Defining ecological and physiological traits of phytoplankton with metatranscriptomics

by

Harriet Alexander

B.S., Wellesley College (2010)

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

February 2015

Author .....
Joint Program in Applied Ocean Science & Engineering
Massachusetts Institute of Technology
& Woods Hole Oceanographic Institution
December 17, 2015

Certified by .....
Sonya T. Dyhrman
Associate Professor of Earth & Environmental Science
Columbia University
Thesis Supervisor

Certified by .....
Elizabeth B. Kujawinski
Associate Scientist with Tenure
Woods Hole Oceanographic Institution
Thesis Supervisor

Accepted by .....
Lauren Mulineaux
Chair, Joint Committee for Biological Oceanography
Massachusetts Institute of Technology
Woods Hole Oceanographic Institution

Defining ecological and physiological traits of phytoplankton with metatranscriptomics

by

Harriet Alexander

Submitted to the MIT-WHOI Joint Program in Oceanography and Applied Ocean Science and Engineering on December 17, 2015, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biological Oceanography

Abstract

Marine phytoplankton are central players in the global carbon cycle, responsible for nearly half of global primary production. The identification of the factors controlling phytoplankton ecology, physiology, and, ultimately, bloom dynamics has been a central problem in the field of biological oceanography for the past century. Molecular approaches enable the direct examination of species-specific metabolic profiles in mixed, natural communities, a task which was previously intractable. In this thesis, I developed and applied novel analytical tools and bioinformatic pipelines to characterize the physiological response of phytoplankton at various levels of taxonomic grouping (strain, species, and functional group) to their environment. An *in silico* Bayesian statistical approach was designed to identify stable reference genes from high-throughput sequence data for use in RT-qPCR assays or metatranscriptome studies. Using this tool and a metatranscriptomic approach, the role of resource partitioning in the coexistence of two closely related diatom species in the same estuarine system was examined. This study demonstrated that co-occurring diatoms in a dynamic coastal marine system have apparent differences in their capacity to use nitrogen and phosphorus, and that these differences may facilitate the diversity of the phytoplankton. The second field study focused on the diatom, haptophyte, and dinoflagellate functional groups, using simulated blooms to characterize the traits that govern the magnitude and timing of phytoplankton blooms in the oligotrophic ocean. The results indicated that blooms form when phytoplankton are released from limitation by resources and that the mechanistic basis for the success of one functional group over another may be driven by how efficiently the transcriptome is modulated following a nutrient pulse. The final study looked at the sub-species level, examining the balance of phenotypic plasticity and strain diversity in the success of the coccolithophore *Emiliania huxleyi*. Results indicated strong control of nitrogen on the species complex and showed that nutrient resupply shifted the strain composition as well as transcript abundance of key biogeochemical genes involved in nutrient acquisition and the life stage of the population. Together, these studies demonstrate the breadth of information that can be garnered through the integration of molecular approaches with traditional biological oceanographic surveys, with each illuminating fundamental questions around phytoplankton ecology and bloom formation.

Thesis Supervisor: Sonya T. Dyhrman
Title: Associate Professor of Earth & Environmental Science
Columbia University

Thesis Supervisor: Elizabeth B. Kujawinski
Title: Associate Scientist with Tenure
Woods Hole Oceanographic Institution

For my grandparents, Betty and Sherwood

Acknowledgments

This thesis, like so many of my adventures to date, would not have come to fruition without the support, help, and care of the many incredible people in my life.

First and foremost, I must thank my advisor, Sonya Dyhrman. Over the last five years, Sonya has nurtured my scientific curiosity, while simultaneously challenging me to delve deeper and work harder than I knew I could. She also went well beyond many advisors waking before dawn to help with harvests and traveling to the *literal* ends of the earth to make science happen. It has been a great privilege to work with such a talented scientist, whose integrity and devotion to science and her colleagues I hope to emulate in my career.

I am also grateful to my thesis committee, Penny Chisholm, Scott Doney, Mick Follows, and Bethany Jenkins, for their guidance and thoughtful input during this process. I especially wish to thank my co-advisor, Elizabeth Kujawinski, who generously folded me into her lab group following Sonya's relocation. Additionally, I would like to thank my co-authors on some of these chapters: Tatiana Rynearson, Mak Saito, Dave Karl, and Sam Wilson.

I am happy to have overlapped with a group of wonderful and supportive people during my time in the Dyhrman lab group. Sheean Haley, it goes without saying that you are the engine that keeps the lab running seamlessly. Your levity, indefatigably, and upper body strength, both on land and at sea, kept me grounded during my thesis and made all my science possible. To my fellow Dyhrman Lab graduate students and postdocs, Louie, Abby, Monica, Solange, Colleen, and Kyle, thank you for your camaraderie, help, and general awesomeness.

The Woods Hole community has ***I also wish to thank all my WHOI and non-WHOI friends for helping me find and maintain balance in my life. Thanks especially to Isabela Le Bras for all your support over the last few years.

Finally, I would like to thank my mother and family for their tireless love and support not only during my Ph.D., but throughout my life. I wish to dedicate this thesis to the memory of my grandparents Betty and Sherwood Finley, to whom I owe my work ethic, love of nature, and creativity.

Throughout graduate school I have been supported by the MIT Presidential Fellowship, the National Defense Science and Engineering Graduate (NDSEG) Fellowship, the Ocean Life Institute Fellowship, and the WHOI Academic Programs Office.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The upper, sunlight region of the pelagic ocean, or "euphotic" zone, is home to microscopic plants, phytoplankton, which in this well-lit environment are able to thrive and photosynthesize. Though individually quite small, the combined net primary production (NPP) of these organisms in the marine system is estimated to be 48.5 Pg of carbon per year, nearly 50% of global NPP (Field et al., 1998). Consequently, the identification of the major factors controlling phytoplankton ecology, physiology, and, ultimately, bloom dynamics has been a central problem in the field of biological oceanography for the past century. From physical explanations (Sverdrup's critical depth hypothesis (Sverdrup, 1953)), to chemical rationale (Redfield ratio(Redfield, 1958)), to ecological theory (Margalef's mandala (Margalef, 1978)), the field has been constantly reevaluating evidence to answer the question: What drives phytoplankton blooms? Despite the advancement seen in the field, major knowledge gaps remain as to what factors are most influential to the dynamics of the complex and diverse phytoplankton community. (Iglesias-Rodriguez et al., 2008)

In this chapter, I present a brief overview to some of the overarching ecological and biogeochemical questions surrounding the phytoplankton as well as some elegant work which has been done to address these questions through both culture and field studies. I then will present some Molecular approaches enable the direct examination of species-specific metabolic profiles in mixed, natural communities, a task which was previously intractable. In this thesis, I developed and applied novel analytical tools and bioinformatic pipelines to characterize the physiological response of phytoplank- ton at various levels of taxonomic grouping (strain, species, and functional group) to their environment.

1.1 Phytoplankton are diverse and abundant

While light sets the vertical limits to phytoplankton distributions, nutrients, both macro- (e.g. nitrogen (N), phosphorus (P))

1.2 I hate my thesis

A diverse community of marine microorganisms are at the heart of the global carbon cycle, creating and transforming dissolved organic matter, Diatom acclimation to elevated CO_2 via cAMP signalling and coordinated gene expression

Photosynthesis by marine phytoplankton (small floating photosynthetic organisms) represents nearly 50% of the total primary production on earth (1). Climate change is predicted to alter the biogeochemistry of the world's oceans (2) as well as the distribution and nature of primary production in the global surface ocean (3). Despite being of central importance, fundamental uncertainties remain regarding the basic ecology and biology of phytoplankton, such as: 1) How do phytoplankton respond to changing (light, nutrient, temperature)

environments? 2) What enables the maintenance of diverse phytoplankton communities? 3) How does community structure influence ecosystem function and biogeochemical cycling?

1.3 Molecular tools to shed light on mixed communities

As with the field of medicine, advances in sequencing and mass spectrometry over the last decade have accelerated the field of biological oceanography. The burgeoning technologies fueling the "-omics"[1] revolution allow a unique glimpse at the previously hidden molecular world of the microbes, enabling the tracking of species-specific metabolism in the environment. Capitalizing upon this revolution to address biogeochemical or ecological questions, however, is becoming increasingly difficult, as our ability to make measurements has surpassed our ability to analyze, visualize, and compare the data produced. My background and training in oceanography, microbiology, bioinformatics, and art make me uniquely poised to work on these informatics problems. During my PhD, I developed pipelines (4, 5) and bioinformatic tools (6) for the analysis of meta -transcriptomic datasets to interrogate the ecology and biogeochemical function of microbes in marine systems. With the University of California Postdoctoral Presidential Fellowship, I would be able to expand upon my existing computation skill set by integrating with the Lab for Data Intensive Biology headed by C. Titus Brown at UC Davis. Dr. BrownâĂŹs lab works to develop tools and approaches to extract meaningful information from large, unwieldy datasets. My research goal is to integrate multiple environmental and âĂIJmeta-omicâĂİ datasets to examine the species-specific physical and chemical drivers of phytoplankton bloom formation.

[1] "-omics" is a catch-all suffix typically used to describe large, molecular datasets (e.g. genomics, the study of the genome, and transcriptomics, the study of the complete set of RNA transcripts produced under certain conditions). The use of this suffix, however, is expanding to other fields (see #badomics).

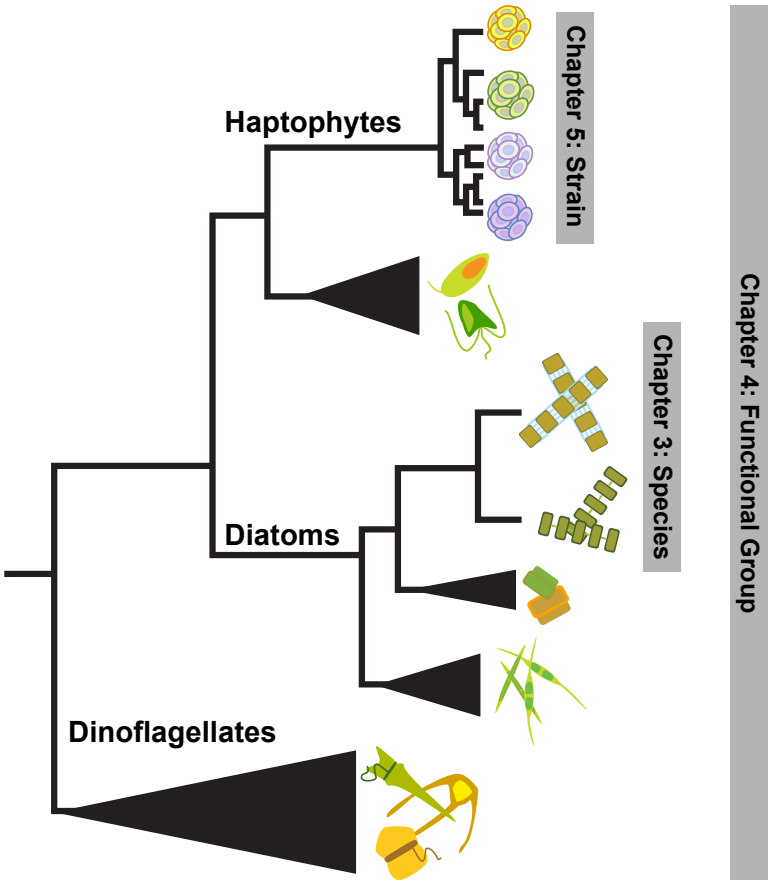


Figure 1.1: Conceptual overview of the levels of diversity explored in chapters 3, 4, and 5 of this thesis.

Chapter 2

Identifying reference genes with stable expression from high throughput sequence data

This chapter was originally published as Alexander, H., Jenkins, B.D., Rynearson, T.A., Saito, M.A., Mercier, M.L., and Dyhrman, S.T. (2012). Identifying reference genes with stable expression from high throughput sequence data. *Front. Microbiol.* 3, 385.

H.A., B.D.J., T.A.R., M.L.M., M.A.S., and S.T.D. performed research; H.A. and S.T.D. analyzed data; H.A. and S.T.D. wrote the paper; and B.D.J., T.A.R., M.L.M., M.A.S contributed to the writing of the paper.

2.1 Abstract

Genes that are constitutively expressed across multiple environmental stimuli are crucial to quantifying differentially expressed genes, particularly when employing quantitative reverse transcriptase polymerase chain reaction (RT-qPCR) assays. However, the identification of these potential reference genes in non-model organisms is challenging and is often guided by expression patterns in distantly related organisms. Here, transcriptome datasets from the diatom *Thalassiosira pseudonana* grown under replete, phosphorus-limited, iron-limited, and phosphorus and iron co-limited nutrient regimes were analyzed through literature-based searches for homologous reference genes, k-means clustering, and Analysis of Sequence Counts (ASC) to identify putative reference genes. A total of 9759 genes were identified and screened for stable expression. Literature-based searches surveyed 18 generally accepted reference genes, revealing 101 homologs in *T. pseudonana* with variable expression and a wide range of mean tags per million. K-means analysis parsed the whole transcriptome into 15 clusters. The two most stable clusters contained 709 genes but still had distinct patterns in expression. ASC analyses identified 179 genes that were stably expressed (posterior probability, post-$p < 0.1$, for 1.25 fold change). Genes known to have a stable expression pattern across the test treatments, like actin, were identified in this pool of 179 candidate genes. ASC can be employed on data without biological replicates and was more robust than the k-means approach in isolating genes with stable expression. The intersection of the genes identified through ASC with commonly used reference genes from the literature suggests that actin and ubiquitin ligase may be useful reference genes for *T. pseudonana* and potentially other diatoms. With the wealth of transcriptome sequence data becoming available, ASC can be easily applied to transcriptome datasets from other phytoplankton to identify reference genes.

2.2 Introduction

Quantitative reverse transcriptase polymerase chain reaction (RT-qPCR) facilitates rapid, accurate, high-throughput analyses of gene expression, greatly enhancing and expanding molecular biological studies in a variety of organisms. This method has moved beyond the realm of model organisms *Adib2004* (Antonov et al., 2005; Caldwell et al., 2005; Marionneau et al., 2005; Flatt et al., 2008) to be employed for the examination of ecological and physiological characteristics of marine microbes in both culture and the environment (Zehr and Turner, 2001; Nicot et al., 2005; Maldonado et al., 2006; Mock et al., 2008; Zhao et al., 2009; Whitney et al., 2011; Wurch et al., 2011a; Allen et al., 2008; Kustka et al., 2007; Lin et al., 2009). There are two primary methods of gene expression analysis for single genes: 1) absolute quantification, whereby the copy number of a gene is determined through comparison of the PCR signal to a standard curve, and 2) relative gene expression, in which the expression of the gene of interest is determined through comparison to a reference gene (or internal control gene), often employing the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001; Pfaffl, 2001; Schmittgen and Livak, 2008).

Inherent in the $2^{-\Delta\Delta CT}$ method is the selection of a reference, or "housekeeping," gene to act as an endogenous control. Ideally, the expression levels of the selected reference gene should remain stable across the treatments being examined. Genes like GAPDH, actin, and rRNA are often targeted as possible reference genes and tested for consistency in expression across treatments (Vandesompele et al., 2002; Pfaffl et al., 2004; Radonic et al.,

2004). However, both Czechowski et al. (2005) and de Jonge et al. (2007) demonstrated that canonical reference genes were often widely differentially regulated. In fact, de Jonge et al. (2007) noted that commonly used reference genes were not represented in the fifty most stably expressed genes in the human genome. Results from RT-qPCR studies using improper reference genes (e.g. genes that are not constitutively expressed) can be significantly different from results obtained with a proper reference gene (Dheda et al., 2005; Lanoix et al., 2012). Considering that previously established reference genes were not among the mostly stably expressed genes in model organisms, basing the selection of candidate genes for non-model organisms solely upon known reference genes may not prove the best method (de Jonge et al., 2007; Czechowski et al., 2005).

Application of RT-qPCR has proven particularly fruitful in the study of marine phytoplankton, illuminating transcriptional responses to physical stressors (Rosic et al., 2010b,b), nutrient limitation (Davis et al., 2006; Moseley et al., 2006; Davis and Palenik, 2008; Stuart et al., 2009; Whitney et al., 2011; Wurch et al., 2011a; Bender et al., 2012; Berg et al., 2008), and the diel cycle (Whitney et al., 2011; Bender et al., 2012), as well as highlighting the modulation and activity of many metabolic pathways (Moseley et al., 2006; McGinn and Morel, 2008a; Mock et al., 2008; Bender et al., 2012). The success of these studies hinged upon the selection of a stably expressed reference gene. While there is often extensive literature characterizing the dynamics of suites of genes expressed under different conditions in studies of model organisms, similar breadth is lacking for non-model organisms, such as marine phytoplankton. With few genome sequences available, the selection of reference genes for eukaryotic phytoplankton is a challenge, and researchers must often choose candidate genes (e.g. actin (Nicot et al., 2005), GAPDH (Czechowski et al., 2005)) based on the literature from model organisms that are distantly related to the study organism. Selecting and validating potential reference genes is a difficult task that consequently slows the development and application of targeted gene expression studies for phytoplankton.

Screening the wealth of sequence data produced by modern ultra high-throughput sequencing technologies may advance and broaden the search for candidate reference genes in non-model organisms. this is particularly true of transcriptome datasets whereby genes with stable expression can be identified between treatment conditions. two statistical techniques, k-means clustering (Hartigan and Wong, 1979) and analysis of sequence counts (ASC) (Wu et al., 2010), usually used to investigate patterns of differential expression in transcriptome datasets, show promise in this regard. The k-means algorithm is a partition-based, non-hierarchical clustering method, which divides sequence tags into the specified k-number of clusters, while minimizing the intra-cluster spread based on the specified distance metric (Hartigan and Wong, 1979; Tavazoie et al., 1999; Gerstein and Jansen, 2000; Quackenbush, 2001; D'haeseleer, 2005). ASC is a novel empirical Bayes method (estimating the prior distribution from the data, itself) to detect differential gene expression generated from quantifiable gene expression counts (as generated by Illumina Digital Gene Expression tag profiling, RNA-seq or similar high-throughput sequencing technologies) (Wu et al., 2010). When applied to transcriptome data these tools cannot only be used to identify genes with differential expression, they can be used to identify genes with highly stable expression patterns.

Here, literature-based searches, k-means clustering, and ASC are compared as tools for reference gene selection using a transcript sequence dataset collected from the centric diatom *Thalassiosira pseudonana*, grown under nutrient replete, phosphorus-limited (P-limited), iron-limited (Fe-limited), and phosphorus and iron co-limited (co-limited) treatments.

2.3 Materials and Methods

2.3.1 Culturing and transcriptome data collection

Axenic *T. pseudonana* CCMP1335 was grown at 14°C under 24 hour light (120 μmol photons $m^{-2}s^{-1}$) after Dyhrman et al. (2012) in f/2 plus silica chelated media made from surface Sargasso Sea water. Nitrate, silica, vitamins, and trace metals were at f/2 concentrations (Guillard and Ryther 1962), while iron and phosphate were modified across treatments. In brief, triplicate cultures of replete (36 μM PO_4, 400 nM Fe), P-limited (0.4 μM PO_4, 400 nM Fe), Fe-limited (36 μM PO_4, 40 nM Fe), and co-limited (0.4 μM PO_4, 40 nM Fe) treatments were harvested when growth deviated from the replete control. Growth was monitored by cell counts. Biomass was harvested onto 0.2 μm filters and flash frozen in liquid nitrogen and total RNA was extracted as described in (Dyhrman et al., 2012). Tag-seq sequencing of the transcriptome was performed by Illumina with a polyA selection and NlaIII digestion, resulting in 21 base pair sequence reads or tags (Dyhrman et al., 2012). Libraries were of varied sizes as follows: replete (∼12 million), P-limited (∼13 million), Fe-limited (∼23 million), and co-limited (∼26 million). Tags were mapped to gene models (predicted protein coding regions) with a pipeline designed by Genesifter Inc., requiring 100% identity and covering 9759 genes. Tag counts within a gene were pooled and normalized to the size of the library, with resulting data expressed in tags per million (TPM). Genes with normalized tag counts less than 2.5 TPM for each of the four treatments were excluded (Figure A.1), leaving 7380 genes in the analysis. The data discussed in this publication have been deposited in NCBI Gene Expression Omnibus (GEO) (Edgar, 2002) and are accessible through GEO Series accession number GSE40509.

2.3.2 Reference gene identification

The current, relevant literature from algae and plant-based studies was queried for reference genes used as endogenous controls for relative gene expression assays. Stably expressed genes reported in the literature were compared using BLASTn (Altschul et al., 1997) against the *T. pseudonana* genome in NCBI (AAFD00000000.2) to find homologs (e-value $< 1.0e-1$). A loose e-value cutoff was used to be inclusive and enhance our collection of all potential reference gene candidates. In addition, the Eukaryotic Orthologous Group (KOG) definitions for the genes found via BLAST were identified, and subsequent genes located in the KOG definition families were included in the analysis.

For the k-means analysis, tag counts from the four treatments corresponding to the 7380 genes with reads greater than 2.5 TPM were clustered using the k-means algorithm under the Pearson correlation coefficient. The distance was measured with a Pearson correlation as it has been found to perform as well or better than other similar distance measures for non-ratio or count-based data (Gibbons and Roth, 2002), such as the *T. pseudonana* transcriptome dataset. The number of clusters (k) was determined via a figure of merit (FOM) estimation, which is an approximation of the predictive power of the clustering method (Yeung et al., 2001). FOM analysis was performed by predicting the FOM value for values of k ranging from $k = 1$ (one cluster) to $k = 50$ (fifty clusters). The FOM value decreases as the within-cluster similarity increases, thus the FOM value was minimized to determine the optimal k-value. All clustering analyses were performed using the MultiExperiment Viewer (MeV) version 4.7 (Saeed et al., 2003, 2006). Possible reference gene targets were identified by isolating clusters of genes that exhibited similarly stable expression patterns across the four

treatments.

Using ASC (Wu et al., 2010), the statistical significance of an observed fold change was determined in pairwise comparisons between each of the limited treatments and the replete control. The posterior probability (post-p) was calculated by computing the posterior mean of the log ratio of proportions over each of the P-limited, Fe-limited, and co-limited treatments relative to the replete treatment for a fold change of 1.10, 1.25, and 1.50. Possible constitutively expressed genes were identified by selecting genes for which the post-p of each of the nutrient-limited treatments relative to the replete treatment for each of the fold change values was less than a specified cutoff. Posterior probability cutoffs between 0.01 and 0.20 were assessed across each of the fold changes (Table 2.1). Ultimately, a post-p of 0.10 was selected for further analyses (meaning that genes selected had less than a 10% chance of having the specified fold change between treatments), for it yielded genes across all of the fold change bins examined and demonstrated a broader range of mean normalized tag counts than seen for a post-p of 0.05 or 0.01. All ASC analyses were made using ASC 0.1.5 in R.

2.4 Results

Transcript sequence data was generated from *T. pseudonana* CCMP1335, grown in four different treatments (replete, P-limited, Fe-limited, and co-limited). Potential reference genes were identified through 1) querying the data to identify expression of common reference genes based on literature searches, 2) a pattern-driven analysis using k-means clustering (Hartigan and Wong, 1979) and 3) a quantitative analysis based the probability of fold change using ASC.

Selection of reference genes often falls upon those used in previous relative expression studies. The literature was surveyed for RT-qPCR expression studies employing the $2^{-\Delta\Delta CT}$ method for the following algae and plants: *T. pseudonana* (Maldonado et al., 2006; McGinn and Morel, 2008a,b; Mock et al., 2008; Park et al., 2008; Carvalho and Lettieri, 2011; Whitney et al., 2011), *Thalassiosira weissflogii* (Davis et al., 2006; McGinn and Morel, 2008a; Park et al., 2008; Whitney et al., 2011), *Phaeodactylum tricornutum* (Siaut et al., 2007; McGinn and Morel, 2008a), *Emiliana huxleyi* (Bruhn et al., 2010; Richier et al., 2011), *Micromonas pusilla* (McDonald et al., 2010), *Chlamydomonas reinhardtii* (Moseley et al., 2006; Zhao et al., 2009), *Alexandrium* spp. (Lee et al., 2009; Moustafa et al., 2010), *Symbiodinium* sp. (Rosic et al., 2010b,a; Leggat et al., 2011), *Prorocentrum minimum* (Guo and Ki, 2012), *Aureococcus anophagefferens* (Berg et al., 2008; Wurch et al., 2011a), *Solanum tuberosum* (Nicot et al., 2005), and *Arabidopsis thaliana* (Avonce et al., 2004). Results from the current literature survey yielded a list of 18 key reference genes frequently employed in the study of gene expression for eukaryotic phytoplankton and plants: actin, calmodulin, cyclin dependent kinase, cyclophilin, cytochrome c, G-protein beta subunit, ferric enterobactin binding periplasmic protein precursor, histones, elongation factors, GAPDH, heat shock protein 90, poly(A) polymerase, ribosomal protein large subunit, ribosomal protein small subunit, SAM, α-, β-, γ-tubulin, and ubiquitin conjugating enzymes (Data Sheet 2-1). It is important to note that as more reference genes are validated as stable, the selection of putative reference genes may expand. The 101 genes identified as homologous to these reference genes across the four treatments in *T. pseudonana* had variable expression patterns and a wide range of mean normalized counts (0.08 to 1087.8 TPM) (Figure 2.1). Genes within a specific gene family (e.g. the five actin genes) had different mean counts as well as variable coefficients of variation (CV), which is indicative of variable expression (Data Sheet 2-1). For example,

ACT 1 (NCBI: 7449411) had a mean expression of 1024.1 TPM and a CV of only 12.3%, where as ACT 5 (NCBI: 7445819) had a lower mean expression of 23.95 and a higher CV of 35.5% (Data Sheet 2-1).

