Syst. Biol. 72(5):984–997, 2023

https://doi.org/10.1093/sysbio/syad038
Advance Access Publication June 19, 2023

Resolving Marine–Freshwater Transitions by Diatoms Through a Fog of Gene Tree Discordance

Wade R. Roberts, Elizabeth C. Ruck, Kala M. Downey, Eveline Pinseel, and Andrew J. Alverson*

Department of Biological Sciences, University of Arkansas, 1 University of Arkansas, Fayetteville, AR, 72701, USA
Correspondence to be sent to: Department of Biological Sciences, University of Arkansas, 1 University of Arkansas, Fayetteville, AR, 72701, USA; E-mail: aja@uark.edu.

Received 12 August 2022; reviews returned 2 June 2023; accepted 16 June 2023
Associate editor: Stacey Smith

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 and extinction. Diatoms are a lineage of ancestrally marine microalgae that have diversified throughout freshwater habitats worldwide. We generated a phylogenomic data set of genomes and transcriptomes for 59 diatom taxa to resolve freshwater transitions in one 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. Evidence from gene trees, protein alignments, and diatom life history together suggest that habitat transitions were largely the product of homoplasy rather than hemiplasy, a condition where transitions occur on branches in gene trees not shared with the species tree. Nevertheless, we identified a set of putatively hemiplasious genes, many of which have been associated with shifts to low salinity, indicating that hemiplasy played a small but potentially important role in freshwater adaptation. 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 further distinguish different sources of adaptive mutation in freshwater diatoms. [hemiplasy; homoplasy; phylogenomics; salinity, Thalassiosirales.]

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; Lee 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 (Pease et al. 2016).

Although phylogenomic data sets 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. Systematic errors that lead to gene tree discordance can be caused by biological sources, such as incomplete lineage sorting (ILS) and hybridization (Maddison 1997; Degnan and Rosenberg 2006), or methodological sources, such as character sampling, compositional heterogeneity, and gene tree error (Foster 2004; Philippe et al. 2011; Xi et al. 2015; Molloy and Warnow 2018). Each of these challenges our ability to resolve species relationships and impacts downstream analyses, such as estimation of divergence times (Smith et al. 2018). Multiple strategies have been proposed to overcome different sources of error, such as excluding third-codon positions from DNA data sets (Sanderson et al. 2000), using site-heterogeneous models for amino acid (AA) 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. Focus 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 encoded by genes with tree 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 AA 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). In conditions where low gene concordance is coupled with short branch lengths and shallow time scales, hemiplasy is expected to have a higher impact on trait reconstruction (Hahn and Nakhleh 2016).

Adaptation to low salinity is a complex trait, so the genomic changes associated with successful freshwater colonizations are multifaceted (Artemov et al. 2017; Cabello-Yeves and Rodriguez-Valera 2019; Rogers et al. 2021) and generally involve mutations in multiple genes and pathways (Jones et al. 2012; 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 data set 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 provide insight into how discordance impacted trait reconstruction.

Materials and Methods

Detailed methods are provided in Supplementary File S1. Briefly, diatom cultures were isolated from natural plankton or acquired from the National Center for Marine Algae and Microbiota or Roscoff Culture Collection. 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 AA 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 (CDS) alignments against AA alignments. To account for possible saturation at third-codon positions, we removed them from the CDS alignments and, separately, used Degen (Regier et al. 2010; Zwick et al. 2012) to recode synonymous sites with corresponding nucleotide ambiguity codes. For example, all leucine codons (CTN, TTR) were degenerated and replaced with YTN. In total, we analyzed data sets consisting of AAs, 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. 2013). We generated summary statistics for final alignments and gene trees with AMAS (Borowiec 2016) and PhyKit (Steenwyk et al. 2021) (Supplementary Table S3). Correspondence analysis of AA frequencies across taxa was performed using GCUA (McInerney 1998).

Species trees were inferred using gene-partitioned 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 gene partitions that violated assumptions of stationarity, reversibility, and homogeneity (SRH; Naser-Khdour et al. 2019). Briefly, stationarity refers to the assumption of constant AA or nucleotide frequencies across time, reversibility implies that substitutions between AAs or nucleotides occur equally, and homogeneity implies that the instantaneous substitution rates are constant along a tree or branch (Felsenstein 2004). 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). Branch support was estimated with local posterior probability values (Sayyari and Mirarab 2016). In addition to heterogeneity in gene histories, compositional heterogeneity can also make species tree inferences difficult at deep time scales (Lartillot and Philippe 2004). To account for this possibility, we estimated a species tree from the concatenated AA 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). Discordance was characterized using gene and site concordance factors (Minh et al. 2020a) and quartet concordance factors (Pease et al. 2018). We tested gene tree quartets for the multispecies coalescent (MSC) model using the R package *MSCquartets* (Rhodes et al. 2021; Allman et al. 2022). We tested the support for competing backbone topologies in our species tree using the approximately unbiased (AU) test (Shimodaira 2002) implemented in IQ-TREE. Relative gene tree support for the same set of backbone topologies was further evaluated using gene genealogy

Table 1. Summary of data sets used in the current study

Data type	Data set name	# genes	Total sites	Parsimony informative sites (%)	Avg bootstrap support (%)	# genes passing SRH[1]	***Thalassiosira*** **grade topology**	
						P < 0.05	IQ-TREE	ASTRAL
Amino acids (AA)	Complete	6,262	3,777,062	60	80	5,522	Topology 4	Topology 1
	Top-PI	1,588	996,027	76	84	1,314	Topology 4	Topology 1
	Top-Taxa	1,574	808,477	64	82	1,374	Topology 5	Topology 1
	Top-PI-top-Taxa + PMSF[2] model	488	246,300	77	85	488	Topology 5	Topology 1
1st + 2nd codon positions (CDS12)	Complete	6,262	7,554,124	55	80	3,259	Topology 4	Topology 5
	Top-PI	1,570	1,950,228	71	82	661	Topology 4	Topology 3
	Top-Taxa	1,574	1,616,954	59	82	782	Topology 5	Topology 3
Degenerate codons (DEGEN)	Complete	6,262	11,311,186	33	81	3,788	Topology 4	Topology 3
	Top-PI	1,569	2,884,728	44	83	771	Topology 4	Topology 2
	Top-Taxa	1,574	2,425,431	36	83	903	Topology 5	Topology 3

[1] Matched-pairs test of symmetry, reversibility, and homogeneity.
[2] Posterior mean site frequency model.

interrogation (GGI) (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 with MCMCtree (Yang 2007; Reis and Yang 2011), using five fossil calibrations (Supplementary File S1), autocorrelated rates, and the approximate likelihood approach. We performed marginal ancestral state reconstruction for marine and freshwater habitat on species trees using hidden state speciation and extinction models in the R package *hisse* (Beaulieu and O'Meara 2016). Lastly, we explored the potential for hemiplasy in habitat reconstructions by calculating: 1) hemiplasy risk factors using the R package *pepo* (Guerrero and Hahn 2018); 2) the differences between the number of marine–freshwater transitions optimized on unconstrained versus species-tree-constrained gene trees; and 3) the number of alignment site changes that differed between unconstrained and species-tree-constrained gene trees.

Phylogenetic trees were plotted using the R Bioconductor packages *ggtree* and *treeio* (Yu et al. 2017; Wang et al. 2019). Assembled genomes and transcriptomes have been deposited at NCBI under BioProject PRJNA825288. Supplementary files, figures, and tables are available from the Dryad Digital Repository (https://doi.org/10.5061/dryad.7m0cfxpxp). Voucher images, proteomes, alignments, trees, log files, and code have been deposited in Zenodo (https://doi.org/10.5281/zenodo.7713227).

