Genus-Wide Transcriptional Landscapes Reveal Correlated Gene Networks Underlying Microevolutionary Divergence in Diatoms

Nathan G. Walworth,†1,2 Josh L. Espinoza,†2 Phoebe A. Argyle,[3] Jana Hinners,[4,5] Naomi M. Levine,[1] Martina A. Doblin,[3] Chris L. Dupont,[2] and Sinéad Collins*,[4]

[1]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-0371, USA

[2]J.Craig Venter Institute, La Jolla, CA 92037, USA

[3]Climate Change Cluster, University of Technology Sydney, Sydney, New South Wales 2007, Australia

[4]School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, UK

[5]Helmholtz-Zentrum Hereon, 21502 Geesthacht, Germany

†These authors contributed equally to this work.

***Corresponding author:** E-mail: s.collins@ed.ac.uk.

Associate editor: Emily Josephs

Abstract

Marine microbes like diatoms make up the base of marine food webs and drive global nutrient cycles. Despite their key roles in ecology, biogeochemistry, and biotechnology, we have limited empirical data on how forces other than adaptation may drive diatom diversification, especially in the absence of environmental change. One key feature of diatom populations is frequent extreme reductions in population size, which can occur both in situ and ex situ as part of bloom-and-bust growth dynamics. This can drive divergence between closely related lineages, even in the absence of environmental differences. Here, we combine experimental evolution and transcriptome landscapes (t-scapes) to reveal repeated evolutionary divergence within several species of diatoms in a constant environment. We show that most of the transcriptional divergence can be captured on a reduced set of axes, and that repeatable evolution can occur along a single major axis of variation defined by core ortholog expression comprising common metabolic pathways. Previous work has associated specific transcriptional changes in gene networks with environmental factors. Here, we find that these same gene networks diverge in the absence of environmental change, suggesting these pathways may be central in generating phenotypic diversity as a result of both selective and random evolutionary forces. If this is the case, these genes and the functions they encode may represent universal axes of variation. Such axes that capture suites of interacting transcriptional changes during diversification improve our understanding of both global patterns in local adaptation and microdiversity, as well as evolutionary forces shaping algal cultivation.

Key words: **diatom, transcriptome, transcriptome landscape, transcriptome variation, population bottleneck.**

Introduction

Diatoms are among the most diverse and ubiquitous eukaryotic microbes in aquatic environments (Field et al. 1998; Armbrust et al. 2004; Bowler et al. 2008). They play crucial roles in aquatic food webs and global nutrient cycling, yet our understanding of how genetic, environmental, and demographic factors can influence trait diversity in diatoms, or indeed the stability of diatom traits within lineages, remains in its infancy. The ecological and biogeochemical roles of phytoplankton are often studied through their functional traits (e.g. cell size) or metabolic capabilities (e.g. nitrogen fixation). Patterns of variation in diatom traits and metabolism underlie many differences in function, but our understanding of rapid trait diversification is limited and often dependent on linking differences in traits with environmental differences (Allen et al. 2008; Hennon et al. 2015; van de Waal and Litchman 2020; Wang et al. 2020). In particular, we lack an exploration of patterns of potential trait diversification in diatoms in the absence of environmental change, which is needed to establish expectations for the functional diversity in interrelated traits that can be accounted for by migration and mutation alone (Ward et al. 2021). This can be addressed experimentally by using chance events to drive rapid diversification in an unchanging laboratory environment. Diversification under these conditions gives us information about potential trait variation available for natural selection or environmental sorting to act on when environments vary and could also inform studies of how dispersal affects diversity in phytoplankton.

Trait diversification and local adaptation depend on the interplay of deterministic selective forces and random events (Ward et al. 2021). One common type of chance event is an

Received: April 28, 2023. **Revised:** August 24, 2023. **Accepted:** September 21, 2023

Open Access

extreme reduction in population size (a population bottleneck), where variation can be fixed by chance rather than by natural selection. As a result, when independent bottlenecked populations are compared, they will often differ from each other in terms of the random mutations that have been fixed by chance. In addition, because most mutations are deleterious, bottlenecked populations will, on average, have lower fitness than their own ancestors. Thus, parallel bottlenecks can produce populations with different genetic starting points for future evolution, and adaptation (as fitness recovery) can occur even in the absence of any environmental change. Following bottlenecks, populations can later be acted on by natural selection if population size recovers. During adaptation, different genetic starting points can drive adaptive differences between populations. This can be the case for populations that exist in different environments or be the basis for divergence when populations are separate but experience similar environments. Chance events are thus important in determining the heritable variation that natural selection has to act upon and can be used to explore different possible adaptive solutions to set growing conditions (Salverda et al. 2017; Windels et al. 2021).

In addition, metabolic pathways that can accumulate more variants by chance may respond more rapidly to natural selection, simply because they will have more genetic variation on which natural selection can act. For example, nitrate metabolism is both modular and somewhat redundant in marine diatoms (Smith et al. 2019), suggesting that variation in nitrate metabolism could build up by chance with little or no effect on fitness, and that a relatively high number of viable nitrate metabolism mutants could exist in bottlenecked populations. In this case, one would expect a range of solutions to functional nitrate usage across subpopulations, rather than "a" single solution, since a range of genetic starting points could exist for fitness recovery after bottlenecks or for adaptation to changes in environmental nitrate concentrations. This would also mean that rapid genetic adaptation to changes in nitrate is possible. This expectation affects how we interpret differences between populations in -omic and trait-based surveys, as well as our expectations about the generality of outcomes from laboratory evolution experiments.

Chance events, and the patterns of variation generated by them, are rarely used to generate variation during laboratory studies in marine phytoplankton, but these events are likely a common feature of evolving populations of marine microbes. In particular, open ocean and blooming phytoplankton can experience extreme fluctuations in population size, during which there is a transient increase in the role of chance events (such as population bottlenecks) relative to natural selection (adaptation to new environments) in evolving populations. For example, diatom population sizes can fluctuate by orders of magnitude over the course of a bloom, and migration between ocean currents continually introduces small subpopulations into new environments (Ruggiero et al. 2017; Behrenfeld et al. 2021). The potential for repeated founder events suggests that chance can play an important and relatively common role in the extent and nature of trait variation in natural populations.

We can experimentally investigate the contribution of chance events to molecular and trait variation in many microbes, including diatoms, with standard bottleneck experiments. Here, we generated trait variation in globally distributed marine diatom strains as a function of chance events in a constant environment using population bottlenecks (see Materials and Methods; samples from the same experiment were also taken by Hinners et al. (2022) to measure whole-cell physiological traits only). Physiological assays for trait values reported in Hinners et al. (2022) were previously developed and published in Argyle et al. (2021a, 2021b). In this manuscript, we describe global shifts in transcript values and identify the expression of specific genes and pathways that drive most of the observed variation.

We selected strains of the order Thalassiosirales (Argyle et al. 2021b) as our model due to its well-studied global distribution with respect to environmental and ecological selection, particularly in pelagic environments where bottlenecking events can regularly occur (Whittaker and Rynearson 2017) (see Alverson et al. 2011 for a detailed phylogeny of Thalassiosirales relative to other diatom taxa). Additionally, Thalassiosirales has one of the most well-characterized diatom genomes (Armbrust et al. 2004) and is one of the most abundant and diverse diatom lineages worldwide relative to other model diatoms with well-characterized genomes like *Phaeodactylum* sp. (Malviya et al. 2016). Prior studies have focused on the influence of environmental and ecological selection, as well as biogeography on diatom evolution and biogeochemistry. This sets the stage for our work on the generation of diversity in the absence of environmental selection. Our work fills a critical knowledge gap in diatom research where population dynamics can generate diversity for natural selection to act upon. Using Thalassiosirales as our model both enables the results of this study to be contextualized with extensive prior research examining environmental selection while also exploring how population dynamics affect trait diversification in a way that is relevant to other globally significant diatoms and microalgae.

We subjected replicate populations of 6 diatom strains to a series of population bottlenecks during batch culture propagation (transfer size of <8 cells) (Fig. 1). Our rationale for using bottlenecks to diversify populations is as follows. Experimental studies in microbial and viral populations have shown that regular population size fluctuations can profoundly affect patterns of adaptation (LeClair and Wahl 2017). Population bottlenecks transiently reduce the supply of beneficial mutations, which can reduce total fitness gain during adaptation (Schoustra et al. 2009). However, bottlenecked populations can also better explore a higher number of alternate adaptive solutions in cases where several high-fitness phenotypes exist (Windels et al. 2021), which results in more genetic and phenotypic diversification over repeated rounds of adaptation. In our experiment, the growth rates of bottlenecked populations initially dropped as expected, most likely due to the

