photosynthesis, and storage compound biosynthesis (Bussard et al. 2017; Nakov et al. 2020; Downey et al. 2022; Pinseel et al. 2022). Diatoms exhibit high levels of inter- and intraspecific variation in gene expression in response to reduced salinity (Nakov et al. 2020;
Pinseel et al. 2022), providing numerous targets for natural selection to optimize gene expression for a particular salinity environment (López-Maury et al. 2008; Bedford and Hartl 2009; Gomez-Mestre and Jovani 2013). Finally, differences in codon usage (Prabha et al. 2017), nucleotide substitution rates (Mitterboeck et al. 2016), transposable element activity (Yuan et al. 2018), epigenetic responses (Artemov et al. 2017), and epistatic interactions (Stern et al. 2022) have been implicated in adaptation to freshwaters in other groups. The genomic resources and phylogenetic framework presented here represent an important advance towards identifying the genes and evolutionary processes underpinning freshwater adaptation by diatoms.

The impact of discordance on placement of freshwater clades

Comparative analyses require a strong phylogenetic framework (Felsenstein 1985), so a major goal of this study was to establish a robust phylogenetic hypothesis for Thalassiosirales. Our primarily genome-based analysis of 6262 nuclear orthologs provided better resolution and, superficially, increased support across most of the tree compared to previous analyses based on a few genes (Alverson et al. 2007). Across character types, methods of inference, and different criteria for including characters or taxa, the placement of one of the two principal freshwater clades, Cyclotella, was consistent across species trees, with an estimated origin in the late Cretaceous. The placement of the second major freshwater lineage, the cyclostephanoids, was less certain.

The freshwater cyclostephanoid clade was placed within a grade of marine species, most of which belong to the polyphyletic genus Thalassiosira. These marine Thalassiosira were divided among four clades, but the arrangement of these clades and the freshwater cyclostephanoids varied across analyses. Uncertainty in the backbone relationships for this part
of the tree was likely caused by a combination of gene tree error and high levels of incomplete
lineage sorting. First, many individual genes contained too little information to confidently
resolve deep splits separated by short branch lengths—a finding that is not unique to this dataset
(Chan et al. 2020; Arcila et al. 2021). Divergence time estimates suggest that these splits
occurred in relatively quick succession, as few as 5 million years. Although many of these
bipartitions had consistently weak support, gene concordance factors increased when we
analyzed orthologs with the highest phylogenetic signal, implicating gene tree error as one of the
sources of instability (Chan et al. 2020; Vanderpool et al. 2020). In other phylogenomic studies
impacted by low phylogenetic signal and high gene tree error, gene genealogy interrogation
(GGI) has been used to overcome noise and identify majority gene support for a single
hypothesis (Hughes et al. 2018; Tea et al. 2021). In our case, no single resolution was supported
by a majority of genes, indicative of nodes that are difficult to resolve even with hundreds of
genes (Nesi et al. 2021). The best-supported constraint tree identified by GGI differed between
amino acid and codon-based gene trees, highlighting conflicting signal even within genes.
Second, in addition to gene tree error, the large number of alternative topologies among gene
trees along the Thalassiosira grade is also consistent with ILS (Arcila et al. 2017). Gene tree
summary methods such as ASTRAL outperform concatenation when ILS is the major source of
discordance (Kubatko and Degnan 2007; Roch and Warnow 2015), but summary methods
perform poorly when gene tree error is high (Roch and Warnow 2015; Xi et al. 2015). Following
Arcila et al. (2017), we attempted to eliminate some of the noise in our dataset by restricting
ASTRAL analyses to the top-ranked constrained gene trees and in doing so recovered a
backbone topology congruent with one of the few originally recovered by both gene tree
summary and concatenation methods (topology 5; Table 1 and Fig. 2). Taken together, these
results suggest that gene tree error negatively impacted our ASTRAL analyses. After identifying and removing some of that error, we were able to recover a stronger hypothesis for the placement of the freshwater cyclostephanoids within the marine Thalassiosira grade.