Results

We combined 42 newly sequenced draft genomes and 50 newly sequenced transcriptomes with publicly available genome or transcriptome data for a final data set of 87 taxa representing 59 species. A total of 17 transcriptomes were used directly in phylogenomic analyses, and the other 33 were used for genome annotation. From the combined data set of genomes and transcriptomes, we generated alignments and gene trees for 6262 orthologs, with each taxon represented in an average of 3275 (52%) orthologs.

Compositional Heterogeneity and Data Set Construction

Relative composition variability (Phillips and Penny 2003) indicated greater compositional heterogeneity in third codon positions compared to AAs 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*. In addition, nucleotide alignments including all three codon positions had significantly higher levels of saturation compared to AAs (Supplementary Fig. S3). To minimize potentially misleading signal in the nucleotide data sets, we removed third codon positions and, to examine the effects of saturation and GC heterogeneity in CDSs, we minimized synonymous signal by recoding CDS with degenerate codons (Zwick et al. 2012). Based on these results, we created three data sets to estimate phylogenetic relationships: AA, CDS12, and DEGEN.

Data set and alignment characteristics are summarized in Table 1 and detailed in Supplementary Table S3. Each data set initially contained 6262 orthologs (Table 1). Gene trees constructed from all data sets were well supported (80 ± 7% average bootstrap; Table 1; Supplementary Table S3). To reduce systematic errors due to model misspecification, we removed orthologs that failed SRH assumptions ($P < 0.05$), which retained 5522 (AA), 3259 (CDS12), and 3788 (DEGEN) orthologs in the data sets (hereafter referred to as the "complete" data sets; Table 1). We then subsetted the complete data sets to maximize phylogenetic 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 data sets by the number of taxa and subsetted these to include the top 25% ranked orthologs with the highest taxon occupancy (“top-Taxa”; Table 1). Orthologs in the top-Taxa data sets contained an average of 76 ± 8% of the total taxa.

Species Tree Inference and Placement of Freshwater Clades

We initially estimated 18 species trees using AAs, codon positions 1 and 2, or degenerate codon sequences, with different cutoffs for taxon occupancy or proportion of PI sites and using both concatenation and summary quartet approaches (Table 1). Correspondence analysis of AA frequencies separated taxa principally by habitat (marine vs. freshwater) rather than phylogeny (Supplementary Fig. S4), which led us to explore whether compositional heterogeneity affected phylogenetic reconstructions. To do so, we estimated an additional species tree using the PMSF mixture model, which can accommodate heterogeneity in AA composition between species at each site (Wang et al. 2018). Due to the computational demands of implementing this model, we applied it only to a reduced AA data set with orthologs that met both the top-PI and top-Taxa filtering criteria (“AA-top-PI-top-Taxa”; Table 1).

Previous phylogenetic analyses of this group 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 to 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 data sets that maximized signal (top-PI) or minimized missing data (top-Taxa) (Supplementary Fig. S5). The phylogenetic position of *Cyclotella* was strongly supported and robust to differences in data type (codons vs. AAs) 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) was recovered only by ASTRAL analysis of AA gene trees and 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 AA 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 in relation to *Cyclostephanos* in 18 of the 20 species trees. Considering all of these results, we chose topology 5 from the PMSF analysis as the reference species tree (Fig. 1).

Discordance Underlies Topological Uncertainty

Incongruence among gene trees and alignment sites is an important factor impacting phylogenetic reconstruction (Degnan and Rosenberg 2006; Mallet et al. 2016). We characterized discordance by calculating gene, site, and quartet concordance factors for each branch (Figs. 1 and 2a; Supplementary Figs. S6 and S7). Gene and site concordance factors (gCF and 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 33 to 100, with values near 33 indicative of no signal for that branch (Minh et al. 2020a). Quartet concordance factors (QC) provide a 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. S6 and S7). Despite having maximum bootstrap and local posterior probability support, however, concordance factors were low for many of the backbone branches (Fig. 1; Supplementary Figs. S6 and S7), including ones 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. S6 and S7). 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), which had low site concordance (Fig. 2a) and near-zero quartet

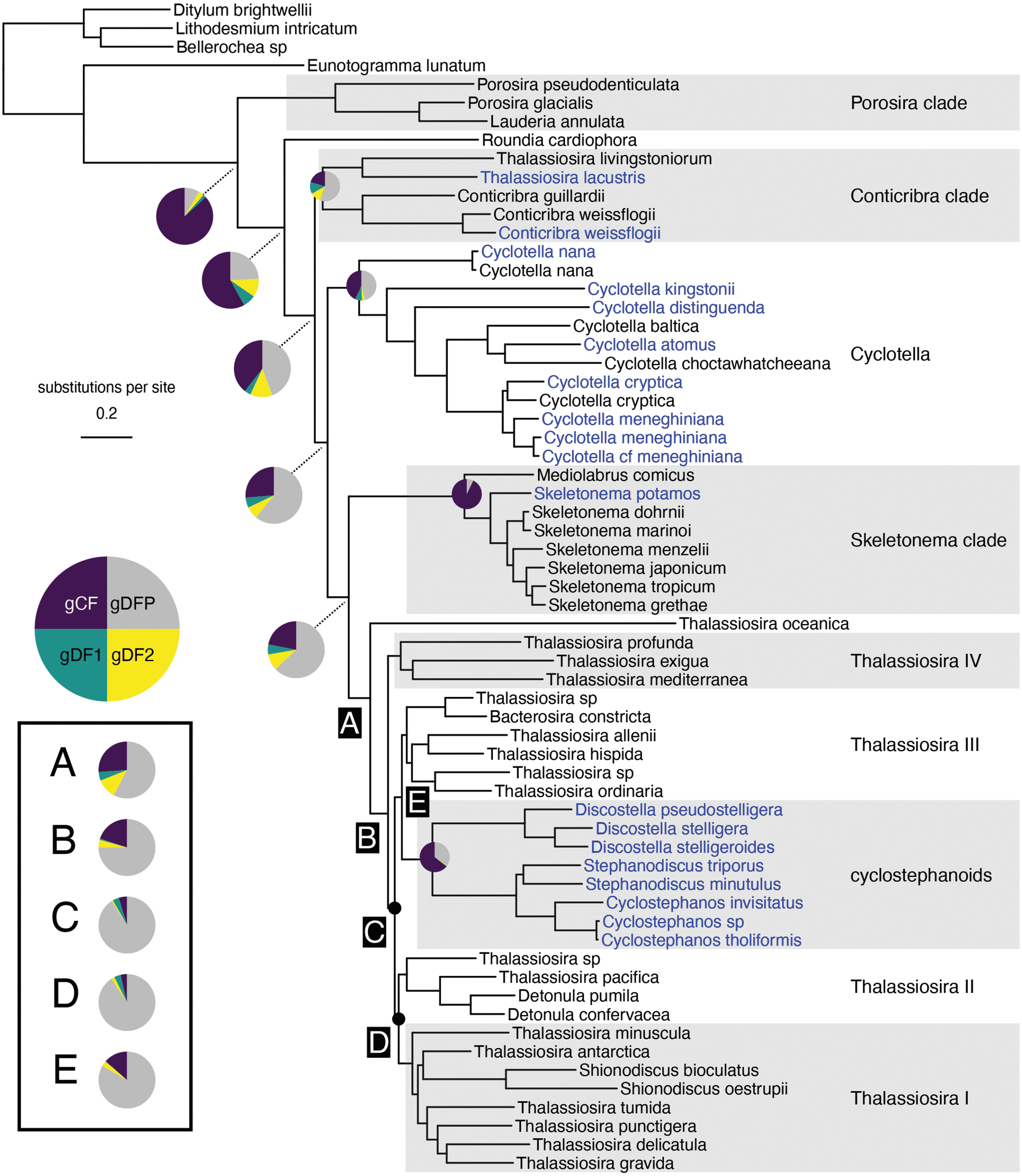

FIGURE 1. Phylogram based on maximum likelihood analysis of amino acids using the posterior mean site frequency (PMSF) model and a data set of 488 loci with the highest proportions of taxa and informative sites ("AA-top-PI-top-Taxa" data set; Table 1). Backbone nodes of the *Thalassiosira* grade are 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 supports the clade (gCF), the proportion that supports both discordant topologies (gDF1, gDF2), and the proportion that are discordant due to paraphyly (gDFP). Size of the pie charts is for clarity only. Color of taxon names indicate ecological habitat: marine (black) or freshwater (blue).

