PLOS BIOLOGY

SHORT REPORTS

Diatom abundance in the polar oceans is predicted by genome size

Wade R. Roberts*, Adam M. Siepielski, Andrew J. Alverson*

Department of Biological Sciences, University of Arkansas, Fayetteville, Arkansas, United States of America

* wader@uark.edu (WRR); aja@uark.edu (AJA)

Citation: Roberts WR, Siepielski AM, Alverson AJ (2024) Diatom abundance in the polar oceans is predicted by genome size. PLoS Biol 22(8): e3002733. https://doi.org/10.1371/journal.pbio.3002733

Academic Editor: Andrew J. Tanentzap, University of Cambridge, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: February 12, 2024

Accepted: July 3, 2024

Published: August 8, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3002733

Data Availability Statement: Sequencing reads and genome assemblies are available from NCBI BioProject PRJNA825288. Datasets and code are available from Zenodo (DOI:10.5281/zenodo.

Abstract

A principal goal in ecology is to identify the determinants of species abundances in nature. Body size has emerged as a fundamental and repeatable predictor of abundance, with smaller organisms occurring in greater numbers than larger ones. A biogeographic component, known as Bergmann's rule, describes the preponderance, across taxonomic groups, of larger-bodied organisms in colder areas. Although undeniably important, the extent to which body size is the key trait underlying these patterns is unclear. We explored these questions in diatoms, unicellular algae of global importance for their roles in carbon fixation and energy flow through marine food webs. Using a phylogenomic dataset from a single lineage with worldwide distribution, we found that body size (cell volume) was strongly correlated with genome size, which varied by 50-fold across species and was driven by differences in the amount of repetitive DNA. However, directional models identified temperature and genome size, not cell size, as having the greatest influence on maximum population growth rate. A global metabarcoding dataset further identified genome size as a strong predictor of species abundance in the ocean, but only in colder regions at high and low latitudes where diatoms with large genomes dominated, a pattern consistent with Bergmann's rule. Although species abundances are shaped by myriad interacting abiotic and biotic factors, genome size alone was a remarkably strong predictor of abundance. Taken together, these results highlight the cascading cellular and ecological consequences of macroevolutionary changes in an emergent trait, genome size, one of the most fundamental and irreducible properties of an organism.

Introduction

The abundance of species in nature is a central feature of all life. Because of this centrality, a principal goal of ecology is to understand what determines organismal abundance [1–3]. Theoretical studies have developed an extensive body of work to understand how demographic parameters (e.g., birth and death rates) affect species abundances [4–6], while observational and experimental studies have identified key abiotic (e.g., nutrient supply) and biotic factors (e.g., species interactions such as competition and predation) that shape the abundances of organisms from local to global scales [7–11]. Another equally large body of literature has sought to identify the key intrinsic features of organisms that shape their abundance [12–14].

12608914). Accession numbers for additional public datasets used in this study are available in S1 Table.

Funding: This work was supported by the National Science Foundation (DEB 1651087 to A.J.A.). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: EPA, evolutionary placement algorithm; ESS, effective sample size; NCMA, National Center for Marine Algae and Microbiota; OTU, operational taxonomic unit; PGLS, phylogenetic generalized least square; RCC, Roscoff Culture Collection.

Among these efforts, the size of an organism has emerged as a fundamental and repeatable predictor of abundance—smaller organisms occur in greater numbers than larger ones [15]. This relationship occurs across unicellular and multicellular lineages, and in terrestrial and aquatic ecosystems [15–17]. A biogeographic component, known as Bergmann's rule, describes an association between body size and temperature, wherein larger-bodied organisms are found in colder environments and smaller organisms in warmer ones [18,19]. Thus, body size and temperature are frequently woven together as key explanations for organismal abundance. The repeatability of these associations, which link a fundamental organismal trait to its abundance and thermal environment, are heralded as a widespread feature of life on Earth [20,21]. But key questions remain, such as what determines size and whether size alone is the most basic intrinsic, ecologically determinant feature of an organism. For multicellular species, size is a complex trait confounded by tissue differentiation, life history, and development [22–24]. For unicellular organisms, which constitute the bulk of life on Earth, their size may be fundamentally shaped by a single intrinsic feature, the size of their genome [23,25]. Across eukaryotes, genome size varies by many orders of magnitude and is correlated with numerous traits of ecological importance, including body size, metabolism, and life history [16,26,27]. As a result, genome size may have important cascading effects on organismal abundance and, ultimately, ecosystem function [28–30].

To test this hypothesis, we asked whether genome size can predict patterns of diatom abundance across the world's oceans. Diatoms are single-celled primary producers that account for 20% of global primary production and are keystone species in marine food webs [31]. We traced the history of genome evolution in one of the most diverse and abundant lineages of marine planktonic diatoms, Thalassiosirales [32,33], to characterize the determinants of genome size on evolutionary timescales. Although a simple association between genome size and body size (cell volume) seems intuitive, a longstanding question is whether genome size drives cell volume, or whether cell volume—an ecologically important and putatively adaptive trait—drives changes in genome size [22,24,34]. We used phylogenetic path analysis to test competing directional hypotheses about the relationship between these 2 traits, which have the potential to shape key population demographic parameters that should, in turn, shape species abundances in accord with basic population ecology theory [4–6]. We then used a large metabarcoding database to determine whether genome size predicts geographic patterns of diatom abundance and temperature associations in the global ocean. Our results identified genome size as a strong predictor of global patterns of phytoplankton species abundance. Thus, in the absence of any additional information, this single, emergent property of an organism can help us understand species abundance in the wild.

Results

Repetitive DNA underlies broad variation in genome size

We characterized the genomes of 67 newly ($n = 46$) and previously sequenced ($n = 21$) diatom strains, representing 51 species of Thalassiosirales (S1 Table). Haploid genome size varied by nearly 50-fold, from 33 Mb in *Cyclotella nana* to 1.5 Gb in *Thalassiosira tumida* (Fig 1) and showed strong phylogenetic signal (Pagel's $\lambda = 0.998$, $P < 0.001$). Estimates of haploid genome size based on *k*-mer counting and sequencing coverage were similar and strongly correlated (Spearman's $\rho = 0.984$, $P < 0.001$) (S1 Fig). Our estimates of genome size were similar for the 3 strains in our dataset with genome size estimates from flow cytometry (S1 Table). For example, our estimate for *Cyclotella nana* CCMP1335 was 33 Mb, while flow cytometry estimated it at 36 Mb (S1 Table). The *k*-mer-based method was unable to estimate genome sizes for 5 taxa, so our results are based on the coverage-based dataset unless stated otherwise. Thalassiosirales

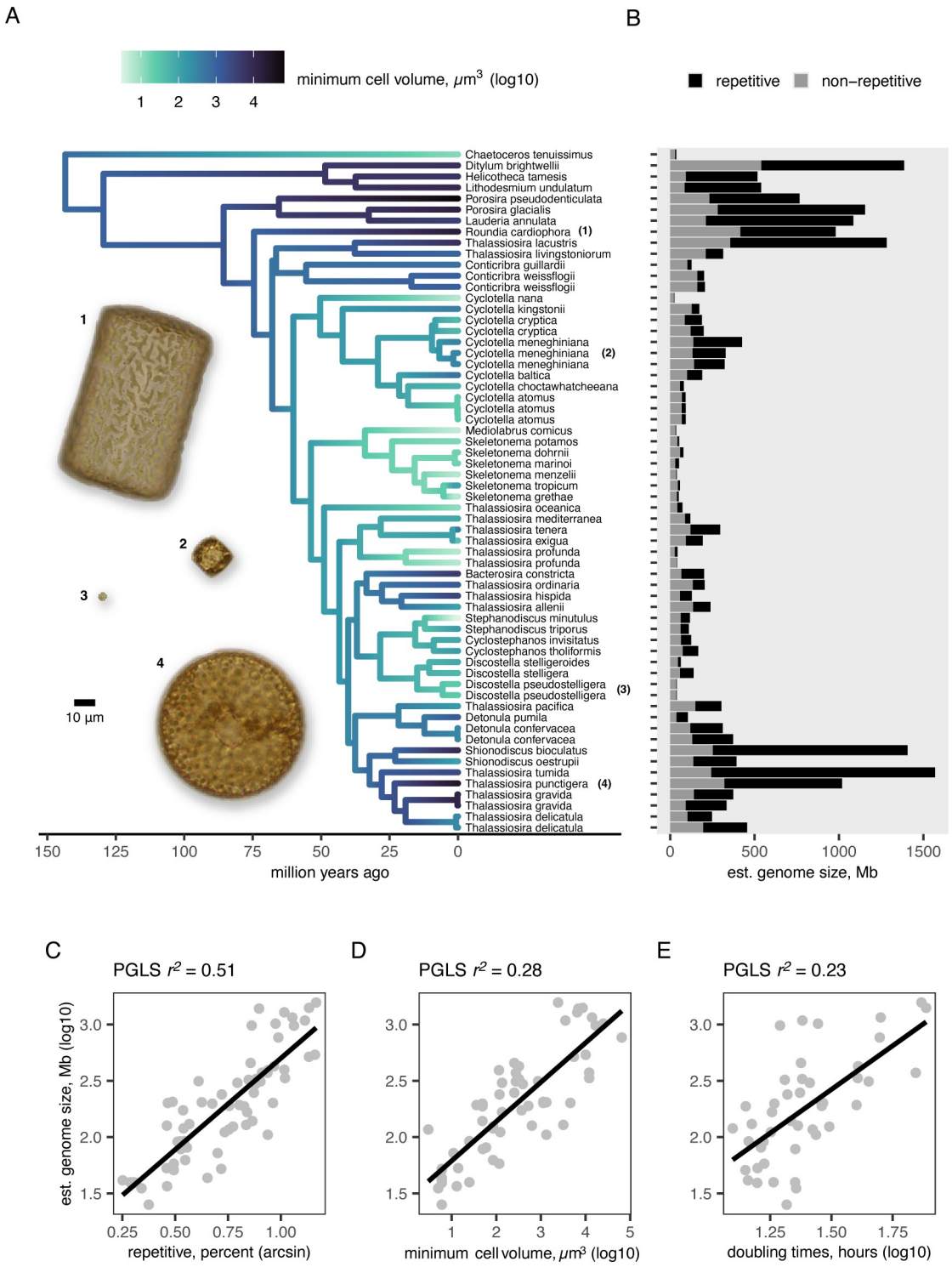

Fig 1. Cell volume and genome size vary widely across diatoms. (A) A time-calibrated phylogeny of the diatom order Thalassiosirales, modified from [35] and with branches colored by minimum cell volume. Light micrographs of live cells illustrate the broad variation in cell volume across the lineage. **(B)** Bar plots show the estimated genome size and proportions of non-repetitive and repetitive DNA in each genome. Panels C–E show PGLSs models predicting genome size with **(C)** percentage of repetitive DNA, **(D)** minimum cell volume, and **(E)** measured cell doubling time. Black lines show the estimated regression coefficients. The data and code needed to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.

https://doi.org/10.1371/journal.pbio.3002733.g001

includes marine and freshwater species [35], but there was no significant difference in genome size between diatoms from the 2 environments (Wilcoxon rank sum test, $P = 0.125$) (S2 Fig).