The anomaly zone describes an especially vexing phylogenetic problem in which short branch lengths are unresolvable, resulting in gene trees that differ from the species more frequently than they agree (Degnan and Rosenberg 2006). Within the Thalassiosira grade, the most common GGI gene tree topologies either did not match the reference species tree or were uninformative for these short branches. This implies that unresolved gene trees (i.e., those with polytomies) are more probable than resolved ones in the Thalassiosira grade, as branch lengths that exceed the boundaries of the anomaly zone should produce resolved gene trees (Huang and Knowles 2009). Gene tree error like that identified in our dataset can lead to underestimation of coalescent branch lengths, which define the anomaly zone boundaries (Linkem et al. 2016; Forthman et al. 2022), resulting in the mistaken identification of an anomaly zone where none exists. Although not entirely clear in our case, the Thalassiosira grade appears to fall outside of the anomaly zone, so the large number of conflicting gene histories is more likely due to ILS and gene tree error. Challenges remain in anomaly zone detection due to the limited application and computational costs of multispecies coalescent methods applied to large genomic datasets and the inclusion of additional biological factors other than ILS (Flouri et al. 2018, 2020).

Factors that are more poorly known in diatoms might also have contributed to the high levels of discordance. We attempted to minimize technical sources of gene tree error during dataset construction (Supplemental File S1), but other sources are more difficult to discern (Cai et al. 2021). These include hybridization and polyploidy, gene duplication and loss, or recombination. Although hybridization is not well documented in diatoms (Casteleyn et al. 2009;
Koester et al. 2010; Tanaka et al. 2015), there is evidence for an ancient allopolyploidy event early on in the evolutionary history of Thalassiosirales (Parks et al. 2018). Many methods that identify hybridization based on gene trees (Edelman et al. 2019; Vanderpool et al. 2020) or site patterns in an alignment (Blischak et al. 2018) can be confounded by ancestral population structure, substitutional saturation, or unsampled ghost lineages (Slatkin and Pollack 2008; Tricou et al. 2022), all of which are poorly characterized in diatoms. High GC content regions have been linked to higher recombination rates (Kent et al. 2012; Lartillot 2013) and can lead to increased discordance (Pease and Hahn 2013). Despite high levels of variation in GC content across Thalassiosirales, intralocus recombination was detected in a vanishingly small (<0.001%) percent of the 6262 orthologs in our analysis (results available on Dryad). In addition to showing whether any of these factors contribute to gene tree discordance, a more thorough exploration of each one will fill important gaps in our understanding of diatom evolution.

Codon bias and amino acid composition in freshwater diatoms

Thalassiosirales includes divergence times across a timescale ranging from thousands (Theriot et al. 2006) to tens of millions of years ago (Fig. 3a), which led us to explore the utility of both amino acid and nucleotide characters for resolving phylogenetic relationships. Amino acids are less susceptible to saturation and useful for resolving deep relationships (Philippe et al. 2011; Rota-Stabelli et al. 2012), whereas nucleotides contain more information to resolve more recent divergences (Simmons et al. 2002; Townsend et al. 2008). Both data types recovered the vast majority of relationships consistently and with strong support, while at the same time revealing similar patterns of discordance along the backbone of the tree. In many cases, however, they differed in their resolutions of the most recalcitrant parts of the tree. Disagreements between
character types within the same dataset, such as those within the *Thalassiosira* grade here, have been found in other groups as well (Gillung et al. 2018; Skinner et al. 2020).

Almost every analysis of the amino acid dataset—including species trees, concordance factors, and AU tests—supported the placement of cyclostephanoids as sister to the remaining *Thalassiosira* clades, but nucleotide analyses placed them with *Thalassiosira* III (Fig. 2).

Discordance caused by codon usage bias and differences in amino acid composition might account for this discrepancy. An association between codon bias and ecology has been demonstrated in a broad diversity of microbes, where species that share an ecological niche have similar codon usage, independent of phylogeny (Botzman and Margalit 2011; Roller et al. 2013; Arella et al. 2021). Similar patterns of codon usage within marine and freshwater habitats have been described in prokaryotes (Cabello-Yeves and Rodriguez-Valera 2019), and we discovered differences in both codon usage and amino acid composition between marine and freshwater diatoms. The amino acid compositions of distantly related freshwater lineages might be sufficiently similar to cause the amino acid characters to support the “sister to the rest” placement of cyclostephanoids. Protein sites with different structural, functional, or selective constraints can lead to differences in amino acid composition between species (Villar and Kauvar 1994; Youssef et al. 2021), something that is not accounted for by standard empirical protein models and may have led to artifacts in our gene and species tree inferences (Wang et al. 2018). When we applied the PMSF model, which accounts for substitution heterogeneity in amino acid sites, cyclostephanoids were placed as sister to *Thalassiosira* III, in agreement with the codon datasets. The similarity in codon usage and amino acid composition between distantly related freshwater diatoms merits further study into the causes and functional significance, if any.
Conclusions