The high-throughput transcript dataset was analyzed with k-means clustering. Prior to performing k-means cluster analysis, FOM optimization was run and found to be minimized at $k = 15$. Thus, k-means analysis was run under the Pearson correlation coefficient for $k = 15$, yielding 15 clusters, for which the intra-cluster variation was minimized (Figure **??**). Of the 15 clusters produced (ranging in size from 162 to 954 genes), Cluster 4 (433 genes), Cluster 9 (243 genes), and Cluster 14 (466 genes) had candidate reference genes based on a low magnitude of change associated with the expression patterns in those clusters (Figure **??**). However, Cluster 4 showed a clear pattern of differential regulation (downregulated in the replete and upregulated in the co-limited), and as such it was not considered to be an optimal candidate cluster and was excluded from additional analyses. Both Cluster 9 and Cluster 14 consisted of genes with a wide range in mean TPM values (1.74 to 4191.91 TPM), with relatively small deviations from the mean value (Figure 2.1; Data Sheet 2-2), which stands in contrast to other clusters that had definite treatment driven expression patterns (Figure **??**). Despite the relatively small deviations from the mean value, genes in Clusters 9 and 14 displayed both clear patterns of regulation, as demonstrated by the average change in tag count relative to the mean (Figure 2.2) and the presence of "outlier" genes with differential expression such as NCBI: 7451632, which was downregulated in the co-limited treatment for Cluster 14 (Figure 2.1; Data Sheet 2-2).

Adapting ASC to examine stable expression patterns, genes for which the post-p was less than 0.1 (e.g. had less than a 10% chance of equaling or exceeding the fold change cutoff) were plotted in three low fold change bins: 1.10, 1.25, and 1.50. A post-p of 0.1 was selected as it optimized the dataset for a wide range of mean gene expression values and provided coverage for each of the fold change bins examined (Table 2.1). The number of genes in each of the fold change bins increased with increasing value of fold change. For example, two genes passed the 1.10 cutoff, 179 genes passed the 1.25 cutoff, and 1375 genes passed the 1.50 cutoff. With the increase in the number of genes came an increase in the variation from the mean of the normalized tag counts (Figure 2.1; Data Sheet 2-3).

The bin with the 1.10 cutoff had two genes (NCBI: 7446346 and 7452192), which are both hypothetical proteins (Figure 2.1). A BLASTn search of 7446346 against the nr NCBI database yielded 69% identity over 251 base pairs (e-value, $1e-13$) to a hypothetical protein (NCBI: CP000544.1) from *Halorhodospira halophila*, a salt-tolerant purple bacterium, and 69% identity over 232 base pairs (e-value, $1e-12$) to a hypothetical protein (NCBI: CP001905.1) from *Thioalkalivibrio* sp. K90mix, also a salt-tolerant chemolithoautotrophic bacteria. BLASTp searches of 7452192 showed the highest identity hits to hypothetical proteins from *Aureococcus anophagefferens* (NCBI: EGB11506.1; 31% identity; e-value, $2e-21$) and from *Chlorella variablis* (NCBI: EFN56803.1; 24% identity; e-value, $7e-11$).

The 1.25 fold change bin was used for the identification of candidate reference genes as it offered a larger selection than the 1.10 fold change bin without including genes with increased deviations from the mean, as was the case with the 1.50 fold change bin. Thus, the 1.25 fold change category was the focus of the rest of the analyses (Data Sheet 2-3). Genes in the 1.25 fold change bin showed a broad range of mean normalized tag counts ranging from 7 to over 1200 TPM with a median of 41.94 TPM, providing for the selection of genes with different levels of constitutive expression in the cell (Figure 2.1). Notably, the median of the average tag counts of the genes in the ASC 1.25 fold change bin was 41.94 TPM, which is much higher than that of both Cluster 9 and Cluster 14 with median values

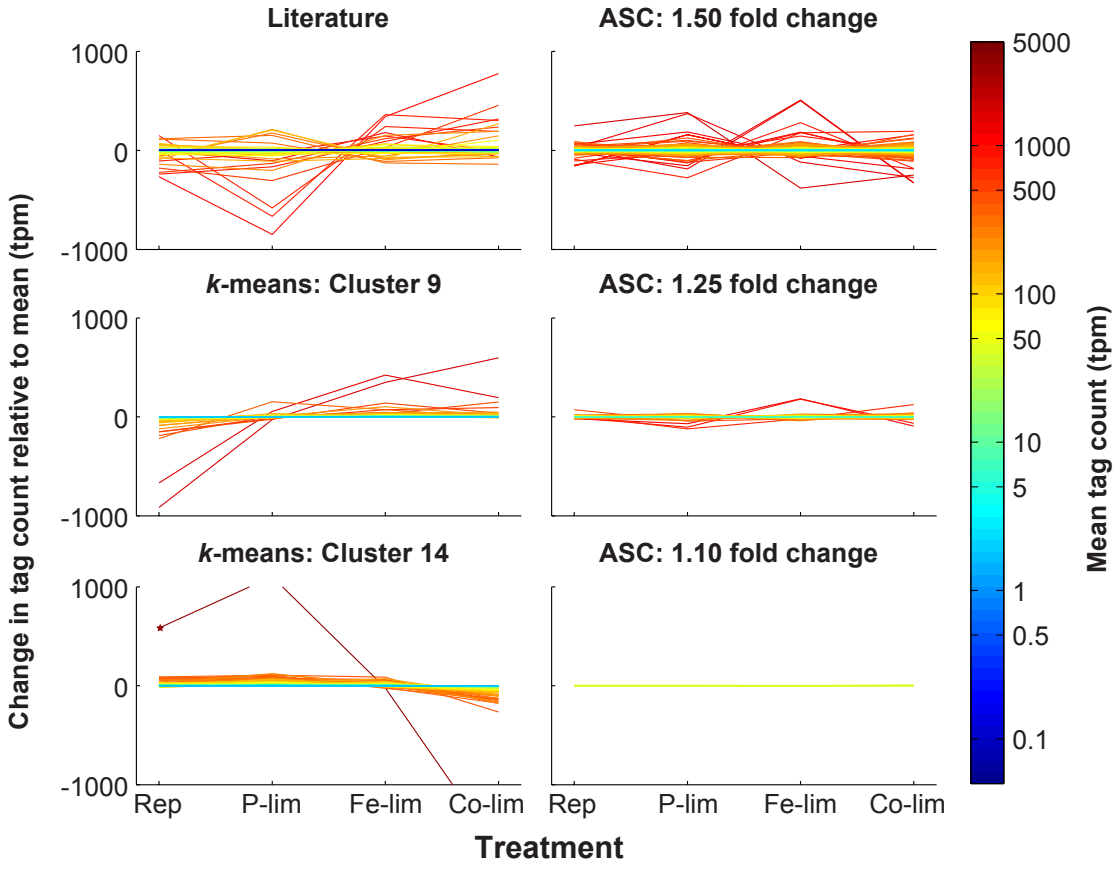


Figure 2.1: Expression patterns of putative reference genes identified through literature-based searches, *k*-means clustering, and ASC analysis. Through literature-based searches, a total of 101 genes homologous to reference genes from previous studies on plants and algae were identified in *T. pseudonana* and plotted to indicate deviation and mean TPM (Literature). *K*-means clustering was applied to the 7380 genes and Cluster 9 (243 genes) and Cluster 14 (466 genes) possessed the genes with the most stable expression pattern across the four treatments. Genes from these clusters are plotted to indicate deviation and mean TPM (*k*-means: Cluster 9; $k-$means: Cluster 14). ASC was used to assess statistical significance (post-$p < 0.1$) of fold changes of 1.10, 1.25, and 1.50 for each treatment relative to the replete control. Genes from these fold change bins are plotted to indicate deviation and mean TPM (ASC: 1.50 fold change; ASC: 1.25 fold change; ASC 1.10 fold change). For a fold change of 1.10, two genes, both hypothetical proteins, (NCBI: 7446346 and 7452192) passed the post-$p < 0.1$ cutoff, and represent the most stable genes based on the ASC analysis (Data Sheet 2-3). For each of the six classes of putative reference genes, tag counts were normalized to total library size (in TPM) and are plotted relative to the mean for each of the four treatments: Replete (Rep), P-limited (P-lim), Fe-limited (Fe-lim), and co-limited (Co-lim). The color of the line correlates to the mean normalized tag count. A star marks a gene (NCBI: 7451632) in Cluster 14 that is not on the scale of expression for P-limited (1104.7 TPM) and co-limited (-1664.9 TPM) treatments.

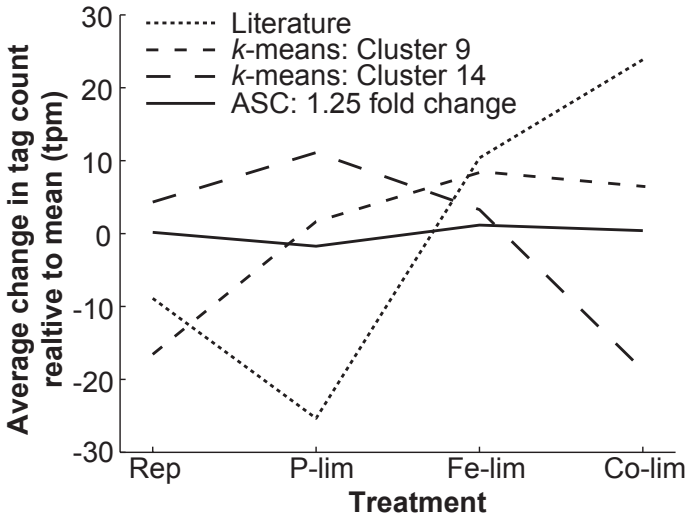


Figure 2.2: Average deviation from the mean level of expression for all genes found with literature-based searches, k-means clustering, and ASC analysis of 1.25 fold change. The average change in tag count from the mean expression (TPM) for all the genes identified through literature-based searches for genes homologous to known reference genes from the literature ($n = 101$), k-means clustering from Cluster 9 ($n = 243$) and Cluster 14 (n = 466), and ASC analysis identifying genes demonstrating a 1.25 fold change with a post-$p < 0.1$ ($n = 179$). The mean standard deviations for the four cases are as follows: Literature (92.62 TPM), Cluster 9 (41.66 TPM), Cluster 14 (43.12 TPM), and ASC (14.24 TPM). The mean TPM is plotted for the four treatments: Replete (Rep), P-limited (P-lim), Fe-limited (Fe-lim), and co-limited (Co-lim).

of 14.18 TPM and 21.93 TPM, respectively.

Table 2.1: Gene counts for the fold change bins of 1.50, 1.25, and 1.10 across posterior probability cutoffs ranging from 0.01 to 0.20.

Fold change	1.50			1.25			1.10		
Posterior probability	Number of genes	Min. TPM	Max. TPM	Number of genes	Min. TPM	Max. TPM	Number of genes	Min. TPM	Max. TPM
post-$p < 0.2$	1649	2.11	1802.38	312	2.83	1281.15	8	20	176.63
post-$p < 0.1$	1375	2.22	1802.38	179	7.06	1281.15	2	51.81	105.73
post-$p < 0.05$	1127	2.83	1802.38	122	20	1281.15	1	105.73	105.73
post-$p < 0.01$	801	5.69	1802.38	62	20	1281.15	0	NA	NA

Underlying differences in the magnitude and pattern of expression variation across treatments were identified by examining the average tag count change for each reference gene detection method (Figure 2.2). If all genes in a group were perfectly constitutively expressed, the average change in tag count relative to the mean observed would be 0 TPM (e.g. the TPM values across all treatments for each of the genes within a group were the same). The average variation from the mean observed in the literature (ranging from -25.34 to 23.84 TPM) highlighted the differential expression across treatments. The average change in tag count relative to the mean in both Cluster 9 (ranging from -16.56 to 8.47) and Cluster 14 (ranging from -18.72 to 11.11 TPM) clearly demonstrated patterns of regulation across treatments (e.g. the upregulation under P-limitation and downregulation under co-limited observed in Cluster 14). In contrast, the average change in tag count relative to the mean observed in the genes identified through ASC (1.25 fold change with post-$p < 0.1$), which showed a low magnitude of variation (ranging from -1.732 to 1.613 TPM) and a small mean standard deviation across the four treatments (14.24 TPM). Ultimately, the expression patterns of the majority of the genes identified through literature-based searches and k-means clustering were more variable across the *T. pseudonana* test treatments, than those genes identified with ASC.

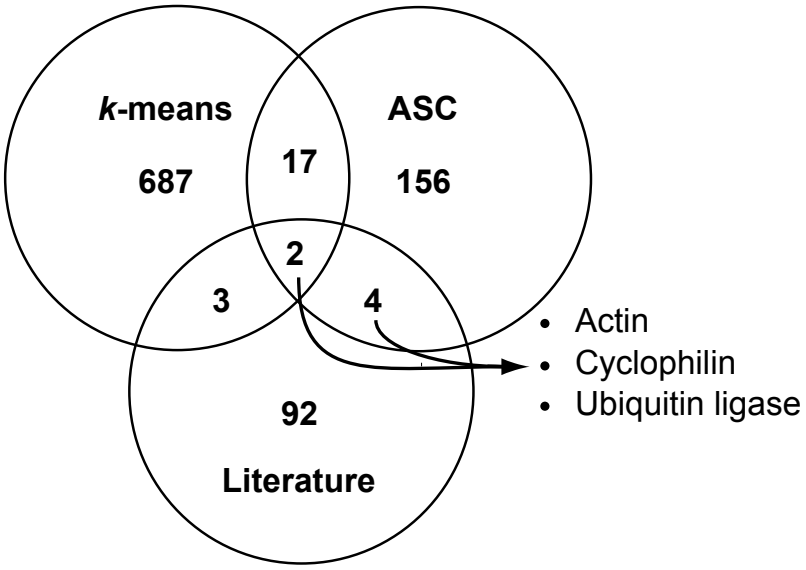


Figure 2.3: Comparison of possible reference genes found with literature-based searches, k-means clustering, and ASC analysis of 1.25 fold change. Venn diagram analysis was used to compare genes identified as candidate reference genes through literature-based homolog searches (totaling 101 genes), with the k-means clustering method (genes in Cluster 9 and 14, totaling 709 genes), and with quantitative exclusion by ASC (based on genes demonstrating a 1.25 fold change with a post-$p < 0.1$, totaling 179 genes). The number of genes in each region is reported. The intersection of all ASC and literature-based searches yielded six total genes representing three different gene families: actin (NCBI: 7449411), cyclophilin (NCBI: 7445376), and ubiquitin ligase (NCBI: 7448637, 7450639, 7446724, and 7451971).

A comparison of the three techniques: literature-based searches, k-means cluster selection, and ASC cutoff at 1.25 fold change revealed comparatively few genes in common between the techniques (Figure 2.3). Of the 709 genes identified through k-means clustering

and the 179 genes found through ASC analysis (genes which pass the 1.25 fold change cutoff for post-$p < 0.1$), 21 genes are shared (Figure 2.3), of which six lacked GO annotations or KOG definitions (Data Sheet 2-2). Between the genes identified through literature and ASC analysis, six genes were held in common; these genes were representative of the general gene classifications: actin (NCBI: 7449411), cyclophilin (NCBI: 7445376), and ubiquitin ligases (NCBI: 7448637, 7450639, 7446724, and 7451971). Only two genes (NCBI: 7448637 and 7446724) were found in common amongst all three methods of reference gene selection, both of which were annotated as putative ubiquitin ligases (Data Sheet 2-1).

2.5 Discussion

Prior to the availability of high-throughput molecular datasets, reference genes for non-model organisms were selected based on literature reports of stably expressed genes in model organisms. With non-model organisms such as eukaryotic phytoplankton this task is particularly difficult, as stably expressed genes are not readily apparent in the relatively limited molecular literature specific to these organisms. Often the selection of a reference gene relies on information from distantly related organisms under dissimilar conditions, leading to extensive validation work (McDonald et al., 2010; Whitney et al., 2011). Herein, we compared the efficacy of reference gene selection based on the literature as compared to verifiable selection through k-means clustering and ASC analysis of high-throughput transcriptome data in *T. pseudonana* across four nutrient treatments (replete, P-limited, Fe-limited, and co-limited). These treatments are of environmental relevance as both P and Fe are major drivers of diatom physiological ecology and consequently carbon fixation (Moore, 2004). Additionally, P and Fe often occur concurrently at very low concentrations in marine systems and have been found to be independently co-limited, or mutually exclusive biochemically (Saito et al., 2008).

Our literature-based search of relative gene expression studies from 12 algae and plants yielded 18 general reference gene categories, for which 101 homologs in the *T. pseudonana* genome were identified (Data Sheet 2-1). While some of these genes demonstrated stable expression (e.g. actin, cyclophilin, and ubiquitin conjugating enzymes), the vast majority displayed some form of differential expression in the treatments examined herein. Furthermore, there was considerable heterogeneity of expression among the different gene copies of actin, cyclophilin, and ubiquitin conjugating enzymes, demonstrating that not all genes within a gene family are stably expressed. These data underscore that a literature-based selection of reference genes necessitates validation across all treatments of interest (Vandesompele et al., 2002; Pfaffl et al., 2004).

Differential expression patterns in high-throughput datasets are often analyzed with clustering methods, such as hierarchical or k-means clustering (D'haeseleer, 2005). Rather than using a clustering method for the identification of differential expression patterns, here it is applied to identify constitutively expressed genes. The k-means clustering algorithm was chosen as it is a top-down or partition-based approach to gene clustering that is not hierarchical and requires few assumptions about the data (Hartigan and Wong, 1979). Several of the 709 putative reference genes identified by k-means analysis (from Clusters 9 and 14) were clearly differentially regulated, with large deviations from the mean expression level. The presence of outliers is to be expected using the k-means method, for it is a pattern-based method and all genes must be placed into one of the partitioned $k = 15$ clusters. Thus, optimal placement of a gene is not always guaranteed. As with a finite number of

clusters, the assignment of a gene is often forced. For example, even genes in Cluster 9 and 14 were subject to strong patterns of regulation, with both clusters demonstrating large average changes in tag count relative to the mean tag count. Arguably, it is better to select a reference gene from a pool of genes that do not share the same pattern of regulation. Therefore, genes uncovered via k-means clustering must be manually surveyed to exclude genes with large deviation prior to the selection of a candidate reference gene.

In lieu of clustering approaches, other studies have used statistical parsing of ESTs in tomato plants (Coker and Davies, 2003) and Affymetrix whole-genome GeneChip data from *A. thaliana* (Czechowski et al., 2005) and humans (de Jonge et al., 2007) to identify reference genes that have small deviations from the mean of replicated treatments. In contrast to these and other statistical methodologies typically applied to high-throughput sequence data with replication, the Bayesian approach to gene expression analysis, ASC, allowed for selection of candidate genes based on a statistical cutoff rather than cardinality. Though typically used for the identification of differentially expressed genes, the function of ASC was reversed in this study by lowering the post-p cutoff. Genes for which post-$p < 0.1$ for a specified fold change were targeted, meaning that genes that were unlikely to have made that fold change were selected. The 1.25 fold change bin yielded the most options for candidate reference genes without sacrificing stability of expression (as was seen in the 1.50 fold change bin).

ASC provides a method of identifying reference genes with expression levels similar to those of target genes. For example, the mean normalized tag counts of genes identified using ASC were broad (from 7 to over 1200 TPM), providing the opportunity for reference gene expression to be generally matched with target gene expression. Current studies frequently employ reference genes for endogenous control that have very high levels of expression across all treatments, such as ACT1 (NCBI: 7449411) in *T. pseudonana* (which has a mean expression value of 1024.1 TPM in this data set), yet these highly expressed genes might not be optimal for studies of genes with low levels of expression or when multiplexing targets in probe-based RT-qPCR analysis.

High-throughput transcript datasets also allow the selection of reference genes to move beyond the confines of gene annotation and previously identified reference genes. In fact, the two genes with the most stable expression in the 1.10 fold change bin are hypothetical, with no clear annotation. Of the 179 genes that passed the 1.25 fold change cutoff with ASC, 44 lacked both GO and KOG annotations. A large percentage of the 11,390 genes in the *T. pseudonana* genome are annotated as hypothetical proteins (Armbrust et al., 2004; Mock et al., 2008), and here we show a number of them are stably expressed across the target conditions. This has been seen with model organisms, where a good majority of constitutively expressed genes fall outside the bounds of preconceived "housekeeping" genes (Czechowski et al., 2005; de Jonge et al., 2007). By using a Bayesian approach such as ASC, hypothetical proteins can be chosen as reference genes.

Comparison of the putative reference genes recovered using ASC to previous studies served to cross-validate the ASC approach. Actin (ACT1, NCBI: 7449411) has been validated in the literature as a suitable reference gene for relative expression studies of *T. pseudonana* under Fe-limitation (Whitney et al., 2011), a treatment considered in this study, and was one of the 179 genes passing the ASC 1.25 fold change cutoff. Additionally, only five of the 179 genes with stable expression found with ASC were identified as differentially expressed in a study of *T. pseudonana* under additional treatments to those described here (e.g. nitrogen limitation, silica limitation, etc.) (Mock et al., 2008) (Data Sheet 2-4). Of the five, only one gene (NCBI: 7451974) was identified as differentially expressed under Fe-limitation, a condition examined in this study. Taken together, this validates the genes

identified with ASC using alternative data and methods, and suggests that the ASC-detected genes are globally stable across many different conditions for *T. pseudonana*. However, one of the two genes identified in the 1.10 fold change bin (NCBI: 7446346) was identified as significantly down-regulated under nitrogen limitation by Mock et al. (2008). This highlights the importance of validating genes across all treatments of interest prior to their use as reference genes.

Notably, the k-means and ASC dataset revealed only 21 genes in common. The 179 genes found through ASC were, in fact, distributed fairly evenly across all of the 15 clusters. The lack of intersection observed between the two datasets is likely related to the parsing ability inherent in k-means clustering. The k-means approach is highly driven by patterns of differential regulation, but does not consider the significance of that regulation (e.g. genes that are not significantly upregulated are placed in a cluster with genes that are significantly upregulated). Thus, the stably expressed genes that were identified by ASC, though not displaying major patterns of regulation, were clustered based on minor patterns in variation of gene expression. Therefore, while k-means clustering provides a global view of commonalities in gene expression patterns, ASC is more robust at identifying reference genes.

Eight genes were common between the ASC and literature-based searches, which were distributed across three general gene classes: actin (NCBI: 7449411), cyclophilin (NCBI: 7445376), and ubiquitin ligases (NCBI: 7448637, 7450639, 7446724, and 7451971). For those interested in identifying suitable reference genes for studies in *T. pseudonana* but lack transcriptome datasets across the treatments of interest, these eight genes may serve as good tentative reference genes as they are verified in this study and have been identified as stable in many other organisms under many conditions. In particular, ubiquitin ligases/conjugating enzymes have been used as reference genes in several studies involving other algae, namely, *Aureococcus anophagefferens*, *Phaeodactylum tricornutum*, and *Prorocentrum minimum* (Siaut et al., 2007; McGinn and Morel, 2008a; Guo and Ki, 2012; Wurch et al., 2011a; Berg et al., 2008), and with further analysis may represent particularly good reference genes in the phytoplankton.

Sequence-based transcriptome profiling has become an increasingly useful method for gene discovery and differential expression analysis. Yet, RT-qPCR is still valuable for the examination of detailed trends in expression in both culture and field studies. Here we show that the application of ASC and, to a lesser extent, k-means clustering can be used to successfully screen transcriptome data for potential reference genes. The isolation of candidate reference genes using ASC with the 1.25 fold change cutoff for post-$p < 0.1$ was more robust and stringent at excluding differentially expressed genes than both the literature-based searches and k-means clustering. Based on these data for *T. pseudonana*, it was shown that ACT 1 and ubiquitin ligase may be useful reference genes. Yet, in addition to these common reference genes, the data demonstrate that there are many more stably expressed genes (both annotated and hypothetical) to choose from for expression studies in this and potentially other diatoms. Notably, this survey focused only on variation in P and Fe supply, so these genes may not transfer to studies of other nutritional drivers or other physical forces, such as light intensity or temperature. As more transcriptome data are generated for phytoplankton, ASC can be employed without sequence replicates, to identify reference genes for other phytoplankton under various conditions. Additionally, the suite of genes identified through these analyses might allow for better multi-gene normalization analysis that would provide for the detection of smaller fold changes with certainty (Vandesompele et al., 2002; Czechowski et al., 2005).

Chapter 3

Metatranscriptome analyses indicate resource partitioning between diatoms in the field

This chapter was originally published as Alexander, H., Jenkins, B.D., Rynearson, T.A., and Dyhrman, S.T. (2015). Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proc. Natl. Acad. Sci. U. S. A.* 112:17, E2182-E2190.

H.A., B.D.J., T.A.R., and S.T.D. designed research; H.A., B.D.J., T.A.R., and S.T.D. performed research; H.A. contributed new reagents/analytic tools; H.A. analyzed data; and H.A., B.D.J., T.A.R., and S.T.D. wrote the paper.