Genome size was strongly correlated with repetitive DNA content (phylogenetic generalized least squares [PGLS] $r^2 = 0.51$, $P < 0.001$) (Figs 1 and S3). The percentage of the genome composed of repetitive DNA ranged from 6% in *Thalassiosira profunda* (genome size: 41 Mb) to 85% in *Thalassiosira tumida* (genome size: 1.5 Gb) (S1 Table). Among the different classes of repetitive DNA, unclassified repetitive elements constituted the largest fraction of most genomes (S4 Fig). These are repetitive sequences that could not be classified into known repeat classes, likely due to the paucity of large diatom genomes that have been sequenced to date. The different classes of repetitive elements increased more-or-less proportionally in larger genomes, such that no single class of repetitive DNA disproportionately drove increases in genome size (S4 Fig). There was no association between haploid genome size and the average length of genes, exons, or introns (S3 Fig), nor the presence of polyploidy (S1 Table). Previous studies have linked GC content to genome size variation in both multicellular and unicellular organisms [36], but genome size was weakly negatively correlated with average genome-wide GC content in these diatoms (PGLS $r^2 = 0.08$, $P = 0.013$) (S3 Fig).

Genome size affects cell size and growth rate

Genome size is strongly correlated with body size, measured as cell volume, in microbial eukaryotes [23]. Although the extent to which increases in genome size require commensurate increases in nuclear and cell volumes is unclear [22], genome size should exert its greatest influence on the minimum volume of a cell. To test whether genome size predicts cell volume in diatoms, we compiled minimum and maximum volumes for the 51 species in our dataset. Maximum cell volume varied by 5 orders of magnitude across species and minimum cell volume varied by 4 (Figs 1 and S5 and S2 Table). Increased genome size was associated with increases in both minimum (PGLS $r^2 = 0.28$, $P < 0.001$) and maximum cell volume (PGLS $r^2 = 0.53$, $P < 0.001$) (Figs 1 and S5). We measured maximum growth rates for the species in our study (S1 Table) to test whether genome size is a predictor of cell division rate and found that species with larger genomes did indeed have longer doubling times (i.e., slower growth rates) (PGLS $r^2 = 0.42$, $P < 0.001$) (S6 Fig). Temperature has profound effects on cellular metabolism and growth rate in both multicellular and unicellular organisms [37], and the addition of temperature to genome size as a predictor of growth rate led to substantial improvement in model fit (PGLS $r^2 = 0.73$, $P < 0.01$) (S6 Fig). Here, lower temperatures and larger genomes were both associated with decreased growth rate (S6 Fig).

Across the tree of life, genome size and body size are strongly correlated with growth rate, nutrient usage, and other life history traits, but causal relationships and trade-offs among these and other correlated traits are not always clear [22,34,38,39]. Although causality cannot be inferred directly from comparative analyses of observational data, we can test the relative support for alternative models. To that end, we used phylogenetic path analysis—a type of structural equation modeling that allows for the evaluation of causal hypotheses from empirical data—to test competing hypotheses about the effects of 4 variables on growth rate: genome size, body size (minimum cell volume), temperature, and genomic GC content. We generated 14 alternative hypotheses (i.e., sets of directional relationships) to test whether genome size has no effect (null models), direct effects (direct models), or indirect effects (indirect models) on growth rate (cell doubling time) (S7 Fig). Using coverage-based genome size estimates, 3 models (direct1, direct2, and direct4 in S3 Table) were equally supported, with ΔCICc values <2 and P values >0.2, indicating good fit to the data (S3 Table). In the best-fit model, direct4, genome size directly affects cell volume and doubling time, and temperature directly affects

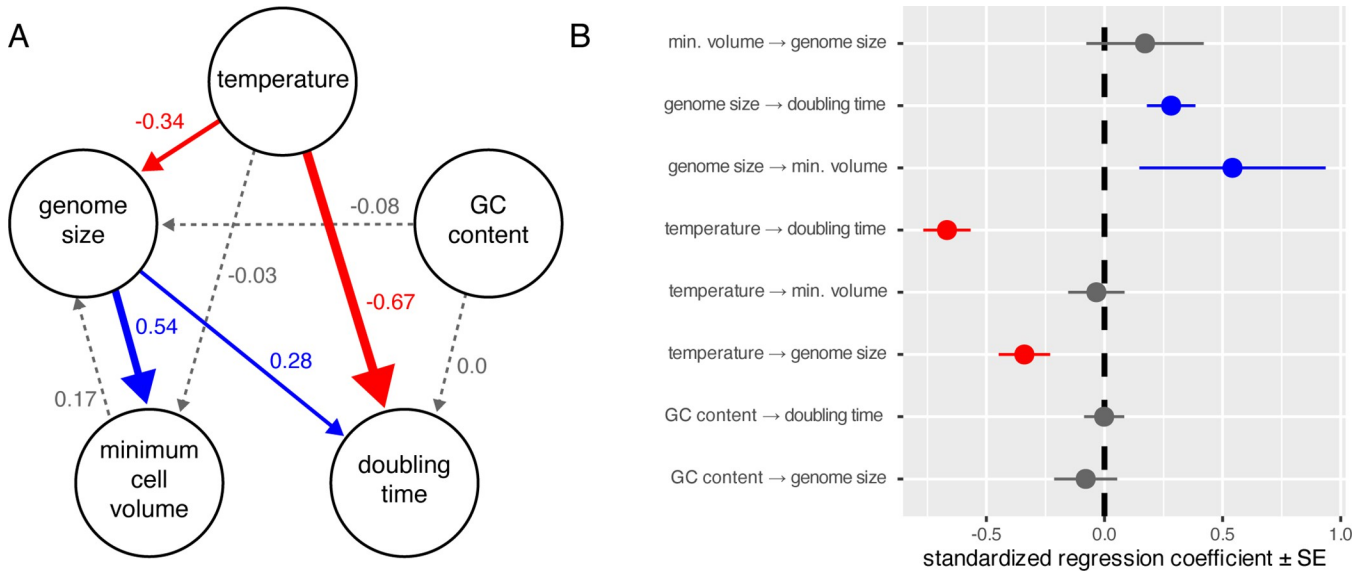

Fig 2. Genome size affects cell volume and doubling time in diatoms. (A) Average model from phylogenetic path analysis using coverage-based genome size estimates. Arrow color and width represent the direction and magnitude of regression coefficients, indicated by numeric labels (positive: blue; negative: red; nonsignificant: gray). Full lines show coefficients that differ significantly from 0, whereas dotted lines overlap with 0. **(B)** Standardized regression coefficients and their standard errors (SE) for paths in the model. The data and code needed to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.

https://doi.org/10.1371/journal.pbio.3002733.g002

genome size and doubling time (S8 Fig). The direct1 and direct2 models remove the effect of GC content on genome size (S7 Fig). The main difference between the top 2 models (direct4 and direct2) is whether genome size impacts cell volume (direct4) or vice versa (direct2). Averaging the top 3 models resulted in a larger path coefficient for genome size affecting cell volume (0.54 versus 0.17) (Fig 2). Finally, testing the same 14 models with the *k*-mer instead of coverage-based genome size estimates gave 6 models (including direct4) with equally strong support (S3 Table). Importantly, all 6 models support genome size directly impacting cell volume, adding further support for the hypothesis that genome size influences cell volume, not the reverse (S7 and S8 Figs). The best-fit (indirect2) and average models using *k*-mer-based genome sizes both suggested that genome size affects doubling times but only indirectly via effects on cell volume, rather than the direct effect of genome size on doubling time supported by coverage-based genome size estimates (S8 Fig).

Genome size, biogeography, and temperature impact diatom abundance in the ocean

Taking advantage of the global metabarcoding database from the *Tara* Oceans expedition [40], we built Bayesian models to test whether genome size influences relative species abundance in the ocean (Fig 3). Using 2 taxonomic assignment methods for operational taxonomic unit (OTU) sequences, we identified 28 species from our study that were also present in $\geq$10 samples of the *Tara* Oceans database. This allowed us to test whether latitude, ocean region, and/or ocean temperature interact with genome size to affect species abundance. Latitude had a significant nonlinear interaction with genome size on species abundance (Fig 3C)—species with larger genomes were more abundant at high latitudes, and species with smaller genomes were more abundant at lower latitudes (Fig 3C). The 2 coldest ocean regions, the Arctic and Southern Oceans, were the only ones with a significant positive regression coefficient relating genome size to abundance, whereas all other regions had either no effect or a negative effect