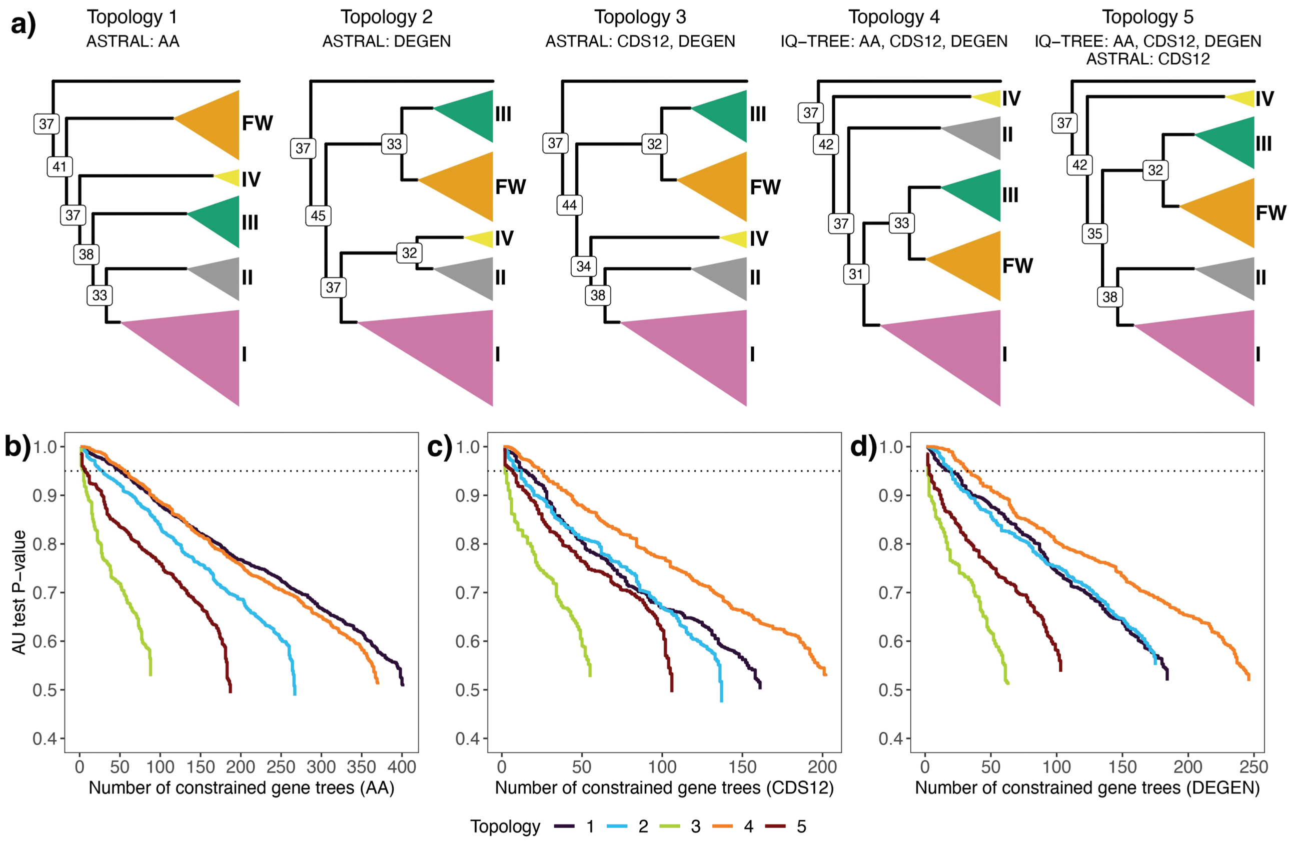

FIGURE 2. a) Phylogenetic hypotheses of the *Thalassiosira* grade inferred using concatenation and summary methods on the amino acid (AA), codon positions 1 and 2 (CDS12), and recoded codon (DEGEN) data sets. Nodes are labeled with the percentage of amino acid sites concordant with the branch (site concordance 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. Panels b–d show results of gene genealogy interrogation tests of alternative hypotheses of relationships within the *Thalassiosira* grade. These tests used data sets 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 panel b shows that there were a total of 88 genes that best-supported topology 3, though only four of those genes were above the dotted line and were significantly better supported than the other four alternative topologies.

concordance factors (Supplementary Fig. S7), consistent with expectations for little to no signal. Site concordance factors did not change appreciably with the CDS12 data set. However, repeating quartet concordance factor calculations with the CDS12 data set resulted in minor switches in support for branches C ($\mathrm{QC}_C = 0.004 \rightarrow -0.006$) and E ($\mathrm{QC}_E = -0.04 \rightarrow 0.07$) (Supplementary Fig. S7). These patterns of shifting support, though minor, combined with the lowest site concordance factor for branch E ($\mathrm{sCF}_E = 32$) 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).

For some branches on the species tree, there were more discordant than concordant genes and sites (Fig. 1; Supplementary Figs. S6 and S7), 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. S6 and S7). 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 paraphyly (Supplementary Fig. S6), indicative of a widespread lack of signal in gene trees. Taken together, these analyses highlighted extensive gene and site discordance, some well-supported and much of it not, across Thalassiosirales.

There are both biological (e.g., introgression) and technical causes (e.g., gene tree error) of gene discordance, but the proportion attributable to each factor can be difficult to distinguish (Morales-Briones et al. 2020; Cai et al. 2021). Less than 1% of gene tree quartets rejected the MSC model tested with MSCquartets (model T3, Holm–Bonferroni α = 0.001; Supplementary Fig. S8), which suggests introgression was relatively unimportant in this data set. To better assess the importance of gene tree error—whether genes with the low phylogenetic signal produced inaccurate gene trees—we recalculated gene and site concordance using just the 1588 AA orthologs with the highest percentage of PI sites (top-PI data set; Table 1). Average gene concordance increased modestly, from 56.6% to 62.1%, but site concordance was unchanged (Supplementary Fig. S9). 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 (Supplementary Fig. S9). 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 data set construction (Supplementary File S1). We found slight negative correlations between node age and both gene concordance ($R^2 = 0.10$, $P < 0.01$) and site concordance ($R^2 = 0.26$, $P < 0.001$) (Supplementary Fig. S10), 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 assess the relative support for topologies 1–5 within the *Thalassiosira* grade (Fig. 2a). Like the original species tree inferences, results of AU tests on concatenated alignments for each data set in Table 1 largely reflected data type (Table 1), with AA characters supporting topology 1 ($P = 0.01$, AU test) and the codon and degenerate codon data sets supporting topology 4 ($P = 0$, AU test) (Supplementary Table S4). We used GGI to look for a secondary signal supporting one or more of the five competing topologies (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 a 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 data sets, implying similar levels of support (Fig. 2b–d; Supplementary Table S5). Support from the AA data set 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 data sets corresponded to topology 4 (Fig. 2c, d), which was originally recovered by concatenation only (Table 1).