3.1 Abstract

Diverse communities of marine phytoplankton carry out half of global primary production. The vast diversity of the phytoplankton has long perplexed ecologists, as these organisms coexist in an isotropic environment while competing for the same basic resources (e.g. inorganic nutrients). Differential niche partitioning of resources is one hypothesis to explain this "paradox of the plankton," but it is difficult to quantify and track variation in phytoplankton metabolism *in situ*. Here we use quantitative metatranscriptome analyses to examine pathways of nitrogen (N) and phosphorus (P) metabolism in diatoms that co-occur regularly in an estuary on the east coast of the US (Narragansett Bay). Expression of known N and P metabolic pathways varied between diatoms, indicating apparent differences in resource utilization capacity that may prevent direct competition. Nutrient amendment incubations skewed N:P ratios, elucidating nutrient responsive patterns of expression, and facilitating a quantitative comparison between diatoms. The resource-responsive (RR) gene sets deviated in composition from the metabolic profile of the organism, being enriched in genes associated with N and P metabolism. Expression of the RR gene set varied over time and differed significantly between diatoms, resulting in opposite transcriptional responses to the same environment. Apparent differences in metabolic capacity and the expression of that capacity in the environment suggest that diatom-specific resource partitioning was occurring in Narragansett Bay. This high-resolution approach highlights the molecular underpinnings of diatom resource utilization and how co-occurring diatoms adjust their cellular physiology to partition their niche space.

3.2 Introduction

The stability and primary productivity of ecosystems has long been linked to the diversity of primary producers (Elton, 1958; Cardinale et al., 2012). This is well documented in terrestrial systems (Naeem et al., 1994; Tilman et al., 2001; Cadotte, 2013; Balvanera et al., 2006; Tilman et al., 1996) and is increasingly being established for marine systems (Behl et al., 2011; Striebel et al., 2009; Steiner et al., 2005; Ptacnik et al., 2008). Marine phytoplankton generate roughly half of global primary production (Nielsen, 1960; Strickland, 1965; Field et al., 1998) and play a critical role in oceanic ecosystem structure and function. Within the phytoplankton, the diatoms generate an estimated 40% of primary production (Nelson et al., 1995). Thus diatoms alone exert profound influence over marine primary production and global carbon (C) cycling, particularly in coastal margins and estuaries.

Phytoplankton are extremely diverse, with estimates of over 200,000 extant species (Sournia et al., 1991; Tett and Barton, 1995; Mann and Droop, 1996). This dramatic level of taxonomic diversity in the plankton is difficult to resolve with the apparently limited number of niches in the pelagic habitat, as these organisms compete for the same two basic resources: light and nutrients. As was highlighted by Hutchinson (1961), the phytoplankton violate Gause's law of competitive exclusion, which posits that two organisms competing for the same resources cannot coexist. Much thought has gone towards identifying the cause of the "paradox of the plankton" and include explanations such as "contemporaneous disequilibrium" of patchy phytoplankton distributions (Richerson et al., 1970), life history differences (Huisman et al., 2001), species oscillations (Huisman and Weissing, 1999), environmental fluctuation (Roy and Chattopadhyay, 2007), intra-specific variation (Menden-Deuer and Rowlett, 2014), and differential niche partitioning (Connel and Connell, 1980). Of these

potential factors, one of the most difficult to directly observe in the plankton is niche partitioning. Different species may have unique strategies that allow them to specialize on certain resources or nutrient forms, and species may have different responses to resource shifts that allow them to avoid competition. Such specialization in eco-evolutionary strategy may underlie the "winner-loser" dynamics observed in productive estuaries and coastal systems, yet resolving patterns of species-specific resource metabolism in the field remains a central challenge.

It is accepted that the macronutrients N and P are central to the structuring of phytoplankton communities across large spatial and temporal scales (Margalef, 1963; Follows et al., 2007; Johnson et al., 2006), and that phytoplankton compete for nutrients in the natural environment (Sommer, 1983, 1985). Studies focused on nutrient geochemistry, and phytoplankton quotas or uptake have emphasized the importance of nutrients to community dynamics, but these do not generally examine resource partitioning between individual species (Hutchins et al., 1999; Zubkov et al., 2003). Transcriptional studies provide species-specific resolution, but few studies have examined the global expression of nutrient metabolism pathways in the field (Marchetti et al., 2012) or in organisms lacking a fully sequenced genome (Frischkorn et al., 2014; Moustafa et al., 2010), and as a result, the mechanistic underpinnings of phytoplankton resource metabolism *in situ* are not well understood. In situ global gene expression analyses (metatranscriptome profiling) are a means for elucidating a species' metabolic capacity and examining patterns in resource utilization potential through time by tracking the expression of species' resource-responsive genes. When simultaneously applied to multiple species in a sample, this can resolve differences in the expressed gene compliment and how it is modulated, which may reflect resource partitioning of phytoplankton niche space (Gifford et al., 2013). For example, this approach has uncovered species-specific expression of genes for the transport of organic compounds (Poretsky et al., 2010; Rinta-Kanto et al., 2012; Gifford et al., 2011), highlighting potential differences in resource partitioning. Although increasingly critical for identifying resource utilization in the bacterioplankton, metatranscriptome profiling has only recently been used to examine resource utilization in coastal eukaryotic phytoplankton populations (Dupont et al., 2015), largely due to challenges in quantifying a transcriptional response in a mixed population and until recently, the lack of reference genomes and transcriptomes for determining the origin of the transcriptional response. Co-occurring phytoplankton may possess different metabolic capabilities and responses to resource availability, which may then enable resource partitioning and the segregation of the fundamental niche or the realized niche. Knowledge of if and how these organisms modulate their niche space would allow predictive models to better resolve species distribution and ecosystem structure and function in the future ocean (Follows et al., 2007).

Herein we examined pathways of resource metabolism between two co-occurring diatoms from the genera *Thalassiosira* and *Skeletonema*, sampled from a time-series site in Narragansett Bay. Narragansett Bay is a highly productive and dynamic estuarine environment on the east coast of the United States with an estimated bay-wide average net production of 269 $gC\ m^2\ yr^{-1}$ (Oviatt et al., 1981). Quantitative metatranscriptomic techniques were developed and used to: 1) assign taxonomic designation, 2) assess and track changes in known metabolic capacity through quantitative molecular fingerprinting (QMF), 3) statistically identify the resource responsive gene set, and 4) proportionalize the expression of resource-responsive genes to track species-specific responses through time, using standardized transcriptional differentiation scores (STD). This multifaceted computational approach enabled the unprecedented resolution of the unique strategies these two diatoms use for re-

source acquisition.

3.3 Materials and Methods

3.3.1 Experimental set up and sample collection

Surface seawater was collected and sampled for total community RNA at the long-term sampling site in Narragansett Bay, RI (41°34′07″ N, 71°23′31″ W) during 2012 (16 May, 21 May, 30 May, 4 June, and 8 June, here called S1 through S5) in conjunction with the weekly time-series sampling effort. To diminish the influence of diel signals, samples were collected and processed between 0830 and 0900 local time. Near surface water was collected in an acid-washed carboy and then filtered onto polycarbonate filters (5.0 μm pore size, 47mm) using a peristaltic pump. Filters were then placed in cryovials and stored in liquid nitrogen until RNA extraction. In this manner all samples were preserved within 15 minutes of collection. In addition to sampling for total community RNA, phytoplankton abundance was measured as part of the long-term weekly survey (Furnas, 1983, 1982).

A nutrient amendment incubation experiment was performed on 30 May 2012, with S3 representing the t = 0 of the experiment. Water collected in conjunction with S3 was pre-filtered through 200Âţm mesh to remove large zooplankton grazers and placed into acid washed 2.5 L bottles. Triplicate bottles were then amended with nutrients to create five treatments: +N, +P, -N, -P, and ambient control. The +N and +P treatments were designed to eliminate the N and P stress signals, respectively, while the -N and -P treatments were supplemented with everything except the nutrient in question (e.g. the -N treatment was amended with P, Si, Fe, and vitamins), to force the draw down of N and P, respectively (Table B.2). N and P amendment concentrations were selected to be approximately 10x the seasonal average ambient N and P concentrations in the surface waters of Narragansett Bay measured at Station II. Silica, Fe, and f/5 vitamin amendments were made in proportion to the f/5 media ratios (Guillard, 1975). Bottles were placed in a flow-through incubator at ambient temperatures and PAR to mimic the collection depth. The incubation was run for 48 hours, at which point all treatments were sampled for total community RNA as described above by filtering and snap- freezing 2L of biomass from each replicate bottle.

3.3.2 RNA extraction and sequencing

Filters from triplicate bottles, representing approximately 6 L of water, were pooled by treatment and extracted for each of the *in situ* and incubation experiment samples. RNA was extracted from individual filters with the RNAeasy Mini Kit (Qiagen), following a modified version of the yeast protocol. Briefly, lysis buffer and RNA-clean zircon beads were added to the filter and samples were vortexed for 1 minute, placed on ice for 30 seconds, and then vortexed again for 1 minute. Samples were then processed following the yeast protocol. The resulting RNA was eluted in water and then treated for possible DNA contamination using TURBO DNA-free Kit (Ambion) following the Rigorous Dnase protocol. RNA from each triplicate was then pooled by sample or treatment, using the RNA Cleanup Protocol from the RNAeasy Mini Kit (Qiagen). The total RNA ($>$ 1000 ng for each sample) was then enriched for eukaryotic mRNA through a poly-A pull down onto oligo-dT beads. The resulting enriched RNA sample then went through library preparation with the Illumina TruSeq RNA Prep Kit (Illumina). Libraries were sequenced at the Columbia University Genome Center (New York, New York) with an Illumina HiSeq2000. Each sample was

sequenced to produce 60 million, 100 base pair, paired end reads (Table B.1). Raw sequence data quality was visualized using FastQC (Andrews) and then cleaned and trimmed using Trimmomatic v 0.27 (paired end mode; 6-base pair wide sliding window for quality below 20; minimum length 25 base pair) (Lohse et al., 2012). Environmental sequence reads are available at the NCBI under accession number SRP055134.

3.3.3 Transcriptome and genome mapping

To assign taxonomic identification to the reads a database was created from transcriptomes made publicly available as of 17 March 2014 through the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP). In total, 401 transcriptomes from 209 species or cultured isolates were collected. Like-species transcriptomes were combined (regardless of strain or condition) using CD-HIT-EST (98% identity; word size of 9). The resulting clustered set of transcripts was considered to be the representative transcriptome for the species or cultured isolate. The 209 transcriptomes created in this manner were concatenated to form a comprehensive species-level transcriptome database from the MMETSP library. Due to the large size of the resulting MMETSP database, trimmed reads were mapped to the MMETSP using the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010) and then counted using the HTSeq 0.6.1 package (Anders et al., 2014).

Transcriptomes from two ecologically relevant diatom species in Narragansett Bay were selected: *Skeletonema* costatum RCC1716 (MMETSP0013, accessed from the publicly available transcriptome databases, Moore Foundation Marine Microbiology Initiative-supported Marine Microbial Eukaryote Transcriptome Sequencing Project, National Center for Genome Resources) and *Thalassiosira rotula* CCMP3096 (a custom assembly available at EBI, accession number Hx2000045970). These transcriptomes were individually clustered using CD-HIT-EST (parameters: -c 0.98, -n 9) (Li and Godzik, 2006). The resulting clustered set of transcripts was then concatenated to form a reference transcriptome database. Trimmed reads from the field and incubation samples were mapped to this transcriptome database using Bowtie2 v 2.2.1 (parameters: -a -sensitive) (Langmead and Salzberg, 2012). As a point of comparison, reads were also mapped using Bowtie2 v2.2.1 under the same parameters to the genome of the model centric diatom species, Thalassiosira pseudonana CCMP1335 (v3.0), an organism not known to be abundant in Narragansett Bay. Mapped reads were then counted by transcript using the HTSeq 0.6.1 python package (parameters: -m union âĂŞs no) (Anders et al., 2014). Reads aligning to more than one full transcript were not counted. KEGG pathways were assigned to the assembled sequences using the online KEGG Automatic Annotation Server (KAAS), using the bi-directional best-hit (BBH) method to obtain KEGG Orthology annotations. In this study, only genes with a normalized count (NC) (raw count / total number of genes mapped to an organism) of at least 2 tags per million (TPM) in at least one of the field or incubation samples were included, thus limiting the sample set to 4318 genes for *T. rotula* (19.3% of the transcriptome) and 20921 genes for *Skeletonema* spp. (75.6% of the transcriptome). This difference in coverage is directly related to their relative abundance in the population.

3.3.4 Transcriptome clustering

To assess relatedness of genes within *Skeletonema* spp. and *T. rotula*, the transcriptomes were translated using ORF Predictor using a reference BLASTx alignment against the NCBI database with an 1e-5 cutoff (Min et al., 2005). These translated peptide sequences were

then combined with the translated proteins from the diatom genomes *Fragilariopsis cylindrus* CCMP1102 v1.0, *Phaeodactylm tricornutum* CCMP632 v2.0, *Pseudo-nitzschia multiseries* CLN-47 v1.0, and *Thalassiosira pseudonana* CCMP1335 v3.0, which were collected from the Joint Genome Institute JGI database. A protein similarity network was then created using EGN, a software program that automates the reconstruction of gene networks from protein sequences through reciprocal BLASTp analysis (e-value <1e-5, hit identity threshold: 20%, best-reciprocal threshold of best e-value: 5%, minimal match coverage threshold: 90%) (Halary et al., 2013, 2010). Networks were then visualized and manipulated using Cytoscape 3.0, where the layout of the network was produced using an edge-weighted spring-embedded model based on e-value, meaning that genes that are closer together are more similar (Smoot et al., 2011; Saito et al., 2012). Known RR genes from previous transcriptome studies of the diatom species, *T. pseudonana*, were selected for analysis: 1) the P-responsive gene, Thaps_24435, a NPT (Dyhrman et al., 2012) and 2) the N-responsive gene, Thaps_25299, an assimilatory nitrate reductase (Bender et al., 2014).

3.3.5 Identification of stable and nutrient-responsive genes

Intercomparison of nutrient-incubation experiments enabled the identification of both nutrient-responsive genes and stably expressed reference genes for *T. rotula* and *Skeletonema* spp.. For each organism, RR genes were identified by comparing the counts for that organism in +N to the -N incubation and the +P to the -P incubation, respectively, using Analysis of Sequence Counts (ASC), an empirical Bayes method, which estimates the prior distribution from the data, itself *Wu2010*. ASC analyses were run using raw count data from each species separately. Genes were considered to be differentially regulated between treatments if for a fold change of 2.0 the posterior probability (post-p) was greater than 0.95 (Dyhrman et al., 2012). After surveying the output of several different post-p cutoffs (Figure B.11), stable genes were identified using ASC as described by Alexander et al. (2012) through pairwise comparisons of each of the incubation treatments (fold change of 1.25, post-$p < 0.1$).

3.3.6 Normalization of metatranscriptome data

Counts from the field were first normalized to the sequences belonging to the species in the library (Equation 3.1). For a particular species, c, the number of reads mapping to a gene g, $c_{i,g}$, was normalized to the sum of all the counts across all genes for that organism yielding the normalized count, $NC_{i,g}$, similar to normalization techniques used for metatranscriptome data (Marchetti et al., 2012; Ottesen et al., 2011).

$$NC_{i,g} = \frac{c_{i,g}}{\sum\limits_{g \epsilon G} c_{i,g}} \tag{3.1}$$

From hence forth, only genes for which $NC > 2$ TPM in at least one sample (incubation or field) were considered. To facilitate interspecies comparisons, the NC was normalized to the geometric mean of the set of stable reference genes, R, yielding a stable gene normalized count ($SGNC$). The calculation of $SGNC$ (Equation 3.2) for transcriptome data, while a novel application to metatranscriptome analysis, was designed to emulate the normalization used in qRT-PCR studies (Vandesompele et al., 2002).

$$SGNC_{i,g} = \frac{NC_{i,g}}{\left(\prod^{R} NC_{i,g}\right)^{1/R}} \tag{3.2}$$

The nutrient responsive genes identified as differentially expressed in the nutrient incubations (Table B.2) were then selected for investigation in the field metatranscriptomes (S1 through S5). The $SGNC$ from the field for these nutrient-related genes were bounded by the $SGNC$ from like nutrient incubations to calculate the standardized transcriptional differentiation score for N (STD_N) (Equation 3.3) and P (STD_P) (Equation 3.4).

$$STD_N = \frac{SGNC_{field} - SGNC_{+N}}{SGNC_{-N} - SGNC_{+N}} \tag{3.3}$$

$$STD_N = \frac{SGNC_{field} - SGNC_{+P}}{SGNC_{-P} - SGNC_{+P}} \tag{3.4}$$

For example, in calculating STD_N, the $SGNC_{field}$ is put in the range of the $SGNC_{+N}$ and $SGNC_{-N}$. By consequence, if the STD_N for a gene in the field equals zero it is more similar in expression to the +N treatment and if it equals one it is more similar in expression to the -N treatment. As such, a plot STD_N against STD_P, can divide the space into two main theoretical quadrants N:P > Redfield ($STD_P > 1$ and $STD_N < 0$) and N:P < Redfield ($STD_N > 1$ and $STD_P < 0$) (Figure B.8). The total number of genes falling into each of the quadrants were counted by varying the bounds considered: the N:P > Redfield ratio quadrant ($STD_P > C$; $STD_N < C$, for $0.25 < C < 0.75$) and the N:P < Redfield ratio quadrant ($STD_P < C$; $STD_N > C$, for $0.25 < C < 0.75$). To conservatively approximate variation, the value of C was varied over 10 different values and the average and standard deviation for the percentages of genes falling into each of the quadrants was quantified. Similarity of data between species by quadrant was assessed using an analysis of variance (ANOVA) with a generalized linear model. The results from a post hoc Tukey test show the divergence of species across time ($p < 0.05$).

3.4 Results and Discussion

3.4.1 Samples and sequencing

Narragansett Bay has seasonal blooms of diatoms which have been monitored through weekly cell counts for over 50 years at a long-term time series station (Borkman and Smayda, 2009; Li and Smayda, 1998). Five eukaryotic surface metatranscriptome samples were taken from surface seawater collected during May and June of 2012 at the time-series site yielding over 358 million 100 base pair, paired end cDNA reads from the field (S1-5) (Table B.1). In conjunction with these field-based surveys, a nutrient amendment incubation experiment was performed with natural communities on 30 May 2012 (S3) to drive the community towards opposite extremes in the nitrogen (N): phosphorus (P) ratio (Redfield ratio) (Table B.2). Eukaryotic metatranscriptomes from the five incubation treatments produced over 264 million 100 base pair, paired end cDNA reads (Table B.1).

To assign taxonomic designation, sequences from the time series were conservatively mapped (such that if a read mapped to more than one gene it was discarded) to a sequence library containing all assembled sequences and annotations generated through the Marine

Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) (Keeling et al., 2014) which were public as of 17 March 2014. The custom sequence library contained 401 transcriptomes across 209 species or cultured isolates. Between 62 to 71% of reads from the *in situ* samples mapped to the MMETSP database with diatoms dominating the libraries, representing 30 to 46% of the total mapped reads (Figure 3.1). The peak in diatom representation coincided with a bloom of *Skeletonema* spp. detected in time-series cell counts (Figure B.1), and a period of historical overlap between the *Skeletonema* and *Thalassiosira* genera. *Skeletonema* and *Thalassiosira* were well represented during the time period studied in both mapped RNA (Figure 3.1) and cell counts (Figure B.1). *T. rotula* was present but at low abundance during the time-series, while *Skeletonema* spp. was abundant, with sampling spanning a bloom of *Skeletonema* (>10,000,000 cells L^{-1}), with peak cell densities in S2 (21 May 2012) (Figure B.1). As such, subsequent analyses were focused on these two groups by remapping the data to representative transcriptomes: *T. rotula* and *S. costatum* (Table B.1). *S. costatum* was chosen as it was the transcriptome from the genus *Skeletonema* that recruited the most hits in the MMETSP database. Because *Skeletonema* is known to include morphologically cryptic species that can only be identified by scanning electron microscopy (Sarno et al., 2005; Zingone et al., 2005; Smayda, 2011), it is referred to here as *Skeletonema* spp. for clarity. Up to 17.5 and 54.9% of reads from a single sample mapped to *T. rotula* and *S. costatum*, respectively. As a point of comparison, reads were also mapped to the genome of a second Thalassiosirid, *T. pseudonana*, a diatom that is not known to be abundant in Narragansett Bay (Table B.1). Though displaying high identity with the 18S rDNA to *T. rotula* and *S. costatum* (96% and 93% identity, respectively), less than 1% of the metatranscriptome reads mapped *T. pseudonana* (Table B.1), highlighting the specificity of the approach.

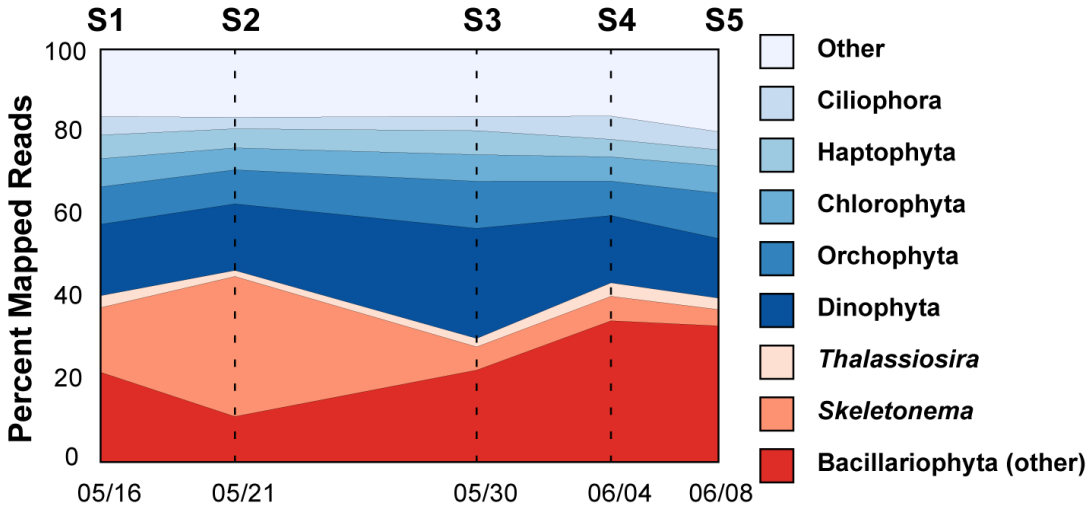


Figure 3.1: Taxonomic classification of RNA-seq paired end reads across the five field samples. Classification was determined by mapping to a database comprised of all publicly available transcriptomes through the Marine Microbial Eukaryotic Transcriptome Project (MMETSP) as of March 17, 2014.

3.4.2 Temporal plasticity in expressed metabolic capacity

Metatranscriptome short reads were mapped to transcriptomes that had been annotated with Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) (Data Sheet 3-1), allowing the expression of KO gene families within a KEGG module (higher-level groupings of KO gene families into pathway or functional classifications) to be examined over time. Normalizing the expression of KEGG modules to the total KEGG annotated reads for each organism across time yielded the Quantitative Metabolic Fingerprint (QMF), which highlighted differences between the two species and differences across time for each species (Figure 3.2). A comparison of the total number of annotated genes falling into each of the KEGG modules revealed a close to one-to-one linear relationship (slope of 1.0948, $R^2 = 0.9123$) (Figure B.2), indicating that the observed differences are not an artifact of gene distribution between organisms. The QMFs of the two organisms were distinct and there were significant shifts in the QMF of each species over time reflecting considerable plasticity in the expressed metabolic capacity (Figure B.3). Central carbohydrate metabolism, carbon fixation, and other carbohydrate metabolism were some of the most highly expressed KEGG modules in the field for both *Skeletonema* spp. and *T. rotula*, though higher for *Skeletonema* spp., where expression of these pathways peaked during S4 representing over 84% of mapped KEGG reads (Figure 3.2). The largest global shift in KEGG module expression was seen in *Skeletonema* spp. on S2 (Figure B.3), when its density peaked at 11,520,000 cells L^{-1}. The S2 time point for *Skeletonema* spp. had increased QMF signals in ATP synthesis, proteasome, and ubiquitin systems and decreased QMF signals in photosynthesis and carbon metabolism relative to other time points. For example, reads mapping in photosynthesis dropped by over an order of magnitude from 0.3-2.2% of annotated transcripts in S1, S3-5 to 0.03% during S2 (Figure 3.2). The temporal plasticity of transcript allocation to different aspects of metabolism for both species was striking and likely reflects the dynamic environment which they inhabit: an estuary, where the geochemistry is highly variable (Nixon et al., 1995).

Temporal plasticity in the KEGG module expression patterns, including a shift away from the expression of carbon fixation and photosynthesis suggests that the elevated *Skeletonema* spp. cell numbers observed in S2 may have been after this diatom reached peak bloom biomass. A significant proportion of the KEGG modules expressed were classified as ribosomes (5-45% for *Skeletonema* spp. and 5-9% for *T. rotula*). Gifford et al. (2013) suggested that ribosomal protein expression correlates with growth rate. Applying this principle to these eukaryotic data suggests the growth rates for both *Skeletonema* spp. and *T. rotula* fluctuated, with peaks in growth rate occurring during S1 and S3 for *Skeletonema* spp.. This pattern for *Skeletonema* spp. did not track with the relative abundance of the organism, which peaked in the S2 sample, again suggesting that this sample was taken during the bloom decline. These growth dynamics cannot be fully resolved without a more detailed sample set.