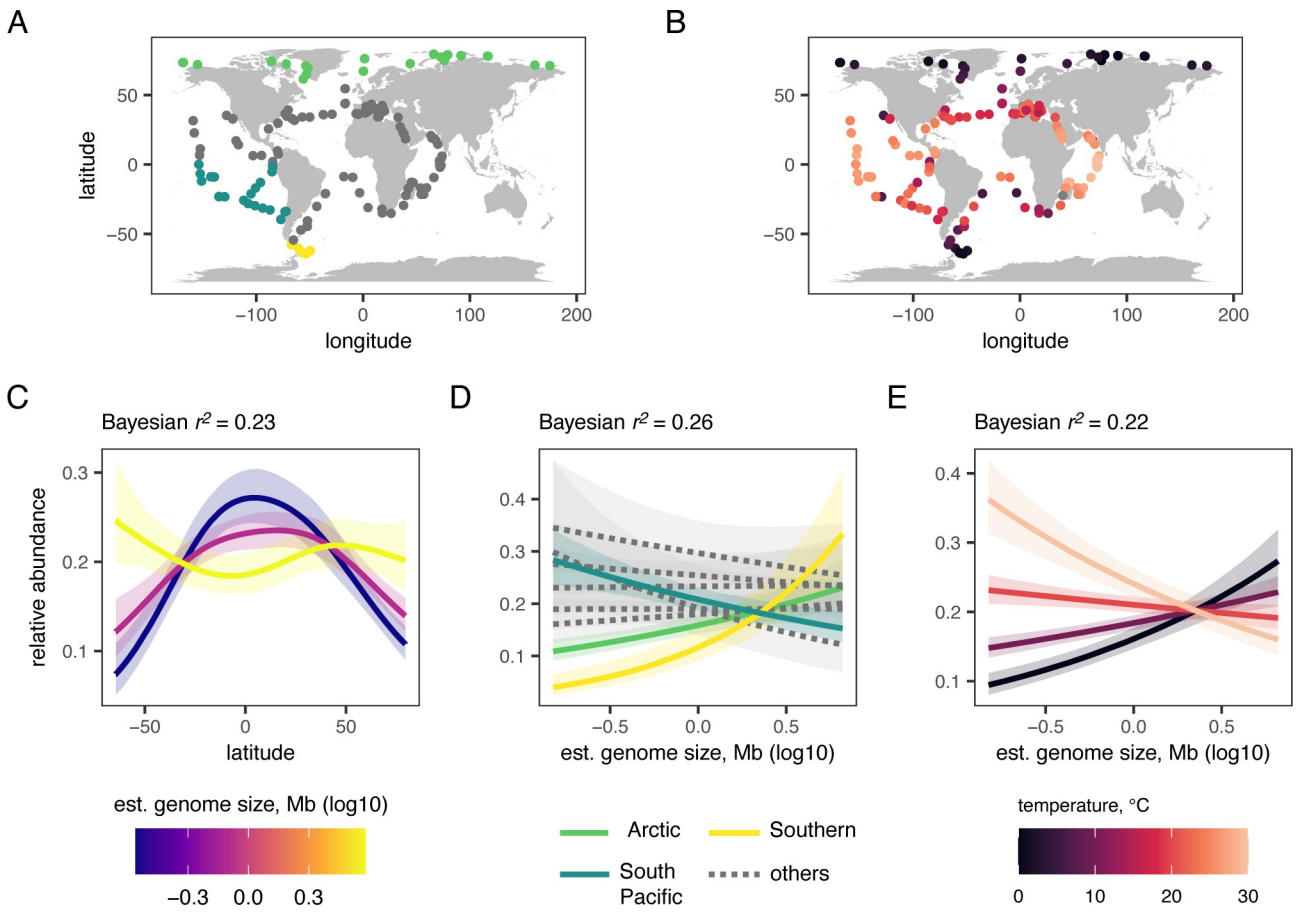

Fig 3. Latitude, ocean region, and temperature interact to shape genome size—abundance relationships. **(A, B)** Maps showing the locations of the 210 sampling stations from the *Tara* Oceans expedition, with points colored to highlight locations within the **(A)** Arctic, Southern, and South Pacific Oceans or **(B)** the temperature of each location at the time of collection. The base layer for the maps is from https://cran.r-project.org/web/packages/maps/index.html. Panels C–E show Bayesian multilevel regression models predicting relative species abundance by the interaction of genome size with **(C)** latitude, **(D)** ocean region, or **(E)** temperature. Nonlinear effects of latitude were modeled in **(C)** using a generalized additive model. Significant estimates for the Arctic, Southern, and South Pacific Oceans are shown with solid lines in **(D)**. Nonsignificant estimates for the other ocean regions are shown with dotted lines in **(D)**. Although all predictors are treated as continuous in **(E)**, we used the model to predict the interactive effect of 4 temperatures (0, 10, 20, and 30˚C) with genome size on species abundance. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.

https://doi.org/10.1371/journal.pbio.3002733.g003

(South Pacific Ocean) of genome size on species abundance (Fig 3D). We tested the effects of temperature directly and found a significant interaction with genome size to predict abundance, in which species with larger genomes were more abundant in colder temperatures (Fig 3E). These results were replicated with a smaller dataset (22 species) of OTUs classified using an alternate method of taxonomic assignment (S9 Fig).

Discussion

One of the most basic and defining ecological properties of a species is its abundance in the environment, which is shaped by numerous interacting abiotic and biotic factors [2,6,9]. As a result, considerable attention has been paid to identifying key ecological processes that define a simple, sufficient, and generalizable ecological model explaining species abundance [11]. Rather than focusing on ecological processes, we sought instead to determine whether a fundamental intrinsic property of an organism—the size of its genome—can explain abundance and an associated vital rate, population growth [15,41].

Trait-based models of phytoplankton ecology use the functional traits of individual species or entire communities to understand the biogeography, seasonal dynamics, and future responses of phytoplankton to environmental change [42,43]. Across taxonomic groups, major ocean regions, and marine and freshwaters, nearly all traits of ecological importance scale allometrically with cell size [38,44,45]. Despite its broad predictive power, however, theoretical and empirical studies have revealed complex interactions between cell size and environmental gradients such as temperature and nutrient supply, two of the principal abiotic factors structuring phytoplankton communities [16,46]. In general, large cells tend to dominate in the cold, nutrient-rich waters of high latitudes, and smaller cells are more abundant in lower latitudes, where temperatures are warmer and nutrient supplies are lower [16,47]. Other factors, such as grazing pressure, can interact with temperature and nutrients to modify size–abundance relationships [46]. Amidst a sea of trait correlations, the extent and complexity of these interactions make it difficult to infer causal relationships and develop a simple ecological model of abundance [39,41,42].

Across the tree of life, cell size is also correlated with the size of both the genome and the cell nucleus [24,34,48]. This relationship is commonly assumed to reflect simple packaging constraints, suggesting that over evolutionary timescales nucleus and cell sizes ebb and flow nonadaptively in response to changes in genome size [22,24,34]. Although intuitive, the mechanisms by which these 3 size components of the cell exert their influence on one another is unclear [22]. Alternatively, the strong associations between cell size and fitness-related traits, such as nutrient acquisition and growth rate, suggest cell size is an adaptive trait [23,24,34]. If larger cells require larger nuclei to balance space requirements for RNA synthesis in the nucleus and protein synthesis in the cytoplasm, then changes in the amount of bulk DNA are a means of modulating the size of the nucleus to maintain an optimal nuclear:cytoplasmic ratio (the "karyotopic ratio") [23]. Our novel approach to these questions—combining phylogenomics, empirical growth rates, and a global DNA metabarcoding database—highlighted a central role for genome size in the cellular and ecological properties of marine diatoms.

Although previous flow cytometry studies found correlations between genome size and cell size in diatoms [49,50], the genome sequences analyzed here identified repetitive DNA as the principal driver of genome size evolution. Nucleotypic effects describe the phenotypic changes that occur in response to changes in genome size [51]. In the diatoms studied here, repeat-driven changes in genome size over the past 100 million years had strong nucleotypic effects on 2 fitness-related traits—cell size and maximum growth rate (Figs 1 and 2). The same nucleotypic effects operate on microevolutionary timescales in diatoms as well. A comparison of 2 populations of the marine planktonic diatom, *Ditylum brightwellii*, with 2-fold difference in genome size and 4-fold difference in maximum cell volume, showed that the population with a smaller genome and cell size had a higher growth rate, and that genome size had a significantly greater (negative) impact on growth than cell size [52]. Larger genomes take longer to replicate, lengthening mitosis and cell doubling time [22,51]. Larger genomes also require additional investments of N and P to replicate and maintain, so in species with large genomes, these 2 essential nutrients cannot be allocated to RNA, ribosomes, and proteins, reducing growth rates and, over longer timescales, selecting for smaller genomes under conditions of nutrient limitation [41].

Although cell size is often considered a "master" phytoplankton trait, our results highlight the ecological importance of genome size as well. Genome size was not driven by increases in the amount of functional DNA, either through gene or genome duplications, but instead through changes in the amount of nonfunctional sequences, highlighting bulk DNA content as a phenotype with far-reaching consequences for phytoplankton physiology and ecology. Although genome size could be interpreted as an adaptive trait in this context [24], this must

be weighed against the deleterious effects of excess DNA, including mobile elements that can disrupt functional genes [53] and the metabolic burden of noncoding DNA [41]. Although evidence for this hypothesis is mixed, [54–56], the inclusion of population genetic parameters in our models might have shown whether diatoms with smaller effective population sizes are potentially more susceptible to nonadaptive genome expansions due to genetic drift [53]. Whatever the cause, our results provide support for a simple model in which many ecologically important traits, though perhaps more proximally related to cell size, are perhaps ultimately attributable to the size of the genome.

Ecologists have identified numerous biotic and abiotic factors that explain organismal abundance and geographic distributions [7–11]. Indeed, diatom abundance is shaped by abiotic factors such as temperature, along with both bottom-up effects such as nutrient supply, and top-down effects such as grazing pressure [57–59]. The data presented here showed that diatoms with larger genomes and, by extension, larger cell volumes are more abundant in regions and latitudes that experience colder temperatures, supporting Bergmann's rule and reinforcing a broader biogeographic trend of larger phytoplankton in colder seas [42]. This pattern has been attributed to temperature and numerous covarying factors [18]. For example, larger genomes and cells require more nutrients, which are generally in greater supply at higher latitudes [16,47]. In addition, grazing marine copepods have larger body sizes in colder temperatures [60], which might select for increased genome and cell size in colder parts of the ocean. Although including these and other factors in our models undoubtedly would have explained more of the variation in abundances, genome size predicted abundance remarkably well. Like most studies, the strong effect of latitude on the association between genome size and abundance reflects the context-specific nature of this association, which is typical of many ecological patterns [61]. Finally, a field study of freshwater benthic diatoms from geothermally heated streams found no evidence for Bergmann's rule [62], suggesting possible differences in diatom size–abundance relationships across ecosystem types or phylogenetic lineages.