The vast differences between marine and freshwaters result in strong selective pressures on freshwater colonists. Indeed, some of these transitions have become some of our most powerful model systems for studying convergent evolution (Elmer and Meyer 2011). Low salinity provokes a broad range of physiological and metabolic responses in diatoms (Nakov et al. 2020; Downey et al. 2022; Pinseel et al. 2022), but the current genetic architectures of freshwater adaptation reflect tens of millions of years of optimization and change since the earliest transitions. As a result, it may not be possible to distinguish the derived alleles that currently allow these diatoms to thrive in freshwaters from the ancestral ones that made the first transitions possible. Nevertheless, the phylogenomic analyses presented here highlighted a key role for hemiplasy in charting the genetic trajectory for freshwater colonizations in this group of diatoms, indicating that some of the necessary alleles were already present in the ancestral marine populations. Some of the subsequent mutations that built upon this scaffold of hemiplasy might be identifiable by the different evolutionary outcomes in which some taxa became irreversibly locked into freshwaters (e.g., cyclostephanoids) while others were able to transition back to the ocean or became salinity generalists (e.g., Cyclotella). The vast new genomic resources and phylogenetic framework presented here represent an important step forward in addressing these types of questions to better understand how diatoms have made this complex ecological transition appear to be so simple superficially.
SUPPLEMENTAL MATERIAL

Datasets and scripts are available from the Dryad Digital Repository:

http://dx.doi.org/10.5061/dryad.[NNNN]

FUNDING

This work was supported by the National Science Foundation (DEB 1651087 to A.J.A).

This research used resources available through the Arkansas High Performance Computing Center, which is funded through multiple NSF grants and the Arkansas Economic Development Commission.

ACKNOWLEDGEMENTS

We thank Ed Theriot for sharing several culture strains, and we thank Matt Ashworth and Jeffery Stone for help with scanning electron microscopy.


Botzman M., Margalit H. 2011. Variation in global codon usage bias among prokaryotic
organisms is associated with their lifestyles. Genome Biol. 12:R109.


Copetti D., Búrquez A., Bustamante E., Charboneau J.L.M., Childs K.L., Eguiarte L.E., Lee S.,
M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North

in freshwater adaptation in threespine sticklebacks (Gasterosteus aculeatus). Evolution.

Degnan J.H., Rosenberg N.A. 2006a. Discordance of species trees with their most likely gene


environments provide new clues about the origins of multicellular plants and algae. J.
Phycol. 53:731–745.

hypoosmotic stress reveals distinct stages of freshwater acclimation by a euryhaline diatom.
bioRxiv.:2022.06.24.497401.

Edelman N.B., Frandsen P.B., Miyagi M., Clavijo B., Davey J., Dikow R.B., García-Accinelli
G., Van Belleghem S.M., Patterson N., Neafsey D.E., Challis R., Kumar S., Moreira G.R.P.,
Salazar C., Chouteau M., Counterman B.A., Papa R., Blaxter M., Reed R.D., Dasmahapatra
K.K., Kronforst M., Joron M., Jiggins C.D., McMillan W.O., Di Palma F., Blumberg A.J.,


Kent C.F., Minaei S., Harpur B.A., Zayed A. 2012. Recombination is associated with the


Linkem C.W., Minin V.N., Leaché A.D. 2016. Detecting the Anomaly Zone in Species Trees and Evidence for a Misleading Signal in Higher-Level Skink Phylogeny (Squamata:


Roberts et al.


Simmons M.P., Ochoterena H., Freudenstein J.V. 2002. Amino acid vs. nucleotide characters:


Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic
986 71.

988 and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics.


992 Youssef N., Susko E., Roger A.J., Bielawski J.P. 2021. Shifts in amino acid preferences as
993 proteins evolve: A synthesis of experimental and theoretical work. Protein Sci. 30:2009–
994 2028.

996 analysis of 52 fish species suggests differential associations of repetitive elements with their
997 living aquatic environments. BMC Genomics. 19:141.