Skeletonema spp., the dominant diatom during the study period (Figure 3.1), had a higher proportion of transcripts related to growth relative to *T. rotula*, such as those encoding aspects of carbon metabolism, N metabolism, sulfur metabolism, and lipid metabolism (Figure 3.2). Conversely, several KEGG modules were more highly expressed in *T. rotula* compared to *Skeletonema* spp., particularly those for glycan metabolism, phosphate and amino acid transport systems, and repair system modules (Figure 3.2). The majority of highly expressed KO modules (e.g. N metabolism) were based on moderate to high expression across several KO gene families, but, in some cases, the differences in expres-

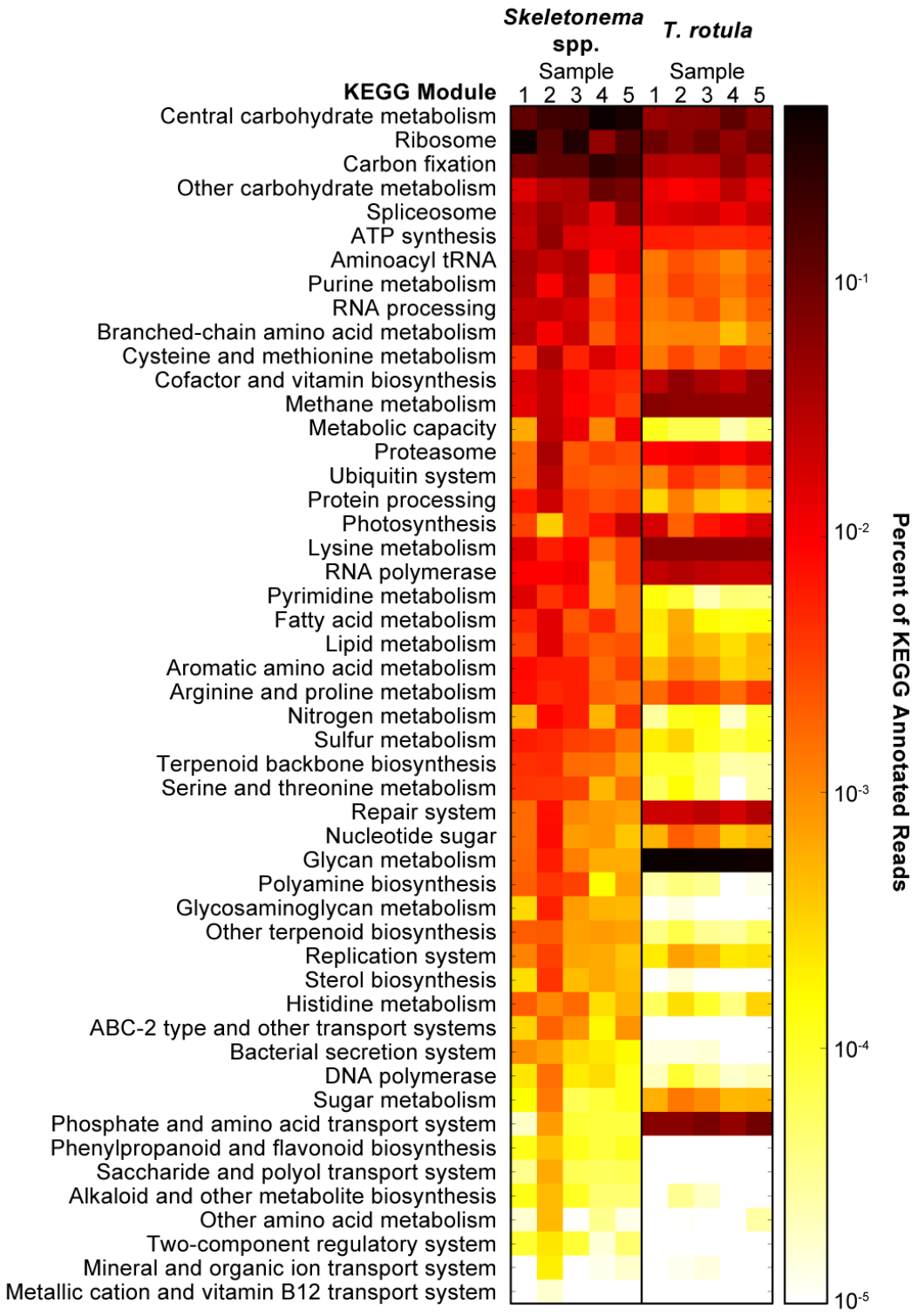


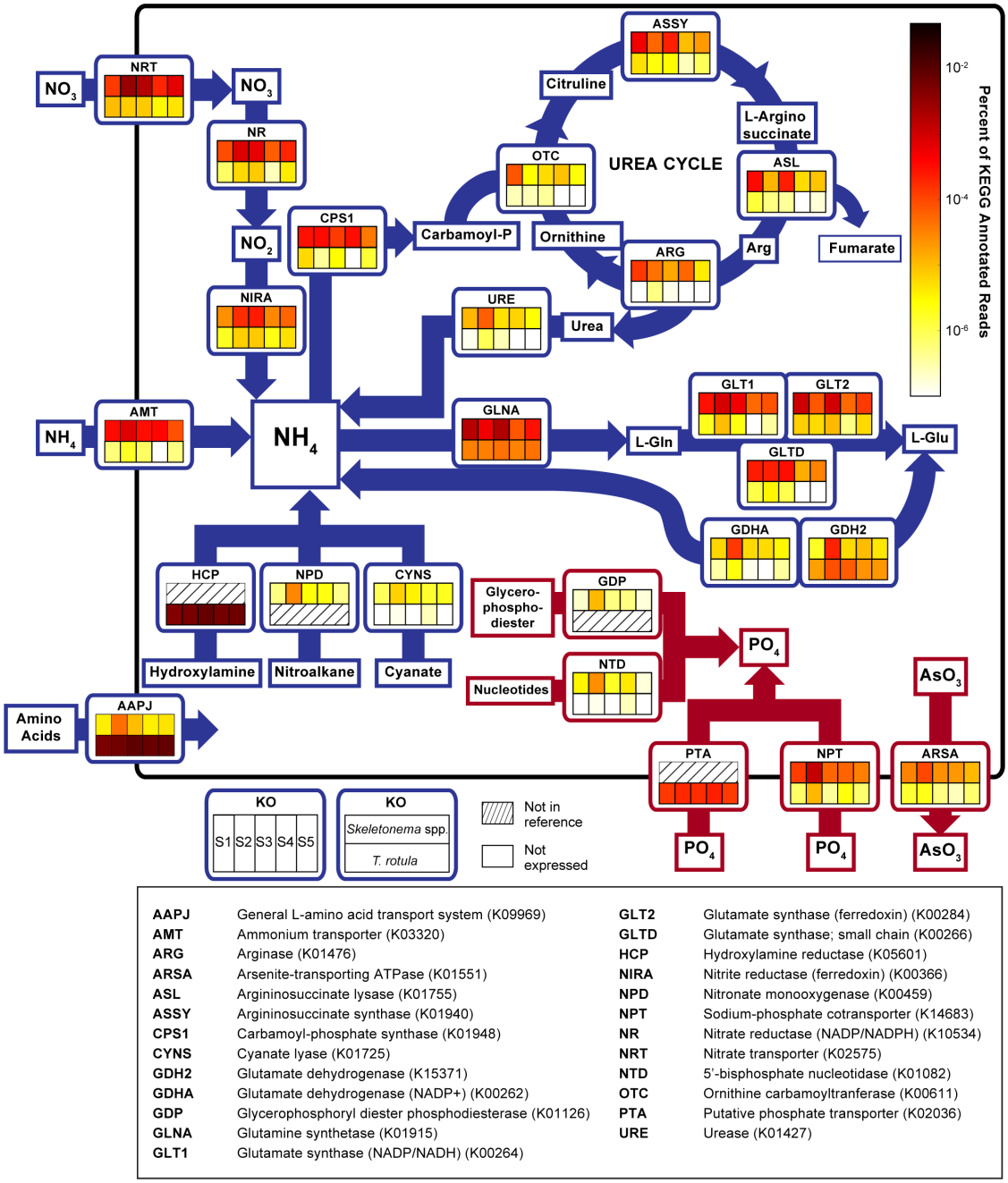
Figure 3.2: Quantitative metabolic fingerprint (QMF) depicting the relative expression of KEGG modules for *Skeletonema* spp. and *T. rotula* in Narragansett Bay across the five sampling time points (S1-S5). Color indicates the proportion of total reads mapping to each KEGG module relative to all KEGG annotated reads.

sion at the module level were due to differences in the expression of a single KO gene family within the KEGG module. For example, the driver of the difference in the expression of glycan metabolism, which represented upwards of 41% of all KEGG annotated reads for *T. rotula* compared to less than 0.6% for *Skeletonema* spp., was primarily associated with the high expression of a putative UDP-N-acetylglucosamine-dolichyl-phosphate N-acetylglucosaminephosphotransferase (K01001). This was identified as a silaffin-like response gene associated with silica polymerization (Shrestha et al., 2012). Differences in silica metabolism may in part drive how the fundamental niche is segregated between these two diatoms. Taken together, the contrast in QMF between the two diatoms underscores the fundamental differences in expressed metabolic capacity that are present in these two co-occurring diatoms and highlights traits of a successful competitor (e.g. high expression of carbon metabolism).

3.4.3 Species-specific resource utilization underpins physiological ecology

KO gene families related to N and P metabolism were examined in the field samples to identify species-specific patterns in resource utilization. *Skeletonema* spp. and *T. rotula* both possess and express core pathways of N and P metabolism (such as the ornithine-urea cycle) (Figure 3.3). Expression of these individual KO gene families was temporally variable, as was observed with the expression of KEGG Modules, but related enzymes in a pathway exhibited a coordinated response (Figure 3.3). For example, the nitrate transporter (K02575), nitrate reductase (K10534), and nitrite reductase (K00366), in *Skeletonema* spp. all had peak expression in S2 (Figure 3.3). *Skeletonema* spp. and *T. rotula* share pathway homologs, including the same suite of N transporters (ammonium: AMT, nitrate: NRT, amino acid: AAPJ), but these genes often had disparate patterns of expression between the two species (Figure 3.3). *Skeletonema* spp., the more abundant diatom, had high expression of KO gene families associated with the acquisition of nitrate and ammonia that were particularly amplified during the S2 bloom event. *T. rotula* had low expression of both of those transporters but high expression of a general amino acid transporter (Figure 3.3). Amino acid transport (North and Stephens, 1972) and nitrate transport (Serra et al., 1978) has previously been found to inversely correlate with intracellular nitrate concentration in the cell or the presence of ammonia in the media. Yet, here, two closely related diatoms, existing in the same parcel of water and the same nutrient environment, are expressing genes to access different pools of dissolved N. Similar to nitrate transport, there was high expression of nitrate/ nitrite reductase KO gene families in *Skeletonema* spp., whereas *T. rotula* appears to possess a different N reduction metabolism. This is observed in a KO gene family that is absent from the reference transcriptome of *Skeletonema* spp.: hydroxylamine reductase (Figure 3.3). This gene has been found in the genomes of both *T. pseudonana* and *P. tricornutum*, and is thought to have been acquired via lateral transfer from bacteria (Bowler et al., 2008). The enzyme may potentially aid redox balancing and electron cycling from nitrate reduction (Allen et al., 2008). While the absence of this gene in *Skeletonema* spp., has not been definitively shown, the marked high expression of this gene in *T. rotula* suggests that this gene product represents a potential point of segregation in the metabolic capacity of these two species. Together, these data suggest that these species have disparate strategies for acquiring N and this may in part drive the relative success of *Skeletonema* spp. over the sample period.

While N has been observed to be a primary nutritional driver in Narragansett Bay (Nixon et al., 1995; Smayda, 1974; Sakshaug, 1977), P may also drive the dynamics of these



AAPJ	General L-amino acid transport system (K09969)	**GLT2**	Glutamate synthase (ferredoxin) (K00284)
AMT	Ammonium transporter (K03320)	**GLTD**	Glutamate synthase; small chain (K00266)
ARG	Arginase (K01476)	**HCP**	Hydroxylamine reductase (K05601)
ARSA	Arsenite-transporting ATPase (K01551)	**NIRA**	Nitrite reductase (ferredoxin) (K00366)
ASL	Argininosuccinate lysase (K01755)	**NPD**	Nitronate monooxygenase (K00459)
ASSY	Argininosuccinate synthase (K01940)	**NPT**	Sodium-phosphate cotransporter (K14683)
CPS1	Carbamoyl-phosphate synthase (K01948)	**NR**	Nitrate reductase (NADP/NADPH) (K10534)
CYNS	Cyanate lyase (K01725)	**NRT**	Nitrate transporter (K02575)
GDH2	Glutamate dehydrogenase (K15371)	**NTD**	5'-bisphosphate nucleotidase (K01082)
GDHA	Glutamate dehydrogenase (NADP+) (K00262)	**OTC**	Ornithine carbamoyltranferase (K00611)
GDP	Glycerophosphoryl diester phosphodiesterase (K01126)	**PTA**	Putative phosphate transporter (K02036)
GLNA	Glutamine synthetase (K01915)	**URE**	Urease (K01427)
GLT1	Glutamate synthase (NADP/NADH) (K00264)		

Figure 3.3: Schematic cell model depicting the relative expression of KO gene families associated with nitrogen (N) and phosphorus (P) metabolic pathways for *Skeletonema* spp. and *T. rotula* in Narragansett Bay across the five sampling time points (S1-S5). Color indicates the proportion of total reads mapping to each KEGG module relative to all KEGG annotated reads.

two diatoms. *Skeletonema* spp. shows elevated expression of a sodium phosphate cotransporter (NPT), again with peak expression during S2 (bloom). *T. rotula* does express the NPT as highly, but by comparison has a much higher transcript count for a putative P transporter (PTA), that is not detected in *Skeletonema* spp. (Figure 3.3). These transporters may have different kinetic properties that allow the two species to diverge in their PO_4 uptake strategies. Genes associated with the scavenging of P from organic molecules, such as glycerophosphoryl diester phosphodiesterase (GDP), also suggest differences in expressed metabolic capacity between the two species. GDP may be associated with exogenous metabolism of dissolved organic P (DOP) or internally in the cleaving of P from lipids (Van Mooy et al., 2009; Dyhrman et al., 2012). The expression of GDP by *Skeletonema* spp., with peak around S2, and apparent absence of this transcript in *T. rotula* suggests *Skeletonema* spp. may be accessing a pool of DOP that is not being utilized by *T. rotula*. In *T. pseudonana*, related transcripts are tightly linked to concomitant changes in the proteome and biochemical activities (Dyhrman et al., 2012). If these transcriptional patterns are linked to similar changes in activities, then these insights suggest that there is a fundamental difference in the metabolic capacity being expressed in the same environment by the two diatoms. *Skeletonema* spp. is both actively taking up PO_4 and hydrolyzing organic sources, where as *T. rotula* is not hydrolyzing DOP and is taking up inorganic PO_4 by a different mechanism. In summary, these data suggest that these two diatoms have unique metabolic capacity for the utilization of specific forms of N and P. Such disparate resource utilization potential could be a niche-defining feature that underpins diatom diversity as well as the "winner-loser" dynamic observed here with the differences in cell abundance between the species.

3.4.4 Identification and modulation of resource responsive genes *in situ* highlights species-specific differences

To identify and quantitatively track resource responsive (RR) genes *in situ*, incubation experiments were used to examine species-specific transcriptional responses to shifts in N:P ratios. Comparing the expression patterns between like nutrient treatments (+N versus -N and +P versus -P) for each of the organisms enabled the statistical identification of a suite of RR genes (Wu et al., 2010) and stable reference genes (Alexander et al., 2012). RR gene counts were normalized to the stable reference genes (Figure B.4) resulting in stable gene normalized counts ($SGNC$). Calculation of a $SGNC$ is similar in concept to reference gene normalization done in qRT-PCR (Bustin, 2000) or metatranscriptome studies of prokaryotes (McCarren et al., 2010), with the added value of not having to rely on reference genes from model diatoms.

Of the transcripts expressed at greater than 2 tags per million (TPM) for at least one treatment, 24.5% and 17.9% were identified as RR by being significantly up or down-regulated between N or P treatments for *Skeletonema* spp. and *T. rotula*, respectively (Data Sheet 3-2, Table B.3). As is common with phytoplankton studies (Marchetti et al., 2012), the majority of the RR genes do not have a KEGG annotation (Figure B.4A). The portion of the RR gene set annotated with KEGG ontology for *Skeletonema* spp. and *T. rotula* revealed that, relative to the full KEGG profile, genes comprising genetic information processing associated with replication (encompassing ribosomes, nucleotide replication and processing) were underrepresented for both organisms in the RR set (Figure B.5). By contrast, the RR sets were enriched for energy, carbohydrate, and lipid metabolism, which encompass pathways known to be associated with the metabolism of N and P (Figure 3.4A,

Figure B.5). Specific genes in this set included, but were not limited to, those associated with N assimilation (e.g. glutamate dehydrogenase, glutamine synthase, nitrate reductase), DON utilization (e.g. urease, aminopeptidase, amino-acid transport system), P scavenging (e.g. phosphate transporter, sodium phosphate cotransporter) and DOP utilization (e.g. phosphatases) (Data Sheet 3-2). A number of these genes have been shown to be N or P responsive in transcriptional studies with cultures of the diatom *T. pseudonona* (Dyhrman et al., 2012; Bender et al., 2014), and transporters, and enzymes for the processing of organic N or P, as observed here, are well known to be resource responsive in many phytoplankton (Dyhrman et al., 2012; Wurch et al., 2011a; Dyhrman et al., 2006; Bruhn et al., 2010). Overall, these genes demonstrated patterns of regulation *in situ* (Figure 3.4B, Figure B.6) similar to what has been observed in culture (Dyhrman et al., 2012; Bender et al., 2014). In the incubations, the sodium-phosphate cotransporter (NPT) was significantly up-regulated in the -P treatment for both species (Figure 3.4B), which is consistent with P regulation of a *T. pseudonana* NPT homolog (Thaps_24435) observed in culture experiments (56). Nitrate reductase, which has been shown to be regulated by N in *T. pseudonana* (Thaps_25299) (Bender et al., 2012), was up-regulated in -N for *T. rotula*, but not *Skeletonema* spp. (Figure B.6). In fact, members of this large gene family (Figure B.7) showed disparate regulation in both species (Figure B.6). These data demonstrate that the use of nutrient amendments is robust for normalizing and identifying N and P responsive genes in the field that are consistent with known signals, but also point to the value of *in situ* analyses, as application of a priori knowledge about regulation from model diatoms could lead to misinterpretations.

Of the RR gene sets for *Skeletonema* spp. and *T. rotula*, only 17.7 and 12.7% of the genes, respectively, were annotated with KEGG ontology (Figure 3.4A). Identifying differentially regulated genes *in situ* through experimental manipulations allowed the expression patterns of genes to be tracked even when their function was unknown. As an example, two RR gene families were identified with homologs in *Skeletonema* spp. and *T. rotula* (Figure 3.4B, Figure B.7). RR1 was up-regulated in -P compared to +P for both species (Figure 3.4B). Homologs from RR1 were also identified in other diatom genomes (Fracy_268075, Phatr_19661, Psemu_246578, Psemu_319824, Thaps_32459) (Figure B.7). Annotations for these genes were limited, though Fracy_268075 was identified as possibly involved in intracellular trafficking, secretion, or vesicular transport; suggesting these proteins may be involved in poly-P metabolism (Ogawa et al., 2000). RR2 demonstrated significantly different patterns of regulation in the two species: up-regulated in -N compared to +N for *T. rotula* but down-regulated in -N compared to +N for *Skeletonema* spp. (Figure 3.4B). A homolog from RR2 was identified in *T. pseudonana* (Thaps_22648) (Figure B.7) and was poorly characterized, with the best BLAST hit to a human dentin sialophosphoprotein. This suggests RR2 could be associated with biomineralization.

To enable cross-comparison of the RR genes between species, their expression was put into a greater metabolic context by proportionalizing the expression in the field to the transcriptional range observed in the incubations with extremes in the N:P ratio. This technique is similar in concept to targeted assays using qRT-PCR to compare expression patterns between species in culture (Kang et al., 2009). Briefly, the $SGNC$ of a gene in the field was bounded by the $SGNC$ from each of the nutrient treatments to yield the standardized transcriptional differentiation score for both N (STD_N) and P (STD_P) (Figure 3.4C). The STD score was used to directly compare expression relative to its maximum and minimum capacity where values of $STD \geq 1$ indicate signals are similar to the deplete condition, and values of $STD \leq 0$ indicate similarity to the replete condition. The STD_N and STD_P were plotted for genes from the NPT and the two highlighted RR gene families,

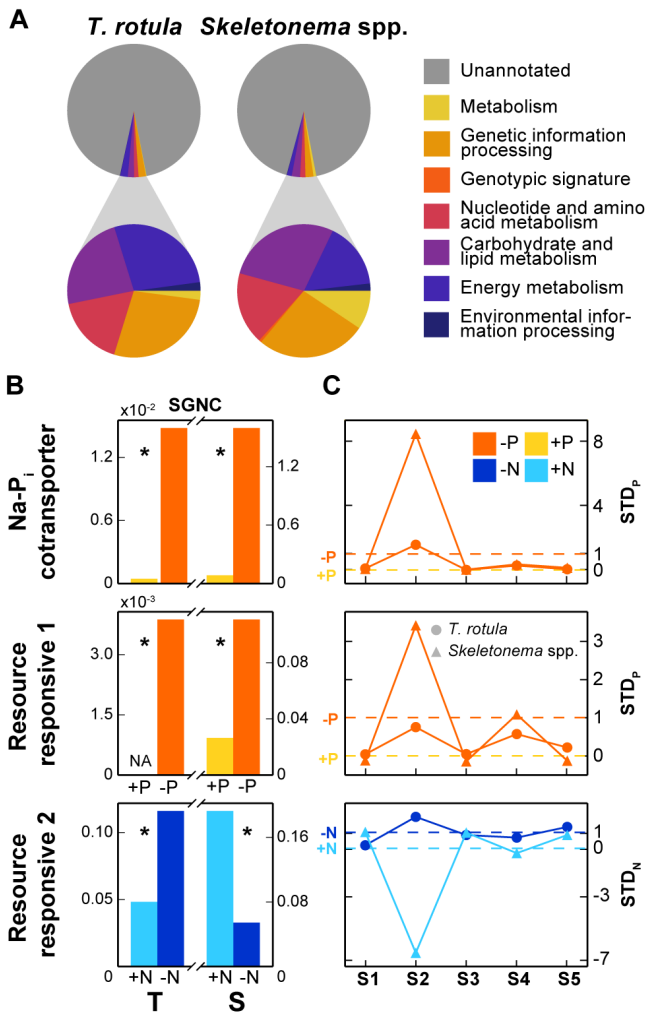


Figure 3.4: Functional composition of the resource-responsive (RR) gene sets for *T. rotula* and *Skeletonema* spp. (A), the relative expression in the incubation samples (B) and standardized transcriptional differentiation (STD) scores (C) for a known P-responsive gene, sodium-phosphate cotransporter, and two novel RR gene families. (A) The RR gene sets were identified through cross comparison of like-nutrient incubations (i.e. +N vs. -N and +P vs. -P), using ASC (fold change = 2, post-$p > 0.95$) (57). The relative functional categorization of the RR gene set for *T. rotula* and *Skeletonema* spp. based on KEGG ontology as assigned by KAAS is depicted at the module-level relative to the portion unannotated with KEGG. (B) Expression pattern in stable gene normalized counts ($SGNC$) of the genes from the associated gene cluster from *T. rotula* (T) and *Skeletonema* spp. (S) plotted in related incubations (i.e. novel P-responsive shows expression from +P and -P incubations). Asterisk indicates significance between pair-wise comparisons (fold change = 2, post-$p > 0.95$) (57). (C) STD scores plotted across the five sample points showing STD_P for the P-significant genes and STD_N for the N-significant genes. Dashed horizontal lines at 0 and 1 indicate the +P/+N and -P/-N for corresponding significant genes.

over the time-series (Figure 3.4C). The NPT for both *Skeletonema* spp. and *T. rotula* showed elevated expression during S2. RR1, which was also identified as significantly expressed in -P compared to +P, also showed elevation during S2 (the bloom day). The expression of RR1, however, was also elevated on S4 for both diatoms, which was not seen for the NPT. However, $STD_P > 1$ for *Skeletonema* spp. indicating a far more P deficient response in *Skeletonema* spp. compared to *T. rotula*, which never demonstrated P-sensitive expression in the field comparable to that observed in the -P incubations (Figure 3.4C). RR2 showed different patterns of expression across time for both species. Most interestingly, perhaps, was the low STD_N score for *Skeletonema* spp. during its bloom period indicating that the RR2 expression was more similar to the +N treatment, whereas the STD_N for *T. rotula* was greater than one suggesting that expression was more similar to the -N treatment (Figure 3.4C). These three, targeted examples suggest that during the large bloom of *Skeletonema* spp., *Skeletonema* spp. was expressing genes in pattern more similar to the -P and +N treatments, while the expression of *T. rotula* was more similar only to the -N treatment. Notably, these are orthogonal patterns associated with the same environment.

The STD_N and STD_P for all of the RR genes were calculated (Data Sheet 3-2) to expand upon the single gene analyses above. The RR genes were plotted based on the $STD_N : STD_P$ (Figure B.8) to examine how similar the pattern was to the incubation N:P ratio (Figure 3.5A, Figure B.9). Redfield regimes have historically been used to characterize different aquatic environments based on the ratio of nutrient resources required for growth. For example, a Redfield ratio of N:P = 16, here called âĂIJRedfieldâĂİ, would predict neither P nor N limitation. As expected, RR genes identified as N-regulated genes fall primarily into the N:P < Redfield quadrant and P-regulated genes fall primarily into the N:P > Redfield quadrant for both *Skeletonema* spp. and *T. rotula* (Figure 3.5A). Observing patterns in distribution of these genes across time, S2 stands out amongst the time points, where a significant (88%) proportion of the P-regulated genes from *Skeletonema* spp. move far into the N:P > Redfield quadrant (Figure 3.5A). This N:P > Redfield physiology is consistent with the single gene analyses (Figure 3.4C) and suggests P availability may have constrained *Skeletonema* spp. populations during the bloom sample (S2). Conversely, a large proportion (59%) of the N-regulated genes in *T. rotula* move into the N:P < Redfield quadrant (Figure 3.5A) consistent with the divergent responsiveness of RR2 observed for *T. rotula* compared to *Skeletonema* spp. (Figure 3.4C). In fact, with the exception of S4 and S5 where *T. rotula* had even distribution between the N:P > Redfield and N:P < Redfield quadrants, the two species always showed statistically significant (Tukey HSD analysis ($p < 0.05$)) orthogonal responses in the distribution of the RR gene set across the two quadrants (Figure 3.5B, Figure B.10). These patterns combined with the temporal variability in gene expression patterns indicate a finely tuned response to the environment, which is distinctive for each diatom species. Although there are many potential controls on diatom dynamics in Narragansett Bay, including top-down processes like predation (Martin, 1970; Lawerence and Menden-Deuer, 2012), these patterns of resource responsive gene expression suggest the presence of bottom-up nutrient control on diatom population dynamics in Narragansett Bay.