Documenting abundance associations at a global scale is not without challenges. For example, in many cases the *Tara* Oceans samples represented a single snapshot in time of abundance at a location, precluding estimates of sampling error and potentially missing rapid seasonal changes in species abundance. Although these types of temporal limitations are common in spatial datasets that are global in scale, they have nevertheless proven to be extremely powerful in revealing broad ecological trends [40,63]. The diatom lineage studied here, Thalassiosirales, was well represented throughout the *Tara* Oceans samples and allowed us to uncover strong evidence linking genome size, temperature, maximum population growth rate, and species abundance [32,33]. Our results are consistent with size–abundance relationships found in large-scale phytoplankton studies [16,46], which might also be driven ultimately by genome size.

Similar associations have been found in multicellular organisms as well, suggesting genome size may shape patterns of species abundance broadly across the tree of life. Genome size has been linked to the distribution of flowering plants along a temperature gradient in the British Isles [64] and was positively correlated with regional abundance in 436 herbaceous plant species across Europe [28]. Although not related to a temperature gradient, salamanders are among the most abundant animal groups in many terrestrial ecosystems [65], and they have among the largest known genomes in the vertebrate lineage [66]. Like diatoms, much of the variation in genome size in these and other groups is attributable to noncoding sequences. Notably, despite the inherent difficulty in estimating abundance at a global scale, the amount of variation in abundance explained by genome size in our study was substantially greater than the typical range of variation accounted for in many ecological studies [67], highlighting the seemingly outsized role of genome size in the ecology of unicellular organisms.

In addition to the ecological consequences, our results highlight the unique cellular trade-offs imposed by changes in genome size in diatoms. Diatoms reproduce asexually throughout most of their life history and are unusual in that one of the 2 daughter cells following a mitotic event is smaller than the parent, leading to a reduction in the average diameter of a cell lineage over time, eventually triggering sexual reproduction and restoring the maximum cell size [68]. Although not measured here, nucleus size is positively correlated with genome size across the tree of life [24,69,70], and the same correlation likely exists for diatoms. With a fixed genome size that constrains the size of the nucleus, diatoms must optimize their surface area:volume ratio as cell size decreases across generations. Diatoms have vacuoles that function in buoyancy control, nutrient storage, and optimization of the surface area:volume ratio. Vacuoles occupy as much as 90% of the cell volume [57], and vacuole size can be modulated in response to environmental conditions [71]. The strong correlation between vacuole size and cell volume has led to the hypothesis that the vacuole has played a key adaptive role in diatom evolution by facilitating increases in cell size as a way to escape grazing pressure [57,58]. This hypothesis does not account for the parallel influence of genome size on cell size confirmed here.

Just as the discovery here of a directional effect of genome size on cell size highlights its lack of consideration from previous models, it likewise highlights the absence of several traits from our study. Although suggested by our models, increases in genome size may not be the proximal cause of increased cell volume. Genome size might affect cell volume indirectly, via upward pressure on nuclear volume or another latent character. In addition, although the functions of the vacuole as they relate to cell size are clear [58], vacuole size is a more labile trait, and it is unclear whether the vacuole exerts a causal influence on cell size or vice versa. The genome, nucleus, and vacuole have different functional roles in relation to cell volume, and all incur costs to maintain [57,72], so with a fixed genome size that presumably constrains the minimum size of both the nucleus and the cell, diatoms probably rely primarily on adjustments to the size and contents the vacuole as the ecological setting (e.g., rates of nutrient uptake, sinking rate, susceptibility to grazers) changes in response to decreases in the volume of a cell lineage over time.

Overall, the results presented here advance our understanding of species abundance by showing that a single emergent trait fundamental to all life, the size of the genome, can predict population abundance at a global scale. Moreover, the geographic variation in this pattern is entirely consistent with longstanding ideas regarding size–abundance associations in relation to the thermal environment. The addition of ecological information and other trait data to genome size estimates would likely generate a more informative model of species abundance, and this remains an important next step. Integrative approaches such as the one developed here, combining the seemingly disparate subdisciplines of phylogenomics and population ecology, may prove useful in forecasting widespread changes in the abundance of diatoms in response to ongoing climate change, especially in polar regions.

Materials and methods

Strain collection and culturing

Strains were collected from a variety of locales in the United States and isolated into monoclonal cultures. Additional strains were acquired from the National Center for Marine Algae and Microbiota (NCMA) in the United States or the Roscoff Culture Collection (RCC) in France. Marine strains were grown in L1 medium [73] and freshwater strains were grown in WC medium [74]. Cells were grown in batch culture at varying temperatures from 5 to 21˚ C on a 12–12 hour light–dark cycle. Cultures newly isolated in the Alverson Lab have been submitted

to the public culture collections at NCMA and The University of Texas Culture Collection of Algae (UTEX) (S1 Table).

Draft genome sequencing, assembly, and gene prediction

See Supplementary File S1 of [35] for a detailed description of DNA extraction, sequencing, draft genome assembly, and gene model prediction. We additionally downloaded short read files for additional Thalassiosirales and Lithodesmiales (outgroups) strains that were available from the NCBI Sequence Read Archive (S1 Table) [75–78]. For these reads, we used a similar workflow as outlined in [35]. Briefly, we trimmed the reads using *Trimmomatic* v.0.36 [79], corrected the trimmed reads with *BayesHammer* [80], assembled the corrected reads with *SPAdes* v.3.12.0 [81], and removed contaminant contigs using *Blobtools* v.1.1.1 [82]. We removed contigs that had taxonomic assignment to bacteria, archaea, or viruses, or were shorter than 1 kb.

We also downloaded the reference genomes for *Chaetoceros tenuissimus* v.1 [83], *Cyclotella cryptica* v.2 [84], *Cyclotella nana* v.4 [85] (DOI: 10.5683/SP2/ZDZQFE), *Skeletonema marinoi* RCC75 v.1 [86], *Skeletonema marinoi* RO5AC v.1.1.2 [87] (DOI: 10.5281/zenodo.7786015), and *Thalassiosira oceanica* v.2 [88] (DOI: 10.5281/zenodo.4589594). When available, we also downloaded the predicted gene models and associated short reads from NCBI that were used in the genome assembly (S1 Table). For every genome with predicted gene models, we also calculated the average lengths of the genes, exons, and introns (S1 Table).

Phylogenetic tree estimation

We used *BUSCO* v.5.1.3 [89,90] to estimate the completeness of each draft assembly using the stramenopiles_odb10 orthologs (n = 100) (S1 Table). We used the detected single-copy orthologs to estimate a phylogenetic tree for all strains included in this study. First, we aligned the amino acid sequences of each ortholog using *MAFFT* v.7.505 [91] and the L-INS-i algorithm ("—localpair—maxiterate 1000"). Next, we trimmed the resulting alignments using *ClipKIT* v.1.3.0 [92] with the default smart-gap mode ("-m smart-gap"). We then concatenated all trimmed alignments into a supermatrix using the *pxcat* command in *Phyx* [93]. Finally, we estimated a phylogenetic tree from the supermatrix using *IQ-Tree* v.2.2.0.3 [94], partitioning the alignment by gene, using the LG+G substitution model, constraining the backbone topology to match the reference tree from [35], and calculating branch support with 10,000 ultrafast bootstrap replicates [95].

To generate an ultrametric, time-calibrated phylogeny, we estimated divergence times using *MCMCtree* in *PAML* v.4.9e [96]. We used nucleotide data from 4 loci: the nuclear *18S* and *28S*, and the plastid *rbcL* and *psbC*. We used the approximate likelihood approach [97], the GTR+G5 substitution model, and the independent rates clock model. Priors in the analysis were kept at their defaults. To ensure convergence in age estimates, we ran 2 independent MCMC chains of 2.5e6 generations, sampling every 200, with the first 5e5 discarded as burn-in. We checked that the effective sample size (ESS) of each parameter estimate was above 200 using *Tracer* v.1.7.2 [98].

We applied 8 calibrations to the reference tree in *MCMCtree* based on previous age estimates or fossil evidence. First, we placed upper and lower bounded constraints on the root (lower: 136 Ma; upper: 156 Ma) and the crown of Thalassiosirales–Lithodesmiales (lower: 115 Ma; upper: 141 Ma) based on estimates in [99]. We then placed a skewNormal prior distribution on the crown age of Thalassiosirales with a minimum age of 75 Ma based on the *Thalassiosiropsis* fossil [100,101]. Based on first fossil appearances in the Neptune marine micropaleontology database [102], we also placed lower bounds for the stem ages of

Lithodesmium undulatum (29.96 Ma), *Porosira glacialis* (9 Ma), and *Bacterosira constricta* (8.35 Ma). For *Cyclostephanos*, we placed a lower bound of 5 Ma on the crown age based on the fossil species *Cyclostephanos undatus* [101,103]. Lastly, we placed a lower bound of 6.15 Ma on the crown age of *Shionodiscus* based on the fossil species *Shionodiscus praeoestrupii* [101,104].

Genome size estimation

We estimated haploid genome sizes using contaminant-free paired-end Illumina reads that aligned to the filtered draft genome assemblies. Briefly, we aligned the reads to the cleaned assembly using *minimap2* v.2.10 [105] using the presets for short reads and outputting the alignments in BAM format ("-ax sr"). We then extracted and kept only the aligned and correctly paired reads using the tool *bam2fastq* (https://github.com/jts/bam2fastq) with options "—aligned—no-unaligned—no-filtered." We estimated the genome size of each strain using 2 bioinformatic approaches based on *k*-mer [106] and read coverage histograms [107] (S4 Table).

For the *k*-mer-based genome size estimates, we used the script *kmercountexact.sh* from BBtools (https://sourceforge.net/projects/bbmap/). This script counts the number of unique *k*-mers in each pair of read files, estimates the genome size, and outputs a *k*-mer frequency histogram. We specified a range of *k*-mer lengths (17, 19, 21, 23, 25, 27, 29, and 31) because genome size estimates are highly dependent upon the chosen *k*-mer. For example, longer *k*-mers will collapse fewer short repeats and genome size estimates will be larger. We averaged the BBtools estimates for each *k*-mer length to produce a final genome size estimate for each strain. Because *kmercountexact.sh* can sometimes incorrectly identify the locations of the heterozygous and homozygous peaks in the histogram, we verified the accuracy manually in R [108]. We plotted the *k*-mer histograms in R to identify the peaks and calculate the estimated haploid genome size (*GS*) using the formula: $GS = N / (C * p)$, where N is the total number of genomic *k*-mers, C is the *k*-mer coverage, and p is the ploidy [106].