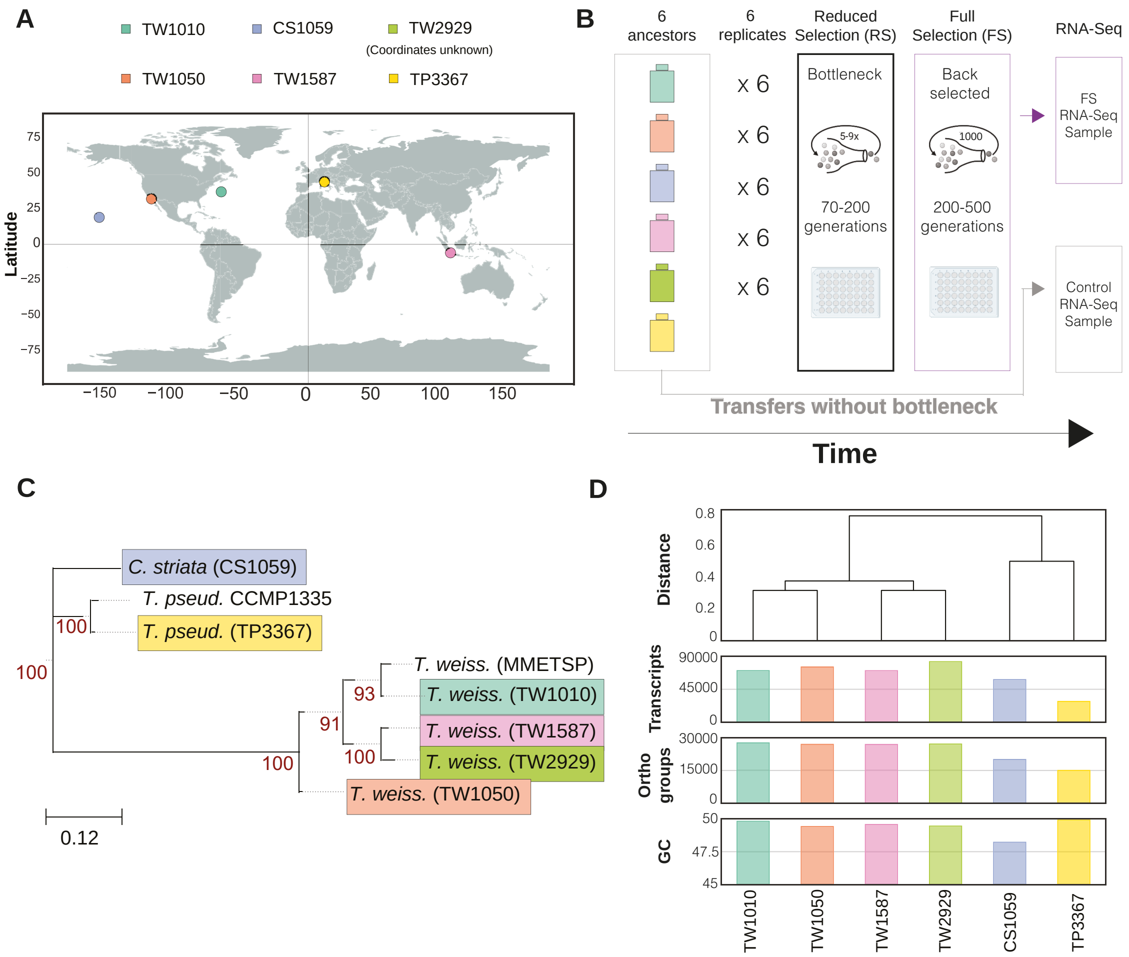

Fig. 1. Strain information and experimental design. a) Geographic coordinates of study strains. Strain TW2929 did not have coordinates provided. b) Experimental design where 6 ancestral strains with 6 replicates each were bottlenecked and then full selected prior to RNA-Seq; b) was adapted from Fig. 1 from Hinners et al. (2022) to indicate that this RNA-Seq data were generated from the same evolution experiment as the trait data in Hinners et al. (2022). c) Phylogenetic tree of diatoms in study and in MMETSP (Keeling et al. 2014) based on concatenated alignment of BUSCO *Protista_83.hmm* marker set. Blue: *C. striata* CS1059. Yellow: *T. pseudonana* TP3367. Turquoise: *T. weissflogii* TW1010. Pink: *T. weissflogii* TW1587. Green: *T. weissflogii* TW2929. Orange: *T. weissflogii* TW1050 (see Fig. S3 for uncollapsed tree). d) Hierarchical clustering of Jaccard distance based on shared orthogroups for study strains.

accumulation of deleterious mutations (Heilbron et al. 2014; Kraemer et al. 2017), but then recovered once we removed the bottleneck and grew large populations (minimum transfer size of 10^3) in standard batch cultures (Fig. S1). We then measured changes to biogeochemically important whole-cell trait values such as cell size, before, during, and after recovery of bottlenecks (Hinners et al. 2022) (see overview for a synopsis of traits and findings in Hinners et al. 2022). This approach enabled the examination of trait divergence without relying on (or allowing) divergent selection because of environmental differences. During fitness recovery, these different genetic starting points increased the chance of seeing diverse trait changes associated with fitness recovery.

Patterns of trait diversity depend on the variation of organism-level traits that are the expression of a poorly annotated and interacting group of transcribed genes. This means that considerable transcriptomic variation may underlie a consistent trait value—e.g. there are many ways to have any given cell size. Because of this, measuring whole-cell trait values (as in Hinners et al. 2022), such as cell size or elemental composition, may underestimate the genetic, epigenetic, or transcriptional variation present following bottlenecks and recovery. It is also necessarily biased in that only a small number of possible functional traits can be assessed. However, by examining transcriptomic patterns directly, we can better understand the potential for variation in metabolic pathways, including those underlying functional traits, at least in cases where the link between transcripts and function is known and annotated.

In this study, we conducted RNA-Seq on sample populations in the absence of bottlenecks (control populations) and in populations subjected to bottlenecks and allowed to subsequently recover. This methodology allowed us to investigate general patterns of transcriptional diversification in these marine diatom strains. We used a unique analytical approach to examine shifts across the pan-transcriptome in evolved populations relative to controls that were maintained as large populations over the same time frame. We examined patterns of transcriptional divergence within lineages in the absence of environmental selection, including whether global gene expression patterns remained or returned to their ancestral (control) configuration. Here, we seek to understand global shifts in transcription and identify the expression of specific genes and pathways that drive most of the observed variation. We find that many transcripts previously only associated with environmental change can be attributed to general demographic change due to population shifts alone, suggesting them to play more universal roles in diatom evolution.

Results and Discussion

Overview

A detailed description of the experimental design (Fig. 1b) can be found in Hinners et al. (2022). The transcriptomics study here only analyzed populations full selected in the ancestral temperature of 20 °C following the bottleneck. In contrast, Hinners et al. (2022) subjected full-selected populations to 2 different temperatures (20 and 24 °C) and measured phenotypes when growth rates stabilized after fitness recovery. Six replicates of each of the 6 ancestral populations ($n = 36$ cultures) were subjected to this bottleneck phase at 20 °C followed by the full-selected phase also at 20 °C. Trait measurements (population growth rate, cell size, cell complexity, relative chlorophyll a content, particulate organic carbon and nitrogen, polar lipid content, silicic acid uptake, and relative reactive oxygen content) including growth rates were performed at the beginning of the experiment (Fig. S1, filled gray circle), throughout the bottleneck if possible, and at the end of the full-selected phase. The experiment was divided into 2 main phases where phase I consisted of an initial 3-molong reduced selection (RS) (i.e. bottleneck) phase (corresponding to 70 to 200 generations, transfer size of up to 8 cells) followed by an 8-mo full selection (FS) (i.e. full-selected) phase (200 to 500 generations, transfer size of 1,000 to 2,000 cells); this full-selection phase was extended to 18 mo for the study described here. This extension was due to work restrictions during the covid pandemic, so RNA was extracted for this experiment approximately 10 mo later than the trait measurements in Hinners et al. (2022) were made. Thus, while many of these transcriptional changes in this study may be associated with observed trait changes in the Hinners et al. (2022) study, we cannot directly link transcript values to trait values. Hence, we comment on commonalities in general patterns observed in both studies, and note where annotated transcripts of known function cooccur with shifts to traits involved in that function across the 2 studies, but do not tie transcription directly to function.

Whole-cell multitrait phenotypes were represented using reduced axes similarly to the transcripts here (see Hinners et al. 2022 and Argyle et al. 2021a, 2021b for detailed descriptions of trait assays). The main conclusions of the whole trait study are that both traits and trait correlations evolve in the absence of environmental change, and that there is considerable variation among replicate populations in how this happens. We then show that despite this variation, change in multiple related traits can be reliably captured using reduced axes to identify patterns and constraints in trait change in diatom populations, including the evolution of novel multitrait phenotypes.

The experiment here captures patterns of viable transcriptomic variation that are fixed by chance due to one type of random event and analyzes how variation in global transcription patterns can emerge after fitness recovery in the absence of environmental change.

Growth Rates

During the bottleneck phase, or RS, population growth rates decreased as expected due to the accumulation of deleterious mutations (Kraemer et al. 2017). This decline in growth rates is consistent with that seen in other microbial experiments that use bottleneck transfers. Summaries of how population bottlenecks, mutation accumulation, and fitness are linked in microbial evolution experiments can be found in Estes and Lynch (2003) and Halligan and Keightley (2009). At the end of the RS phase, population growth rates of viable populations were reduced by an average of 45% compared to ancestral population growth rates (Fig. S1). Following the RS phase, populations were then propagated in batch culture with large transfer sizes in the same environment, until postbottleneck population growth rates had stabilized (Fig. S2); these are referred to as “full-selected” populations. In the FS phase of the experiments, populations adapt and partially or completely recover fitness. Because replicate populations accumulate different mutations during the RS phase, they adapt from different genetic starting points during the FS phase. This increases the chance of divergence between populations of the same initial genotype relative to cases where no population bottlenecks occurred. Because adaptation (increase in fitness) occurs mainly in response to the loss in fitness during the RS phase rather than due to a drop in fitness as a result of an environmental change, we can attribute divergence between populations of the same initial genotype to chance events alone rather than to selection imposed by environmental change. Control populations were maintained as large populations through the entire experiment. Transcriptomes were then generated across control populations and in populations after FS at 20 °C (see Materials and Methods for full experimental design and Hinners et al. 2022 for whole-cell level trait measurements).