This work addresses fundamental knowledge gaps in how phytoplankton species are able to co-occur while they compete for the same basic resources. Co-occurring diatoms appear to have different functional capabilities in N and P metabolism, and this metabolic potential is modulated in field populations in a distinctive way for each diatom. These findings suggest that differential resource partitioning is occurring between these two diatoms *in situ*. Such resource partitioning could facilitate the vast diversity of the phytoplankton and the structure, function, and productivity of aquatic ecosystems. In culture studies, resource-

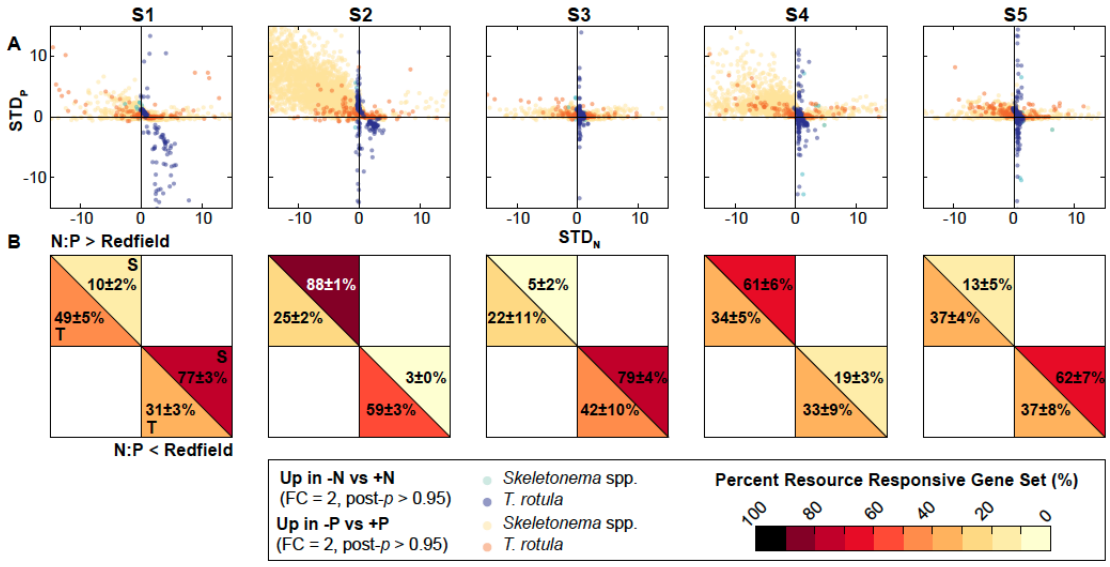


Figure 3.5: Evolution of resource-responsive (RR) gene partitioning over time in Narragansett Bay for *T. rotula* and *Skeletonema* spp.. (A) The stable gene normalized field signal for each gene identified as significantly (2-fold change, post-p $>$ 0.95) up-regulated in -P vs. +P for *Skeletonema* spp. (yellow) and *T. rotula* (orange) and in -N vs. +N for *Skeletonema* spp. (cyan) and *T. rotula* (dark blue) were proportionalized relative to the expression for those genes in nutrient incubations, yielding the STD_N and STD_P for each gene. These data are plotted for Sample 1 through Sample 5. (B) The proportion of identified RR genes falling into the N:P $>$ Redfield and N:P $<$ Redfield quadrants for *T. rotula* (T) and *Skeletonema* spp. (S).

related transcriptional changes have been shown to be tightly choreographed with changes in proteins, activities, and biochemical pools (56, 62, 69). If further work were able to similarly link the transcriptional patterns observed here with changes in enzymatic activities or uptake rates, then shifts in the RR gene sets may reflect aspects of the realized niche and how it differs between these two species. These detailed, *in situ* transcriptional comparisons would not have been possible without proportionalization to metabolic capacity (STD), which provides a quantitative means to directly compare transcriptional patterns between species. This approach could be applied to other systems, organisms, or environmental parameters to identify responsive genes and proportionalize their expression, with the aim of answering similar questions about how co-occurring species adjust their cellular physiology to partition their niche space.

Chapter 4

Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean

This chapter was originally published as Alexander, H., Róuco, M., Haley, S.T., Wilson, S.T., Karl, D.M., and Dyhrman, S.T. (2015). Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proc. Natl. Acad. Sci. U. S. A.* 112:44, E5972-E5979.

H.A. and S.T.D. designed research; H.A., M.R., S.T.H., S.T.W., and S.T.D. performed research; H.A., D.M.K., and S.T.D. analyzed data; H.A. and S.T.D. wrote the paper; and M.R., S.T.H., S.T.W., and D.M.K. contributed to the writing of the paper.

4.1 Abstract

A diverse microbial assemblage in the ocean is responsible for nearly half of global primary production. It has been hypothesized and experimentally demonstrated that nutrient loading can stimulate blooms of large eukaryotic phytoplankton in oligotrophic systems. Although central to better balancing biogeochemical models, knowledge of the metabolic traits that govern the dynamics of these bloom-forming phytoplankton is limited. We employed eukaryotic metatranscriptomic techniques to identify the metabolic basis of functional group-specific traits that may drive the shift between net heterotrophy and autotrophy in the oligotrophic ocean. Replicated blooms were simulated by deep seawater addition to mimic nutrient loading in the North Pacific Subtropical Gyre and the transcriptional responses of phytoplankton functional groups were assayed. Responses of the diatom, haptophyte, and dinoflagellate functional groups in simulated blooms were unique, with diatoms and haptophytes significantly (95% confidence) shifting their quantitative metabolic fingerprint from the *in situ* condition, while dinoflagellates showed little response. Significantly differentially abundant genes identified the importance of co-limitation by nutrients, metals, and vitamins in eukaryotic phytoplankton metabolism and bloom formation in this system. The variable transcript allocation ratio, used to quantify transcript reallocation following DSW amendment, differed for diatoms and haptophytes, reflecting the long-standing paradigm of phytoplankton r- and K-type growth strategies. While the underlying metabolic potential of the large eukaryotic phytoplankton was consistently present, the lack of a bloom during the study period suggests a crucial dependence on physical and biogeochemical forcing, which are susceptible to alteration with changing climate.

4.2 Introduction

It has been postulated that the net oxygen state of oligotrophic systems is controlled by irregular blooms of large eukaryotic phytoplankton (Karl et al., 2003), which have been shown to respond to nutrient input in open ocean ecosystems (McAndrew et al., 2007). Nutrient loading shifts the community away from a tightly regenerating microbial loop, based on small phytoplankton (typically bacterioplankton), towards a community with a higher proportion of larger phytoplankton cells (typically eukaryotic phytoplankton) and, consequently, more carbon export (McAndrew et al., 2007). Data from Station ALOHA in the oligotrophic North Pacific Subtropical Gyre (NPSG) during the summer period have identified increased export of organic carbon, diatomassociated biogenic silica (BSi), and, to a lesser extent, particulate inorganic carbon (PIC) associated with calcification (Karl et al., 2012). These increases are potentially a result of nutrient-driven community shifts. Blooms of eukaryotic phytoplankton in oligotrophic environments are often dominated by the diatom functional group (Villareal et al., 2012) and may determine the balance between net heterotrophic and net autotrophic conditions (Karl et al., 2003). Such blooms, however, are thought to be under-sampled and may often go undetected in satellite ocean color records (Villareal et al., 2011). Consequently, the drivers of bloom formation and concomitant carbon export remain poorly understood, particularly in critical oligotrophic systems.

Though central to better balancing global biogeochemical models of GPP (López-Urrutia et al., 2006), knowledge of the biogeochemical drivers that govern the dynamics of these bloom-forming organisms in oligotrophic systems is limited. Nutrient environments are integral to the structuring of phytoplankton communities (Margalef, 1978; Tilman, 1977;

Cavender-Bares et al., 2001) and initiating blooms. Originally thought to be a stable low-fluctuating habitat, long-term monitoring at Station ALOHA has demonstrated that within the constraints of an oligotrophic ecosystem the nutrient regime can be quite dynamic, alternating in controlling factors over many time scales (Karl et al., 2001). These oscillations may be driven by biological (e.g. bloom of N2-fixing cyanobacteria, which draw down P or Fe, while injecting new nitrogen into the system (Karl, 2002)), physical (e.g. nutrient supply from transient eddies (Benitez-Nelson et al., 2007)), or anthropogenic (e.g. atmospheric deposition of nutrients (Kim et al., 2014)) forcing. Regardless of the source, these nutrient-loading events could act to stimulate blooms in the large eukaryotic phytoplankton community. Historically, the distributions of key eukaryotic phytoplankton function groups have been tracked relative to nutrient stoichiometries to examine how nutrients influence the physiological ecology of different functional groups like diatoms and haptophytes (Cortés et al., 2001; Villareal et al., 2012). Although valuable, it is still difficult to relate stoichiometries to the dynamics of bloom formation and, in particular, to patterns in key groups like diatoms without specific measures of functional group physiology.

Molecular-level tools that can track transcripts, proteins, or even metabolites and biochemicals in a taxon-specific way are increasingly being employed in cultures and field populations to track metabolic capacity and physiological responses (Marchetti et al., 2012; Alexander et al., 2015a; Dupont et al., 2015; Pearson et al., 2015; Ottesen et al., 2014; Hennon et al., 2015; Dyhrman et al., 2012; Amin et al., 2015). Molecular assessment of physiology for eukaryotic populations is most tractable in coastal systems with high biomass (Alexander et al., 2015a; Dupont et al., 2015), so, in oligotrophic ocean regions, molecular studies of physiology have typically been limited to the numerically abundant members of the microbial community: picoplankton (cyanobacteria, heterotrophic bacteria, and small picoeukaryotes). Groundbreaking studies have demonstrated the responsiveness of the dominant picoplankton community to pulses of deep seawater (DSW) (Shi et al., 2012) and of DOM (McCarren et al., 2010), as well as the diel synchrony of the photosynthetic and heterotrophic communities (Ottesen et al., 2014), yet were limited in their coverage of the rare, large eukaryotic fraction due both to the low biomass of their standing stock and the lack of reference sequences. A recent world ocean survey has highlighted the diversity encompassed within the eukaryotic planktonic community (de Vargas et al., 2015) and suggested that the structuring of these communities may be driven by physical and chemical oceanographic features (Villar et al., 2015). Yet again, the responsiveness of these keystone eukaryotic phytoplankton communities to changing environmental conditions remains poorly described and understood. New resources, including a new marine microbial eukaryote sequencing initiative (Keeling et al., 2014), in combination with the bioinformatic pipeline used here are making possible the study of oligotrophic eukaryotic phytoplankton with unprecedented mechanistic resolution. Here we examined potential biogeochemical controls on phytoplankton bloom formation in the NPSG and the metabolic traits that govern the distribution and responses of unique functional groups.

4.3 Materials and Methods

4.3.1 Sample collection

Seawater for the *in situ* eukaryote community mRNA analysis was collected at the HOT, Station ALOHA (22°45′ N, 158°00′ W) from a depth of 25 m at 1400 hrs (local time) on three occasions during August and September 2012 as part of the HOE-DYLAN research

expedition aboard the R/V Kilo Moana. Hydrocasts for sampling were performed using a conductivity-temperature-depth (CTD) rosette water sampler equipped with 24 Scripps 12-L sampling bottles. Water was collected in acid-washed 20-L carboys and approximately 60-L of seawater was prescreened through $200\mu m$ mesh and then filtered onto polycarbonate filters ($5.0\mu m$ pore size, $47mm$, Whatman) by way of peristaltic pump. This size fraction was targeted to sample large eukaryotic phytoplankton, which are known to form blooms and contribute to export flux in the NPSG. Filters were changed every 20 minutes or when flow rate decreased. Filters were placed in cryovials and stored in liquid nitrogen until mRNA extraction. The total length of filtration time did not exceed 3 hours. Nutrient delivery in the ocean is highly variable; here we chose to model a single nutrient pulse. The incubation experiments were performed with two treatments: a DSW treatment, which included 10% 0.2 μm filtered seawater collected from 700 m added to whole seawater collected at 25 m and a control treatment with no addition (Table C.1). Triplicate 20-L carboys of each treatment were incubated at 30% surface light levels using on-deck incubators for 7 days and processed as described above on the final day at 1400 h (local time). Nutrient concentrations for phosphate (PO_4), nitrate and nitrite [NO_2+NO_3], and silicate [SiO_4] (Table C.1) were measured by filtering 125 mL of seawater through a 0.2 μm, 47-mm polycarbonate filter, and stored frozen (20°C) in acid washed bottles until analysis at the Chesapeake Bay Lab at the University of Maryland according to the facilityâĂŹs protocols. Chlorophyll a was measured on whole water samples collected onto GFF filters ($25mm$, Whatman) using a 90% acetone extraction and assayed by fluorescence using the AquaFluor Turner TD700 (Parsons et al., 1984). There was little difference between the *in situ* sample and the control treatment both with regards to total chlorophyll a concentration and transcriptional profile (Figure C.1, Figure C.7). There was no significance difference in QMF between the control treatment at the final time point and the *in situ* sample taken at the start of the incubation (Figure C.7). To most conservatively compare non-bloom and bloom scenarios, analyses thus focused on the comparison of the *in situ* community to the DSW amended samples.

4.3.2 RNA extraction and sequencing

RNA was extracted from individual filters with the RNeasy Mini Kit (Qiagen), following a modified version of the yeast protocol. Briefly, lysis buffer and RNA-clean zirconia/silica beads was added to the filter and samples were vortexed for 1 minute, placed on ice for 30 seconds, and then vortexed again for 1 minute. Samples were then processed following the yeast protocol. The resulting RNA was eluted in water and then treated for possible DNA contamination using TURBO DNA-free Kit (Ambion) following the Rigorous DNase protocol. RNA from individual filters was then pooled by sample, using the RNA Cleanup Protocol from the RNeasy Mini Kit (Qiagen). The resulting RNA sample thus represented approximately 60 L of total seawater for the *in situ* sample. Filters were pooled across like triplicate bottles by treatment, totaling 60 L from each of the incubation treatments. The total RNA sample was then enriched for eukaryotic mRNA through a poly-A pull down. The resulting enriched mRNA sample then went through library preparation with the Illumina TruSeq mRNA Prep Kit (Illumina). Libraries were sequenced with the Illumina HiSeq2000 at Columbia Genome Center (New York, NY). Each sample was sequenced to produce a targeted 60 million, 100 base pair, paired end reads. These environmental sequence data are deposited in the Sequence Read Archive (SRA) through the National Center for Biotechnology Information (NCBI) under accession number SRP056385. Raw sequence data quality was visualized using FastQC and then cleaned and trimmed using Trimmomatic v

0.27 (paired end mode; 6-base pair wide sliding window for quality below 20; minimum length 25 base pair).

4.3.3 Genome database creation and mapping

Environmentally relevant algal genomes (*Aureococcus anophagefferens* CCMP1984 v1.0, *Emiliania huxleyi* CCMP1516 v1.0, *Fragilariopsis cylindrus* CCMP1102 v1.0, *Micromonas pusilla* RCC 299 v3.0, *Ostreococcus lucimarinus* v2.0, *Ostreococcus tauri* v2.0, *Phaeodactylum tricornutum* CCMP632 v2.0, *Pseudo-nitzschia multiseries* CLN-47 v1.0, *Thalassiosira pseudonana* CCMP1335 v3.0) were collected from the Joint Genome Institute (JGI) database and concatenated. Trimmed, paired-end reads from each of the samples were mapped to this concatenated genome library using the Burrows-Wheeler Aligner (Li and Durbin, 2010) (BWA-mem, parameters: -k 10 -aM) and then counted using the HTSeq 0.6.1 package (Anders et al., 2014).

4.3.4 MMETSP database creation and mapping

Transcriptome sequences and annotations generated through the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) that were made public as of 17 March 2014 were collected, representing 401 transcriptomes across 209 species or cultured isolates. Transcriptomes from like species (regardless of strain or condition) and cultured isolates were pooled and clustered using CD-HIT-EST (98% identity; word size of 9). The resulting clustered set of transcripts was considered to be the representative transcriptome for the species or cultured isolate. Transcriptomes were annotated with KEGG Orthology annotations using the bi-directional best hit (BBH) method through the KEGG Automatic Annotation Server (KAAS) (Moriya et al., 2007). It is worth noting that the KEGG module names are human defined and some genes are artificially grouped in the context of phytoplankton metabolism. For example, this is the case with the "methane metabolism" module, which includes a suite of genes related to carbon not specifically methane metabolism. The 209 transcriptomes created in this manner were concatenated to form a comprehensive species-level transcriptome database from the MMETSP library. Due to the large size of the resulting MMETSP database, trimmed reads were mapped to the MMETSP using the Burrows-Wheeler Aligner (Li and Durbin, 2010) (BWA-mem, parameters: -k 10 -aM) and then counted using the HTSeq 0.6.1 package (Anders et al., 2014).

4.3.5 Differential expression analysis

Counts obtained from HTSeq were pooled for like-KEGG orthologs across all species in a functional group. The quantitative metabolic fingerprint (QMF) was assessed by normalizing global patterns of expression at the module-level to the total mapped reads, an approach similar to those used in several metagenomic and metatranscriptomic studies focused on prokaryotes (Shi et al., 2011; Ottesen et al., 2014; Shi et al., 2012). PCA and confidence ellipses of the QMF signals by functional group and sample type (*in situ*, no addition control, and DSW addition) were calculated and visualized using FactorMineR package in R (Figure 4.2, Figure C.7). No significant difference was seen between the *in situ* and no addition control samples (Figure C.7). For each functional group, the pooled KEGG counts from the *in situ* samples (S1-S3) were compared to those from the corresponding DSW amendment (E1-E3) using Analysis of Sequence Counts (ASC), an empirical Bayes method (Wu et al., 2010). Genes were considered to be differentially abundant between treatments if for a

fold change of 2.0 the posterior probability (post-p) was greater than 0.95 (Dyhrman et al., 2012). Patterns of differential abundance were visualized using Circos (Krzywinski et al., 2009). Global shifts in the expression of genes independent of functional group were assessed with TMM normalization using the Microbial Assemblage Normalized Transcript Analysis package (MANTA, v. 1.12.0)(Marchetti et al., 2012).

4.3.6 Variable transcript allocation modeling

Variable transcript allocation following DSW amendment was calculated for each functional group. Though there was a normal distribution of log fold change across all functional groups, the means were off-set for the diatoms and the haptophytes (Figure C.4). From the set of all genes (G), the genes which had statistically significant increased transcript abundance (Equation 4.1) and decreased transcript abundance (Equation 4.2) as identified with ASC (2 fold change, post-$p > 0.95$) (Wu et al., 2010) in DSW amended treatments (E) relative to the *in situ* sample (S) were considered.

$$U = \{g : post - p(\frac{T_{E,g}}{T_{S,g}} > 2) > 0.95, g \epsilon G\} \tag{4.1}$$

$$D = \{g : post - p(\frac{T_{S,g}}{T_{E,g}} > 2) > 0.95, g \epsilon G\} \tag{4.2}$$

A variable transcript allocation score (VTA) was calculated for both the set of genes with both increased (Equation 4.3) and decreased (Equation 4.4) abundance, taking the ratio of the summed transcripts per million (T) for the *in situ* (S) and experimental DSW amended treatments (E) of every gene (u or d) within the set of significantly differentially abundant genes (U or D). VTA scores were calculated so as to always be greater than one, thus the VTA_{Dn} sums the reciprocal of the ratio summed in VTA_{Up} (Equations 4.3 and 4.4). VTA_{Dn} is the magnitude of the decreased transcriptional response following DSW addition, while the VTA_{Up} is the magnitude of the increased transcriptional response following DSW addition. VTA_{Up} and VTA_{Dn}, as ratios, focus on the total transcript pool shifted between S and E rather than the number of genes with differential abundance. As such, we can directly compare these two ratios with $VTAR$ (Equation 4.5) and assess the metabolic trait of reallocation efficiency. If $VTAR > 1$, there was a larger transcript pool (TPM) in genes that were increased than were decreased, indicating an efficient reallocation of the transcript pool. By contrast, if $VTAR < 1$, less TPM was increased than was decreased. This model defines the total metabolic responsiveness as a trait that can be compared between the functional groups.

$$VTA_{Up} = \sum_{u \epsilon U} \frac{T_{E,u}}{T_{S,u}} \tag{4.3}$$

$$VTA_{Dn} = \sum_{d \epsilon D} \frac{T_{S,d}}{T_{E,d}} \tag{4.4}$$

$$VTAR = \frac{VTA_{Up}}{VTA_{Dn}} \tag{4.5}$$

4.4 Results and Discussion

We leveraged increases in reference sequence availability (Keeling et al., 2014) to identify the expressed metabolic capabilities of different phytoplankton functional groups. Total RNA from the eukaryotic community ($> 5\mu m$) of the surface mixed layer at Station ALOHA was sequenced (60 million 100 base pair, poly-A selected, paired end reads) from three samples collected during August-September 2012 (S1-S3). To better understand the underlying functional group-specific metabolism associated with blooms, three blooms were simulated via on-deck incubations. While there are many potential nutrient inputs or bloom drivers (e.g. biologically fixed nitrogen), these blooms were simulated through the addition of deep seawater (DSW) to mimic nutrient loading by upwelling. These incubation experiments were modeled after McAndrew et al. (2007) and were performed in conjunction with each of the in situ samples, amending surface water with a 10% mixture by volume of water from below the nutricline (700 m). Sequence reads from the in situ samples (S1-S3) and the bloom simulations (E1-E3) were mapped to 1) non-symbiotic microalgal genomes and 2) a custom database comprised of all publically available transcriptomes as of 17 March 2014 from the Marine Microbial Eukaryotic Transcriptome Project (MMETSP) (Keeling et al., 2014).

The MMETSP provides a 50x expansion of molecular data across the eukaryotic tree of life, which both better reflects the broad diversity within the protists and adds higher-definition coverage for well-studied groups such as diatoms and dinoflagellates (Keeling et al., 2014). Our leveraging of this database for the pipeline developed herein enabled unprecedented identification of taxonomic composition (58-62% of reads identified) compared to mapping the same dataset to genomes (12-14%), which have been used in previous studies (Marchetti et al., 2012). Due to the taxonomic bias of the non-symbiotic algal genomes, the majority of reads from the in situ eukaryote community mRNA samples were annotated as diatoms or haptophytes. When mapped to the custom MMETSP database, however, dinoflagellates dominated, constituting 36-40% of the mapped reads, with the diatoms and haptophytes the next most highly represented functional groups (Figure 4.1A). The dominance of dinoflagellates in the 5 Îijm size fraction, confirmed in historic surface pigment analyses at Station ALOHA (Letelier et al., 1993), stands in contrast to previous eukaryotic metatranscriptome studies in the oligotrophic ocean where they accounted for $< 5\%$ of reads (Marchetti et al., 2012). Although clearly present and important to the community, dinoflagellate read abundance here and in sediment trap data collected during the same cruise series (Fontanez et al., 2015) may be magnified by their large transcript pool (Moustafa et al., 2010; Hackett et al., 2004).