GGI can also be used to explore the effects of gene tree error on summary quartet methods by filtering the input trees for ASTRAL to include only the highest-ranking constrained genes (Arcila et al. 2017; Mirarab 2017). For each of the three data sets, we performed two ASTRAL analyses using as input either all the top scoring (rank 1) constrained gene trees ($n_{\mathrm{AA}} = 1588$, $n_{\mathrm{CDS12}} = 1570$, and $n_{\mathrm{DEGEN}} = 1569$) or just the subset that had statistical support ($P < 0.05$) above the AU-based rank 2 topology ($n_{\mathrm{AA}} = 142$, $n_{\mathrm{CDS12}} = 56$, and $n_{\mathrm{DEGEN}} = 71$). 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 AAs 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 a so-called “anomaly zone” (Degnan and Rosenberg 2006). However, all calculations of the anomaly zone boundary $a(x)$ on our species trees were below zero (Degnan and Rosenberg 2006), indicating that no internode branch pairs were likely to produce anomalous gene trees. Additionally, polytomy tests in ASTRAL using each data set rejected the null hypothesis that any of these branches was 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 (95% CI: 96–120 Ma) (Fig. 3a; Supplementary Fig. S11). One of the two main freshwater lineages, *Cyclotella*, originated in the late Cretaceous (95% CI: 66–86 Ma) and the other freshwater lineage, the cyclostephanoids, originated later in the Eocene (95% CI: 36–48 Ma) (Fig. 3a; Supplementary Fig. S11). Radiation of the *Thalassiosira* grade lineages occurred during the Paleocene, from 57 Ma (95% CI: 48–63 Ma) to 73 Ma (95% CI: 61–80 Ma) (Fig. 3a; Supplementary Fig. S11). The overlapping confidence intervals allow for the possibility that these lineages diverged in much more rapid succession than suggested by their mean ages.

Marine–Freshwater Transitions on Gene versus Species Trees

We used HiSSE to estimate the number of independent marine–freshwater transitions on the time-calibrated species tree (Fig. 3a). The best-fit HiSSE model was a character-independent model (CID-4; Supplementary Table S6), which indicates that shifts in diversification rate occurred independently of marine–freshwater transitions. Using parameter estimates from the CID-4 model, we inferred a total of

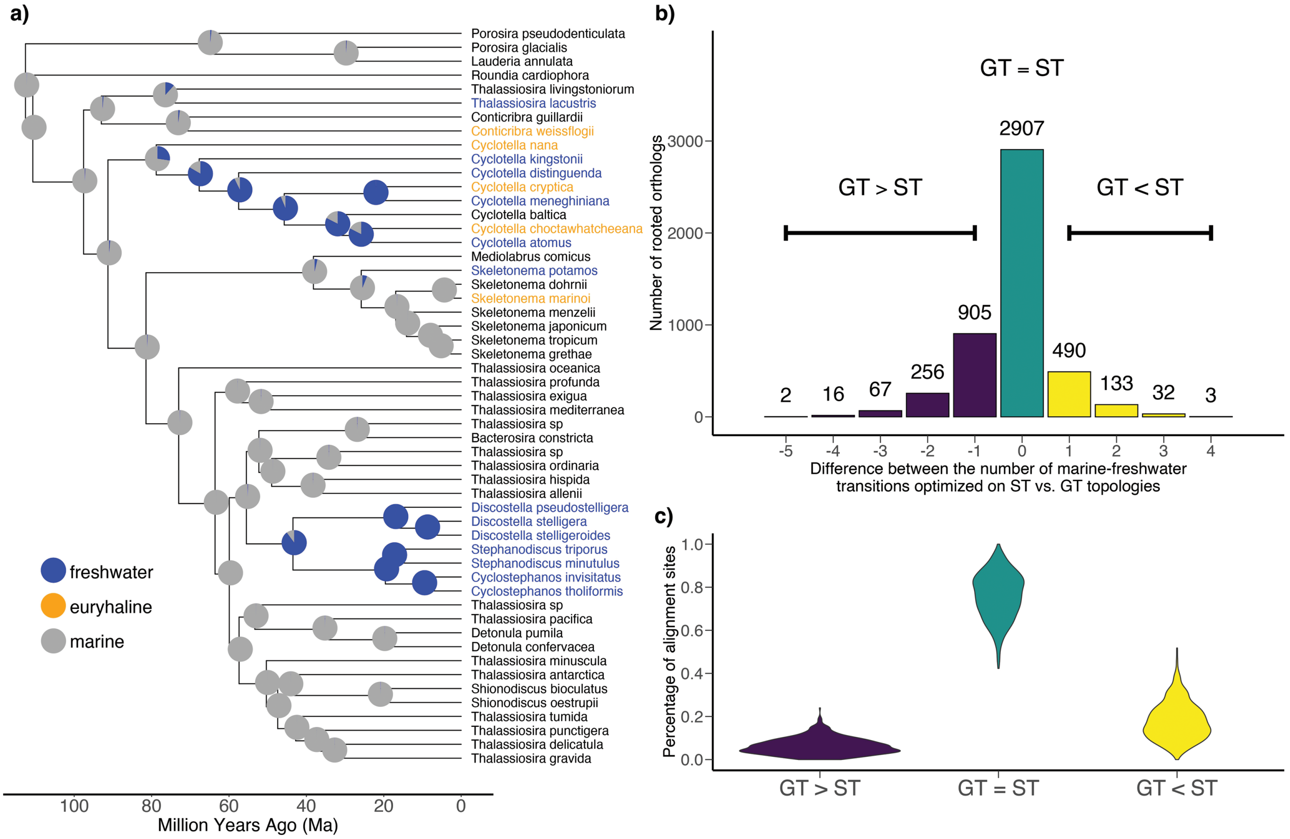

FIGURE 3. a) Divergence times and ancestral state reconstruction of marine and freshwater habitat in Thalassiosirales. Conspecific taxa were removed prior to ancestral state reconstruction, leaving one tip per species. Pie charts denote the probability of each node reconstructed as either marine (grey) 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. S11. b) Summary of the difference between the number of parsimony optimized marine–freshwater transitions on rooted orthologs with the estimated gene tree topology (GT) versus the topology constrained to match the species tree (ST). Bar colors denote whether a GT had greater (purple), equal (green), or fewer (yellow) numbers of transitions relative to the ST. Numbers above bars are the number of rooted orthologs in each bar. c) Summary of the percentage of aligned amino acid sites that have greater (purple), equal (green), or fewer (yellow) numbers of state transitions on the GT versus ST.

six transitions from marine to freshwaters (Fig. 3a). Given a large number of discordant gene trees (Fig. 1), we next sought to determine whether our reconstructions were impacted by hemiplasy, that is, marine–freshwater transitions that occurred on branches in a gene tree not shared with the species tree. Narrowly defined, hemiplasy describes discordance due to ILS (Avise and Robinson 2008), but we use the term hemiplasy broadly to describe discordant gene trees with fewer overall habitat transitions than the species tree, regardless of the underlying cause (see Hahn and Nakhleh 2016).