Strain Phylogeny and Orthology

To assess the relatedness of the strains used in this study, we conducted phylogenomic analysis using a set of highly conserved diatom proteins (e.g. Keeling et al. 2014). Our phylogeny (Fig. 1c; Fig. S3) generally agrees with the ITS2 phylogeny in Argyle et al. (2021b) demonstrating concordance between ITS2 and multiprotein sequence conservation. It is worth noting that strain CCMP1059 obtained from the Provasoli-Guillard National Centre of Marine Phytoplankton is classified as *Thalassiosira* sp. (NCMA, https://ncma.bigelow.org/CCMP1059). However, upon sequence analysis of the ITS2 region by Argyle et al. (2021b), it mapped most closely to *Cyclotella striata* (94.17% identity; Table S1). Since *C. striata* is still closely related to the other strains and within the order Thalassiosirales, this updated classification does not change the interpretations or general conclusions of this manuscript. We next identified global orthologs across publicly available diatom genomes and our transcriptomes, which showed that unique ortholog abundance scaled linearly with unique transcript abundance ($R^2 = 0.95$; Fig. S4). Hierarchical clustering based on the presence/absence of global orthologs, resulted in 2 main clusters with one cluster composed of the *Thalassiosira weissflogii* strains and the other of *Thalassiosira pseudonana* and *C. striata* (Fig. 1d). Hence, TP3367 and CS1059 share more similar numbers of global orthologs and are more phylogenetically related to each other than the TW strains, although they were isolated from different locations (Fig. 1a). These phylogenetic and ortholog differences may have ancient origins, which could have been followed by subsequent ortholog and protein sequence divergence driven by environmental divergence among these diatom taxa. For example, this similarity could be driven by adaptation to warm temperature by TP3367 and CS1059 following the evolutionary divergence of TP3367 from TW strains. However, further research is needed to understand potential reasons for differences across strains. In summary, these data from our globally distributed diatom isolates reveal the diatoms in this study to be diverse both in phylogeny and genome characteristics. The different strains represent different high-fitness phenotypes in that they were sufficiently abundant in situ at time of sampling to have been sampled, have growth rates comparable to other temperate diatoms under similar culturing conditions, and have reasonably stable growth rates and phenotypes in the laboratory when propagated as large populations (see Hinners et al. 2022). It is thus unlikely that these phenotypes represent fundamentally maladaptive trait combinations or are less informative than other phenotypes evolved in laboratory experiments, even though it is entirely possible that they do not represent the most fit phenotype possible under laboratory or field conditions.

Transcriptome Landscapes

To understand the genetic variation produced in our experiment, we consider the integrated phenotype composed of many interdependent genetic relationships (Malcom et al. 2014). Here, we use transcriptome data to assess how global transcriptional relationships shift in response to fitness recovery in a constant environment. We define a transcriptome landscape (from here t-scape) which is similar to a trait-scape used in previous studies (Walworth et al. 2021; Argyle et al. 2021b, 2021a). The t-scape uses principal coordinate analysis (PCoA) of center log-ratio (CLR) transformed expression values to collapse global transcript levels onto several axes of variation to reveal those transcripts that drive most of the global transcriptional variation. Using this approach, t-scapes can reveal expression patterns across metabolic pathways that are associated with population diversification. Furthermore, location in the t-scape (i.e. PCoA plot) can be affected by both transcript levels and correlations between transcripts. Accordingly, Fig. 2a and b show 3 different high-fitness regions in the t-scape based on core ortholog expression from diverse diatom strains. To robustly assess the prevalence of orthogroups across the different strains at a range of stringency levels, we chose 2 cutoffs using orthogroups detected in at least 3 samples. The rationale for this was to both minimize the influence of singletons and/or sequencing artifacts and to examine how strain relationships changed across orthogroup thresholds in PCoA space. We first only considered orthogroups in at least 3 samples with minimum $N = 10$ counts as our relaxed threshold. Then, we increased stringency by considering orthogroups in at least 3 samples with minimum $N = 300$ counts as our strict threshold. A lower bound of 10 counts and upper bound of 300 counts were chosen because this is when strains began to segregate and converge in PCoA space, respectively. The key finding here was that the overall ordination (strain clustering and their relative positions [i.e. relationships] in PCoA space) remained conserved across orthogroup thresholds demonstrating widespread orthogroup expression, and thus core metabolic pathways, to be underlying strain-specific divergence.

In general, the clustering of strains based on core ortholog expression was consistent with both the phylogenetic and global ortholog abundance clusters (Fig. 1c and d). For example, strains TW1050, TW2929, TW1587, and TW1010 core ortholog expression converge into a single region (Fig. 2a and b). One full-selected population, CS1059-4 (Fig. 2a and b—blue triangle in lower left quadrant), evolved a strategy where the evolved population occupied an area of t-scape near the TW strains, rather than near its own ancestor. Interestingly, Hinners et al. (2022) found that most phenotypic outlier populations originated from CS1059, suggesting that this strain may have more flexibility in trait values or correlations than other strains, at least for whole organism traits. The transcriptional divergence of CS1059-4 (Fig. 2) may indicate early molecular divergence from CS1059 replicate populations prior to more pronounced trait divergence. Next, we analyzed each strain-specific cluster to explore transcripts driving intraspecific variation at each peak.

Strain-Specific Transcriptional Landscapes

We then assessed whether populations returned to the ancestral location in the t-scape (Fig. 3, circles) or if they moved to another location during the backselection phase of the

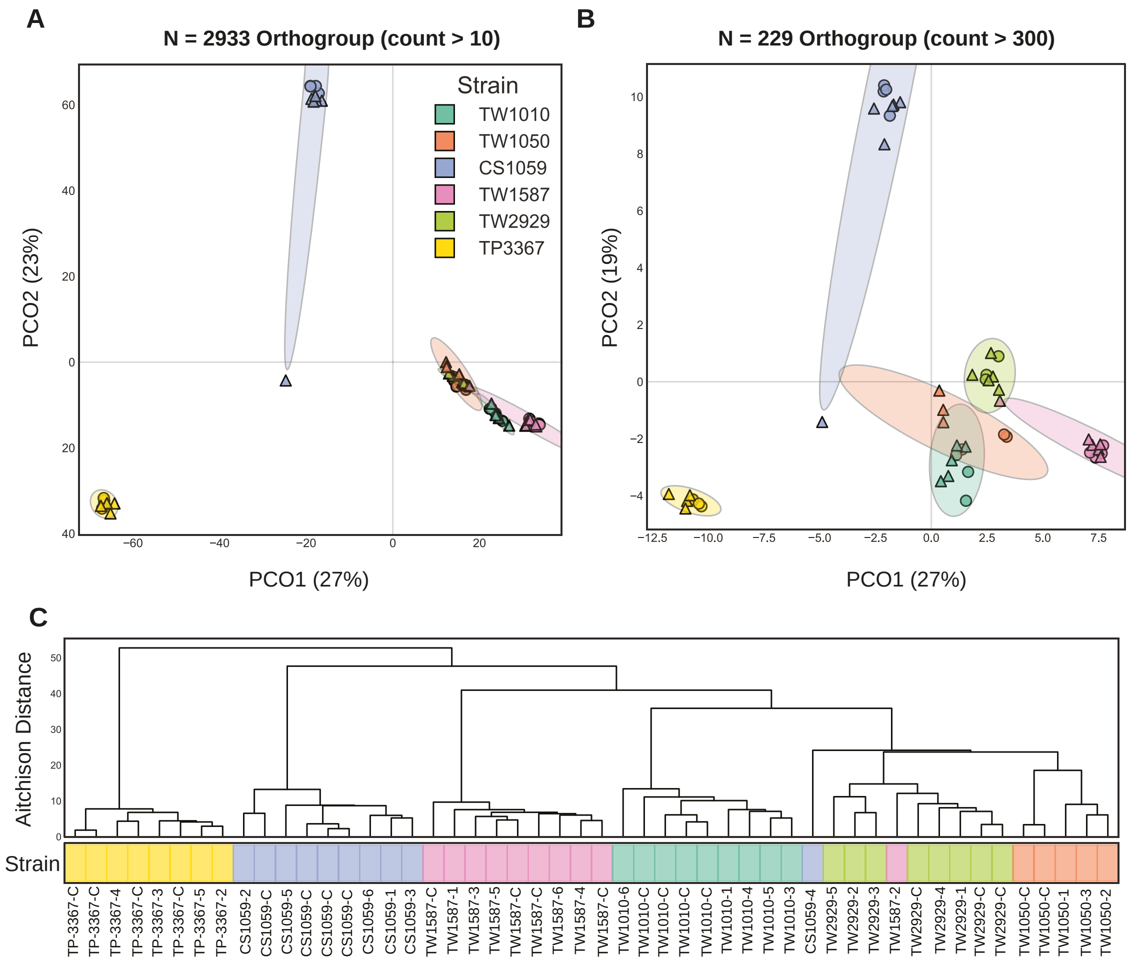

Fig. 2. Global analysis of ortholog expression among diatom strains Aitchison PCoA of conserved orthogroups that had a prevalence of a) 10 counts per sample ($N = 2{,}933$ orthogroups) and b) 300 counts per sample ($N = 229$ orthogroups). c) Hierarchical clustering using Aitchison distance of conserved orthogroup expression with at least 300 counts per sample. Circles represent ancestral controls, and triangles represent bottlenecked strains followed by FS recovery (i.e. full selected). In c), strain labels are followed by either "C" (e.g. TP-3367-C) to designate a control population or a number (e.g. TP-3367-4) to designate the replicate population number for that strain. The colored circles within each strain grouping represent 3 SD from the centroid.

experiment. In each strain, at least one backselected population moved to a new location in the t-scape. This is useful for assessing whether most mutations change transcriptional relationships, once fitness is allowed to recover. This affects whether we expect the transcriptional relationships of strains to diversify rapidly over relatively short timescales (months) even in the absence of environmental drivers, or whether ancestral transcript relationships are essentially stable on short timescales in constant environments. More importantly, this approach investigates which transcripts and transcript relationships diversify following periods when the action of natural selection may be reduced.

The strain-specific PCoA plots capture between 49% and 99% of total transcriptional variance. Figure 3 shows strain-specific t-scapes constructed from the global transcription of each replicate population within each strain. Control populations (circles) formed clusters reflecting minimal movement on the t-scape when populations were not subjected to bottlenecks or to environmental change; this supports the assumption that most of the control populations were already well adapted to the laboratory environment, and that transcriptomes for these strains were stable when the cultures were maintained as large populations. This is consistent with expecting that culture collection strains of a given genotype have broadly similar phenotypes over time and across laboratories under standard conditions and is the basis for the use of reference strains. However, TP3367 control populations exhibited more movement in the t-scape than other strains suggesting that this strain may have been adapting to some aspect of the laboratory environment. While this TP3367 trend is interesting, it is out of scope for this study, which focused on evaluating the generation of variation in strains where the assumption that that particular strain is already well adapted to the laboratory environment is met. While the data from TP3367 are included for completeness, we do not expect it to show the same pattern of changes as the other strains where control populations clustered.