For read coverage-based estimates, we aligned the filtered read files to the draft assembly using *minimap2* and exported the alignments in BAM format. We then used the script *pileup.sh* from BBtools to calculate the total number of mapped base pairs and export a table of per-contig average coverage estimates. Next, we plotted the distribution of per-contig coverages in R and identified the mode of the distribution using the *asselin* function from the R package *modeest* [109]. To estimate the genome size, we used the Lander–Waterman formula [110]: $GS = LN / C$, where L is the read length, N is the total number of mapped reads, and C is the mode of the coverage distribution.

For *Lauderia annulata*, *Roundia cardiophora*, and *Thalassiosira punctigera*, the draft genome assemblies were too fragmented and incomplete to use for genome size estimates using the above approaches. Instead, we used the transcriptomes from these strains to estimate genome size in an approach adapted from [111]. Our approach to transcriptome assembly and contig filtering is detailed in Supplementary File S1 of [35]. In this approach, the filtered read files were aligned against the transcriptome assembly using *minimap2*, the resulting BAM file was parsed using *pileup.sh*, and the mode of the coverage distribution was estimated in R. The same formula for the read coverage-based approach is then used to estimate genome size.

Ploidy estimation

We used *Smudgeplot* v.0.2.5 [112] to estimate the ploidy of each strain (S1 Table). This method uses short reads to disentangle genome structure and estimate ploidy. For this method, we used the sets of aligned and contaminant-filtered reads that were used in genome size

estimation. We used *jellyfish* v.2.3.0 [113] to count *k*-mers of length 21 from both forward and reverse reads. *Smudgeplot* then extracted the heterozygous *k*-mer pairs and calculated their coverages in order to estimate the ploidy of the sample.

Repeat content estimation

We estimated the percentage of each genome that is composed of repetitive elements using the pipeline *DNApipeTE* v.1.3 [114] (S1 Table). This method allows for the fast assembly, quantification, and annotation of repeat sequences from a low-coverage sampling of reads. For each strain, we provided *DNApipeTE* with a single file of combined forward and reverse reads, the estimated genome size (in bp), and a custom repeat library for repeat annotation. The read files consist of the aligned and contaminant-filtered reads extracted previously for genome size estimation. We generated the custom repeat libraries for the genome assemblies using *RepeatModeler2* v.2.0.1 [115]. *DNApipeTE* performs sampling of the reads to produce low-coverage datasets to use during analyses. After initial testing of the pipeline using different coverages (0.01×, 0.05×, 0.1×, 0.25×), we determined that a coverage of 0.25× would be used. For some lower quality genomes with fewer available reads, a lower coverage (0.1×) was used. We estimated the repeat content for each strain using both the *k*-mer- and read coverage-based genome size estimations.

Growth measurements

We calculated the maximum growth rates of strains using the relative chlorophyll *a* fluorescence (RF) measured on a Trilogy Fluorometer (Turner Designs, California, United States of America). In triplicate, we grew each strain at their maintenance temperature (5˚, 15˚, or 21˚ C) and measured their RF daily or every other day. We input the RF values into the R package *growthrates* [116] to calculate the maximum growth rate (μ) during exponential growth using the linear method [117]. We calculated the doubling time (in hours) of each strain by dividing μ by the natural logarithm of 2. We averaged the triplicate doubling times to get a strain average (S5 Table).

Cell volume measurements

We collected information about minimum and maximum observed cell diameters and heights from biovolume databases and primary literature (S2 Table). For many marine species, we compiled cell volume data from the Helsinki Commission Phytoplankton Expert Group (HELCOM PEG) dataset (https://helcom.fi/helcom-at-work/projects/peg/) or [118]. For the remaining species, cell volume was calculated using size ranges reported in the primary literature. The height of cells is rarely reported, making it difficult to calculate cell volume using real measurements. We therefore used an approach to calculate cell heights using a ratio, *height = diameter * 0.5* [118]. For *Skeletonema*, we made a simplifying assumption that cell height equals cell diameter [119,120]. We then calculated the minimum and maximum cell volumes using appropriate equations for the approximate geometric shape of each species (S2 Table).

Taxonomic assignment of OTUs

We downloaded the assembled and clustered V9-*18S* rDNA OTU sequences and read count abundances from the *Tara* Oceans expedition from Zenodo (DOI: 10.5281/zenodo.3768510) [121,122]. We initially filtered the OTUs to only those that had previous taxonomic assignments to the Thalassiosirales and Lithodesmiales. We aligned these filtered OTU sequences to

a set of reference *18S* rDNA sequences using *ssu-align* (http://eddylab.org/software/ssu-align/). We estimated the *18S* reference tree using *IQ-Tree* with the TIM2+F+I+G4 model. We then used the evolutionary placement algorithm (EPA) [123] in *RAxML* v.8.2.11 [124] to place the OTU sequences onto the reference phylogenetic tree of our *18S* sequences.

After OTU placement, we used *Gappa* v.0.8.0 [125] to parse the resulting EPA jplace file and assign a taxonomy to each OTU using the computed likelihood weights. We used 2 approaches to assign taxonomy. First, we used "*gappa examine assign*" to compute a majority taxonomy for each internal node based on a consensus of its descendants. For example, if the consensus threshold is set to 0.5 and 4 descendants are labeled "A;B;C" and 3 are labeled "A;B;D," the inner label will get labeled "A;B;C." We chose a consensus threshold of 0.7 for this first approach. This first approach resulted in OTUs for 22 species that overlapped with our genome size dataset. Second, we used "*gappa edit accumulate*" which adds together the likelihood weight ratio of each placement downward toward the root until the accumulated mass at the basal branch reaches the defined threshold. This approach is useful to assess placements distributed across nearby branches of the reference tree when the reference contains multiple representatives of the same species. We chose a threshold of 0.5 for this second approach. The second approach resulted in OTUs for 28 species that overlapped with our genome size dataset.

OTU abundances and associated metadata

Following the taxonomic assignments of the OTUs, we filtered the original OTU abundance count table down to only those assigned to the species in our dataset. To account for differences in read depth between the *Tara* Oceans samples, we transformed the OTU counts to proportions prior to statistical analyses [126]. We used the R package *dplyr* [127] to aggregate the abundances for each species and calculate the sum of each species at each *Tara* sampling location. This generated a table with a single relative abundance measurement per species at each sampling locale. We then downloaded 2 tables containing the associated environmental metadata of each sample taken in the *Tara* Oceans expeditions from PANGAEA (DOI: 10.1594/PANGAEA.858201, DOI: 10.1594/PANGAEA.875576) [128]. We combined the 2 metadata tables and merged the combined data with the abundances. This produced a final table containing abundances of each species and the metadata for each sampling locale.

Statistics

The following variables were $\log_{10}$-transformed before statistical analyses: genome size, average exon length, average intron length, doubling times, and cell volumes. We applied arcsin transformation to the percent repeat content. No transformation was applied to average gene length, GC percent, temperature, and latitude.

We calculated the phylogenetic signal of genome size with Pagel's lambda [129] with the *phylosig* function in the R package *phytools* [130]. Ancestral state reconstruction of minimum cell volume was estimated using maximum likelihood under an Ornstein–Uhlenbeck model with the *anc.ML* function in the R package *phytools*. We calculated Spearman's rho between the *k*-mer- and read coverage-based genome size estimates using the R stats function *cor.test*. To test for the correlation between variables, we performed linear regression using the R stats function *lm* and PGLS using the functions *comparative.data* and *pgls* from the R package *caper* [131]. The phylogenetic tree used in PGLS was the ultrametric tree estimated above.

We performed phylogenetic path analysis [132] using the function *phylo_path* in the R package *phylopath* [133]. Fourteen models were designed and tested to determine if genome size had no effect (null models), direct effects (direct models), or indirect effects (indirect

models) on doubling time (S7 Fig). We used the default evolutionary model of "lambda" and calculated averaged models using the "full" approach [132].

We tested 3 hypotheses that species abundances are predicted by the interaction of genome size and ocean region, temperature, or latitude. Prior to model fitting, we centered all continuous variables in these models, except latitude and relative abundance. We fit these models using Bayesian phylogenetic multilevel models, estimating the posterior distributions of each model using *rstan* [134] via the *brms* package [135]. Species abundances (the response variable) were modeled using a gamma distribution with a log link function. Prior sensitivity was assessed by running models with uninformative or weakly informative priors. To account for phylogenetic non-independence between species in our models, we calculated a covariance matrix from the ultrametric phylogenetic tree using the function *vcv.phylo* in the R package *ape* [136]. We calculated the mean and 95% credible intervals for each model parameter and each derived quantity, from the joint posterior distribution of the models. Model effects were considered significant if the means and 95% credible intervals did not overlap zero [137]. Predictions for the interactive effect of latitude, temperatures, or ocean region with genome size on abundances were made from the models using the *predictions* function in the R package *marginaleffects* [138]. For latitude effects, we fit generalized additive models to account for nonlinear effects of latitude on species abundance. Each model in *brms* was run using 2 Markov chains for 10,000 iterations. We assessed chain convergence using potential scale reduction factors [139] and model fit using Bayesian r^2 [140]. In addition, we assessed model fits using posterior predictive checking.

We plotted figures, phylogenetic trees, and maps using the R packages *ggplot2* [141], *aplot* [142], *ggtree* [143], *maps* [144], and *ggmap* [145]. Figure edits were made using Adobe Illustrator.