To assess the potential for hemiplasy in our trait reconstructions, we calculated hemiplasy risk factors (HRFs), which are the ratio of the probabilities of hemiplasy to homoplasy in different parts of the species tree (Guerrero and Hahn 2018). HRFs were calculated using 12 different sets of coalescent branch lengths that represent a range of plausible effective population size (N_e) and generation times for diatoms (Supplementary Fig. S12). These estimates of coalescent branch lengths using taxon-specific life history information should be more accurate than estimates from gene tree summary methods such as ASTRAL, which are prone to underestimating branch lengths in the presence of high gene tree estimation error (Sayyari and Mirarab 2016; Forthman et al. 2022). The risk of hemiplasy ranged from low to nonexistent in nearly all scenarios involving realistic generation times for diatoms (1 day, 7 days, and 1 month), whereas elevated hemiplasy risk was found only under scenarios of unrealistically long generation times (6 months) (Supplementary Fig. S12).

We next assessed whether there was any empirical evidence for hemiplasy in the alignments and gene trees. For each ortholog, we separately optimized, using parsimony, the number of marine–freshwater transitions on the estimated gene tree (GT) and the gene tree constrained to match the species tree topology (ST). Alternative model-based approaches to assess hemiplasy require a nucleotide substitution rate as proxy for the character transition rate, thereby assuming that character transitions are controlled by single nucleotide changes (Hibbins et al. 2020). Although our understanding of the genetic architecture of salinity tolerance in diatoms is limited, this complex trait is likely controlled by multiple loci (Pinseel et al. 2022; Downey et al. 2023), with possible epistatic interactions (Stern et al. 2022). We, therefore, used parsimony as a simple, empirical approach because it does not rely on unknown character transition rates and assumes that the number of gene trees containing hemiplasies is a reasonable proxy for the probability of hemiplasy. We calculated the difference between the number of transitions on the ST versus GT and found a small percentage of orthologs (13.7%) with at least one instance of hemiplasy, broadly defined to include all instances with fewer transitions on the gene versus species tree topology (GT < ST; Fig. 3b). This set of orthologs likely includes topologies affected by hemiplasy due to ILS (Avise and Robinson 2008), homoplasy (convergent or parallel substitutions), and gene tree error. Manual inspection of these gene trees showed that many of them were paraphyletic and occurred between *Thalassiosira lacustris*, *Conticribra weissflogii*, *C. nana*, and the remaining *Cyclotella* taxa, whereas a smaller number were polyphyletic and involved *Cyclotella, Skeletonema potamos*, and the cyclostephanoid clade (Supplementary Table S7). The remaining majority of orthologs either had identical numbers of marine–freshwater transitions (GT = ST), matching the species tree, or had more transitions on the gene tree (GT > ST), which we attribute to error (Fig. 3b).

To determine whether the set of putatively hemiplasious (GT < ST) genes might be related to salinity adaptation, we compared them against a set of genes that were differentially expressed in a low-salinity acclimation experiment using the marine/euryhaline diatom *Skeletonema marinoi* (Pinseel et al. 2022). A total of 532 of the 658 rooted orthologs with GT < ST were present in the *S. marinoi* genome, and 203 of these (~38%) were differentially expressed in response to salinity change (Supplementary Table S8). This included genes involved in key pathways and functions important in acclimation to low salinity, such as carotenoid biosynthesis, the xanthophyll cycle, light-harvesting activities, protein chaperones, nitrogen assimilation, regulation of transcription and translation, and the metabolism of polyamines, AAs, and fatty acids (Supplementary Table S8) (Cheng et al. 2014; Bussard et al. 2017; Pinseel et al. 2022; Downey et al. 2023).

For the same set of 658 orthologs, we then calculated the difference in the number of AA transitions that occurred at each alignment site using the inferred GT versus constrained ST topologies (Fig. 3c). Although agnostic to marine–freshwater transitions, this approach measured the impact of gene tree discordance, irrespective of the underlying cause, to overall molecular homoplasy. Using this approach, we found that an average of 18% of AA sites within an ortholog had fewer transitions on the GT versus the ST (ST < GT; Fig. 3c).

Discussion

Transitions to Freshwaters

Marine–freshwater transitions have been key events in the diversification of lineages across the tree of life (Jamy et al. 2022), including diatoms where freshwater taxa have experienced increased rates of both speciation and extinction compared to their marine ancestors (Nakov et al. 2019). Our study focused on a 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). Data from the 42 genomes presented here greatly expand the phylogenetic 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 marine–freshwater transitions in Thalassiosirales and tested whether these transitions were fully independent, owing to separate parallel or convergent mutations (homoplasy) or whether some of the transitions were attributable to hemiplasy, that is, the sorting of shared ancestral polymorphisms in discordant gene trees (Avise and Robinson 2008). Even in the face of extensive gene tree discordance, the probability of hemiplasy influencing trait reconstructions was virtually nonexistent across the backbone of the tree (Supplementary Fig. S12). The risk of hemiplasy was only elevated under scenarios with unrealistically long generation times for diatoms (Supplementary Fig. S12), many of which can undergo daily cell divisions (Smayda 1969; Tuchman et al. 1984). The large effective population sizes of some microbes, which increase the probability of hemiplasy, may be offset by their rapid generation times, reducing the risk of hemiplasy and highlighting the need for better empirical estimates of N_e and μ in groups like diatoms (Guerrero and Hahn 2018). The relatively few empirical studies of hemiplasy so far have focused on vascular plants or animals with generation times that are years or decades in length (Copetti et al. 2017; Wu et al. 2018; Vizueta et al. 2019). These studies also focused on much younger lineages, making it more likely that shared ancestral polymorphisms could still be detected.

Longer branches subtending separate clades with a shared trait make it more likely that homoplasy, rather than hemiplasy, has occurred (Hahn and Nakhleh 2016). In our case, each of the clades with freshwater taxa has sufficiently long coalescent branch lengths to have achieved monophyly and have had multiple intervening speciation events leading to non-freshwater taxa (Fig. 3a). Other properties of the habitat transitions were also indicative of homoplasy rather than hemiplasy: ancestral state reconstructions were unambiguous, the six freshwater transitions were not paraphyletic, and gene concordance was high on the long internal stem branches subtending the two main freshwater lineages (Hahn and Nakhleh 2016; Wu et al. 2018).

Empirical analyses of gene trees and alignments revealed relatively small numbers of genes and sites with putative hemiplasies (Fig. 3b,c). We identified approximately three-fold fewer AA sites with potential hemiplasies than was found in a much younger (10 Ma) radiation of saguaro cacti (Copetti et al. 2017). Given the length of time that has elapsed since the two major freshwater transitions occurred, the probability that shared ancestral polymorphisms have persisted decreases and the probability that lineage-specific adaptive mutations have evolved increases (Suh et al. 2015; 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; Pinseel et al. 2022; Downey et al. 2023), 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, any hemiplasious alleles were most likely overwritten in the extant freshwater descendants, leaving a shared ancestral phenotype (e.g., enhanced transport of sodium ions) as the only remaining evidence of the hemiplasy. The divergent evolutionary outcomes also suggest a greater role for homoplasy over hemiplasy. 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. 2023). Similarly, most freshwater transitions at the tips of the tree (Fig. 1) involve 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, narrowly adapted 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. Moreover, salinity fluctuations are common in brackish and marine systems, such as coastlines influenced by precipitation and river discharge. A substantial number of the putatively hemiplasious genes identified here are also involved in acclimation to low salinity conditions in one of the species in our analysis, *S. marinoi* (Pinseel et al. 2022), suggesting a possible role of hemiplasy in freshwater adaptation for this group of diatoms. In addition, adaptation to low or changing salinity may be linked to modifications of gene expression (Bussard et al. 2017; Nakov et al. 2020; Pinseel et al. 2022; Downey et al. 2023), 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). The genomic resources and phylogenetic framework presented here represent an important advance toward identifying the genes and processes underpinning freshwater adaptation by diatoms.