Across all strains, at least one full-selected population (triangles) found another transcriptional peak in a region

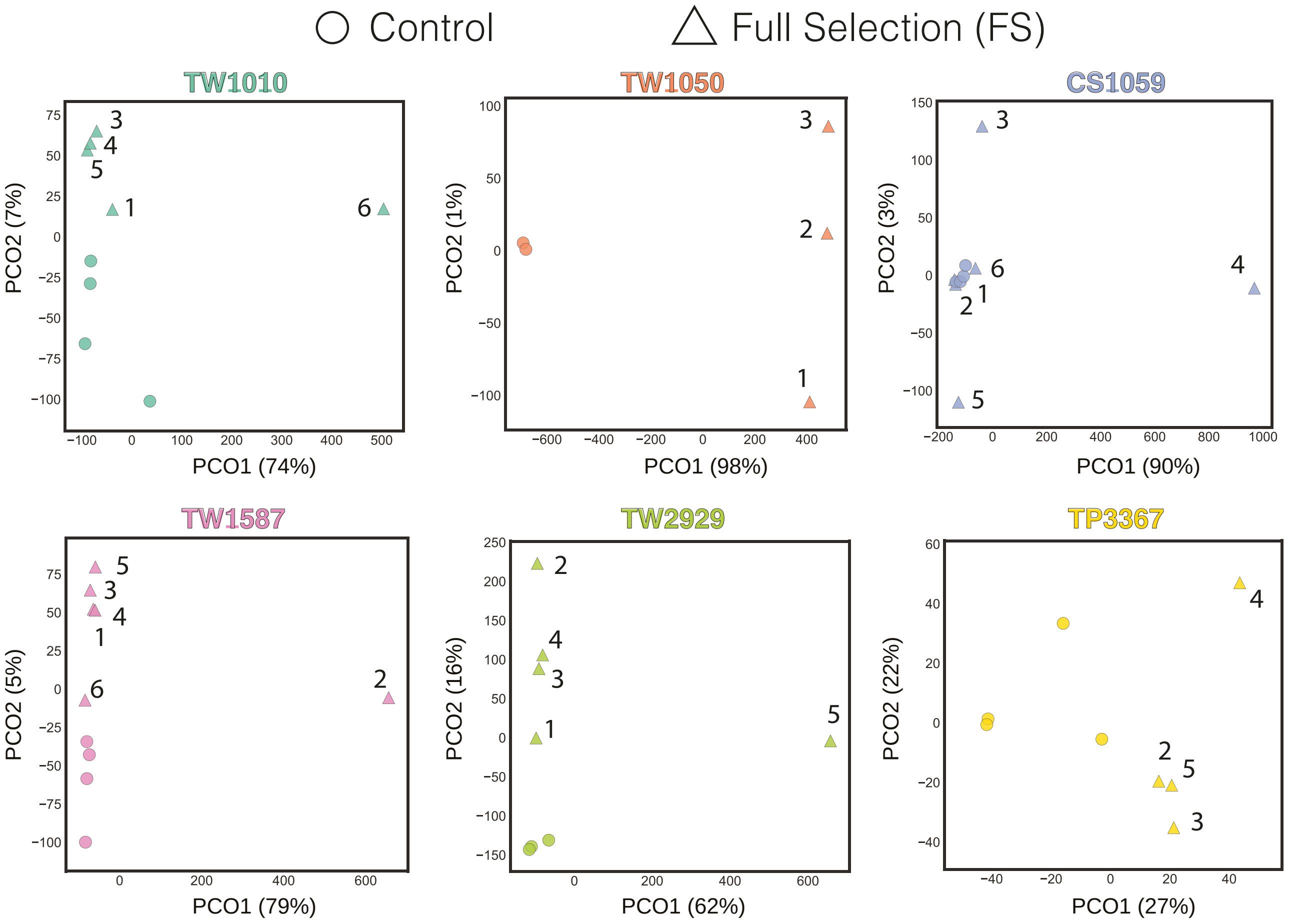

Fig. 3. Aitchison PCoA of transcript expression for study strains. Circles represent ancestral controls, and triangles represent bottlenecked strains followed by FS recovery (i.e. full selected). Numbers next to triangles denote full-selected replicate number for each strain (see Table S2 for inputs to these analyses).

of the t-scape not occupied by the control populations of that strain, indicated by movement along PCO1 (Fig. 3). The fact that at least one full-selected population per lineage (>17% of populations) moved to a new location in the t-scape demonstrates that changes to global transcriptional circuitry are common in the order, Thalassiosirales, even in constant environments. Most of the new peaks were differentiated along PCO1—the axis which captured the majority of the variance in the ancestral populations. While most of the movement and explained variance occurred along PCO1, transcripts driving differentiation along PCO2 could have biological significance in generating phenotypic differences, but they explain less variance and are not explored further in this study.

Some strains produced more variation in transcriptional patterns between replicate populations than did others. Specifically, most full-selected populations from strains CS1059, TW1587, and TW1010 returned to the ancestral region of the t-scape. Conversely, all full-selected populations from strain TW1050 migrated to an alternative region in the t-scape. This suggests that transcriptional diversification in the face of population size fluctuations varies between genotypes. An alternate explanation is that the culture collection isolate of TW1050 is less well adapted to laboratory conditions than the other strains. However, there was no significant mean increase in growth rate of TW1050 populations that were bottlenecked and recovered relative to the ancestor in this experiment (Hinners et al. 2022). This suggests that the control populations did not have significantly lower fitness under our laboratory conditions than the full-selected populations for the genotype. For all strains, more high-fitness regions of the t-scape may be identified if more replicate populations were generated and sequenced. As noted previously, TP3367 control populations exhibited intraspecific transcriptional variation without bottlenecking (Fig. 3f, circles), which could indicate that TP3367 was still adapting to general lab conditions. Additionally, less variation overall was associated with the PCO axes of this strain. Hence, we cannot be sure that diversification of full-selected populations was only due to bottlenecks. So, while TP3367 full-selected populations did seem to find 2 new defined peaks with TP3367-4 residing in one and TP3367-2, TP3367-3, and TP3367-5 residing in the other along PCO2 (22% explained variance), we are cautious in our interpretation due to the variability in control populations.

We next investigated if movement in the t-scape was driven primarily by expression of genes shared among strains (i.e. global orthologs). To do this, we reconstructed the t-scapes using only the global orthologs (Fig. S5), which showed similar relationships between controls (circles) and full-selected populations (triangles) as was found between controls and full-selected populations in t-scapes constructed from all transcripts (Fig. 3). In line with this, Fig. S5 demonstrates that diatoms in the order, Thalassiosirales, can rearrange transcriptional relationships among core diatom genes (i.e. orthologs) following bottleneck events. We next investigated what transcripts and pathways strains explained movement on the trait-scape.

Identifying Transcripts Associated with New Expression Patterns

We first used pathway enrichment analysis (Materials and Methods) on transcripts loaded onto respective strain-level PC axes to identify significantly overrepresented pathways (File S1). These analyses revealed enrichment in numerous central metabolic pathways such as carbon fixation, pyruvate metabolism, glycolysis/gluconeogenesis, amino acid metabolism, porphyrin and chlorophyll metabolism, and pentose phosphate metabolism. The overrepresentation of these pathways is consistent with the observation that orthologs drove the majority of transcriptional variation in t-scapes (Fig. S6). Next, we identified specific transcripts in these pathways.

To examine which transcripts were most associated with the movement of full-selected strains within the t-scape through the discovery of new peaks (Fig. 3), we analyzed transcripts harboring the largest PCoA loading values (both positive and negative) on PCO1 and PCO2 axes, respectively. Each transcript loading value reflects how much a particular transcript contributes to that principal coordinate axis such that transcripts with large loading values are important for explaining the observed variance in the data set. To do this, we identified highly up- and downregulated transcripts driving the clustering patterns in the t-scapes.

Specifically, for PCO1, we first tested different numbers of transcripts harboring the most positive loading values (e.g. $n = 500$, $n = 1{,}000$, and $n = 1{,}500$) for each strain. We then conducted hierarchical clustering of the Euclidean distance among their transcript values. We did the same for genes with the most negative loading values. We then selected the number of transcripts where at least one of these smaller hierarchical dendrograms maintained the same clustering pattern as the global t-scape containing all transcripts in Fig. 3. Of these transcripts with the most extreme loading values, the majority represented global orthologs (Fig. S6). Other than universally expressed core orthologs across all diatoms (e.g. Fig. 2), many of these orthologous sequences are not shared by all diatom strains. This demonstrates that most changes associated with transcriptome diversification within an environment are components of the diatom pangenome. Taken together, these data show that exploration of the t-scape in a stable environment is not primarily driven by strain-specific transcripts but by changes in widely shared pathways.