Supporting information

S1 Fig. Strong correlation between haploid genome size estimates. Scatterplot showing the correlation between estimated haploid genome sizes via *k*-mer- and coverage-based approaches. The black line indicates the regression coefficient. Spearman's rho and associated *P* value are shown above the plot. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S2 Fig. No difference in genome size between marine and freshwater diatoms. Violin plots showing the distribution of coverage-based haploid genome size estimates for marine (black) and freshwater (blue) species. The test statistic and *P* value from a Wilcoxon rank sum test are shown above the plot. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S3 Fig. Genome size is predicted by repetitive elements, gene lengths, and GC content. Phylogenetic generalized least squares (PGLS) models predicting genome size by **(A)** the estimated percentage of repetitive elements in the genome, **(B)** the average gene length, **(C)** the average exon length, **(D)** the average intron length, and **(E)** the genome average GC content. Black lines show the estimated regression coefficient. The PGLS r^2 and associated *P* value are shown above each plot. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S4 Fig. Percentages of different repetitive element classes in each genome. (A) A time-calibrated phylogeny of the diatom order Thalassiosirales, modified from [35]. Strain numbers follow the species names. **(B)** Stacked bar plot showing the percentage of each genome belonging to different repetitive element classes as estimated using *dnaPipeTE*. Colors denote the different repeat classes and gray represents the percentage of the genome that is non-repetitive. Abbreviations: DNA, DNA transposons; LINE, long interspersed nuclear elements; LTR, long terminal retroelements; RC, rolling circles or *Helitrons*. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S5 Fig. Genome size is predicted by minimum and maximum cell volumes. Phylogenetic generalized least squares (PGLS) models predicting genome size by the **(A)** minimum and **(B)** maximum calculated cell volume. Black lines show the estimated regression coefficient. The PGLS r^2 and associated *P* value are shown above each plot. The range of cell volume sizes (minimum to maximum) are shown for each species on **(C)** linear and **(D)** $\log_{10}$-transformed scales. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S6 Fig. Doubling time is predicted by genome size and temperature. Phylogenetic generalized least squares (PGLS) model predicting doubling time by the additive effects of genome size and temperature. The black line shows the estimated regression coefficient. The points are colored according to the growth temperature of the strain. The PGLS r^2 and associated *P* value are shown above each plot. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S7 Fig. Models tested in the phylogenetic path analyses. Directed acyclic graphs showing the 14 models tested in the phylogenetic path analysis. Models were defined to test if genome size had no effect (null models), direct effects (direct models), or indirect effects (indirect models) on doubling time. Temperature, GC percentage, and minimum cell volume are also included as variables in the models. Arrow direction indicates the causal relationship being tested between 2 variables. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S8 Fig. The best and average models from the phylogenetic path analyses. Results of the phylogenetic path analysis. **(A, B)** Best models and **(C, D)** averaged models using coverage- **(A, C)** or *k*-mer-based **(B, D)** genome size estimates. Arrow color and width represent the direction and magnitude of regression coefficients, indicated by numeric labels (positive: blue; negative: red; nonsignificant: gray). Full lines show coefficients that differ significantly from 0, whereas negative lines overlap with 0 and are nonsignificant. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S9 Fig. Latitude, ocean region, and temperature interact with genome size—abundance relationships. Results from alternative taxonomic assignment of barcodes for 22 diatom species across 210 sampling stations from the *Tara* Oceans expedition. Panels show Bayesian multilevel regression models predicting relative abundance by the interaction of genome size with **(A)** latitude, **(B)** ocean region, or **(C)** temperature. Nonlinear effects of latitude were modeled in **(A)** using a generalized additive model. Significant estimates for the Arctic, Southern, South

Pacific, and North Pacific Oceans are shown with solid lines in **(B)**. Nonsignificant estimates for the other ocean regions are shown with dotted lines in **(B)**. Although all predictors are treated as continuous in **(C)**, we used the model to predict the interactive effect of 4 discrete temperatures (0, 10, 20, and 30˚C) with genome size on relative species abundance. The Bayesian r^2 for each model is indicated above each panel. The data and code to generate this figure can be found in https://doi.org/10.5281/zenodo.12608914.
(TIF)

S1 Table. Summary of strain collection, accession numbers, and genome characterization.
(XLSX)

S2 Table. Summary of the calculated minimum and maximum cell volumes.
(XLSX)

S3 Table. Results of the phylogenetic path analyses.
(XLSX)

S4 Table. Results of genome size estimation.
(XLSX)

S5 Table. Results of growth experiments and doubling time calculation.
(XLSX)

Acknowledgments

We thank Jeremy Beaulieu for comments on an earlier version of the manuscript. Simon Tye helped design and implement Fig 1. The used resources were available through the Arkansas High Performance Computing Center, which is funded through multiple NSF grants and the Arkansas Economic Development Commission.

Author Contributions

Conceptualization: Wade R. Roberts, Adam M. Siepielski, Andrew J. Alverson.

Data curation: Wade R. Roberts.

Formal analysis: Wade R. Roberts, Andrew J. Alverson.

Funding acquisition: Andrew J. Alverson.

Investigation: Wade R. Roberts, Adam M. Siepielski, Andrew J. Alverson.

Methodology: Wade R. Roberts, Adam M. Siepielski.

Project administration: Wade R. Roberts, Andrew J. Alverson.

Resources: Andrew J. Alverson.

Supervision: Andrew J. Alverson.

Visualization: Wade R. Roberts.

Writing – original draft: Wade R. Roberts, Andrew J. Alverson.

Writing – review & editing: Wade R. Roberts, Adam M. Siepielski, Andrew J. Alverson.

References

1. Relyea R, Ricklefs RE. The Economy of Nature: Seventh Edition. Macmillan Learning; 2013.

2. Sutherland WJ, Freckleton RP, Godfray HCJ, Beissinger SR, Benton T, Cameron DD, et al. Identification of 100 fundamental ecological questions. J Ecol. 2013; 101:58–67.
3. Levin SA, Carpenter SR, Godfray HCJ, Kinzig AP, Loreau M, Losos JB, et al. The Princeton Guide to Ecology. Princeton University Press; 2009.
4. Murdoch WW. Population regulation in theory and practice. Ecology. 1994; 75:271–287.
5. Case TJ. An Illustrated Guide to Theoretical Ecology. Oxford University Press; 1999.
6. Sibly RM, Hone J. Population growth rate and its determinants: An overview. Philos Trans R Soc Lond B Biol Sci. 2002; 357:1153–1170. https://doi.org/10.1098/rstb.2002.1117 PMID: 12396508
7. Ricklefs RE. Community diversity: Relative roles of local and regional processes. Science. 1987; 235:167–171. https://doi.org/10.1126/science.235.4785.167 PMID: 17778629
8. Levin SA. The problem of pattern and scale in ecology: The Robert H. MacArthur Award Lecture Ecology. 1992; 73:1943–1967.
9. Buckley LB, Roughgarden J. Effect of species interactions on landscape abundance patterns. J Anim Ecol. 2005; 74:1182–1194.
10. Louthan AM, Doak DF, Angert AL. Where and when do species interactions set range limits? Trends Ecol Evol. 2015; 30:780–792. https://doi.org/10.1016/j.tree.2015.09.011 PMID: 26525430
11. McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, et al. Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. Ecol Lett. 2007; 10:995–1015. https://doi.org/10.1111/j.1461-0248.2007.01094.x PMID: 17845298
12. Messier J, McGill BJ, Lechowicz MJ. How do traits vary across ecological scales? A case for trait-based ecology. Ecol Lett. 2010; 13:838–848. https://doi.org/10.1111/j.1461-0248.2010.01476.x PMID: 20482582
13. Pawar S, Woodward G, Dell AI, editors. Trait-Based Ecology–From Structure to Function. 1st ed. Academic Press; 2015.
14. Zakharova L, Meyer KM, Seifan M. Trait-based modelling in ecology: A review of two decades of research. Ecol Model. 2019; 407:108703.
15. White EP, Ernest SKM, Kerkhoff AJ, Enquist BJ. Relationships between body size and abundance in ecology. Trends Ecol Evol. 2007; 22:323–330. https://doi.org/10.1016/j.tree.2007.03.007 PMID: 17399851
16. Marañón E. Cell size as a key determinant of phytoplankton metabolism and community structure. Ann Rev Mar Sci. 2015; 7:241–264.
17. Cermeño P, Marañón E, Harbour D, Harris RP. Invariant scaling of phytoplankton abundance and cell size in contrasting marine environments. Ecol Lett. 2006; 9:1210–1215. https://doi.org/10.1111/j.1461-0248.2006.00973.x PMID: 17040323
18. Sommer U, Peter KH, Genitsaris S, Moustaka-Gouni M. Do marine phytoplankton follow Bergmann's rule sensu lato? Biol Rev Camb Philos Soc. 2017; 92:1011–1026. https://doi.org/10.1111/brv.12266 PMID: 27028628
19. James FC. Geographic size variation in birds and its relationship to climate. Ecology. 1970; 51:365–390.
20. Atkinson D. Temperature and organism Size—A biological law for ectotherms? In: Begon M, Fitter AH, editors. Advances in Ecological Research. Academic Press; 1994. p. 1–58.
21. Atkinson D, Ciotti BJ, Montagnes DJS. Protists decrease in size linearly with temperature: ca. 2.5% degrees C(-1). Proc Biol Sci. 2003; 270:2605–2611. https://doi.org/10.1098/rspb.2003.2538 PMID: 14728784
22. Doyle JJ, Coate JE. Polyploidy, the nucleotype, and novelty: The impact of genome doubling on the biology of the cell. Int J Plant Sci. 2019; 180:1–52.
23. Cavalier-Smith T., Economy speed and size matter: Evolutionary forces driving nuclear genome miniaturization and expansion. Ann Bot. 2005; 95:147–175.
24. Cavalier-Smith T. Skeletal DNA and the evolution of genome size. Annu Rev Biophys Bioeng. 1982; 11:273–302. https://doi.org/10.1146/annurev.bb.11.060182.001421 PMID: 7049065
25. Cavalier-Smith T. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J Cell Sci. 1978; 34:247–278. https://doi.org/10.1242/jcs.34.1.247 PMID: 372199
26. Gregory RT. The Evolution of the Genome. Elsevier; 2011.
27. Hessen DO, Daufresne M, Leinaas HP. Temperature-size relations from the cellular-genomic perspective. Biol Rev Camb Philos Soc. 2013; 88:476–489. https://doi.org/10.1111/brv.12006 PMID: 23551915

28. Herben T, Suda J, Klimesová J, Mihulka S, Říha P, Símová I. Ecological effects of cell-level processes: Genome size, functional traits and regional abundance of herbaceous plant species. Ann Bot. 2012; 110:1357–1367. https://doi.org/10.1093/aob/mcs099 PMID: 22628380

29. Roddy AB, Théroux-Rancourt G, Abbo T, Benedetti JW, Brodersen CR, Castro M, et al. The scaling of genome size and cell size limits maximum rates of photosynthesis with implications for ecological strategies. Int J Plant Sci. 2020; 181:75–87.