The species composition of the functional groups, reflected in rank abundance (Figure 4.1B), was highly conserved across all three in situ samples (S1-S3), underscoring the stability of phytoplankton populations in this well-studied oligotrophic system. Between 18.1 and 20.7% of reads mapped to the MMETSP database were annotated with KEGG orthology, elucidating differences in the mRNA distribution between functional groups at the pathway level. Looking at the module-level, the general distribution of transcripts in proportion to the total was assessed using quantitative metabolic fingerprinting (QMF) (Alexander et al., 2015a). Diatoms have a larger proportion of mRNA in the transport-related system (e.g. metallic cation, B12, phosphate, and amino acid) compared to haptophytes and dinoflagellates (Figure 4.2A). Purine metabolism was consistently a large component of haptophyte QMF (5.6-27%), an order of magnitude higher than diatoms and dinoflagellates (1-2%). Purine nucleotides may represent a source of DON accessible to haptophytes, as haptophytes have been found to grow on purines as their sole N source (Palenik and Henson,

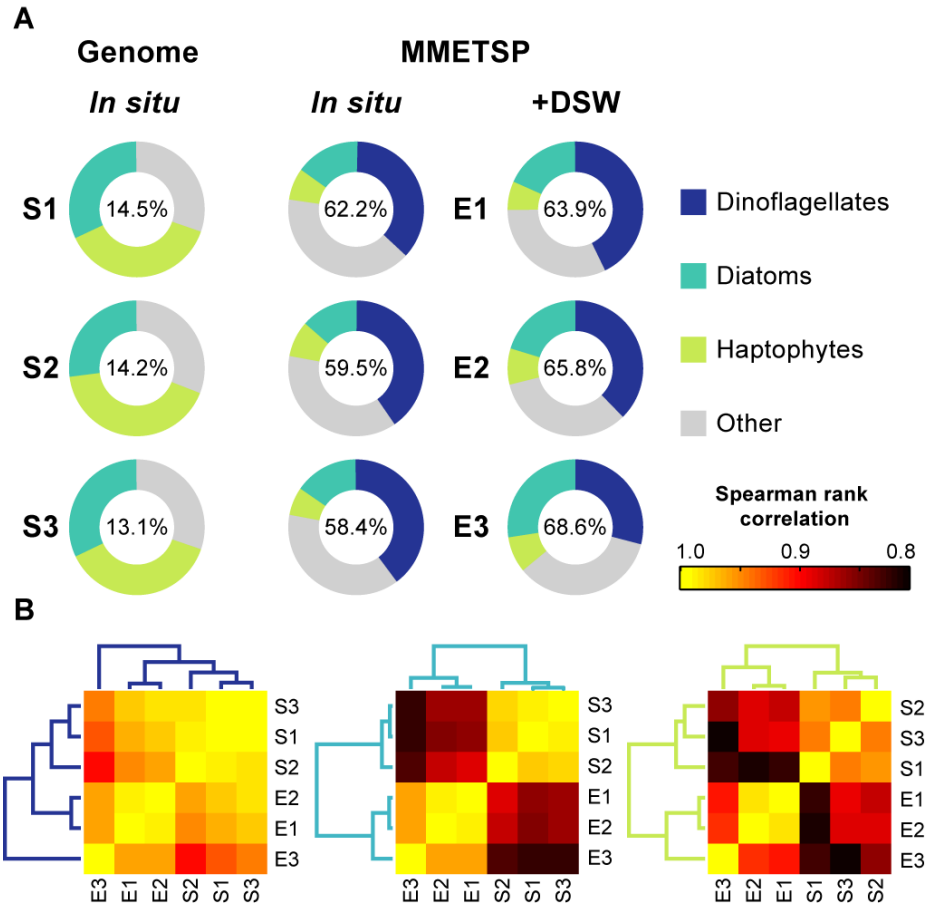


Figure 4.1: Taxonomic distribution in mRNA mapped reads consistent across time but altered by deep seawater (DSW) addition. Sequences collected during the summer of 2012 at Station ALOHA (S1: 6 August, S2: 24 August, S3: 2 September) and corresponding deep seawater (DSW) incubation experiments (E1-E3) were mapped to two custom databases: 1) non-symbiotic microalgal genomes and 2) all freely available transcriptomes from the MMETSP as of 17 March 2014. (A) Taxonomic affiliation of reads across the three most abundant functional groups: dinoflagellates, diatoms, and haptophytes mapped to both the genome and MMETSP databases for S1-S3. The corresponding DSW addition incubations E1-E3 were only mapped the MMETSP database. The percent of total reads mapped is indicated inside each of the circles. (B) Spearman rank correlation for species composition shifts within each of the three functional groups across S1-S3 and E1-E3.

1997). As the precursors for nucleic acid biosynthesis, purine uptake in the ocean has also been attributed to nucleotide salvage (Winn and Karl, 1984). These functional group differences observed at the module-level are underscored by principle component analysis (PCA) with the QMF for each functional group differing with 95% confidence (Figure 4.2B) and are suggestive of metabolic partitioning between functional groups. Within a functional group, the QMF was stable across time (Figure 4.2A), in contrast to the variability observed in coastal systems over similar time scales (Alexander et al., 2015a; Dupont et al., 2015). This stability likely reflects both the unique physiological attributes of oligotrophic phytoplankton as well as the comparatively static geochemical environment.

The static nature of population structure and functional group QMF was altered by replicated nutrient-rich DSW addition experiments, which led to a 7- to 17-fold increase in chlorophyll a, increases not observed in the control treatment (Figure C.1). This increase was consistent with previous studies, which also noted increases in diatoms (McAndrew et al., 2007). Diatom-associated mRNA reads increased in each of the DSW experiments (Figure 4.1A). Although species designations can be influenced by the composition of the database used for read mapping, apparent taxonomic shifts occurred at the species-level for diatoms and haptophytes (Figure C.2). Shifts in taxonomic composition were consistent with DSW addition for diatoms and haptophytes, with rank abundance clustering for all experiments (Figure 4.1B). Taxonomic shifts in diatoms were driven by an increase in the rank abundance of certain species, which in all cases were pennate forms (Figure C.2), including genera known to be present in the NPSG like Pseudo-nitzschia (Silver et al., 2010) and common in oligotrophic nutrient amendment incubation studies (Marchetti et al., 2005, 2012). Although species shifts also occurred within the haptophtyes, *Emiliania huxleyi* was always the most dominant taxon (Figure C.2). The shifts in diatom dominance compared to the consistent dominance of *E. huxleyi* may reflect differences in evolutionary strategies, with metabolic diversity spread across many species in the diatoms and a single species complex with a pangenome, *E. huxleyi*, in the haptophytes (Read et al., 2013). The QMF of DSW addition was significantly different from the QMF of the in situ community for both diatoms and haptophytes, but not dinoflagellates (Figure 4.2B). Following DSW addition, the QMF for both diatoms and haptophytes was characterized by increased expression of modules associated with growth, such as carbon fixation (Figure 4.2A). These shifts are not the result of changes in species composition, as the patterns of expression from individual species tracked the summed community (Figure C.3). The lack of change in the QMF of dinoflagellates (Figure 4.2B) likely reflects their range of life strategies (Hackett et al., 2004) and minimal transcriptional regulation of gene expression as observed in culture-studies (Moustafa et al., 2010).

Variability within the QMF modules was resolved by statistically assessing the changes in abundance of individual genes for each functional group using a Bayesian approach (Wu et al., 2010) (Figure 4.2C, Figure C.4). Statistical significance (2-fold change, posterior probability (post-p) > 0.95) of differential abundance was examined for 4038 KEGG orthologs common to diatoms, haptophytes, and dinoflagellates. As with the QMF (Figure 4.2A), dinoflagellates demonstrated little to no significant changes in gene expression (Figure 4.2C). The suites of transcripts with significantly increased transcript abundance following DSW were highly conserved for both diatoms and haptophytes across the three replicate experiments (40% of 334 and 19% of 490 genes common, respectively) but differed between functional groups (41 of the total 824 significantly increased transcripts genes common) (Figure C.5). Of the genes with increased transcript abundance many of those conserved across all three experiments for diatoms and haptophytes were associated with growth (e.g. ATP

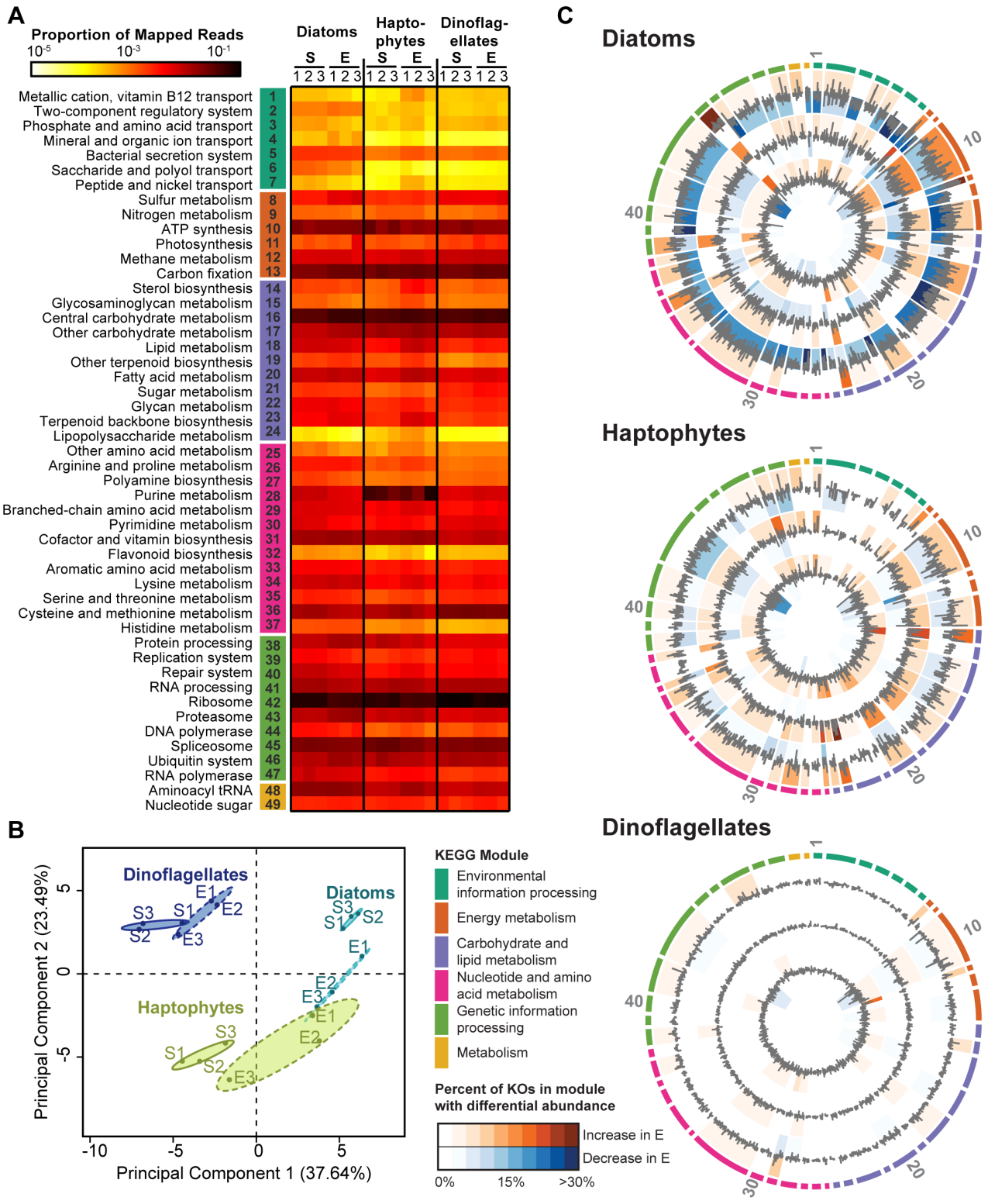


Figure 4.2: Quantitative metabolic fingerprint (QMF) and patterns of differential expression across KEGG orthology following DSW addition underscore functional group traits. Caption continued on following page.

Figure 4.2: Quantitative metabolic fingerprint (QMF) and patterns of differential expression across KEGG orthology following DSW addition underscore functional group traits. (A) The relative metabolic partitioning of the mRNA pool across the three *in situ* samples (S1-S3) and corresponding deep seawater (DSW) incubation experiments (E1-E3) was assessed using QMF. The summed proportion of mapped reads falling into each of the KEGG modules is depicted as a heat map. (B) Principal component analysis of the QMF signals for each of the functional groups across S1-S3 and E1-E3; 95% confidence ellipses are indicated for each of the sample types by functional group. (C) Log fold change and significance of differential expression between deep seawater (DSW) amendments and *in situ* samples for KEGG orthologs is visualized with Circos (Krzywinski et al., 2009) for the diatoms, haptophytes, and dinoflagellates. Outermost ring colors indicate the KEGG super module, with individual wedges of the pie corresponding to KEGG modules as numbered in A. Concentric circles indicate the expression of the three, replicated DSW addition experiments compared to *in situ* samples: E3 (outer), E2 (middle), E1 (inner). The log fold change of individual KEGG orthologs is depicted as a bar plot bounded -3 to 3. The background color of individual KEGG modules identifies the percentage of genes within module that were significantly (2 fold-change, post-$p > 0.95$) increased (orange) or decreased (blue) in abundance, where darker colors indicate that a higher percentage of genes within that module were significantly different.

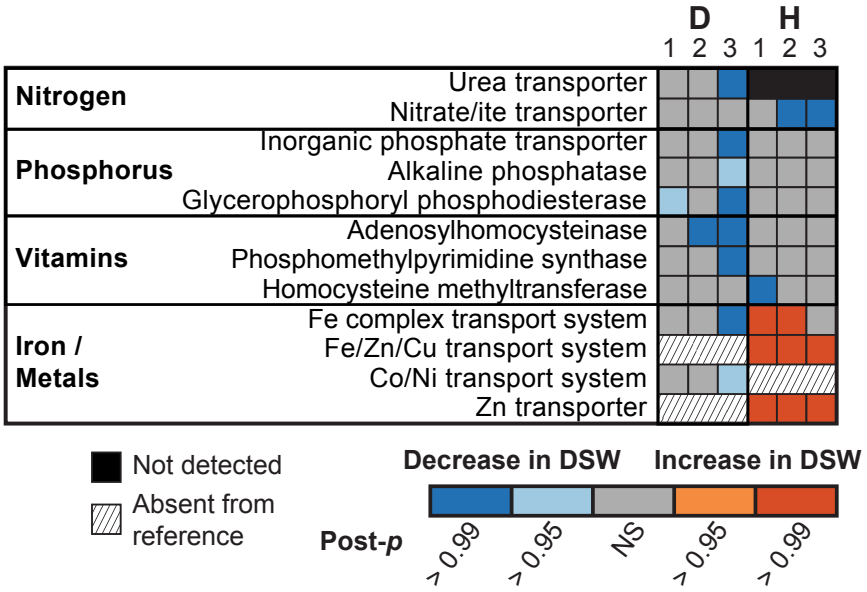


Figure 4.3: Shifts in transcript abundance of genes responsive to biogeochemical forcing. The significance of changes in abundance (2 fold-change, post-$p > 0.95$, or > 0.99) for genes known to be associated with N, P, vitamin, Fe, or other trace metals metabolism for diatoms (D) or haptophytes (H) is indicated as blue (decrease) or orange (increase). Genes present within the reference transcriptome, but not detected in the field were marked in black, and genes absent from the reference are hashed. KEGG IDs are as follows: Urea transporter (K11959), nitrite/ate transporter (K02575), phosphate transporter (K08176), glycerophosphoryl diester phosphodiesterase (K01126), adenosylhomocysteinase (K01251), phosphomethylpyrimidine synthase (K03147), 5-methyltetrahydrofolate-homocysteine methyltransferase (K00548), iron complex transport system (K02013), iron/zinc/copper transport system (K11706), cobalt/nickel transport system (K02006), zinc transporter (K14715).

synthesis (10), photosynthesis (11), and carbon fixation (13)) (Figure 4.2C, Fig S5). Addi-

tionally, following DSW addition, diatoms had signals indicative of the incorporation of both nitrogen (increasing abundance of glutamate and glutamine synthase) and iron (switching from NADPH to ferredoxin sulfate reductase). These changes in transcriptional patterns indicate that both diatoms and haptophytes increase fundamental metabolic processes required for photosynthetic growth in response to DSW.

For diatoms and haptophytes, the consistency in genes with increased transcript abundance stands in contrast to the patterns of genes with significantly decreased transcript abundance following DSW addition (Figure C.5). Diatoms were typified by significant decreases in transcript abundance of many genes following DSW, consisting of a large portion (1389 genes) of their metabolism compared to haptophytes (490 genes) (Figure 4.2C). Genes with decreased transcript abundance were variable across the three experiments (1.5% and 6.9% similar for diatoms and haptophytes, respectively) (Figure C.5) and imply a tailoring of basal metabolism to the change in biogeochemical environment from DSW amendment. Across the KEGG modules, specific genes known to be markers of nutrient limitation significantly decreased in abundance following DSW addition (Figure 4.3), potentially signifying a limitation in the in situ population that was alleviated following resupply. Such genes or their protein products are frequently used as proxies to identify limitation in the field (Saito et al., 2014). In E1 and E2, there was a decrease in diatom-associated glycerophosphoryl phosphodiesterase (Dyhrman et al., 2012) and adenosylhomocysteinase (Bertrand et al., 2012), indicative of phosphorus (P) and vitamin limitation, respectively (Figure 4.3). Silica transporters, though not in KEGG, were surveyed and not found to have significant shifts in abundance. E3 was characterized by a decrease in abundance of many genes indicative of limitation in the diatoms, including but not limited to, a urea transporter (Bender et al., 2012), a phosphate transporter (Dyhrman et al., 2012), and metal transporters (Figure 4.3). This is suggestive of diatom co-limitation in E3, similar to patterns of co-limitation recently observed in picocyanobacteria in the Pacific Ocean (Saito et al., 2014). There was a decrease in transcript abundance for haptophyte-associated nitrate and nitrite transporters in E2 and E3 and homocysteine methyltransferase in E1 (Figure 4.3). This may be indicative of nitrogen and vitamin limitation based on the pattern in other organisms (Bertrand et al., 2012; Bender et al., 2014), but the regulation of these targets in haptophytes is poorly understood. Metal transport proteins were significantly increased for haptophtyes in E1-E3 and indicated metabolic strategies following the addition of resources that differ from diatoms. In contrast to the diatoms, no markers of P limitation, such as the phosphate transporter or alkaline phosphatase in *E. huxleyi* (Dyhrman et al., 2006; Dyhrman and Palenik, 2003; Xu et al., 2006), were significantly decreased for haptophytes, consistent with their known tolerance for P limitation (Lessard et al., 2005). These data evince that diatoms and haptophytes are not under the same biogeochemical controls in situ and employ disparate strategies following DSW addition to capitalize on newly available resources. Being able to identify and characterize multiple markers of limitation in a genera-specific manner for these eukaryotes is of central importance to the modeling of aperiodic blooms of these groups in oligotrophic systems.

The above analyses focus primarily on the gene content and transcriptional patterns; however, the underlying eco-evolutionary metabolic traits for a functional group may be better described by considering the shift in the total transcript pool (TPM). Using the statistically resolved patterns of increases and decreases in transcript abundance, a variable transcript allocation ratio ($VTAR$) was calculated to model functional group transcriptional responsiveness to DSW amendment (Figure 4.4). $VTAR > 1$ indicates an efficient reallocation of the transcriptional potential from genes with decreased abundance to genes with

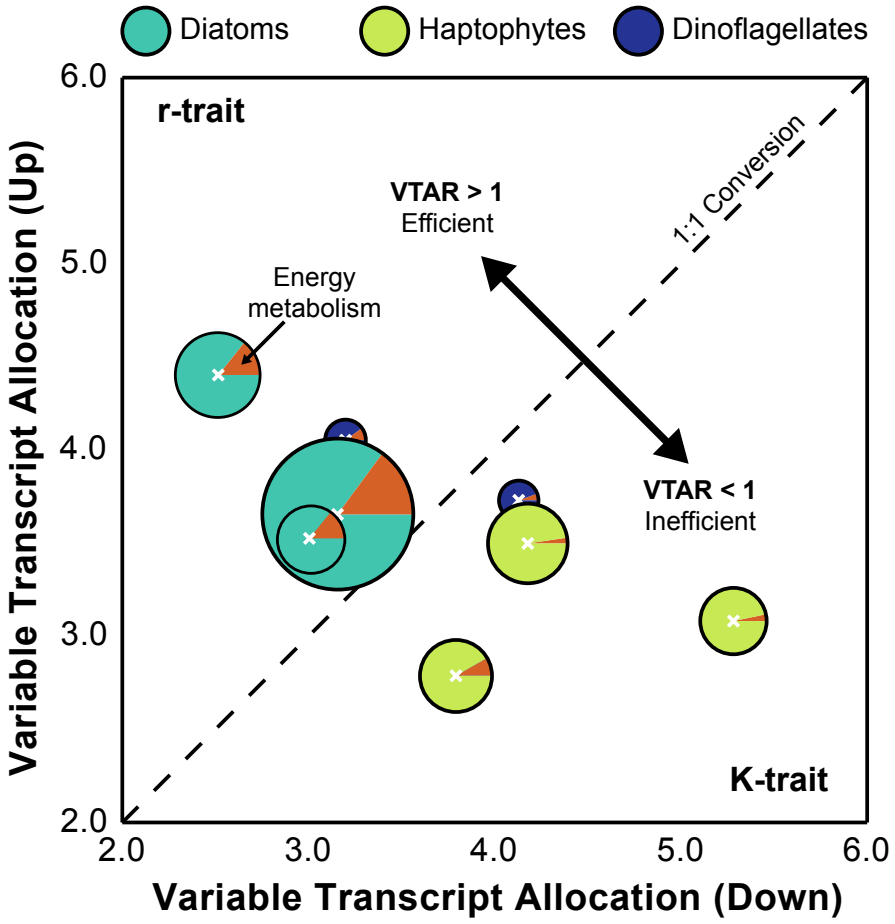


Figure 4.4: Variable transcript allocation space differentiates functional group strategies. The variable transcript allocation score (Equations 4.3 and 4.4) of the genes with significantly increased (VTA_{Up}) or decreased (VTA_{Down}) abundance in the deep seawater (DSW) amendment relative to the *in situ* sample is plotted for diatoms, haptophytes, and dinoflagellates for E1-E3. The size of the pie indicates the total number of genes with significantly different transcript abundances between the *in situ* and DSW amended treatments. The proportion of increased TPM in E within the energy metabolism super group is illustrated as a pie slice in orange.

increased abundance following DSW amendment, while $VTAR < 1$ indicates an inefficient reallocation. The dinoflagellates had variable VTA scores due to their small pool of differentially abundant genes, however the VTA scores consistently resulted in disparate patterns for the diatom and the haptophytes (Figure 4.4). In every experiment, the diatoms fell above the 1:1 line, with a $VTAR > 1$ ($1.15-1.75$), while haptophytes fell below, with a $VTAR < 1$ ($0.58-0.83$) (Figure 4.4). The relative efficiency of reallocation, here defined by $VTAR$, reflects differences in the metabolic traits of these functional groups and aligns with preexisting ecological traits as defined by (Margalef, 1978). Diatoms are r-selected with high maximum uptake rates that enhance their competition under high or fluctuating nutrients (such as a DSW upwelling event). This trait is reflected in the significant decrease in transcript abundance of many genes across a broad metabolic range coupled with the targeted increase of a subset of genes largely falling within the energy metabolism super module (Figure 4.4), a pattern which was also observed with gene-focused analyses using trimmed mean of M value (TMM) normalization (Marchetti et al., 2012; Robinson and Oshlack, 2010) (Figure C.6). Haptophytes are K-selected, possessing a low half-saturation constant that enhances growth under low nutrient conditions, but are unable to capitalize on nutrient pulses like r-selected competitors. Again, this ecological trait is reflected in changes in the haptophyte transcript pool. Though numerically fewer of their genes significantly decreased in abundance with DSW addition, the total TPM represented by those genes with decreased abundance exceeded that of those induced, defined by $VTAR < 1$ (Figure 4.4). This is also reinforced by the gene-specific analysis (Figure C.6). It has been speculated that the mechanistic basis for the r- and K- tradeoff dichotomy in the phytoplankton lies in the disparate investment in growth or resource acquisition machinery (Litchman and Klausmeier, 2008). The large portion of the transcript pool that increased ($VTAR > 1$) for diatoms shows an ability to capitalize on newly available resources, with 14.2-14.9% of the increased TPM in the KEGG energy metabolism super module (Figure 4.4). Haptophytes ($VTAR < 1$), by contrast, do not efficiently reallocate the transcript pool, with only 5.5-10.2% of the increased TPM in energy metabolism (Figure 4.4). In short, haptophytes do not appear to modulate their transcript pool to capitalize on growth processes as efficiently as diatoms (Figure 4.4).

These functional group-specific molecular and metabolic mechanisms underpin the aperiodic eukaryotic phytoplankton blooms in the oligotrophic ocean. Whereas both diatoms and haptophytes, including calcifying groups, likely contribute to a shift towards a net autotrophic condition when there is a nutrient pulse, the ecosystem function of oligotrophic systems may ultimately hinge on the unique trait of the diatoms to more efficiently turn over their scavenger metabolism to one of enhanced production. This finding is consistent with the dominance of diatom-associated BSi export relative to PIC export during summer in the NPSG (Karl et al., 2012). Unlike the preceding 13 years of study at Station ALOHA (Karl et al., 2012), enhanced production and export characteristic of bloom events were not observed during the summer of 2012, which exhibited a period of sustained net-heterotrophy. We demonstrated through simulated blooms that the metabolic capacity for enhanced production is inherent in the large eukaryotic phytoplankton regardless of water mass, suggesting that the lack of bloom in 2012 was variably due to deficiency in macronutrients, vitamins, and metals. As the conditions observed during summer 2012 may be increasingly encountered in a future ocean (Doney et al., 2012), modeling the molecular traits and tradeoffs of these populations will help better predict ecosystem state and metabolic balance of the ocean.

Chapter 5

Physiological response and strain variation of the Emiliania huxleyi species complex under changing nutrient environments

5.1 Abstract

Phytoplankton are well tuned to respond to changing environments, which may happen at the community-level with functional group succession, at the species-level through shifts in strain composition, or at the strain-level through alterations to phenotype. Community-level shifts have been well described; however, strain or phenotypic shifts have been more difficult to identity and describe in the field. Here, we examined the intersecting roles of metabolic plasticity and strain diversity in the response of natural populations of the biogeochemically significant cocolithophore *Emiliania huxleyi* to shifting nutrient regimes in the North Pacific Subtropical Gyre (NPSG). Using a metatranscriptomic approach, field observations were paired with microcosm studies to track the compositional and metabolic responses to shifts in the geochemical environment. The transcriptomes and genome of five strains were clustered based on protein homology to identify the 'core' set of genes common across strains, as well as sets of genes unique to each strain. These strain-specific gene sets were used to track strain composition in the field and microcosms. The strain composition of the *in situ* samples varied little over the sampling period, with transcripts specific to strains CCMP1516, CCMP370 and PLYM219 being the most abundant. Following the addition of nitrogen, however, transcripts specific to strains CCMP374 and CCMP379 exhibited dramatic increases. In addition to the variations in strain diversity observed following nutrient addition, significant changes in transcript abundance were observed for gene pathways involved in nitrogen, and phosphorus metabolism. The data suggest that nitrogen is a major driver of the physiological ecology of *E. huxleyi* in this system, and nitrogen supply may be linked to shifts in the ploidy of the population and changes in both nutrient physiology and calcification state. Together, these data underscore the ecological importance of the "pan genome" of *E. huxleyi*, suggesting that genetic variability within the species complex combined with global metabolic plasticity may be at the heart of its success in a wide variety of marine environments.