Below, we outline the details of how we defined the transcripts included in this analysis. In Fig. 4a, TW1010-6 is replicate population number 6 of the full-selected TW1010 strain. TW1010-6 moved to a new location in the TW1010 t-scape (Fig. 4a, top panel) and formed its own expression cluster relative to the other replicate populations (Fig. 4a, bottom panel). Upon clustering different amounts of transcripts harboring the most negative (Fig. 4b, upper left plot) and most positive (Fig. 4b, upper right plot) loading values on PCO1, we observed that at 2,000 transcripts ($n = 4{,}000$ total transcripts analyzed for PCO1), at least one of the dendrograms (Fig. 4b, upper panels) had a clustering pattern consistent with the pattern seen in the case with all transcripts (Fig. 4a, bottom panel). Observing consistent clustering patterns between the smaller dendrograms and the global dendrograms in Fig. 4a (bottom panel) indicates that the subset of transcripts in the former dendrogram is those primarily driving the clustering pattern in the global one. In this case, the 2,000 most positive loading values (Fig. 4b, upper right plot) reflected the most similar pattern to the global t-scape pattern (Fig. 4a, bottom plot) followed by the 2,000 most negative loading values (Fig. 4b, upper left plot). Clustering beyond 2,000 transcripts introduced less consistent clustering patterns relative to the global transcriptional plots indicating a greater inclusion of transcripts that did not strongly contribute to the clustering of TW1010-6 in its strain-specific t-scape. The strong contrast in expression values observed in the transcriptional dendrograms (Fig. 4b, purple = higher relative expression and blue = lower relative expression) is consistent with their corresponding, extreme loading values. Here, the most negative loading values correspond to reduced transcript levels in TW1010-6 relative to other TW1010 populations while the most positive loading values correspond to greater relative transcription.

To identify metabolic pathways potentially involved in movement on the t-scape, we tested the 2,000 most positive and negative loading values for all other strains for PCO1, using the same method as above (e.g. Fig. 4c and d). We conducted the same analyses for PCO2 for all strains and found that the 1,000 most positive and negative loading values ($n = 2{,}000$ total) yielded clustering consistent with the strain-specific, global t-scape along PCO2 (e.g. Fig. 4b and d, bottom 2 plots). This reduced number of transcripts is consistent with less variance being explained on PCO2 than on PCO1, such that departures from control expression values or patterns may not represent transcriptional patterns that differ very much from the ancestral one. All CLR-transformed transcripts harboring the most extreme loading values per PCO per strain can be found in File S2.

Figure 4 shows 2 different representative cases. In one instance, the TW1010-6 full-selected population moved to a new location in the t-scape whereas all other full-selected

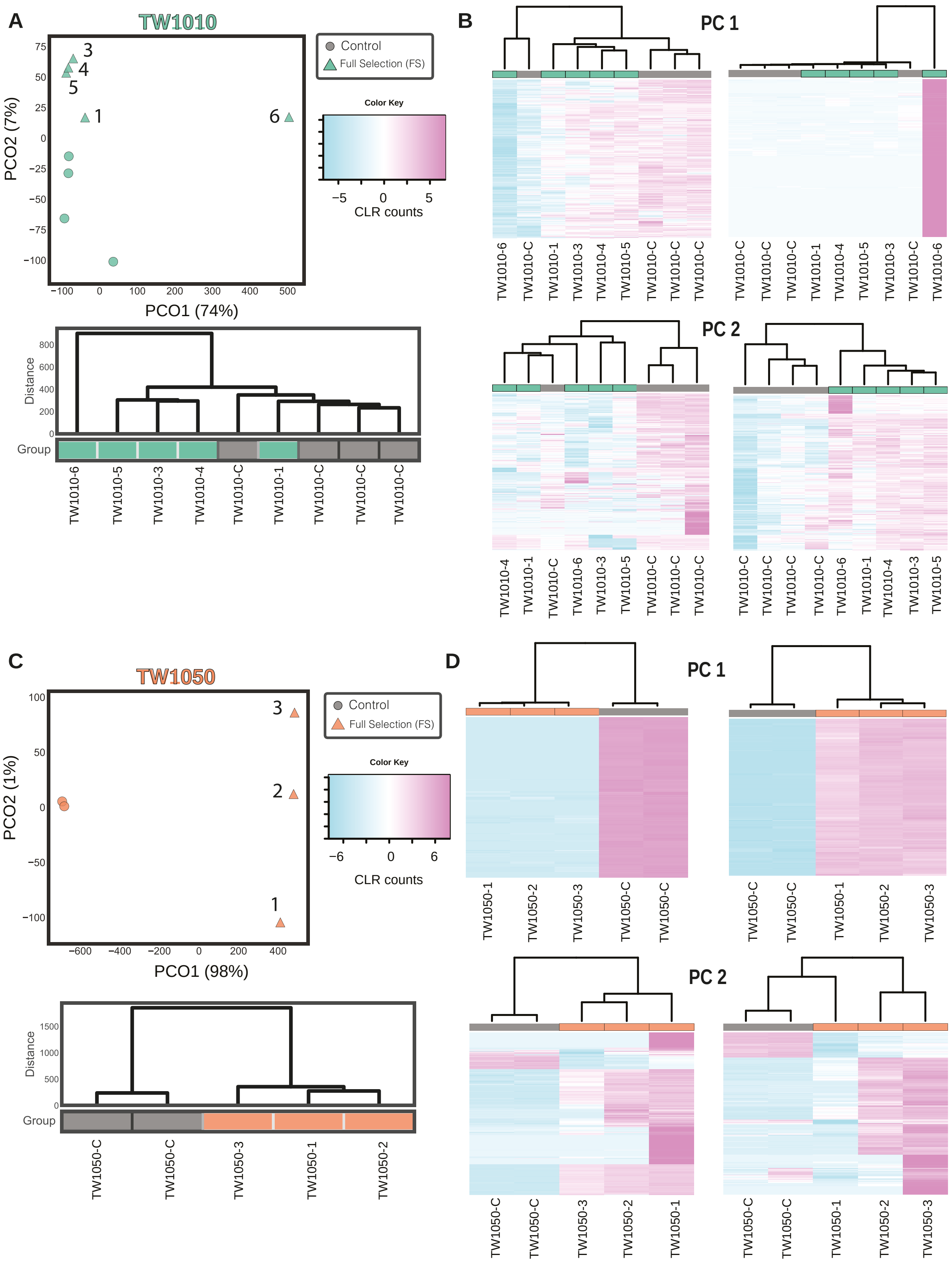

Fig. 4. Representative analysis of transcripts driving the discovery of new peaks in TW1010 and TW1050. a) The upper and lower panels are the TW1010 PCoA plot and hierarchical dendrograms of CLR-transformed transcript counts, respectively. b) The upper row contains hierarchical clusters of TW1010 CLR-transformed transcript counts of the 2,000 most negative (left plot) and 2,000 most positive (right plot) PCoA transcript loading values for PCO1. The lower row contains the same clusters except for PCO2. c) The upper and lower panels are the TW1050 PCoA plot and hierarchical dendrograms of CLR-transformed transcript counts, respectively. d) The upper row contains hierarchical clusters of TW1050 CLR-transformed transcript counts of the 2,000 most negative (left plot) and 2,000 most positive (right plot) PCoA transcript loading values for PCO1. The lower row contains the same clusters except for PC2 using the 1,000 most positive and 1,000 most negative loading values.

populations returned or remained in the region occupied by the TW1010 controls. While TW1010-1, TW1010-3, TW1010-4, and TW1010-5 formed a cluster along PCO2 (Fig. 4a, PCoA plot; Fig. 4b, bottom right plot), it is unclear if this is a clearly defined cluster distinct from the controls due to the low variance explained on PCO2. In the other case, all full-selected replicates of strain TW1050 migrated to a single new location along PCO1, which explains >93% of the variance (Fig. 4c and d). Hence, the TW1050 control populations may have more options for diversification, at least in this standard laboratory environment. Due to the low amount of explained variance on the TW1050 PCO2 axis, no new locations were obviously occupied along PCO2 following backselection.

Metabolic Transcripts Associated with New Peak Discovery

To examine the most influential metabolic transcripts driving movement on the t-scape in this experiment, we focused on transcripts harboring the most positive loading values across PCO1 for all strains (e.g. Fig. 4b and d, upper right plots). These positive loading values reflect transcripts with higher expression values in the full-selected replicates that moved to new locations following bottlenecking and FS (e.g. TW1010-6, TW1050-1, TW1050-2, and TW1050-3). Taken together, these transcripts represent those that uniquely increased in the replicates that found a peak that differed from the control populations of the same genotype.

Across all strains, numerous transcripts related to deleterious physiological changes, reactive oxygen species (ROS), carbon metabolism, and nitrogen metabolism exhibited consistently increased relative expression levels. This suggests that variation in these gene expression patterns may have a fundamental role in generating diatom transcriptome diversity. These transcriptional changes are consistent with trait observations from Hinners et al. (2022) where the largest overall differences between full-selected populations relative to their ancestors and controls were in levels of ROS, particulate organic carbon (POC), particulate organic nitrogen (PON), and lipid content. For example, in the study here, increased expression of numerous heat shock proteins (HSPs), aldehyde dehydrogenases (ALDHs), superoxide dismutases (SOD), aconitases, and glutathione-related transcripts was observed across strains (File S2). These changes are consistent with prior studies in diatoms that detected upregulation of these transcripts during deleterious physiological changes (Allen et al. 2008; Lauritano et al. 2015; Wang et al. 2020). Increased transcripts of trehalose 6-phosphate synthase were detected in TW1010, TW1050, CS1059, and TW1587. Trehalose is an intermediate, disaccharide sugar that can aid in osmotic adjustment through protein stabilization. Trehalose accumulation has been observed under iron limitation in diatoms (Allen et al. 2008) and osmotic stress in red algae (Cao et al. 2020) and can signal changes in glycolytic activity. Here, we observed these shifts not as a result of environmental change but by selecting variation initially fixed by chance events in a constant environment. It is also plausible that general deleterious physiological changes are associated with this decline in organismal function commonly associated with passage through repeated bottlenecks, where deleterious mutations tend to be fixed.