30. Bennett MD. Variation in genomic form in plants and its ecological implications. New Phytol. 1987; 106:177–200.

31. Armbrust EV. The life of diatoms in the world's oceans. Nature. 2009; 459:185–192. https://doi.org/10.1038/nature08057 PMID: 19444204

32. Nakov T, Beaulieu JM, Alverson AJ. Insights into global planktonic diatom diversity: The importance of comparisons between phylogenetically equivalent units that account for time. ISME J. 2018; 12:2807–2810.

33. Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, et al. Insights into global diatom distribution and diversity in the world's ocean. Proc Natl Acad Sci U S A. 2016; 113:E1516–E1525. https://doi.org/10.1073/pnas.1509523113 PMID: 26929361

34. Gregory TR. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc. 2001; 76:65–101. https://doi.org/10.1017/s1464793100005595 PMID: 11325054

35. Roberts WR, Ruck EC, Downey KM, Pinseel E, Alverson AJ. Resolving marine–freshwater transitions by diatoms through a fog of gene tree discordance. Syst Biol. 2023; 72:984–997.

36. Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, Pacini E, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc Natl Acad Sci U S A. 2014; 111:E4096–E4102. https://doi.org/10.1073/pnas.1321152111 PMID: 25225383

37. Brown JH, Gillooly JF, Allen AP, Savage VM, West GB. Toward a metabolic theory of ecology. Ecology. 2004; 85:1771–1789.

38. Litchman E, Klausmeier CA, Schofield OM, Falkowski PG. The role of functional traits and trade-offs in structuring phytoplankton communities: Scaling from cellular to ecosystem level. Ecol Lett. 2007; 10:1170–1181.

39. Raven JA. Small is beautiful: The picophytoplankton. Funct Ecol. 1998; 12:503–513.

40. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, et al. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015; 348:1261605. https://doi.org/10.1126/science.1261605 PMID: 25999516

41. Hessen DO, Jeyasingh PD, Neiman M, Weider LJ. Genome streamlining and the elemental costs of growth. Trends Ecol Evol. 2010; 25:75–80. https://doi.org/10.1016/j.tree.2009.08.004 PMID: 19796842

42. Barton AD, Pershing AJ, Litchman E, Record NR, Edwards KF, Finkel ZV, et al. The biogeography of marine plankton traits. Ecol Lett. 2013; 16:522–534.

43. Litchman E, Klausmeier CA. Trait-based community ecology of phytoplankton. Annu Rev Ecol Evol Syst. 2008; 39:615–639.

44. Chisholm SW. Phytoplankton Size. In: Falkowski PG, Woodhead AD, Vivirito K, editors. Primary productivity and biogeochemical cycles in the sea. NY: Springer New York; 1992. p. 213–237.

45. Edwards KF, Thomas MK, Klausmeier CA, Litchman E. Allometric scaling and taxonomic variation in nutrient utilization traits and maximum growth rate of phytoplankton. Limnol Oceanogr. 2012; 57:554–566.

46. Gjoni V, Glazier DS, Wesner JS, Ibelings BW, Thomas MK. Temperature, resources and predation interact to shape phytoplankton size–abundance relationships at a continental scale. Glob Ecol Biogeogr. 2023; 32:2006–2016.

47. Irwin AJ, Finkel ZV, Schofield OME, Falkowski PG. Scaling-up from nutrient physiology to the size-structure of phytoplankton communities. J Plankton Res. 2006; 28:459–471.

48. Malerba ME, Marshall DJ. Larger cells have relatively smaller nuclei across the Tree of Life. Evol Lett. 2021; 5:306–314. https://doi.org/10.1002/evl3.243 PMID: 34367657

49. Von Dassow P, Petersen TW, Chepurnov VA, Virginia AE. Inter- and intraspecific relationships between nuclear DNA content and cell size in selected members of the centric diatom genus Thalassiosira (Bacillariophyceae). J Phycol. 2008; 44:335–349.

50. Connolly JA, Oliver MJ, Beaulieu JM, Knight CA, Tomanek L, Moline MA. Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). J Phycol. 2008; 44:124–131.

51. Bennett MD. The duration of meiosis. Proc Roy Soc Lond B. 1971; 178:277–299.

52. Sharpe SC, Koester JA, Loebl M, Cockshutt AM, Campbell DA, Irwin AJ, et al. Influence of cell size and DNA content on growth rate and photosystem II function in cryptic species of Ditylum brightwellii. PLoS ONE. 2012; 7:e52916. https://doi.org/10.1371/journal.pone.0052916 PMID: 23300819

53. Lynch M. The Origins of Genome Architecture. 1st ed. Sinauer Associates Inc; 2007.

54. Whitney KD, Baack EJ, Hamrick JL, Godt MJW, Barringer BC, Bennett MD, et al. A role for nonadaptive processes in plant genome size evolution? Evolution. 2010; 64:2097–2109. https://doi.org/10.1111/j.1558-5646.2010.00967.x PMID: 20148953

55. Roddy AB, Alvarez-Ponce D, Roy SW. Mammals with small populations do not exhibit larger genomes. Mol Biol Evol. 2021; 38:3737–3741.

56. Lefébure T, Morvan C, Malard F, François C, Konecny-Dupré L, Guéguen L, et al. Less effective selection leads to larger genomes. Genome Res. 2017; 27:1016–1028. https://doi.org/10.1101/gr.212589.116 PMID: 28424354

57. Hansen AN, Visser AW. The seasonal succession of optimal diatom traits. Limnol Oceanogr. 2019; 64:1442–1457.

58. Behrenfeld MJ, Halsey KH, Boss E, Karp-Boss L, Milligan AJ, Peers G. Thoughts on the evolution and ecological niche of diatoms. Ecol Monogr. 2021; 91:E01457.

59. Sarthou G, Timmermans KR, Blain S, Tréguer P. Growth physiology and fate of diatoms in the ocean: a review. J Sea Res. 2005; 53:25–42.

60. Campbell MD, Schoeman DS, Venables W, Abu-Alhaija R, Batten SD, Chiba S, et al. Testing Bergmann's rule in marine copepods. Ecography. 2021; 44:1283–1295.

61. Catford JA, Wilson JRU, Pyšek P, Hulme PE, Duncan RP. Addressing context dependence in ecology. Trends Ecol Evol. 2022; 37:158–170. https://doi.org/10.1016/j.tree.2021.09.007 PMID: 34756764

62. Adams GL, Pichler DE, Cox EJ, O'Gorman EJ, Seeney A, Woodward G, et al. Diatoms can be an important exception to temperature-size rules at species and community levels of organization. Glob Chang Biol. 2013; 19:3540–3552. https://doi.org/10.1111/gcb.12285 PMID: 23749600

63. Callaghan CT, Nakagawa S, Cornwell WK. Global abundance estimates for 9,700 bird species. Proc Natl Acad Sci U S A. 2021; 118:e2023170118.

64. Grime JP, Mowforth MA. Variation in genome size–an ecological interpretation. Nature. 1982; 299:151–153.

65. Semlitsch RDO'Donnell KM, Thompson FR III. Abundance, biomass production, nutrient content, and the possible role of terrestrial salamanders in Missouri Ozark forest ecosystems. Can J Zool. 2014; 92:997–1004.

66. Gregory TR. Genome size evolution in animals. In: Gregory TR, editor. The Evolution of the Genome. Burlington: Academic Press; 2005. p. 3–87.

67. Møller A, Jennions MD. How much variance can be explained by ecologists and evolutionary biologists? Oecologia. 2002; 132:492–500. https://doi.org/10.1007/s00442-002-0952-2 PMID: 28547634

68. Edlund MB, Stoermer EF. Ecological, evolutionary, and systematic significance of diatom life histories. J Phycol. 1997; 33:897–918.

69. Jovtchev G, Schubert V, Meister A, Barow M, Schubert I. Nuclear DNA content and nuclear and cell volume are positively correlated in angiosperms. Cytogenet Genome Res. 2006; 114:77–82. https://doi.org/10.1159/000091932 PMID: 16717454

70. Mueller RL, Gregory TR, Gregory SM, Hsieh A, Boore JL. Genome size, cell size, and the evolution of enucleated erythrocytes in attenuate salamanders. Zoology. 2008; 111:218–230. https://doi.org/10.1016/j.zool.2007.07.010 PMID: 18328681

71. Dell'Aquila G, Zauner S, Heimerl T, Kahnt J, Samel-Gondesen V, Runge S, et al. Mobilization and cellular distribution of phosphate in the diatom *Phaeodactylum tricornutum*. Front Plant Sci. 2020; 11:579.

72. Malerba ME, Ghedini G, Marshall DJ. Genome size affects fitness in the eukaryotic alga *Dunaliella tertiolecta*. Curr Biol. 2020; 30:3450–3456.e3.

73. Guillard RRL, Hargraves PE. Stichochrysis immobilis is a diatom, not a chrysophyte. Phycologia. 1993; 32:234–236.

74. Guillard RRL, Lorenzen CJ. Yellow-green algae with chlorophyllide c 1, 2. J Phycol. 1972; 8:10–14.

75. Nelson DR, Hazzouri KM, Lauersen KJ, Jaiswal A, Chaiboonchoe A, Mystikou A, et al. Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. Cell Host Microbe. 2021; 29:250–266.e8. https://doi.org/10.1016/j.chom.2020.12.005 PMID: 33434515

76. Liu K, Chen Y, Cui Z, Liu S, Xu Q, Chen N. Comparative analysis of chloroplast genomes of Thalassiosira species. Front Mar Sci. 2021; 8:788307.

77. Wang Y, Chen Y, Wang J, Liu F, Chen N. Mitochondrial genome of the harmful algal bloom species Odontella regia (Mediophyceae, Bacillariophyta). J Appl Phycol. 2021; 33:855–868.