5.2 Introduction

Central to the carbon cycle, marine phytoplankton are estimated to constitute nearly half of global primary productivity (Field et al., 1998). Beyond their contributions to primary production (1-10% of total marine carbon fixation), cocolithophores are an important source of particulate inorganic carbon in the form of calcite ($CaCO_3$) and are estimated to comprise about 50% of calcite deposition to sediments (Poulton et al., 2007). Consequently, cocolithophores play a dual role in the cycling of carbon, both in the organic carbon pump, drawing CO_2 out of the atmosphere, and the carbonate counter pump, where CO_3^{2-} removed for calcification increases total alkalinity thus leading to a positive feedback on atmospheric pCO_2 (Zondervan et al., 2002). The ratio of calcification to carbon fixation has been found to vary across environmental factors such as temperature, salinity, light and nutrients (Paasche, 2001; Bollmann and Herrle, 2007; Zondervan, 2007; Feng et al., 2008).

Numerically, *Emiliania huxleyi* is the most abundant cocolithophore species in the modern ocean (Paasche, 2001), known for its cosmopolitan distribution and ability to form large blooms in both eutrophic coastal regions and oligotrophic open ocean regions (Holligan et al., 1993; Brown and Yoder, 1994). Margalef (1978) put forth a hypothesis that such blooms by cocolithophores (K-selected) follow upon the boom-bust dynamics of diatoms (r-selected), where diatoms bloom quickly depleting the environment of key nutrients such as nitrogen (N), phosphorus (P), and silica (Si). In these low nutrient environments, it is frequently *E.*

huxleyi, which is thought to be the most r-selected of the coccolithophores, that thrives and blooms (Litchman et al., 2006). *E. huxleyi* is known to be well-adapted to such low-nutrient environments, having the ability to alter its metabolism to scavenge nutrients from organic compounds (Palenik and Henson, 1997; Dyhrman and Palenik, 2003; Bruhn et al., 2010; Rouco et al., 2013). This metabolic plasticity may be central to its ability to adapt to the environment and form large blooms.

The potential effects of rising atmospheric CO_2 (Raven et al., 2005; Meinshausen et al., 2011) on the production and calcification of this keystone phytoplankton group is debated and has been found to vary both across environmental parameters, such as the nutrient environment (Sciandra et al., 2003; Leonardos and Geider, 2005; Rouco et al., 2013), and amongst strains (Riebesell et al., 2000; Iglesias-Rodriguez et al., 2008; Langer et al., 2009). Beyond the variable response to carbonate chemistry, inter-strain variability has been observed in the ability to grow on various organic substrates (Strom and Bright, 2009) and in enzymatic activities (Steinke et al., 1998; Dyhrman and Palenik, 2003; Alcolombri et al., 2015). The capacity for this variability was largely revealed through the genome sequencing of many *E. huxleyi* strains (Read et al., 2013). Historically believed to be a single species, the genome of *E. huxleyi*, termed a "pan genome", was found to be highly variable across strains, not only in microsatellite regions, but in gene content, with up to 25% coding regions variable across species (Read et al., 2013). Such genomic variability has been described in a cosmopolitan cyanobacterial species (Kashtan et al., 2014) and may be central to the success of these species in diverse environmental conditions (Biller et al., 2014), but had not been previously described for a marine microeukaryote. While significant diversity as described by non-coding microsatellite loci has been observed in the field (Iglesias-Rodriguez et al., 2006), the variability of gene content as seen across various isolates has not been directly observed *in situ*. Global surveys of mixed communities have suggested that the known diversity of *E. huxleyi* may be a cornerstone of its future response to changing ocean conditions (e.g. increased stratification and acidification) (Beaufort et al., 2011).

In spite of the intersecting importance of metabolic plasticity and strain diversity in the ecology of *E. huxleyi*, these dynamics have not yet been assessed in the field, being primarily limited to monoculture studies of individual isolates. Using a metatranscriptomic approach, the relative contribution and activity of different strains of *E. huxleyi* and their combined physiological signature were tracked both *in situ* in the North Pacific Subtropical Gyre and under altered nutrient conditions through replicated microcosm incubations.

5.3 Materials and Methods

5.3.1 Sample collection and shipboard nutrient incubation experiments

Seawater for the *in situ* eukaryote community mRNA analysis was collected at the HOT, Station ALOHA (22°45′ N, 158°00′ W) from a depth of 25 m at 1400 hrs (local time) on six occasions during the summer of 2012 (S1: 6 August, S2: 12 August, S3: 24 August, S4: 30 August, S5: 2 September, S6: 5 September) using a Eulerian sampling scheme as part of the HOE-DYLAN research expedition as per Alexander et al. (2015b). Water was collected in acid-washed 20-L carboys and approximately $60L$ of seawater was prescreened through $200\mu m$ mesh and then filtered onto polycarbonate filters ($5.0\mu m$ pore size, $47mm$, Whatman) by way of peristaltic pump. Filters were changed every 20 minutes or when flow rate decreased. Filters were placed in cryovials and stored in liquid nitrogen until mRNA extraction. The total length of filtration time did not exceed 3 hours.

In conjunction with these field-based surveys, two factorial nutrient amendment incubation experiments focused on the macronutrients N and P were performed with natural communities (T_0 of E1 was S1 and T_0 of E2 was S4) (STable 1). Incubations were modeled after a simulated 10% deep seawater (DSW) upwelling as described in Alexander et al. (2015b) and designed to tease apart the potential nutritional components of DSW upwelling. The concentration of iron was modeled after Marchetti et al. (2012) and vitamin B_{12} was modeled after Bertrand et al. (2007). Triplicate 20-L carboys of each treatment were incubated at 30% surface light-levels using on-deck incubators for 7 days and processed as described above, on the final day at 1400 hrs (local time). Nutrient concentrations for phosphate [PO_4], nitrate and nitrite [NO_2+NO_3] were measured by filtering $125mL$ of seawater through a $0.2\mu m$, $47mm$ polycarbonate filter, and stored frozen (-20°C) in acid washed bottles until analysis at the Chesapeake Bay Lab at the University of Maryland according to the facility's protocols. Samples for alkaline phosphatase activity (APA) were collected by filtering 250-ml of whole seawater onto polycarbonate filters ($0.2\mu m$ pore size, $47mm$, Whatman) and frozen at -20°C. These filters were then resuspended in artificial seawater and assayed for APA fluorometrically using the fluorogenic phosphatase substrate 6,8-difluoro-4-methylumbelliferyl phosphate (diMUF-P, Molecular probes) following established field protocols (Dyhrman et al., 2006). Chlorophyll a was measured on whole water samples collected onto GF/F filters ($25mm$, Whatman) using a 90% acetone extraction and assayed by fluorescence using the AquaFluor Turner TD700 (Parsons et al., 1984).

5.3.2 RNA extraction and sequencing

RNA was extracted from individual filters with the RNeasy Mini Kit (Qiagen), following a modified version of the yeast protocol. Briefly, lysis buffer and RNA-clean zirconia/silica beads was added to the filter and samples were vortexed for 1 minute, placed on ice for 30 seconds, and then vortexed again for 1 minute. Samples were then processed following the yeast protocol. The resulting RNA was eluted in water and then treated for possible DNA contamination using TURBO DNA-free Kit (Ambion) following the Rigorous DNase protocol. RNA from individual filters was then pooled by sample, using the RNA Cleanup Protocol from the RNeasy Mini Kit (Qiagen). The resulting RNA sample thus represented approximately $56L$ of total seawater for the *in situ* sample. Filters were pooled across like triplicate bottles by treatment, totaling $56L$ from each of the incubation treatments. The total RNA sample was then enriched for eukaryotic mRNA through a poly-A pull down. The resulting enriched mRNA sample then went through library preparation with the Illumina TruSeq mRNA Prep Kit (Illumina). Libraries were sequenced with the Illumina HiSeq2000 at Columbia Genome Center (New York, NY). Each sample was sequenced to produce a targeted 60 million, 100 base pair, paired end reads. Raw sequence data quality was visualized using FastQC and then cleaned and trimmed using Trimmomatic v 0.27 (paired end mode; 6-base pair wide sliding window for quality below 20; minimum length 25 base pair).

5.3.3 Community- and strain-specific mapping and expression analysis

Transcriptome sequences and annotations generated through the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) that were made public as of 17 March 2014 were collected and treated as per Alexander et al. (2015b) to track species composition of the metatranscriptomes. Due to the large size of the resulting MMETSP

database, trimmed reads from the metatranscriptome were mapped to the MMETSP using the Burrows-Wheeler Aligner (Li and Durbin, 2010) (BWA-mem, parameters: -k 10 -aM) and then counted using the HTSeq 0.6.1 package (Anders et al., 2014).

The combined transcriptomes (as assembled from the NCGR on 4 September 2013) from unialgal cultures of *Emiliania huxleyi* strains CCMP374 (MMETSP1006-MMETSP1009), CCMP379 (MMETSP0994-MMETSP0997), CCMP370 (MMETSP1154-MMETSP1157), and PLYM219 (MMETSP1150-MMETSP1153). All transcriptome assemblies used are available through the iMicrobe data commons. Additionally, the predicted transcripts from the *E. huxleyi* genome, strain CCMP1516, were used. All transcriptomes were trimmed based on predicted peptide length, requiring sequences be longer than 70 amino acids. The resulting set of genes was considered for subsequent analyses. Peptide sequences were clustered into gene clusters with orthoMCL (Li et al., 2003), using standard parameters: BLASTP with an e-value cutoff of 1e-5, and an inflation value (-I) of 1.5. Initially, the transcripts unique to CCMP1516, here surveyed using the predicted transcripts from the genome, were the most dominant of the subsets of genes in these analyses, representing 50% of the *E. huxleyi* reads in the field (Supplemental Figure Core Changes). Closer inspection demonstrated that many of the most highly represented genes identified as unique to CCMP1516 were associated with metabolic stasis or senescence (e.g. OG1_5_1124, a group of homologous proteins in the *E. huxleyi* genome such as JGI #413698 annotated as putative senescence-related proteins and highly expressed in all field samples). Many of the proteins in the unique set of CCMP1516 were identified as "core" amongst the 13 strains surveyed by Read et al. (2013), yet were absent in some or all of the transcriptomes of the four strains in this study. This absence likely is related to the fact that these strains were largely sampled under exponential growth conditions, limiting the expression of genes that might be associated with stressors or stasis. The lack of 'core' gene representation in some of these transcriptomes underscores the importance of growth condition in transcriptome completeness.

Using this clustering framework, field and incubation samples were mapped to the data set using RSEM, a software package designed to estimate gene and isoform expression values from RNA-seq data. Here we define orthologous groups as genes and individual transcripts (from any strain) as isoforms. Data were mapped using RSEM version 1.2.20 (parameters: –paired-end -p8 -bowtie2 -bowtie2-mismatch-rate 0.2)[1]. Taking a conservative approach, the RNA abundances from like treatments (each consisting of pooled triplicate bottles), which were run with different communities from separate water masses more than two weeks apart, were considered to be biological replicates for differential abundance analysis. These analyses were run with edgeR using default parameters to calculate dispersion and to assess differential abundance of both individual transcripts and orthologous groups of each of the amended incubations compared to the no-addition control. Looking to previous literature, genes thought to be associated with nitrogen and phosphorus metabolism (Dyhrman et al., 2006; Rokitta et al., 2014; McKew et al., 2015) and with calcification and ploidy state (von Dassow et al., 2009; Mackinder et al., 2011; Frada et al., 2012) were compared against the translated proteins comprising the orthologous groups used in this study (tblastn with an e-value cutoff of 1e-20).

[1]A note: RSEM is not yet able to deal with gapped mapping, such as enabled by bwa, which was used for the community-level mapping due to database size constraints.

5.4 Results and Discussion

Total mRNA ($> 5.0\mu m$) from the surface mixed layer at Station ALOHA in the NPSG was deeply sequenced six times during the summer of 2012, following a Eularian sampling scheme. To perturb the nutrient environment of the community, two identical microcosm incubation experiments were conducted with natural populations approximately two weeks apart. These experiments included a no addition control, a 10% (v/v) deep ($700m$) seawater (DSW) treatment to simulate upwelling, and four treatments designed to skew the nitrogen (N) : phosphorus (P) ratio (Supplemental Table Nutrients Added, Supplemental Figure Nutrient Concentrations TF). Sequence reads from the *in situ* and experimental treatments were conservatively mapped to a custom database comprised of all publicly available transcriptomes in the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP) (Keeling et al., 2014). Data were also mapped separately to a curated set of available *E. huxleyi* transcriptomes (Keeling et al., 2014)) and to the genome (Read et al., 2013) encompassing five genetically distinct strains: CCMP1516, CCMP370, CCMP374, CCMP379, and PLYM219 (Supplemental Table, Strain information).

5.4.1 Diatom and haptophyte community structure

The taxonomic composition of the RNA pool was used to examine diatom and haptophyte populations over time and in the microcosms. Population structure, as measured with Spearman rank abundance, was stable across all *in situ* samples for both the haptophytes and the diatoms (Figure 1A). Following N-addition, however, a distinct shift in the population structure of diatoms and haptophytes occurred, with treatments to which N was added (+N, -P, +DSW) clustering separately from those to which no N was added (*in situ*, control, +P, -N) with UPGMA distance between the two clusters of 0.45 and 0.5 for diatoms and haptophytes, respectively (Figure 1A). In diatoms, clustering was also observed between treatments to which trace metals were added in addition to N (-P, +DSW) compared to treatments to which only N was added (UPGMA distance of 0.28) (Figure 1A). This pattern was not observed in haptophytes, where the -P treatments clustered separately from the +N and +DSW treatments (Figure 1A). Following N-addition there was an increase in RNA reads associated with *Emiliania* relative to its abundance in the control, but no other haptophyte became the most abundant (Figure 1C, D). While *E. huxleyi* was consistently the most abundant haptophyte (Figure 1C, D), the dominant diatom changed from Pseudo-nitzschia to Cylindrotheca when N and trace metals were both added (Figure 1C, D). This pattern was strikingly consistent across both microcosms separated in time by several weeks (Fig). The importance of Si and Fe in the structuring of diatom populations is well-documented (Marchetti et al., 2005, 2012), and here treatments containing N and Fe are clearly driving changes in diatom rank abundance at the genus level. A similar pattern was not observed within the haptophytes, where *E. huxleyi* was consistently the most abundant. The strain diversity as manifest in the pan genome of *E. huxleyi* (Read et al., 2013) may be central to the observed dynamics between haptophytes and diatoms, with strains in *E. huxleyi* playing a similar role to the genus in diatoms.

5.4.2 *E. huxleyi* species-complex physiological ecology

Clustering of orthologous proteins was used to isolate the transcript signals from different *E. huxleyi* strains in the field and examine strain-specific physiological ecology. From the

five strains isolated across the world's oceans (Figure 2A) there were a total of 132,888 predicted transcripts (Supplemental Table Strain details). Despite the variable isolation locations, the gene content of these strains covered the same major functional classes with similar relative abundances at the broad level of the KOG class (Supplemental Figure KOG distribution). Grouping these transcripts based on predicted protein homology using OrthoMCL (Li et al., 2003) yielded 56,647 distinct orthologous groups. Orthologous groups varied in size and strain representation, with some groups containing proteins from a single strain and some groups having representative proteins from each of the five strains surveyed here (Supplemental Figure Strain groupings). To better understand the physiology of the *E. huxleyi* species complex, metabolic pathways putatively related to N, P, calcification, and ploidy were tracked for both the 'meta' organism (sum of all strains – orthologous group) and for individual strains.

Taking a conservative approach, the two replicated experiments, which were performed two weeks apart with different initial communities, were considered as biological replicates. Using these biological replicates, the significance of differential abundance was assessed for each of the orthologous groups and individual gene orthologs in each of the amended incubations compared to the no addition control with edgeR. Genes with a two-fold increase (FDR < 0.05) were considered to be differentially abundant. Using this approach, treatments to which nitrate was added (namely, +N, -P, and +DSW) had between 1,212 and 1,466 orthologous groups with FDR < 0.05 compared to treatments with no N added that had at most 2 genes significantly differentially abundant (Supplemental Figure Manta). The significantly differentially abundant genes for three treatments that received N were conserved, with 45% of differentially abundant orthologs common across the three treatments (Supplemental Figure Venn). Far fewer strain-specific transcripts (161 - 918) were identified as significantly regulated (Supplemental Figure Manta). This is likely due to a lack of statistical power, in that the coverage of individual strains was necessarily lower than the whole. Because of this, the following sections focus on the expression of orthologous groups, which consider the global signature of all the strains. However, it is worth noting the wide spread of estimated fold change of individual, strain-specific transcripts compared to the orthologous group (Figure 4).

Nitrogen scavenging and assimilation

Broadly speaking, genes associated with N-metabolism, which were significantly differentially abundant following N-addition, could be broken into two groups: genes associated with the response to N-limitation and genes associated with the response to newly available substrates, such as ammonium (Figure XA). In microcosm treatments that received N, a number of gene families that are known to be regulated by N supply had decreased abundance. For example, transporters of organic and inorganic N sources were significantly decreased (FDR < 0.05) relative to the no addition control (Figure 4). This included a family of urea transporters (UTP), ammonium transporters (AMT), and nitrate transporters (NRT) (Figure 4; Supplemental Table Genes). In addition to the decreases in UTP, urease (URE), which scavenges N in the form of NH_4^+ from urea, was significantly decreased following N-addition. Of the N-metabolism genes surveyed, the largest decrease in abundance was observed in three orthologous groups of amidases (AMD) and formamidases (FMD), which scavenge NH_4^+ from amides and formamides (Figure 4). These patterns were found to persist not only across the two replicated experiments but also across treatments, with many of the genes being significantly regulated in more than one N-amended experiment.

There was a strong correlation between the observed patterns of transcript regulation for each of the aforementioned gene families and prior transcriptomic (Rokitta et al., 2014) and rate-based (Palenik and Henson, 1997; Bruhn et al., 2010) studies focused on N-limitation responses. Most striking, however, is the coordination of each of these markers with a lab-based proteomic study of the physiological response of CCMP1516 to N-limitation (Figure XA) (McKew et al., 2015). The AMT, NRT, and UTP gene families observed to be down in N-amended incubations were each significantly increased in the proteome of N-limited cultures of CCMP1516 (Figure XA). Additionally, the URE and one of the FMD gene families were found to be up in proteomic analyses (Figure XA). This pattern of regulation in URE was also observed at the transcript level in N-limited cultures of *E. huxleyi*, which may serve as a means of accessing N from the ornithine-urea cycle (**?**) or as a means of accessing exogenous urea (Dyhrman and Anderson, 2003). There is a strong literature on the ability of *E. huxleyi* to grow on amides and other organic N sources (Palenik and Henson, 1997), and evidence of transcript and enzymatic induction of AMD/FMD under N-limiting conditions (Bruhn et al., 2010; Palenik and Henson, 1997). The data from this field study combined with many previous lab-based studies suggest that these N-limitation responses are highly conserved across strains (Figure 3), and across environmental and laboratory conditions as well (Figure 4, X).

Beyond the acquisition of N, several markers of changes in N assimilation and energy production were pronounced following the addition of N to the incubations. The NH_4^+ released by FMD, AMD, and URE is ultimately incorporated into biological material through glutamine synthase (GS) (Rokitta et al., 2014). A shift between the smaller, GS type II, under low N conditions to the larger GS type III with a higher N requirement, following N-addition was observed. In culture transcriptome comparisons, Rokitta et al. (2014) noted that in both haploid and diploid stages, *E. huxleyi* induces a malate:quinone-oxidoreductase (MQO) that can bypass malate-dehydrogenase (MDH) in the TCA cycle and feed electrons directly into the electron transport chain to enable the production of ATP, while limiting carbon loss to respiration. MQO was significantly decreased (between 5 and 9 log fold change) in both +N and +DSW (Supplemental Figure CarbonNucleotideAA). Additionally, though not significantly different, MDH increased 6.4 fold under N-addition (FDR = 0.36) (Supplemental Figure TCA). The MQO, absent from diatom genomes but found to be highly expressed both in N-limited cultures and in this oligotrophic setting, may be a unique and niche-defining aspect of the N-limitation response of *E. huxleyi*. The coordination of the N-induced transporters, enzymes used for the scavenging of N from organic molecules, and shifts in energy metabolism strongly suggest that the *in situ* population of *E. huxleyi* was released from N-limitation with the presence of an added N source, here nitrate.

One exception to the coordination observed between the field data with McKew et al. (2015) and other transcriptomic studies (Dyhrman et al., 2006; Rokitta et al., 2014) were a set of transporters that were significantly increased following N-addition. In each treatment where N was added, three groups of ammonium transporters (AMTs) and a nitrite/formamide transporter (NAR) significantly increased. While an increase of AMT under N-repletion has been observed in centric diatoms (Bender et al., 2014), this shift was likely due to the setting of these experiments. Unlike studies done on axenic cultures, these incubations were performed with whole seawater, consisting of mixed communities of heterotrophic, mixotrophic, and autotrophic organisms. Over the seven day incubation, it is likely that there was active nitrogen fixation (Karl et al., 1997) or remineralization (Casciotti et al., 2008) that may have produced ammonium or amides.

These observations provide an unprecedented glimpse at the import of N in controlling *in*

situ molecular physiology of *E. huxleyi* in the NPSG. Moreover, the choreography observed in the patterns of gene regulation between laboratory studies on axenic, monoclonal cultures and field incubation experiments with mixed communities is striking and suggests that these responses are highly conserved within *E. huxleyi*.

Phosphorus scavenging

The N:P ratio was elevated in the +N, +DSW and -P treatments, the latter of which represented the most extreme condition (Figure Nutrients). Genes associated with P-metabolism, and known to be P-regulated in culture (Dyhrman et al., 2006; McKew et al., 2015) showed a global trend towards increased abundance following N-addition. The -P condition had more genes changing in P-metabolism than the other treatments, with significant increases observed in genes associated with P-transport, P-scavenging from organic molecules, and polyphosphate (poly-P) metabolism. A family of vacuolar transport chaperonins (VTC), which are thought to be associated with poly-P metabolism (Ogawa et al., 2000; Hothorn et al., 2009; Dyhrman et al., 2012), had the largest significant fold change in each of the conditions to which N was added (Figure 4). Although poly-P accumulation was thought to be a luxury uptake response (Perry et al., 1976), VTC has been observed to be to be increased in diatoms under P-limitation (Dyhrman et al., 2006, 2012) and may be indicative of internal poly-P cycling consistent with recent observation of bulk poly-P in the NPSG and the Sargasso Sea (Martin et al. 2014; Diaz et al. 2015).

Genes associated with the scavenging of PO_4 from organic molecules were also significantly increased, with two glycerophosphoryl diester phosphodiesterase (GDP) orthologous groups and a 5'-nucleotidase (NTD) orthologous group significantly more abundant following N-addition (Figure 4). Additionally, two gene families whose inductions under P-limitation are well-characterized in *E. huxleyi*, alkaline phosphatase (AP1) (Xu et al., 2006) and phosphate-repressible phosphate permease (PRPP) (Chung et al., 2003; Dyhrman and Palenik, 2003; Dyhrman et al., 2006), were significantly increased only in -P (Figure 4). This suggests that P-cycling from organic molecules such as nucleotides (both internal and external) may be central to the low-P-response in *E. huxleyi*. NTD has also been observed to be significantly increased in diatom and pelagophyte metabolism under P-limitation (Wurch et al., 2011b; Dyhrman et al., 2012), indicating that this metabolic strategy as seen across diverse taxonomic groups may be a deeply rooted response to P-limitation (Martiny et al., 2013).

As with the genes associated with N-limitation, genes associated with P-limitation were well choreographed with the proteomic data from McKew et al. (2015) (Figure XB). Each of the gene families identified as significantly increased in the -P treatment was significantly increased in the P-limited cultures of CCMP1516, with the exception of two gene families, which were likely below their detection limit or lost during the extraction because of membrane association (Figure XB). Notably, NTD, which was not detected by McKew et al. (2015) in CCMP1516, was found to be present in *E. huxleyi* CCMP374 and CCMP373, and induced under P-limitation (Dyhrman and Palenik, 2003). Similarly PRPP has been found to be induced under -P conditions at both the transcript- (Dyhrman et al., 2006) and protein-levels (McKew et al., 2015), as well as in cultures grown on organic nitrogen (Bruhn et al., 2010).

The regulation of *E. huxleyi* AP1 showed particular sensitivity to the presence of PO_4 in the environment (Figure 4B,C). Alkaline phosphatase is a cell surface protein used for scavenging organic P from the environment and its induction in -P is seen in many diverse phy-

toplankton groups (Sakshaug et al., 1984; Dyhrman and Palenik, 1997, 2003; Wurch et al., 2011a). AP1 in *E. huxleyi* has been shown to be increased 1000-fold at the transcript-level when subjected to P-limiting conditions (Xu et al., 2006), and was found to constitute 3% of all spectral counts in a P-limited proteomic data (McKew et al., 2015). In +DSW, where the added P was higher than the no addition control, -P, and +N, there was a slight decrease in AP1 transcript abundance. Additionally, bulk community alkaline phosphatase activity (APA) was significantly increased in the -P condition relative to all other treatments (Tukey HSD, $p < 0.05$) and tracked well with the AP1 transcript abundance (Figure XBC). The difference between bulk community activity and transcript abundance in +DSW condition underscores the power of species-specific methods, suggesting that other organisms in the community might be less sensitive to the presence of inorganic phosphate. This falls in line with previous findings that suggest that *E. huxleyi* may have one of the highest affinities (out of the eukaryotic algae) for P, leading to its success in P-limited competition experiments (Riegman and Stolte, 2000) and potentially enabling its ability to bloom in P-limited environments (Lessard et al., 2005). Unlike the diatoms, which appear to be co-limited by N and Fe (Figure 1), *E. huxleyi* appears to be primarily limited by N, with a secondary limitation of P.