All strains also exhibited increased expression of numerous nitrogen transporters involved in nitrogen acquisition but not the reductases and hydrolases involved in nitrogen assimilation. Increased expression of only various nitrate transporters was detected in strains TW1010, TW1050, and TW1587, while only elevated transcription of ammonium transporters was observed in CS1059 and TW2929. TP3367 highly expressed both nitrate and ammonium transporters. Furthermore, numerous glutamine fructose-6-phosphate transaminases had increased expression across all strains. This enzyme is responsible for the metabolic transfer of nitrogenous groups and is involved in glutamate and amino sugar metabolism. Collectively, increased transcription of core nitrogen metabolism genes, transporters of different nitrogen species, and significant differences in particulate organic nitrogen (Hinners et al. 2022) between full-selected and control populations suggests nitrogen metabolism to be a core pathway involved in diatom diversification, even in the absence of changes to nitrogen availability. The modularity of core nitrogen metabolism pathways (Smith et al. 2019) and redundancy (e.g. urea, aminos, nitrate, and ammonia) can aid in the sustained viability of a handful of cells exploring phenotypic space following chance events or during subsequent adaptation from slower-growing starting points. For example, the key enzymes involved in ammonium metabolism, glutamine synthetase and glutamate synthase (GS-GOGAT), are located in both mitochondria and chloroplast in diatoms (Smith et al. 2019). One testable hypothesis that follows from this work is whether high near-neutral diversity exists in nitrogen-related traits in natural diatom populations. In our experiments, shifts in nitrogen metabolism are not indicative of adaptation to any change in the availability of nitrogen from the environment but could instead be a fundamental adaptive strategy associated with fitness recovery. Additionally, this modularity and redundancy may enable nitrogen metabolism in diatoms to change and still produce viable cells, such that changes to transcripts associated with nitrogen metabolism are associated with movement in the t-scape. As such, fitness recovery may often involve reevolving nitrogen metabolism that allows faster growth from slower-growing but viable intermediates that emerge from chance events. One interesting possibility is thus that nitrogen metabolism is especially prone to diversification in the face of bottlenecks, which is consistent with the unique, flexible nitrogen metabolism observed in diatoms relative to other green lineages (Smith et al. 2019).

In terms of carbon, energy, and core metabolic pathways, numerous carbonic anhydrase (CA) and thioredoxin transcripts also exhibited elevated expression in TW1010, TW1050, TW1587, and TP3367 indicating potential shifts in equilibrium between intracellular CO_2 and HCO_3^-. CAs can catalyze the reversible interconversion of CO_2 and water into HCO_3^- and protons and play a central role in carbon acquisition (Clement et al. 2016), though the exact role is localization dependent (Hopkinson et al. 2016). As with nitrogen acquisition described above, the modularity and redundancy of carbon acquisition through numerous types of CAs may also enable certain carbon acquisition pathways to be compromised during population bottlenecks while still producing viable cells. Concurrent transcriptional changes to different CAs following bottlenecks suggest flexibility of CA-associated carbon acquisition is associated with the generation of transcriptional diversity, and that there is scope for divergence during fitness recovery alone. Increased transcription of cytosolic malate dehydrogenase was observed in TW1010, TW1050, CS1059, TW1587, and TP2929 and is central to both the tricarboxylic acid cycle and gluconeogenesis. Numerous transcripts of clathrin subunits also increased expression. Clathrin is a structural protein that helps deform membranes to facilitate invagination of molecules into vesicles (i.e. clathrin-mediated endocytosis). Although not highly expressed in other eukaryotes, it was found to be the sixth most abundant protein in the *T. pseudonana* proteome and plays central roles in nutrient acquisition, vesicle transport, and segregation of organelles (Nunn et al. 2009). Finally, elevated transcripts of fucoxanthin chlorophyll proteins (FCPs) and other light-harvesting photosystem genes were observed (File S2). FCPs make up the key molecular complex performing light harvesting in diatoms (Gelzinis et al. 2015) and may be fundamental to light-derived energy generation in response to significant demographic change. Changes in expression and trait values of these critical pathways have been primarily observed as a result of environmental change (e.g. Allen et al. 2008; Bertrand et al. 2012; Mock et al. 2017; Bender et al. 2018; Smith et al. 2019). Here, we observe collective shifts in expression across strains that were allowed to diversify in a constant environment, indicating that variation in these transcripts is readily generated during chance events and suggests that these pathways can rapidly diversify during subsequent fitness recovery. The fact that we observe these common shifts across globally distributed, disparate diatom strains in the order, Thalassiosirales, suggests these pathways may be key, conserved players in diatom transcriptomic diversification more broadly. Hence, these transcripts may underlie fundamental axes of variation along which natural selection can act.

Conclusion

Here, we investigate patterns of transcriptional variation in a constant environment in a model diatom genus by using an integrated approach that pairs generating divergent populations with examining underlying global transcription. Our data reveal that variation generated using population bottlenecks in a constant environment can allow populations to find new metabolic configurations following fitness recovery. This informs our understanding of the variation that natural selection may have to act upon following chance events that are relatively common in open ocean diatoms, such as extreme reduction in population size during migration, or between diatom blooms. It is worth noting that chance bottleneck events can be common for other microbial populations in environments such as the human body. For example, Badrane et al. (2023) observed the emergence of mixed populations from clonal but genetically diverse strains of the opportunistic yeast pathogen, *Candida glabrata* bloodstream infection, following bottleneck events. Particularly, they observed emergent, diverse phenotypes including antifungal resistance, which could help explain broader patterns of genetic variation and inform the usage of antimicrobials (Badrane et al. 2023).

Despite the time elapsed between the whole cell measurements (Hinners et al. 2022) and RNA extraction, there were several consistent changes in both whole-cell traits and transcripts across carbon, nitrogen, energy, and oxidative stress pathways indicating that these pathways can shift to produce viable variation during population bottlenecks, although these changes are sometimes also associated with lower population growth. In particular, the Hinners et al. (2022) study identified that in the populations later used for the transcriptomics study here, trait correlations departed most strongly from ancestral values for populations with high particulate carbon (POC) and nitrogen (PON) content. Overall, the strongest departure from ancestral trait correlations was related to large shifts in correlation due to changes in a single trait, often reactive oxygen. While the timing of sampling means that we cannot directly connect transcriptional changes and whole-cell trait values, the concordance between changes in traits and changes in annotated transcripts is striking.

During fitness recovery, some populations evolve transcriptional patterns that differ from control populations consistently propagated without population bottlenecks in the same environment, suggesting that multiple solutions for adaptive trait values and trait combinations exist. Even though many of these changes are also associated with environmental adaptation in other studies, chance events alone are sufficient here to provide starting points from which diverse rapid growth populations evolve. Since a common set of transcripts was involved in diversification across the genus, we suggest that these transcripts, and the relationships between them, are especially likely to be involved in diversification. Given that these transcriptional changes were captured using demographic fluctuations that are likely common in evolving diatom populations, there is also the intriguing possibility that this variation could also be acted on by natural selection in the event of environmental change following a population bottleneck. Plausible scenarios where this might occur are population subdivision and migration into a new environment or environmental change between diatom blooms. We suggest that these patterns of transcriptional change may identify metabolic pathways commonly used during trait diversification. Future studies using more distantly related ancestral diatom strains harboring different transcriptional circuitry would shed further light on the generality of the reduced axes, and the variation in transcript levels and relationships, identified here. It is worth noting that our experiment used nutrient-replete, benign conditions, which revealed the fundamental flexibility of certain metabolic pathways. Future studies could explore the conservation of these trends using more stressful or variable conditions, where viability selection may be more stringent and, at a certain point, may inhibit diversification.

We identified a particular set of pathways associated with rapid diatom diversification in a constant environment. Investigation of reduced axes in other phytoplankton taxa could tease out changes to general transcriptional patterns associated with the generation of diversity versus adaptation, as well as those changes that occur broadly across taxa versus those that are functional group, genotype, or environment specific. Overall, identifying general patterns and transcriptional relationships associated with diversification adds critical knowledge to genotypic and phenotypic limits of phytoplankton diversity and highlights the potential role of common chance events for evolving phytoplankton populations.

Materials and Methods

Diatom Cultures

Six strains of *Thalassiosira* sp. from the Provasoli-Guillard National Center of Marine Phytoplankton (NCMA, formerly known as the CCMP, https://ncma.bigelow.org/) culture collection were used: CCMP 1010, 1050, 1059, 1587, 2929, and 3367 (Table S1). Extensive trait and phenotypic characterization of these strains are described in Argyle et al. (2021a, 2021b) and Hinners et al. (2022). Cultures were grown in sterile f/2 media (Guillard 1975) made from natural seawater (collected in St Abbs, United Kingdom), at 20 °C and approximately 60 μmol photons/m^2/s (measured with a 4-pi sensor) at a 12-h:12-h light:dark cycle. Our rationale for using this light level is that mutation accumulation during bottlenecks can lead to changes in traits including light optima; at this light level, ancestral population growth is within the normally reported range for these species or strains, but high extinction rates in the RS phase of the experiment can be avoided. For the evolution experiment, cultures were maintained in transparent 48-well plates covered with Breathe-Easy breathable plate seals (Sigma-Aldrich). Larger additional cultures (1.5 L) were grown for RNA extraction so that sufficient biomass could be obtained at low cell densities to ensure that harvested cells were not light or nutrient limited.

Evolution Experiment

RS Phase

During the RS phase, bottlenecks were induced every 7 d by transferring ∼8 cells per replicate to new medium to fix variants mainly by chance between growth cycles. As growth rates decreased through time, we extended the bottleneck period to every 14 d toward the end of this phase resulting in an average of 18 generations. All populations were bottlenecked at the same time as long as replicates had reached a minimum cell concentration of 2,000 cell/mL. If cell concentrations were lower, cultures were instead diluted to 500 cell/mL to allow for population recovery before a new bottleneck. Bottlenecks were repeated 5 to 9 times depending on the genotype corresponding to a total RS phase length of 3 mo (70 to 200 generations), depending on population growth rates. Toward the end of this phase, some population growth rates decreased to a degree where growth was no longer observed. In these cases, previously saved transfers were used to induce a new bottleneck. Growth rates were monitored via in vivo fluorescence, and at the end of this phase, fluorescence-based population growth rates were reduced by an average of 45% compared to ancestral growth rates (Fig. S1).