998 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree

1000 Zou Z., Zhang J. 2015. Are Convergent and Parallel Amino Acid Substitutions in Protein

1002 Zwick A., Regier J.C., Zwickl D.J. 2012. Resolving discrepancy between nucleotides and amino
1003 acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-
FIGURE LEGENDS

Figure 1. Phylogram based on maximum likelihood analysis of amino acids using the posterior mean site frequency (PMSF) model and a dataset of 488 loci with the highest proportions of taxa and informative sites (“AA-top-PI-top-Taxa” dataset; Table 1). Backbone nodes of the *Thalassiosira* grade are indicated by the letters A–E. All branches had bootstrap support (BS) values of 100 except for those with black circles which had BS = 90. Pie charts on backbone nodes show the proportion of gene trees that support the clade (gCF), the proportion that support both discordant topologies (gDF1, gDF2), and the proportion that are discordant due to polyphyly (gDFP). The size of the pie charts is only for aesthetics.

Figure 2. **a)** Phylogenetic hypotheses of the *Thalassiosira* grade inferred using concatenation and summary methods on the amino acid (AA), codon positions 1 and 2 (CDS12), and recoded codon (DEGEN) datasets. Nodes are labeled with the percentage of amino acid sites concordant with the branch (site concordance factor). The principal clade of interest, the freshwater cyclostephanoids, is colored orange and labeled ‘FW’. The four focal clades of marine *Thalassiosira* and allies are labeled I–IV. Below, results of gene genealogy interrogation tests of alternative hypotheses of relationships within the *Thalassiosira* grade. These tests used datasets filtered to include only the top 25% of orthologs based on the percentage of parsimony informative sites (top-PI) for **b)** amino acids, **c)** codon positions 1 and 2, and **d)** recoded codons. Lines correspond to the cumulative number of genes (x-axis) supporting topology hypotheses with the highest probability and their *P* values (y-axis) from the Approximately Unbiased (AU) topology tests. Values above the dashed line indicate topological hypotheses that are significantly better than the alternatives (*P* < 0.05). For example, the green line in (b) shows that...
there were a total of 88 genes that best supported topology 3, while only four of those genes were above the dotted line and were significantly better supported than the other four alternative topologies.

Figure 3. a) Divergence times and ancestral state reconstruction of marine and freshwater habitat in the Thalassiosirales. Conspecific taxa were removed prior to ancestral state reconstruction, leaving one tip per species. Tip labels are colored according to their habitat (blue=freshwater, orange=euryhaline, grey=marine). Pie charts denote the probability of each node reconstructed as either marine (grey) or freshwater (blue) using parameters estimated from the HiSSE CID–4 model. Euryhaline taxa were coded as marine for the purposes of ancestral state reconstruction. Divergence times for the full set of taxa can be found in Supplementary Fig. S9. b) Hemiplasy Risk Factors (HRF) on internal branches show increased values on most short internal branches, demonstrating an increased potential for hemiplasy to influence trait reconstruction. Branch lengths in coalescent units were inferred using ASTRAL.
Cyclostephanos tholiformis
Cyclostephanos invisitatus
Stephanodiscus minutulus
Stephanodiscus triporus
Discostella pseudostelligera
Discostella stelligeroides
Discostella stelligera
Skeletonema potamos
Cyclotella atomus
Cyclotella meneghiniana
Cyclotella distinguenda
Cyclotella kingstonii
Thalassiosira lacustris
Conticribra guillardii
Conticribra weissflogii
Cyclotella nana
Cyclotella choctawhatcheeana
Cyclotella atomus
Mediolabrus comicus
Skeletonema marinoi
Skeletonema menzelii
Skeletonema japonicum
Skeletonema tropicum
Skeletonema grothae
Thalassiosira oceanica
Thalassiosira profunda
Thalassiosira exigua
Thalassiosira mediterranea
Thalassiosira sp
Bacterosira constricta
Thalassiosira sp
Thalassiosira ordinaris
Thalassiosira hispida
Thalassiosira allenii
Discostella pseudostelligera
Discostella stelligera
Discostella stelligeroides
Stephanodiscus triporus
Stephanodiscus minutulus
Cyclostephanos invisitatus
Cyclostephanos tholiformis
Thalassiosira sp
Thalassiosira pacifica
Detonula pumila
Detonula confervacea
Thalassiosira minuscula
Thalassiosira antarctica
Shionodiscus bioculatus
Shionodiscus oestrupii
Thalassiosira tumida
Thalassiosira punctigera
Thalassiosira delicatula
Thalassiosira grvidula

a)