The Impact of Discordance on Placement of Freshwater Clades

Comparative analyses require a strong phylogenetic framework, so a major goal of this study was to establish a robust phylogenetic hypothesis for Thalassiosirales. Our data set of 6262 nuclear orthologs provided better resolution and, superficially, increased support across most of the tree compared to previous analyses (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 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 data sets and analyses. Uncertainty in the backbone relationships for this part of the tree was likely caused by a combination of gene tree error and ILS. 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 data set (Chan et al. 2020; Arcila et al. 2021). Divergence time estimates suggest that these splits occurred in 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 most phylogenetic signal, implicating gene tree error as one source of instability (Chan et al. 2020; Vanderpool et al. 2020). In other phylogenomic studies impacted by high gene tree error, GGI has been used to identify the 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 AA and codon-based gene trees, highlighting conflicting signal even within genes. In addition to gene tree error, the large number of alternative topologies among gene trees in 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 cause 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 tried to eliminate noise in our data set 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 summary and concatenation methods. Taken together, these results suggest that gene tree error negatively impacted our ASTRAL analyses. After identifying and removing some of that error, we recovered stronger support for the placement of freshwater cyclostephanoids.

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 tree 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 or weakly supported gene trees are more probable than resolved ones in the *Thalassiosira* grade. Under ILS, 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 data set can lead to underestimation of coalescent branch lengths by summary methods like ASTRAL (Sayyari and Mirarab 2016; Forthman et al. 2022). Coalescent branch lengths define the anomaly zone boundaries (Degnan and Rosenberg 2006; Linkem et al. 2016), so underestimates could result in mistaken identity of an anomaly zone where none exists. After incorporating realistic estimates of N_e and generation times for diatoms in our coalescent branch length estimates, the *Thalassiosira* grade fell outside the calculated boundaries of the anomaly zone (Degnan and Rosenberg 2006).

Codon Bias and Amino Acid Composition in Freshwater Diatoms

The phylogeny of Thalassiosirales includes divergence times across a timescale ranging from thousands (Theriot et al. 2006) to tens of millions of years (Fig. 3a), which led us to explore the utility of both AA and nucleotide characters for resolving phylogenetic relationships. AAs 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 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. Different schemes to mitigate the effects of saturation in the codon data sets also produced disagreements within the *Thalassiosira* grade. Disagreements between character types within the same data set, such as those within the *Thalassiosira* grade here, have also been found in other groups (Gillung et al. 2018; Skinner et al. 2020).

Almost every analysis of the AA data set—including species trees, concordance factors, and AU tests—supported the placement of cyclostephanoids as sisters to the remaining *Thalassiosira* clades, but nucleotide analyses placed them with *Thalassiosira* III (Fig. 2). Discordance caused by codon usage bias and differences in AA 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). Differences in codon usage between marine and freshwater prokaryotes have been described (Cabello-Yeves and Rodriguez-Valera 2019), and we discovered differences in both codon usage and AA composition between marine and freshwater diatoms. The AA compositions of distantly related freshwater lineages might be sufficiently similar to cause AA 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 AA 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 compositional heterogeneity in AA sites, cyclostephanoids were placed as sisters to *Thalassiosira* III, in agreement with the codon data sets. The similarity in codon usage and AA 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. Low salinity provokes a broad range of physiological and metabolic responses in diatoms (Nakov et al. 2020; Pinseel et al. 2022; Downey et al. 2023), 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 be difficult to identify specific alleles—whatever the process that generated them—that currently allow these diatoms to thrive in freshwaters. Nevertheless, the phylogenomic analyses presented here suggest that convergent or parallel substitutions likely played a more important role than hemiplasy in facilitating freshwater colonizations in this group of diatoms. 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 superficially simple.

SUPPLEMENTAL MATERIAL

Supplementary materials, including files, tables, and figures, can be found in the Dryad Digital Repository (https://doi.org/10.5061/dryad.7m0cfxpxp). Voucher images, proteomes, alignments, trees, log files, and code have been deposited in Zenodo (https://doi.org/10.5281/zenodo.7713227).

ACKNOWLEDGMENTS

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

FUNDING

This work was supported by the National Science Foundation (DEB 1651087 to A.J.A) and the Simons Foundation (725407 to E.P.). 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.

References

Allman E.S., Mitchell J.D., Rhodes J.A. 2022. Gene tree discord, simplex plots, and statistical tests under the coalescent. Syst. Biol. 71:929–942.

Alverson A.J., Beszteri B., Julius M.L., Theriot E.C. 2011. The model marine diatom *Thalassiosira pseudonana* likely descended from a freshwater ancestor in the genus *Cyclotella*. BMC Evol. Biol. 11:125.

Alverson A.J., Jansen R.K., Theriot E.C. 2007. Bridging the Rubicon: phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. Mol. Phylogenet. Evol. 45:193–210.

Arcila D., Hughes L.C., Meléndez-Vazquez B., Baldwin C.C., White W.T., Carpenter K.E., Williams J.T., Santos M.D., Pogonoski J.J., Miya M., Ortí G., Betancur-R R. 2021. Testing the utility of alternative metrics of branch support to address the ancient evolutionary radiation of tunas, stromateoids, and allies (Teleostei: Pelagiaria). Syst. Biol. 70:1123–1144.

Arcila D., Ortí G., Vari R., Armbruster J.W., Stiassny M.L.J., Ko K.D., Sabaj M.H., Lundberg J., Revell L.J., Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:1–10.

Arella D., Dilucca M., Giansanti A. 2021. Codon usage bias and environmental adaptation in microbial organisms. Mol. Genet. Genomics 296:751–762.

Armbrust E.V., Berges J.A., Bowler C., Green B.R., Martinez D., Putnam N.H., Zhou S., Allen A.E., Apt K.E., Bechner M., Brzezinski M.A., Chaal B.K., Chiovitti A., Davis A.K., Demarest M.S., Detter J.C., Glavina T., Goodstein D., Hadi M.Z., Hellsten U., Hildebrand M., Jenkins B.D., Jurka J., Kapitonov V.V., Kröger N., Lau W.W.Y., Lane T.W., Larimer F.W., Lippmeier J.C., Lucas S., Medina M., Montsant A., Obornik M., Parker M.S., Palenik B., Pazour G.J., Richardson P.M., Rynearson T.A., Saito M.A., Schwartz D.C., Thamatrakoln K., Valentin K., Vardi A., Wilkerson F.P., Rokhsar D.S. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86.

Artemov A.V., Mugue N.S., Rastorguev S.M., Zhenilo S., Mazur A.M., Tsygankova S.V., Boulygina E.S., Kaplun D., Nedoluzhko A.V., Medvedeva Y.A., Prokhortchouk E.B. 2017. Genome-wide DNA methylation profiling reveals epigenetic adaptation of stickleback to marine and freshwater conditions. Mol. Biol. Evol. 34:2203–2213.

Avise J.C., Robinson T.J. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. Syst. Biol. 57:503–507.

Beaulieu J.M., O'Meara B.C. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Syst. Biol. 65:583–601.

Borowiec M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660.

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

Bussard A., Corre E., Hubas C., Duvernois-Berthet E., Le Corguillé G., Jourdren L., Coulpier F., Claquin P., Lopez P.J. 2017. Physiological adjustments and transcriptome reprogramming are involved in the acclimation to salinity gradients in diatoms. Environ. Microbiol. 19:909–925.

Cabello-Yeves P.J., Rodriguez-Valera F. 2019. Marine-freshwater prokaryotic transitions require extensive changes in the predicted proteome. Microbiome 7:117.