FS Phase

During the FS phase, full-selected populations were propagated in batch culture with transfer sizes of 1,000 cells every 7 d in the ancestral environment. Populations were transferred ∼25 times corresponding to 200 to 500 generations. At this point, whole-cell traits were measured (Hinners et al. 2022). Populations were grown under the same FS conditions for a subsequent 10 mo, after which biomass was harvested for RNA extraction. Maximum growth rates were measured every 5 to 10 transfers to monitor fitness recovery, and before termination of this phase, growth rates were measured over 4 transfers (4 wk) to ensure population growth rates had stabilized, indicating populations were on or near a high-fitness peak (Fig. S2).

Growth Rate

Growth rates were measured through daily in vivo fluorescence with a Tecan Spark plate reader (excitation: 455 nm, emission: 620 nm) (Hinners et al. 2022). Exponential growth rates were calculated for each time step as follows:

$$\mu = \frac{\ln(x_2) - \ln(x_1)}{t_2 - t_1} \tag{1}$$

Measurements were carried out at the same time each day; some of the assays take more than 1 h. Methodology was tested and peer reviewed for Argyle et al. (2021a, 2021b). Maximum growth rates were determined over 4 consecutive time steps. During the bottleneck phase, growth rates were determined on single replicates per population. Final growth rates were determined from 3 replicates per population. Other traits, as well as bacterial counts, were measured via flow cytometry as described in Hinners et al. (2022). Bacterial counts showed no substantial changes over time. Cells for the transcriptomic analysis were harvested at approximately half the maximum cell density attainable and so were not light or nutrient limited during harvest.

Phylogenomics and Geographic Visualization

Phylogenomic trees were inferred with *IQTREE2* (Minh et al. 2020) using concatenated protein alignments constructed via *hmmsearch* for marker detection, *MUSCLE* (Edgar 2004) for marker protein alignment, and *ClipKIT* (Steenwyk et al. 2020) for alignment trimming. The concatenated alignments were based on the BUSCO *Protista_83.hmm* (Simão et al. 2015) marker set available through *Anvi'o* (Eren et al. 2015) using the suggested *E*-value noise cutoff of 1e−25. Phylogenomic trees were analyzed and visualized using ETE 3 (Huerta-Cepas et al. 2016) in Python. This was performed using the *phylogeny.py* module of *VEBA* (Espinoza and Dupont 2022).

Geographic coordinates in relation to strain origin were processed using *GeoPandas* (https://github.com/geopandas/geopandas), *GeoPy* (https://github.com/geopy/geopy), and *Matplotlib* (https://github.com/matplotlib/matplotlib).

Transcriptome Assembly, Gene Modeling, and Orthology

Sequence reads were quality controlled using *KneadData* (Beghini et al. 2021) with the GRCh38.p13 human genome as a reference for potential decontamination. This methodology yielded transcriptomes with depths between 3,433,160 and 26,023,146 reads mapping between 14,761 and 55,849 unique transcripts (refer to File S1 for richness and depth statistics per strain). De novo transcriptomes were grouped by strain (e.g. TW1010 ancestors and bottlenecks) and coassembled using *rnaSPAdes* (Bushmanova et al. 2019).

Following the protocol detailed in Santoro et al. (2021), we used *TransDecoder* (https://github.com/TransDecoder/TransDecoder) for gene modeling in a multistep process to minimize false positives. In particular, we used the following procedure: (i) *TransDecoder.LongOrfs*, with transcript-to-gene mappings assigned by *rnaSPAdes*, to generate putative open reading frames (ORFs); (ii) *hmmsearch* (Eddy 2011) to identify protein domains using the *PFAM* v33.1 and *TIGRFAM* v15.0 databases; (iii) *Diamond blastp* (Buchfink et al. 2021) against all Bacillariophyceae (diatoms) genomes available in NCBI (GCA_000149405.2, GCA_000150955.2, GCA_000296195.2, GCA_001750085.1, GCA_002217885.1, and GCA_900660405.1); and (iv) *TransDecoder.Predict* with the putative ORFs from (i), the protein domains from (ii), and the alignments from (iii) using the --single_best_only argument; this was implemented using the *transdecoder_wrapper.py* script of *VEBA*. Genes were annotated by best-hit *Diamond* blastp alignment to NCBI's nonredundant protein database (accessed on v2021.08.03) using the *annotate.py* module of *VEBA*.

Orthogroups were identified using *OrthoFinder* (Emms and Kelly 2019) with the high-quality proteins generated from our *TransDecoder* procedure and all of the Bacillariophyceae proteins listed previously. Consensus annotations for orthogroups were assigned by using the most common organism-agnostic annotation within the grouping using *UniFunc*, a natural language processing software developed for bioinformatics (Queirós et al. 2021).

Pathway Enrichment Analysis

We performed KEGG pathway enrichment analysis on each strain using the *GSEA's Prerank* rank module (Subramanian et al. 2005) via the *GSEApy* Python package (Fang et al. 2022). To prepare the data for pathway enrichment, we aggregated the counts for transcripts by their BRENDA enzyme representative (e.g. EC:1.1.1.1) and identified conserved enzymes that had at least 300 counts in each sample which were later used for pathway enrichment. The enzyme count matrix (i.e. sample vs. enzymes) was CLR transformed followed by Euclidean distance (i.e. Aitchison distance), and PCoA was performed. The PCoA loadings of the conserved enzymes were used as feature ranks (e.g. weights) in the *Prerank* module using min_size = 5 and permutation_num = 1,000 parameters. Enriched pathways were considered significant if false discovery rate < 0.25 which is recommended by the *GSEA* documentation.

Transcript Analysis

Taking a compositional approach, we used the CLR transformation on raw transcript counts by taking the log of each count and dividing by the geometric mean using the *compositional* Python package (Espinoza et al. 2020). Hierarchical clustering and PCoA ordinations were performed using *SciPy* (Virtanen et al. 2020) and *Soothsayer* (Espinoza et al. 2021) Python packages. PCoA analyses were conducted in the same manner for all PCoA plots where all axes represent collapsed variance of CLR-transformed transcripts. Heatmaps were generated using the *heatmap* function in the R *stats* package (https://www.R-project.org/).

Supplementary material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising from this submission.

Funding

This research was supported by the Gordon and Betty Moore Foundation Marine Microbes Initiative (MMI 7397 to S.C., N.M.L., and M.A.D.) and NSF OCE-1558453 and NSF OCE-2049299 to C.L.D.

Conflict of interest statement. None declared.

Data availability

Raw sequence reads have been deposited to the NCBI Short Read Archive with accession number PRJNA835115.

References

Allen AE, LaRoche J, Maheswari U, Lommer M, Schauer N, Lopez PJ, Finazzi G, Fernie AR, Bowler C. Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proc Natal Acad Sci U S A*. 2008:**105**(30):10438–10443. https://doi.org/10.1073/pnas.0711370105.

Alverson AJ, Beszteri B, Julius ML, Theriot EC. The model marine diatom *Thalassiosira pseudonana* likely descended from a freshwater ancestor in the genus *Cyclotella*. *BMC Evol Biol*. 2011:**11**(1):125. https://doi.org/10.1186/1471-2148-11-125.

Argyle PA, Hinners J, Walworth NG, Collins S, Levine NM, Doblin MA. A high-throughput assay for quantifying phenotypic traits of microalgae. *Front Microbiol*. 2021a:**12**:706235. https://doi.org/10.3389/fmicb.2021.706235.

Argyle PA, Walworth NG, Hinners J, Collins S, Levine NM, Doblin MA. Multivariate trait analysis reveals diatom plasticity constrained to a reduced set of biological axes. *ISME Commun*. 2021b:**1**(1):59. https://doi.org/10.1038/s43705-021-00062-8.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 2004:**306**(5693):79–86. https://doi.org/10.1126/science.1101156.

Badrane H, Cheng S, Dupont C, Hao B, Driscoll E, Morder K, Liu G, Newbrough A, Fleres G, Kaul D, *et al.* Genotypic diversity and unrecognized antifungal resistance among populations of *Candida glabrata* from positive blood cultures. *Nat Commun*. 2023:**14**:5918. https://doi.org/10.21203/rs.3.rs-2706400/v1.

Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 2021:**10**:e65088. https://doi.org/10.7554/eLife.65088.

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**(3):e01457. https://doi.org/10.1002/ecm.1457.

Bender SJ, Moran DM, McIlvin MR, Zheng H, McCrow JP, Badger J, DiTullio GR, Allen AE, Saito MA. Colony formation in *Phaeocystis antarctica* : connecting molecular mechanisms with iron biogeochemistry. *Biogeosciences* 2018:**15**(16):4923–4942. https://doi.org/10.5194/bg-15-4923-2018.

Bertrand EM, Allen AE, Dupont CL, Norden-Krichmar TM, Bai J, Valas R, Saito MA. Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. *Proc Natal Acad Sci U S A*. 2012:**109**(26):E1762–E1771. https://doi.org/10.1073/pnas.1201731109.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C, Maumus F, Otillar RP, *et al.* The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 2008:**456**(7219):239–244. https://doi.org/10.1038/nature07410.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 2021:**18**(4):366–368. https://doi.org/10.1038/s41592-021-01101-x.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* 2019:**8**(9):giz100. https://doi.org/10.1093/gigascience/giz100.

Cao Y, Ashline DJ, Ficko-Blean E, Klein AS. Trehalose and (iso)floridoside production under desiccation stress in red alga *Porphyra umbilicalis* and the genes involved in their synthesis. *J Phycol*. 2020:**56**(6):1468–1480. https://doi.org/10.1111/jpy.13057.

Clement R, Dimnet L, Maberly SC, Gontero B. The nature of the CO2-concentrating mechanisms in a marine diatom, *Thalassiosira pseudonana*. *New Phytol*. 2016:**209**(4):1417–1427. https://doi.org/10.1111/nph.13728.

Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011:**7**(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004:**5**(1):113. https://doi.org/10.1186/1471-2105-5-113.

Emms DM, Kelly S. Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019:**20**(1):238. https://doi.org/10.1186/s13059-019-1832-y.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvi'o: an advanced analysis and visualization platform for 'omics data. *Peerj* 2015:**3**:e1319. https://doi.org/10.7717/peerj.1319.

Espinoza JL, Dupont CL. VEBA: a modular end-to-end suite for in silico recovery, clustering, and analysis of prokaryotic, microeukaryotic, and viral genomes from metagenomes. *BMC Bioinformatics* 2022:**23**(1):419. https://doi.org/10.1186/s12859-022-04973-8.

Espinoza JL, Dupont CL, O'Rourke A, Beyhan S, Morales P, Spoering A, Meyer KJ, Chan AP, Choi Y, Nierman WC, *et al.* Predicting antimicrobial mechanism-of-action from transcriptomes: a generalizable explainable artificial intelligence approach. *PLoS Comput Biol*. 2021:**17**(3):e1008857. https://doi.org/10.1371/journal.pcbi.1008857.

Espinoza JL, Shah N, Singh S, Nelson KE, Dupont CL. Applications of weighted association networks applied to compositional data in biology. *Environ Microbiol*. 2020:**22**(8):3020–3038. https://doi.org/10.1111/1462-2920.15091.

Estes S, Lynch M. Rapid fitness recovery in mutationally degraded lines of *Caenorhabditis elegans*. *Evolution* 2003:**57**(5):1022–1030. https://doi.org/10.1111/j.0014-3820.2003.tb00313.x.

Fang Z, Liu X, Peltz G, McKay A, Fröhlich F, Reese F, Kimmel J, engelsdaniel, Xiaohui L. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics*. 2022:btac757. https://doi.org/10.1093/bioinformatics/btac757.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 1998:**281**(5374):237–240. https://doi.org/10.1126/science.281.5374.237.

Gelzinis A, Butkus V, Songaila E, Augulis R, Gall A, Büchel C, Robert B, Abramavicius D, Zigmantas D, Valkunas L. Mapping energy transfer channels in fucoxanthin–chlorophyll protein complex. *Biochim Biophys Acta*. 2015:**1847**(2):241–247. https://doi.org/10.1016/j.bbabio.2014.11.004.

Guillard RRL. Culture of phytoplankton for feeding marine invertebrates. In: Smith WL, Chanley MH, editors. *Culture of marine invertebrate animals*. New York (NY): Plenum Press; 1975. p. 26–60.

Halligan DL, Keightley PD. Spontaneous mutation accumulation studies in evolutionary genetics. *Annu Rev Ecol Evol Syst*. 2009:**40**(1):151–172. https://doi.org/10.1146/annurev.ecolsys.39.110707.173437.

Heilbron K, Toll-Riera M, Kojadinovic M, MacLean RC. Fitness is strongly influenced by rare mutations of large effect in a microbial mutation accumulation experiment. *Genetics* 2014:**197**(3):981–990. https://doi.org/10.1534/genetics.114.163147.

Hennon GMM, Ashworth J, Groussman RD, Berthiaume C, Morales RL, Baliga NS, Orellana MV, Armbrust EV. Diatom acclimation to elevated CO2, via cAMP signalling and coordinated gene expression. *Nat Clim Change*. 2015:**5**:761–765. https://doi.org/10.1038/nclimate2683.

Hinners J, Argyle PA, Walworth NG, Doblin MA, Levine NM, Collins S. 2022. Multitrait diversification in marine diatoms in constant and warmed environments. Biorxiv:2022.04.08.487611.

Hopkinson BM, Dupont CL, Matsuda Y. The physiology and genetics of CO2 concentrating mechanisms in model diatoms. *Curr Opin Plant Biol*. 2016:**31**:51–57. https://doi.org/10.1016/j.pbi.2016.03.013.

Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016:**33**(6):1635–1638. https://doi.org/10.1093/molbev/msw046.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*. 2014:**12**(6):e1001889. https://doi.org/10.1371/journal.pbio.1001889.

Kraemer SA, Böndel KB, Ness RW, Keightley PD, Colegrave N. Fitness change in relation to mutation number in spontaneous mutation accumulation lines of *Chlamydomonas reinhardtii*. *Evol Int J Org Evol*. 2017:**71**(12):2918–2929. https://doi.org/10.1111/evo.13360.

Lauritano C, Orefice I, Procaccini G, Romano G, Ianora A. Key genes as stress indicators in the ubiquitous diatom *Skeletonema marinoi*. *BMC Genomics* 2015:**16**(1):411. https://doi.org/10.1186/s12864-015-1574-5.

LeClair JS, Wahl LM. The impact of population bottlenecks on microbial adaptation. *J Stat Phys*. 2017:**172**(1):114–125. https://doi.org/10.1007/s10955-017-1924-6.

Malcom JW, Hernandez KM, Likos R, Wayne T, Leibold MA, Juenger TE. Extensive cross-environment fitness variation lies along few axes of genetic variation in the model alga *Chlamydomonas reinhardtii*. *New Phytol*. 2014:**205**(2):841–851. https://doi.org/10.1111/nph.13063.

Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, Wincker P, Iudicone D, de Vargas C, Bittner L, *et al.* Insights into global diatom distribution and diversity in the world's ocean. *Proc Natl Acad Sci U S A*. 2016:**113**(11):E1516–E1525. https://doi.org/10.1073/pnas.1509523113.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020:**37**(5):1530–1534. https://doi.org/10.1093/molbev/msaa015.

Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, Salamov A, Sanges R, Toseland A, Ward BJ, *et al.* Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* 2017:**541**(7638):536–540. https://doi.org/10.1038/nature20803.

Nunn BL, Aker JR, Shaffer SA, Tsai S, Strzepek RF, Boyd PW, Freeman TL, Brittnacher M, Malmström L, Goodlett D. Deciphering diatom biochemical pathways via whole-cell proteomics. *Aquat Microb Ecol*. 2009:**55**:241–253. https://doi.org/10.3354/ame01284.

Queirós P, Novikova P, Wilmes P, May P. Unification of functional annotation descriptions using text mining. *Biol Chem*. 2021:**402**(8):983–990. https://doi.org/10.1515/hsz-2021-0125.

Ruggiero MV, D'Alelio D, Ferrante MI, Santoro M, Vitale L, Procaccini G, Montresor M. Clonal expansion behind a marine diatom bloom. *ISME J*. 2017:**12**(2):463–472. https://doi.org/10.1038/ismej.2017.181.

Salverda MLM, Koomen J, Koopmanschap B, Zwart MP, de Visser JAGM. Adaptive benefits from small mutation supplies in an antibiotic resistance enzyme. *Proc National Acad Sci*. 2017:**114**(48):12773–12778. https://doi.org/10.1073/pnas.1712999114.

Santoro EP, Borges RM, Espinoza JL, Freire M, Messias CSMA, Villela HDM, Pereira LM, Vilela CLS, Rosado JG, Cardoso PM, *et al.* Coral microbiome manipulation elicits metabolic and genetic restructuring to mitigate heat stress and evade mortality. *Sci Adv*. 2021:**7**(33):eabg3088. https://doi.org/10.1126/sciadv.abg3088.

Schoustra SE, Bataillon T, Gifford DR, Kassen R. The properties of adaptive walks in evolving populations of fungus. *Plos Biol*. 2009:**7**(11):e1000250. https://doi.org/10.1371/journal.pbio.1000250.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015:**31**(19):3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Smith SR, Dupont CL, McCarthy JK, Broddrick JT, Oborník M, Horák A, Füssy Z, Cihlář J, Kleessen S, Zheng H, *et al.* Evolution and regulation of nitrogen flux through compartmentalized metabolic networks in a marine diatom. *Nat Commun*. 2019:**10**:4552. https://doi.org/10.1038/s41467-018-07882-8.

Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol*. 2020:**18**(12):e3001007. https://doi.org/10.1371/journal.pbio.3001007.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc National Acad Sci U S A*. 2005:**102**(43):15545–15550. https://doi.org/10.1073/pnas.0506580102.

Van de Waal DB, Litchman E. Multiple global change stressor effects on phytoplankton nutrient acquisition in a future ocean. *Philos Trans R Soc Lond B Biol Sci*. 2020:**375**(1798):20190706–20190708. https://doi.org/10.1098/rstb.2019.0706.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, *et al.* Scipy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020:**17**(3):261–272. https://doi.org/10.1038/s41592-019-0686-2.

Walworth NG, Hinners J, Argyle PA, Leles SG, Doblin MA, Collins S, Levine NM. The evolution of trait correlations constrains phenotypic adaptation to high CO2 in a eukaryotic alga. *Proc Biol Sci*. 2021:**288**(1953):20210940. https://doi.org/10.1098/rspb.2021.0940.

Wang H, Chen F, Mi T, Liu Q, Yu Z, Zhen Y. Responses of marine diatom *Skeletonema marinoi* to nutrient deficiency: programmed cell death. *Appl Environ Microb*. 2020:**86**(3):e02460-19. https://doi.org/10.1128/AEM.02460-19.

Ward BA, Cael BB, Collins S, Young CR. Selective constraints on global plankton dispersal. *P Natl Acad Sci U S A*. 2021:**118**(10):e2007388118. https://doi.org/10.1073/pnas.2007388118.

Whittaker KA, Rynearson TA. Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. *Proc Natal Acad Sci U S A*. 2017:**114**(10):2651–2656. https://doi.org/10.1073/pnas.1612346114.

Windels EM, Fox R, Yerramsetty K, Krouse K, Wenseleers T, Swinnen J, Matthay P, Verstraete L, Wilmaerts D, den Bergh BV, *et al.* Population bottlenecks strongly affect the evolutionary dynamics of antibiotic persistence. *Mol Biol Evol*. 2021:**38**(8):3345–3357. https://doi.org/10.1093/molbev/msab107.