78. Wang Y, Liu S, Wang J, Yao Y, Chen Y, Xu Q, et al. Diatom biodiversity and speciation revealed by comparative analysis of mitochondrial genomes. Front Plant Sci. 2022; 13:749982. https://doi.org/10.3389/fpls.2022.749982 PMID: 35401648

79. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

80. Nikolenko SI, Korobeynikov AI, Alekseyev MA. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics. 2013; 14(Suppl 1):S7.

81. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012; 19:455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

82. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. F1000Res. 2017; 6:1287.

83. Hongo Y, Kimura K, Takaki Y, Yoshida Y, Baba S, Kobayashi G, et al. The genome of the diatom *Chaetoceros tenuissimus* carries an ancient integrated fragment of an extant virus. Sci Rep. 2021; 11:22877.

84. Roberts WR, Downey KM, Ruck EC, Traller JC, Alverson AJ. Improved reference genome for Cyclotella cryptica CCMP332, a model for cell wall morphogenesis, salinity adaptation, and lipid production in diatoms (Bacillariophyta). 2020; G3(10):2965–2974.

85. Filloramo GV, Curtis BA, Blanche E, Archibald JM. Re-examination of two diatom reference genomes using long-read sequencing. BMC Genomics. 2021; 22:379. https://doi.org/10.1186/s12864-021-07666-3 PMID: 34030633

86. Sorokina M, Barth E, Zulfiqar M, Kwantes M, Pohnert G, Steinbeck C. Draft genome assembly and sequencing dataset of the marine diatom Skeletonema cf. costatum RCC75. Data Brief. 2022; 41:107931.

87. Pinseel E, Ruck EC, Nakov T, Jonsson PR, Kourtchenko O, Kremp A, et al. Local adaptation of a marine diatom is governed by genome-wide changes in diverse metabolic processes. bioRxiv. 2023. p. 2023.09.22.559080. https://doi.org/10.1101/2023.09.22.559080

88. Lommer M, Specht M, Roy A-S, Kraemer L, Andreson R, Gutowska MA, et al. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biol. 2012; 13:R66. https://doi.org/10.1186/gb-2012-13-7-r66 PMID: 22835381

89. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717

90. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018; 35:543–548. https://doi.org/10.1093/molbev/msx319 PMID: 29220515

91. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013; 30:772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

92. Steenwyk JL, Buida TJ 3rd, Li Y, Shen X-X, Rokas A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 2020; 18:e3001007. https://doi.org/10.1371/journal.pbio.3001007 PMID: 33264284

93. Brown JW, Walker JF, Smith SA. Phyx: Phylogenetic tools for unix. Bioinformatics. 2017; 33:1886–1888. https://doi.org/10.1093/bioinformatics/btx063 PMID: 28174903

94. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020; 37:1530–1534. https://doi.org/10.1093/molbev/msaa015 PMID: 32011700

95. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013; 30:1188–1195. https://doi.org/10.1093/molbev/mst024 PMID: 23418397

96. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24:1586–1591. https://doi.org/10.1093/molbev/msm088 PMID: 17483113

97. dos Reis M, Yang Z. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol. 2011; 28:2161–2172. https://doi.org/10.1093/molbev/msr045 PMID: 21310946

98. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018; 67:901–904. https://doi.org/10.1093/sysbio/syy032 PMID: 29718447

99. Nakov T, Beaulieu JM, Alverson AJ. Accelerated diversification is related to life history and locomotion in a hyperdiverse lineage of microbial eukaryotes (Diatoms, Bacillariophyta). New Phytol. 2018; 219:462–473. https://doi.org/10.1111/nph.15137 PMID: 29624698

100. Hasle GR, Syvertsen EE. Thalassiosiropsis, a new diatom genus from the fossil records. Micropaleontology. 1985; 31:82–91.

101. Alverson AJ. Timing marine–freshwater transitions in the diatom order Thalassiosirales. Paleobiology. 2014; 40:91–101.

102. Neptune LD. A marine micropaleontology database. Math Geol. 1994; 26:817–832.

103. Theriot E, Kociolek JP. Two new Pliocene species of Cyclostephanos (Bacillariophyceae) with comments on the classification of the freshwater Thalassiosiraceae. J Phycol. 1986; 22:121–128.

104. Shiono M, Koizumi I. Taxonomy of the Thalassiosira trifulta group in Late Neogene sediments from the northwest Pacific Ocean. Diatom Res. 2000; 15:355–382.

105. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018; 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191 PMID: 29750242

106. Li X, Waterman MS. Estimating the repeat structure and length of DNA sequences using L-tuples. Genome Res. 2003; 13:1916–1922. https://doi.org/10.1101/gr.1251803 PMID: 12902383

107. Schell T, Feldmeyer B, Schmidt H, Greshake B, Tills O, Truebano M, et al. An annotated draft genome for Radix auricularia (Gastropoda, Mollusca). Genome Biol Evol. 2017; 9:0. https://doi.org/10.1093/gbe/evx032 PMID: 28204581

108. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/.

109. Poncet P. modeest: Mode Estimation. 2019. Available from: https://CRAN.R-project.org/package=modeest.

110. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics. 1988; 2:231–239. https://doi.org/10.1016/0888-7543(88)90007-9 PMID: 3294162

111. Hu H, Bandyopadhyay PK, Olivera BM, Yandell M. Characterization of the *Conus bullatus* genome and its venom-duct transcriptome. BMC Genomics. 2011; 12:60.

112. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020; 11:1432. https://doi.org/10.1038/s41467-020-14998-3 PMID: 32188846

113. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011; 27:764–770. https://doi.org/10.1093/bioinformatics/btr011 PMID: 21217122

114. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol Evol. 2015; 7:1192–1205. https://doi.org/10.1093/gbe/evv050 PMID: 25767248

115. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020; 117:9451–9457. https://doi.org/10.1073/pnas.1921046117 PMID: 32300014

116. Petzoldt T. growthrates: Estimate Growth Rates from Experimental Data. 2022. Available from: https://CRAN.R-project.org/package=growthrates.

117. Hall BG, Acar H, Nandipati A, Barlow M. Growth rates made easy. Mol Biol Evol. 2014; 31:232–238. https://doi.org/10.1093/molbev/mst187 PMID: 24170494

118. Leblanc K, Arístegui J, Armand L, Assmy P, Beker B, Bode A, et al. A global diatom database–abundance, biovolume and biomass in the world ocean. Earth Syst Sci Data. 2012; 4:149–165.

119. Sarno D, Kooistra WHCF, Medlin LK, Percopo I, Zingone A Diversity in the genus Skeletonema (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species with the description of four new species. J Phycol. 2005; 41:151–176.

120. Sarno D, Kooistra WHCF, Balzano S, Hargraves PE, Zingone ADiversity in the genus Skeletonema (Bacillariophyceae): III. Phylogenetic position and morphological variability of *Skeletonema costatum* and *Skeletonema grevillei*, with the description of *Skeletonema ardens* sp. nov. J Phycol. 2007; 43:156–170.

121. Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci Data. 2017; 4:170093. https://doi.org/10.1038/sdata.2017.93 PMID: 28763055

122. Ibarbalz FM, Henry N, Brandão MC, Martini S, Busseni G, Byrne H, et al. Global trends in marine plankton diversity across kingdoms of life. Cell. 2019; 179:1084–1097.e21. https://doi.org/10.1016/j.cell.2019.10.008 PMID: 31730851

123. Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, et al. EPA-ng: Massively parallel evolutionary placement of genetic sequences. Syst Biol. 2019; 68:365–369. https://doi.org/10.1093/sysbio/syy054 PMID: 30165689

124. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

125. Czech L, Barbera P, Stamatakis A. Genesis and Gappa: Processing, analyzing and visualizing phylogenetic (placement) data. Bioinformatics. 2020; 36:3263–3265. https://doi.org/10.1093/bioinformatics/btaa070 PMID: 32016344

126. McKnight DT, Huerlimann R, Bower DS. Methods for normalizing microbiome data: an ecological perspective. Methods Ecol Evol. 2019; 10:389–400.

127. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A grammar of data manipulation. 2023. Available from: https://CRAN.R-project.org/package=dplyr.

128. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, et al. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data. 2015; 2:150023. https://doi.org/10.1038/sdata.2015.23 PMID: 26029378

129. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999; 401:877–884. https://doi.org/10.1038/44766 PMID: 10553904

130. Revell LJ. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). PeerJ. 2024; 12:e16505. https://doi.org/10.7717/peerj.16505 PMID: 38192598

131. Orme D. The caper package: Comparative analysis of phylogenetics and evolution in R. mirror.rcg.sfu.ca; 2013. Available from: https://mirror.rcg.sfu.ca/mirror/CRAN/web/packages/caper/vignettes/caper.pdf.

132. von Hardenberg A, Gonzalez-Voyer A. Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. Evolution. 2013; 67:378–387. https://doi.org/10.1111/j.1558-5646.2012.01790.x PMID: 23356611

133. van der Bijl W. phylopath: Easy phylogenetic path analysis in R. PeerJ. 2018; 6:e4718.

134. Stan Development Team. RStan: the R interface to Stan. 2023. Available from: https://mc-stan.org/.

135. Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. J Stat Softw. 2017; 80:1–28.

136. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20:289–290. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327

137. Lesaffre E, Lawson AB. Bayesian Biostatistics. Wiley & Sons, Limited, John; 2012.

138. Arel-Bundock V. marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests. 2023. Available from: https://CRAN.R-project.org/package=marginaleffects.

139. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. SSO Schweiz Monatsschr Zahnheilkd. 1992; 7:457–472.

140. Gelman A, Goodrich B, Gabry J, Vehtari A. R-squared for Bayesian regression models. Am Stat. 2019; 73:307–309.

141. Wickham H. ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag; 2016. Available from: https://ggplot2.tidyverse.org.

142. Yu G. aplot: Decorate a "ggplot" with Associated Information. 2023. Available from: https://CRAN.R-project.org/package=aplot.

143. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol. 2017; 8:28–36.

144. Becker RA, Minka TP, Deckmyn A. maps: Draw geographical maps. 2022.

145. Kahle D, Wickham H. ggmap: Spatial visualization with ggplot2. R J. 2013:144–161. Available from: https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.