Cai L., Xi Z., Lemmon E.M., Lemmon A.R., Mast A., Buddenhagen C.E., Liu L., Davis C.C. 2021. The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. Syst. Biol. 70:491–507.

Chan K.O., Hutter C.R., Wood P.L. Jr, Grismer L.L., Brown R.M. 2020. Target-capture phylogenomics provide insights on gene and species tree discordances in Old World treefrogs (Anura: Rhacophoridae). Proc. Biol. Sci. 287:20202102.

Chen M.-Y., Teng W.-K., Zhao L., Hu C.-X., Zhou Y.-K., Han B.-P., Song L.-R., Shu W.-S. 2021. Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. ISME J. 15:211–227.

Cheng R.-L., Feng J., Zhang B.-X., Huang Y., Cheng J., Zhang C.-X. 2014. Transcriptome and gene expression analysis of an oleaginous diatom under different salinity conditions. Bioenergy Res. 7:192–205.

Copetti D., Búrquez A., Bustamante E., Charboneau J.L.M., Childs K.L., Eguiarte L.E., Lee S., Liu T.L., McMahon M.M., Whiteman N.K., Wing R.A., Wojciechowski M.F., Sanderson M.J. 2017. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc. Natl. Acad. Sci. U.S.A. 114:12003–12008.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Dickson B., Yashayaev I., Meincke J., Turrell B., Dye S., Holfort J. 2002. Rapid freshening of the deep North Atlantic Ocean over the past four decades. Nature 416:832–837.

Dittami S.M., Heesch S., Olsen J.L., Collén J. 2017. Transitions between marine and freshwater environments provide new clues about the origins of multicellular plants and algae. J. Phycol. 53:731–745.

Downey K.M., Judy K.J., Pinseel E., Alverson A.J., Lewis J.A. 2023. The dynamic response to hypo-osmotic stress reveals distinct stages of freshwater acclimation by a euryhaline diatom. Mol. Ecol. 32:2766–2783.

Edwards S.V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Emms D.M., Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Forthman M., Braun E.L., Kimball R.T. 2022. Gene tree quality affects empirical coalescent branch length estimation. Zool. Scr. 51:1–13.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Gillung J.P., Winterton S.L., Bayless K.M., Khouri Z., Borowiec M.L., Yeates D., Kimsey L.S., Misof B., Shin S., Zhou X., Mayer C., Petersen M., Wiegmann B.M. 2018. Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids. Mol. Phylogenet. Evol. 128:233–245.

Guerrero R.F., Hahn M.W. 2018. Quantifying the risk of hemiplasy in phylogenetic inference. Proc. Natl. Acad. Sci. U.S.A. 115:12787–12792.

Guillard R.R.L., Ryther J.H. 1962. Studies of marine planktonic diatoms: I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. Can. J. Microbiol. 8:229–239.

Hahn M.W., Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution 70:7–17.

Hibbins M.S., Gibson M.J., Hahn M.W. 2020. Determining the probability of hemiplasy in the presence of incomplete lineage sorting and introgression. Elife 9:e63753.

Huang H., Knowles L.L. 2009. What is the danger of the anomaly zone for empirical phylogenetics? Syst. Biol. 58:527–536.

Hughes L.C., Ortí G., Huang Y., Sun Y., Baldwin C.C., Thompson A.W., Arcila D., Betancur-R R., Li C., Becker L., Bellora N., Zhao X., Li X., Wang M., Fang C., Xie B., Zhou Z., Huang H., Chen S., Venkatesh B., Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. U.S.A. 115:6249–6254.

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

Jombart T., Kendall M., Almagro-Garcia J., Colijn C. 2017. treespace: statistical exploration of landscapes of phylogenetic trees. Mol. Ecol. Resour. 17:1385–1392.

Jones F.C., Grabherr M.G., Chan Y.F., Russell P., Mauceli E., Johnson J., Swofford R., Pirun M., Zody M.C., White S., Birney E., Searle S., Schmutz J., Grimwood J., Dickson M.C., Myers R.M., Miller C.T., Summers B.R., Knecht A.K., Brady S.D., Zhang H., Pollen A.A., Howes T., Amemiya C., Baldwin J., Bloom T., Jaffe D.B., Nicol R., Wilkinson J., Lander E.S., Di Palma F., Lindblad-Toh K., Kingsley D.M.; Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484:55–61.

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

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kenny N.J., Plese B., Riesgo A., Itskovich V.B. 2019. Symbiosis, selection and novelty: freshwater adaptation in the unique sponges of lake Baikal. Mol. Biol. Evol. 36:2462–2480.

Kirst G.O. 1990. Salinity tolerance of eukaryotic marine algae. Annu. Rev. Plant Physiol. Plant Mol. Biol. 41:21–53.

Kirst G.O. 1996. Osmotic adjustment in phytoplankton and macroalgae. In: Kiene R.P., Visscher P.T., Keller M.D., Kirst G.O., editors. Biological and environmental chemistry of DMSP and related sulfonium compounds. Boston (MA): Springer US. p. 121–129.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lee C.E., Downey K., Colby R.S., Freire C.A., Nichols S., Burgess M.N., Judy K.J. 2022. Recognizing salinity threats in the climate crisis. Integr. Comp. Biol. 62:441–460.

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: Scincidae). Syst. Biol. 65:465–477.

Liu L., Wu S., Yu L. 2015. Coalescent methods for estimating species trees from phylogenomic data. J. Syst. Evol. 53:380–390.

Lozupone C.A., Knight R. 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U.S.A. 104:11436–11440.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Mallet J., Besansky N., Hahn M.W. 2016. How reticulated are species? Bioessays 38:140–149.

McCairns R.J.S., Bernatchez L. 2010. Adaptive divergence between freshwater and marine sticklebacks: insights into the role of phenotypic plasticity from an integrated analysis of candidate gene expression. Evolution 64:1029–1047.

McInerney J.O. 1998. GCUA: general codon usage analysis. Bioinformatics 14:372–373.

Mendes F.K., Hahn Y., Hahn M.W. 2016. Gene tree discordance can generate patterns of diminishing convergence over time. Mol. Biol. Evol. 33:3299–3307.

Minh B.Q., Hahn M.W., Lanfear R. 2020a. New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol. 37:2727–2733.

Minh B.Q., Nguyen M.A.T., von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30:1188–1195.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020b. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.

Mirarab S. 2017. Phylogenomics: constrained gene tree inference. Nat. Ecol. Evol. 1:56.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

Mitterboeck T.F., Chen A.Y., Zaheer O.A., Ma E.Y.T., Adamowicz S.J. 2016. Do saline taxa evolve faster? Comparing relative rates of molecular evolution between freshwater and marine eukaryotes. Evolution 70:1960–1978.

Molloy E.K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.

Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F., Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2020. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae s.l. Syst. Biol. 70:219–235.

Nakov T., Beaulieu J.M., Alverson A.J. 2018. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). New Phytol. 219:462–473.

Nakov T., Beaulieu J.M., Alverson A.J. 2019. Diatoms diversify and turn over faster in freshwater than marine environments. Evolution 73:2497–2511.

Nakov T., Judy K.J., Downey K.M., Ruck E.C., Alverson A.J. 2020. Transcriptional response of osmolyte synthetic pathways and membrane transporters in a euryhaline diatom during long-term acclimation to a salinity gradient. J. Phycol. 56:1712–1728.

Naser-Khdour S., Minh B.Q., Zhang W., Stone E.A., Lanfear R. 2019. The prevalence and impact of model violations in phylogenetic analysis. Genome Biol. Evol 11:3341–3352.

Nawaly H., Tsuji Y., Matsuda Y. 2020. Rapid and precise genome editing in a marine diatom, *Thalassiosira pseudonana* by Cas9 nickase (D10A). Algal Res 47:101855.

Nesi N., Tsagkogeorga G., Tsang S.M., Nicolas V., Lalis A., Scanlon A.T., Riesle-Sbarbaro S.A., Wiantoro S., Hitch A.T., Juste J., Pinzari C.A., Bonaccorso F.J., Todd C.M., Lim B.K., Simmons N.B., McGowen M.R., Rossiter S.J. 2021. Interrogating phylogenetic discordance resolves deep splits in the rapid radiation of old world fruit bats (Chiroptera: Pteropodidae). Syst. Biol. 70:1077–1089.

Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2018. Quartet sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am. J. Bot. 105:385–403.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:e1002379.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Phillips M.J., Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. 28:171–185.

Pinseel E., Nakov T., Van den Berge K., Downey K.M., Judy K.J., Kourtchenko O., Kremp A., Ruck E.C., Sjöqvist C., Töpel M., Godhe A., Alverson A.J. 2022. Strain-specific transcriptional responses overshadow salinity effects in a marine diatom sampled along the Baltic Sea salinity cline. ISME J. 16:1776–1787.

Prabha R., Singh D.P., Sinha S., Ahmad K., Rai A. 2017. Genome-wide comparative analysis of codon usage bias and codon context patterns among cyanobacterial genomes. Mar. Genomics 32:31–39.

Price M.N., Dehal P.S., Arkin A.P. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Regier J.C., Shultz J.W., Zwick A., Hussey A., Ball B., Wetzer R., Martin J.W., Cunningham C.W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–1083.

Reis M., Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. 28:2161–2172.

Rhodes J.A., Baños H., Mitchell J.D., Allman E.S. 2021. MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. Bioinformatics 37:1766–1768.

Roberts W.R., Downey K.M., Ruck E.C., Traller J.C., Alverson A.J. 2020. Improved reference genome for *Cyclotella cryptica* CCMP332, a model for cell wall morphogenesis, salinity adaptation, and lipid production in diatoms (Bacillariophyta). G3 10:2965–2974.

Roch S., Warnow T. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst. Biol. 64:663–676.

Rogers R.L., Grizzard S.L., Titus-McQuillan J.E., Bockrath K., Patel S., Wares J.P., Garner J.T., Moore C.C. 2021. Gene family amplification facilitates adaptation in freshwater unionid bivalve *Megalonaias nervosa*. Mol. Ecol. 30:1155–1173.

Roller M., Lucić V., Nagy I., Perica T., Vlahovicek K. 2013. Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res. 41:8842–8852.

Rosenberg N.A. 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.

Rota-Stabelli O., Lartillot N., Philippe H., Pisani D. 2012. Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. Syst. Biol. 62:121–133.

Sanderson M.J., Wojciechowski M.F., Hu J.M., Khan T.S., Brady S.G. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. 17:782–797.

Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.

Sayyari E., Mirarab S. 2018. Testing for polytomies in phylogenetic species trees using quartet frequencies. Genes 9:132.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Simmons M.P., Gatesy J. 2021. Collapsing dubiously resolved gene-tree branches in phylogenomic coalescent analyses. Mol. Phylogenet. Evol. 158:107092.

Simmons M.P., Ochoterena H., Freudenstein J.V. 2002. Amino acid vs. nucleotide characters: challenging preconceived notions. Mol. Phylogenet. Evol. 24:78–90.

Skinner R.K., Dietrich C.H., Walden K.K.O., Gordon E., Sweet A.D., Podsiadlowski L., Petersen M., Simon C., Takiya D.M., Johnson K.P. 2020. Phylogenomics of Auchenorrhyncha (Insecta: Hemiptera) using transcriptomes: examining controversial relationships via degeneracy coding and interrogation of gene conflict. Syst. Entomol. 45:85–113.

Smayda T.J. 1969. Experimental observations on the influence of temperature, light, and salinity on cell division of the marine diatom, *Detonula confervacea* (Cleve) Gran. J. Phycol. 5:150–157.

Smith S.A., Brown J.W., Walker J.F. 2018. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. PLoS One 13:e0197433.

Steenwyk J.L., Buida T.J., Labella A.L., Li Y., Shen X.-X., Rokas A. 2021. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics 37:2325–2331.

Stern D.B., Anderson N.W., Diaz J.A., Lee C.E. 2022. Genome-wide signatures of synergistic epistasis during parallel adaptation in a Baltic Sea copepod. Nat. Commun. 13:4024.

Storz J.F. 2016. Causes of molecular convergence and parallelism in protein evolution. Nat. Rev. Genet. 17:239–250.

Suh A., Smeds L., Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13:e1002224.

Suyama M., Torrents D., Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.

Tea Y.-K., Xu X., DiBattista J.D., Lo N., Cowman P.F., Ho S.Y.W. 2021. Phylogenomic analysis of concatenated ultraconserved elements reveals the recent evolutionary radiation of the fairy wrasses (Teleostei: Labridae: *Cirrhilabrus*). Syst. Biol. 71:1–12.

Terekhanova N.V., Barmintseva A.E., Kondrashov A.S., Bazykin G.A., Mugue N.S. 2019. Architecture of parallel adaptation in ten lacustrine threespine stickleback populations from the white sea area. Genome Biol. Evol. 11:2605–2618.

Theriot E.C., Fritz S.C., Whitlock C., Conley D.J. 2006. Late quaternary rapid morphological evolution of an endemic diatom in Yellowstone Lake, Wyoming. Paleobiology 32:38–54.

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

Townsend J.P., López-Giráldez F., Friedman R. 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol. 67:437–447.

Tuchman M.L., Theriot E., Stoermer E.F. 1984. Effects of low level salinity concentrations on the growth of *Cyclotella meneghiniana* Kütz. (Bacillariophyta). Archiv für Protistenkunde 128:319–326.

Vanderpool D., Minh B.Q., Lanfear R., Hughes D., Murali S., Harris R.A., Raveendran M., Muzny D.M., Hibbins M.S., Williamson R.J., Gibbs R.A., Worley K.C., Rogers J., Hahn M.W. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. PLoS Biol. 18:e3000954.

Villar H.O., Kauvar L.M. 1994. Amino acid preferences at protein binding sites. FEBS Lett. 349:125–130.

Vizueta J., Macías-Hernández N., Arnedo M.A., Rozas J., Sánchez-Gracia A. 2019. Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation. Mol. Ecol. 28:4028–4045.

Wang L.-G., Lam T.T.-Y., Xu S., Dai Z., Zhou L., Feng T., Guo P., Dunn C.W., Jones B.R., Bradley T., Zhu H., Guan Y., Jiang Y., Yu G. 2019. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. Mol. Biol. Evol. 37:599–603.

Wang H.-C., Minh B.Q., Susko E., Roger A.J. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67:216–235.

Wu M., Kostyun J.L., Hahn M.W., Moyle L.C. 2018. Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. Mol. Ecol. 27:3301–3316.

Xi Z., Liu L., Davis C.C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92:63–71.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Y., Smith S.A. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31:3081–3092.

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

Yu G., Smith D.K., Zhu H., Guan Y., Lam T.T.-Y. 2017. Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8:28–36.

Yuan Z., Liu S., Zhou T., Tian C., Bao L., Dunham R., Liu Z. 2018. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics 19:141.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 19:153.

Zou Z., Zhang J. 2015. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol. Biol. Evol. 32:2085–2096.

Zwick A., Regier J.C., Zwickl D.J. 2012. Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. PLoS One 7:e47450.