Shift in life stage and calcification state

Tracking genes associated with ploidy state and calcification suggests that the community in the NPSG and in non-N-amended incubations were haploid and non-calcifying. The addition of N, however, putatively shifted the community to diploid and calcifying states (Figure 4). Genes thought to be associated with calcification (Mackinder et al., 2010) and found to be up-regulated in calcifying cells (Mackinder et al., 2011) were found to be significantly increased following N-addition. These genes included those associated with inorganic carbon transport (e.g. carbonic anhydrases (Îš, Îş, Îť CA) and a group of anion (Cl- / HCO_3^-) exchangers (AE1)), calcium (Ca) acquisition (e.g. voltage-gated Ca2+ channel (CAV), Na^+/Ca_2^+-K^+ exchanger (NCKX), and Ca_2^+/Mg_2^+-permeable cation channel (CX)), and proton transport (e.g. Vacuolar H+-ATPase V0 sector subunits c/c (ATPVcc)). Simultaneously in N amended incubations, there was a significant decrease in the abundance of two genes used as markers of haploid life phase (Frada et al., 2012), dynein heavy chain (DYH) and histone H2A (H2A). The life cycle of *E. huxleyi* is thought to fluctuate between diploid and haploid stages, with calcification only occurring in the diploid stage, and motility during the haploid stage (Paasche, 2001). The coordinated increase of genes associated with calcification with the decrease of genes associated with haploid life stage, particularly DYH, which is integral to the flagella, highlight this known association of calcification and ploidy state (Frada et al., 2012). The pattern of increased calcification following N-addition also falls in line with previously described coordination between nutrient environment and calcification in *E. huxleyi* (Paasche, 2001). P-limitation has been observed to increase Ca content per coccolith and induce calcification in non-calcifying cultures, while N-limitation was found to decrease Ca content per coccolith (Paasche and Brubak, 1994; Paasche, 1998). The link between ploidy or life phase and nutrient concentration is not well understood (Green et al., 1996), though a connection to viral infection has been hypothesized (Frada et al., 2008). Thus, the predominant state of the *in situ* population of *E. huxleyi* in the NPSG was haploid and non-calcifying in E1 and E2, but shifted with N-addition, suggesting either the strong control of N on the life stage of the population or, potentially, consistent pressure from viruses.

5.4.3 Strain variability with altered geochemistry

There were 5,243 genes identified as shared amongst the five strains here, fewer than the nearly 20,000 "core" genes reported in the comparative genome analysis (Read et al., 2013). To encompass the most comprehensive set of common genes, all proteins identified as 'core' by Read et al. (2013) were added to the shared orthologous groups, resulting in 16,914 'core' orthologous groups (Supplemental Figure Core Changes, Figure 2B). This core set accounted for 80% of the reads mapping to *E. huxleyi* in the field and 70% in incubations (Supplemental Figure Core Changes). The sum of the five sets of genes unique to each of the strains accounted for 15% of all transcripts mapped to *E. huxleyi* (Figure 2C). Using the strain-specific orthologous groups as a metric for the relative strain composition in the field (or marker of the most active strain), each of the five strains was detected across the month-long sampling period (Figure 2C). These data are consistent with the finding that single cells of cosmopolitan taxa isolated from the same environment may possess diverse sets of genetic alleles and flexible gene sets (Kashtan et al., 2014). These results also mirror microsatellite studies of *E. huxleyi* that found high diversity at the global-scale and many genetic polymorphisms between individuals collected in the same sample (Iglesias-Rodriguez et al., 2006).

The strain composition of the *in situ* samples varied little over the sampling period. Strains CCMP1516, CCMP370, and PLYM219 had the highest abundances of strain-specific transcripts, while those of CCMP374 and CCMP379 were less abundant (Figure 2C). Although the dominant haptophyte species did not change regardless of nutrient treatment as did the dominant diatom (Figure 1C, D), shifts in the strain-level composition of the *E. huxleyi* were observed. In addition to being dominant in the field, transcripts specific to CCMP1516 and PLYM219 were the most abundant in treatments where N was not added (Control, -N, +P) (Figure 2D, E). By contrast, transcripts specific to both CCMP379 and CCMP374 exhibited dramatic increases (e.g. $< 1\% to 5\%$ of reads for CCMP374) following N-addition in both experiments (Figure 2D, E). These shifts may reflect the different capabilities within the original seed population, such as previously observed in strain variability in growth rate on different organic nitrogen substrates (Strom and Bright, 2009) or enzymatic activity (e.g. FMD) (Palenik and Henson, 1997; Bruhn et al., 2010). Given better sequencing coverage, variability between strains in the expression of genes, such as FMD, might have been statistically distinguishable (Figure 4).

A subtle difference in the N-amended treatments consistent across experiments, however, was that strain-specific transcripts associated with CCMP379 were more abundant in -P, which had been augmented with iron, trace metals, and vitamins (Figure 2D, 2E). Such enhanced abundance of a particular strain with added trace metals is potentially supported by the heterozygosity observed in trace metal-associated genes (Read et al., 2013) and trace metal quotas across strains (Sunda and Huntsman 1992). These results suggest both that multiple strains co-exist in the same environment and that alterations in the geochemical environment can shift the relative abundance and activity of various strains. Such observations support the hypothesis by Read et al. (2013), who postulated that the diversity in the pan genome as observed across cultured isolates may be present in the field and central to the cosmopolitan nature of *E. huxleyi*.

In addition to the observed shifts in strain abundance following changes in the geochemical environment, the expression of the set of 5,243 shared genes (Supplemental Figure Shared Genes) was examined by strain across the nutrient treatments to statistically validate the apparent shifts in strain abundance and transcriptional response. The 5,243 shared gene

set was better annotated than the entire set of orthologous groups, with 73% of the shared orthologous groups annotated with KOG as compared to only 36.9% for the sum (Supplemental Figure KOG annotation), likely because this set contains essential and well-studied metabolic pathways (e.g. carbon fixation). Focusing on this relatively well-annotated fraction, RSEM (Li and Dewey, 2011) was used to estimate the relative contribution of each 'isoform', here defined as the orthologous genes from each of the strains comprising the core orthologous group (Supplemental Figure SharedGeneComp). Principle components analysis (PCA) of the estimated TPM of the 5,243 shared orthologous groups, broken down by the five strains across time and experimental condition, showed separation of strains and conditions. The first three components of the PCA explained 36.2% of the variance (Figure 3). The primary separation occurred along the first component (20.5% of variance), which clearly separated treatments to which N was added from treatments where no N was added (Figure 3C), similar to shifts observed between functional groups following DSW addition (Alexander et al., 2015b). No difference was seen between strains in the first three components in the treatments to which no N was added (Figure 3A,B). The second and third component separated strains in treatments where N was added, with the second component (8.1%) separating CCMP1516 from the four other strains and the third component (7.6%) separating CCMP379 from CCMP374 (Figure 3 B). Interestingly, these data appear tied to the unique gene sets, as the unique gene sets of CCMP379 and CCMP374 were observed to increase in abundance following N-addition (Figure 2). These data suggest that the metabolic profiles of strains based solely on shared genes are similar, at least in the context of this very oligotrophic environment, where haptophytes are thought to be limited (Alexander et al., 2015b). Strikingly, the addition of N appears to change the metabolic profile of all strains in a similar way, e.g. enhanced carbon fixation (Component 1), such as was observed in the strong conservation and choreography of N and P metabolic genes (Figure 4). However, differences between the responses of strains in the shared set of genes exist (Components 2 and 3), paralleling observations of variable response to nutrient pulses across diatom species (Alexander et al., 2015a).

With their description of the pan genome of *E. huxleyi*, Read et al. (2013) speculated that the variability in gene content across strains may be indicative of the Bass-Becking hypothesis, or "everything is everywhere but the environment selects" (Baas-Becking, 1934; de Wit et al., 2006), holding true for the strain diversity observed in *E. huxleyi*. These data from both field observations and incubation experiments demonstrate that a diverse consortium of strains existed simultaneously in the field (Figure 2C). Furthermore, the shift in strain composition or activity in N-amended treatments (Figure 2D,E) may be indicative of the "environmental selection" of strains. Cross comparison of the species-specific modulation of the shared gene set also suggests that while there is a strongly conserved response to N-limitation, individual strains respond differently to the same environmental stimuli.

5.5 Conclusion

Metabolic plasticity in response to environmental change is a current cornerstone to the study of phytoplankton physiology, with much effort being put towards characterizing transcript- and protein-level shifts following perturbation (Dyhrman et al., 2006, 2012; Wurch et al., 2011a; Bertrand et al., 2012; Jones et al., 2013; Bender et al., 2014; Frischkorn et al., 2014). A compliment to the observed physiological response to the environment in some species

has been found to be strain or genomic variability (Kashtan et al., 2014). This study looked at the intersection of these two types of response in field populations of *E. huxleyi* under changing nutrient environments, finding that both strain variability and physiological plasticity are central to the success of *E. huxleyi*. Metatranscriptomic approaches enabled the direct tracking of metabolic shifts in the *E. huxleyi* species complex following changes in the geochemical environment, demonstrating the central role of N in determining the metabolic state of *E. huxleyi* in the NPSG. These analyses suggested that unlike diatoms, which appear to be N/Fe co-limited, P might be secondarily limiting to *E. huxleyi*. In spite of the many factors that separate field and laboratory studies, choreography of the transcriptional response to N- and P-limitation with previous proteomic studies (McKew et al., 2015) was observed, suggesting a highly conserved physiological response to these stressors across the species complex. In concert with these observed shifts in physiology, the co-existence of at least five strains in the same environment based on the variable gene sets of each strain was observed. This combined with the modulation of the strain-specific transcript abundance support the hypothesis that the genomic variability of this species complex may facilitate its ability to thrive in a range of environments (Read et al., 2013), such as in other cosmopolitan phytoplankton (Derelle et al., 2006; Johnson et al., 2006) inherently limits this type of study and highlights the importance of continued isolation and culturing of phytoplankton strains to facilitate this type of strain-specific study both more broadly in *E. huxleyi*, but also across other phytoplankton groups.

Together these direct observations of the shifts in the molecular physiology and the strain composition of natural populations of *E. huxleyi* in response to changes in the geochemical environment suggest that 1) physiological responses to N and P limitation are highly conserved across strains, 2) many strains concurrently exist in the NPSG and 3) fluctuations of environmental parameters may change strain abundances. There is debate as to what level of taxonomic resolution is important in global biogeochemical models (Follows and Dutkiewicz, 2011). The results of this study suggest that while strain variability occurs within *E. huxleyi* species complex, the global physiological response to nutrient environments is highly conserved across strains and may be sufficient for accurately capturing the dynamics of this biogeochemically significant taxonomic group.

Chapter 6

Summary and conclusions

"We cannot cheat on DNA. We cannot get round photosynthesis. We cannot say I am not going to give a damn about phytoplankton. All these tiny mechanisms provide the preconditions of our planetary life."

– Barbara Ward, *Who Speaks for Earth?*

6.1 Thesis summary

Lipsum

6.2 Conclusions and future directions

Capitalizing upon this revolution to address biogeochemical or ecological questions, however, is becoming increasingly difficult, as our ability to make measurements has surpassed our ability to analyze, visualize, and compare the data produced.

Appendix A

Chapter 2 Supplemental Information

A.1 Supplemental Figures

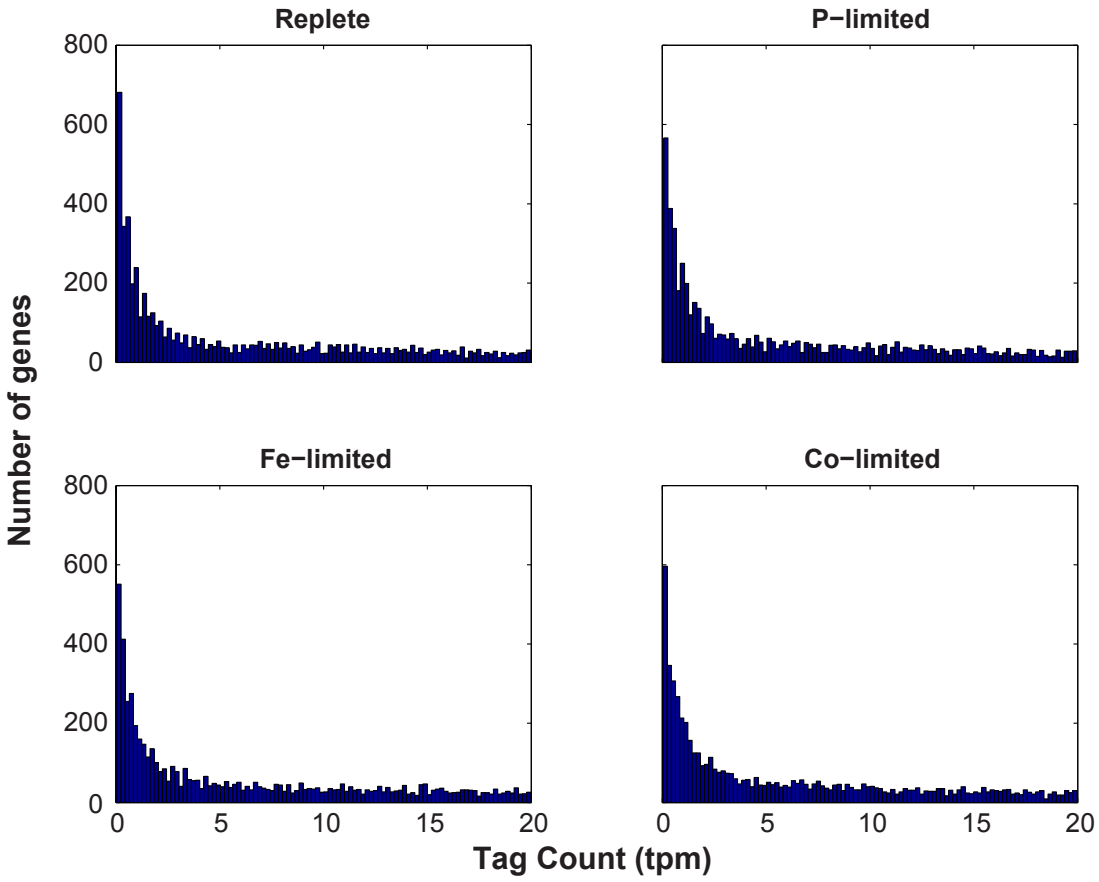


Figure A.1: Histogram analysis of the distribution of normalized tag counts (TPM) for each gene across each of the four treatments (Replete, P-limited, Fe-limited, and co-limited). The abundance of normalized tag counts (TPM) was assessed, tallying the total number of genes with a given tag count. Only tag counts less than 20 are depicted to aid the visualization of the inflection in the data at 2.5 TPM.

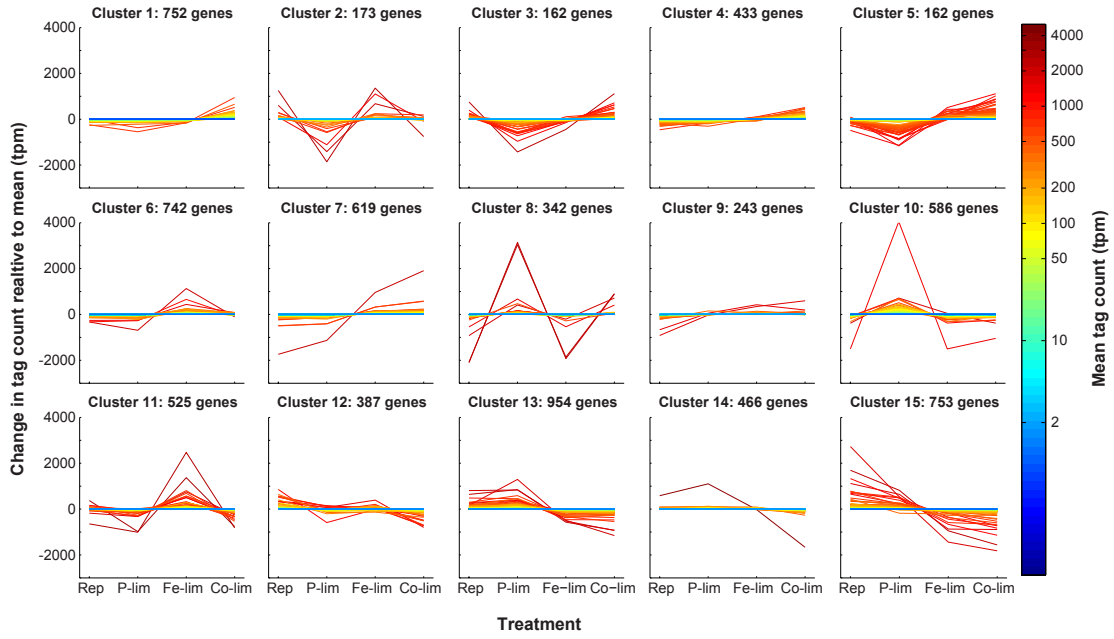


Figure A.2: K-means clustering of normalized genes. The 7380 genes that passed the 2.5 TPM cutoff were clustered into 15 clusters using the k-means algorithm under the Pearson correlation coefficient. Tag counts normalized to total library size (in TPM) for each gene are plotted relative to the mean (indicated by the color of the line) for each of the four treatments: Replete (Rep), P-limited (P-lim), Fe-limited (Fe-lim), and co-limited (Co-lim).

A.2 Supplemental Data

Data Sheet 2-1 : Genes in the *T. pseudonana* genome homologous to reference genes from relative expression studies in algae and plants. Data Sheet 2-1 can be downloaded from the online version of the manuscript of Alexander et al. (2012) through Frontiers in Aquatic Microbiology.

Data Sheet 2-2 : Putative reference genes identified with k-means clustering analysis (Cluster 9 and Clusters 14). Data Sheet 2-2 can be downloaded from the online version of the manuscript of Alexander et al. (2012) through Frontiers in Aquatic Microbiology.

Data Sheet 2-3 : Data Sheet 3. Putative reference genes identified with ASC analysis ($p < 0.1$ for a fold change of 1.25). Data Sheet 2-3 can be downloaded from the online version of the manuscript of Alexander et al. (2012) through Frontiers in Aquatic Microbiology.

Data Sheet 2-4 : The intersection of differentially expressed genes identified by Mock et al. (2008) and stably expressed genes identified through ASC (1.25 fold change bin, $p < 0.1$). Data Sheet 2-4 can be downloaded from the online version of the manuscript of Alexander et al. (2012) through Frontiers in Aquatic Microbiology.

Appendix B

Chapter 3 Supplemental Information

B.1 Supplemental Figures

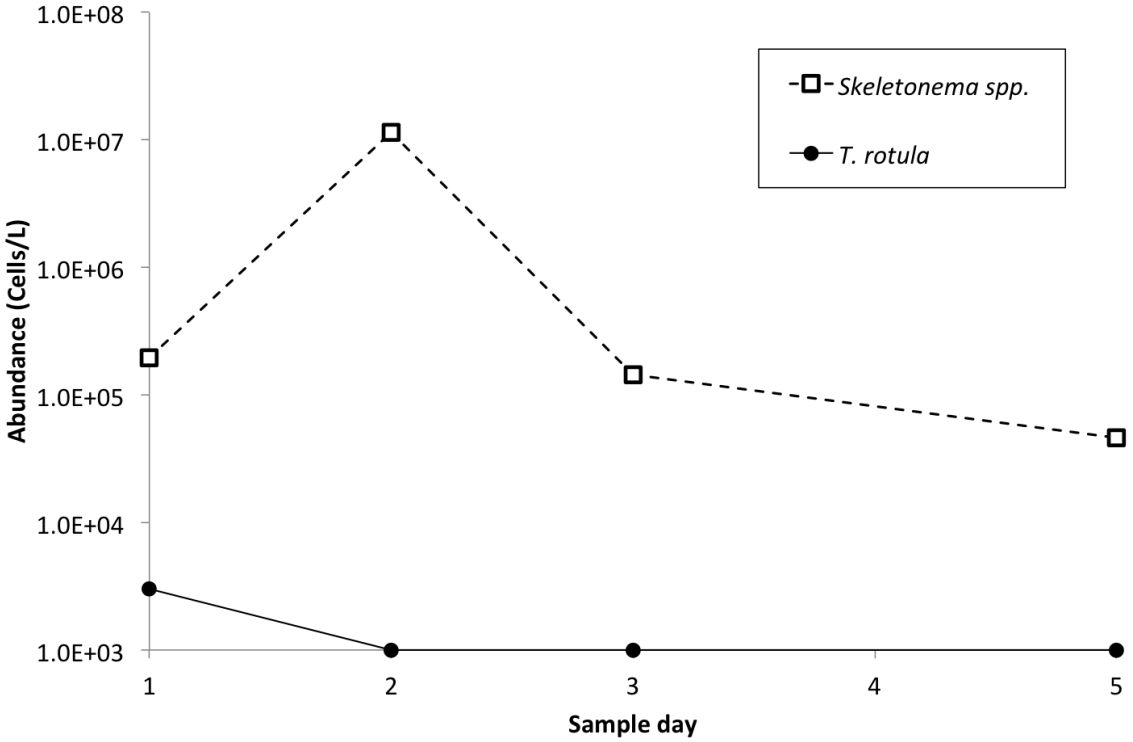


Figure B.1: Abundance estimation from cell counts of *Skeletonema* spp. and *T. rotula* across the five sample points during the spring of 2012.

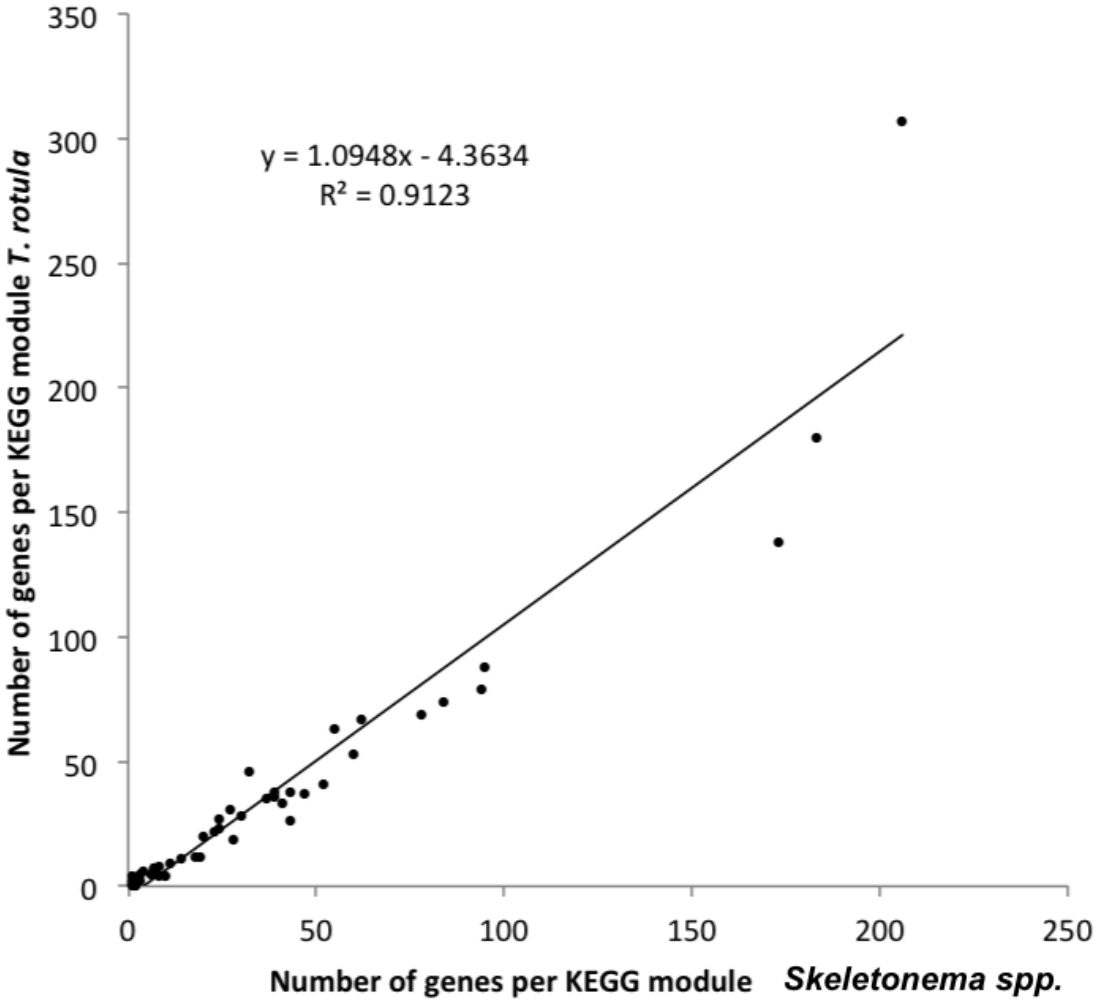


Figure B.2: Total number of genes assigned to each KEGG module for *Skeletonema* spp. and *T. rotula*.

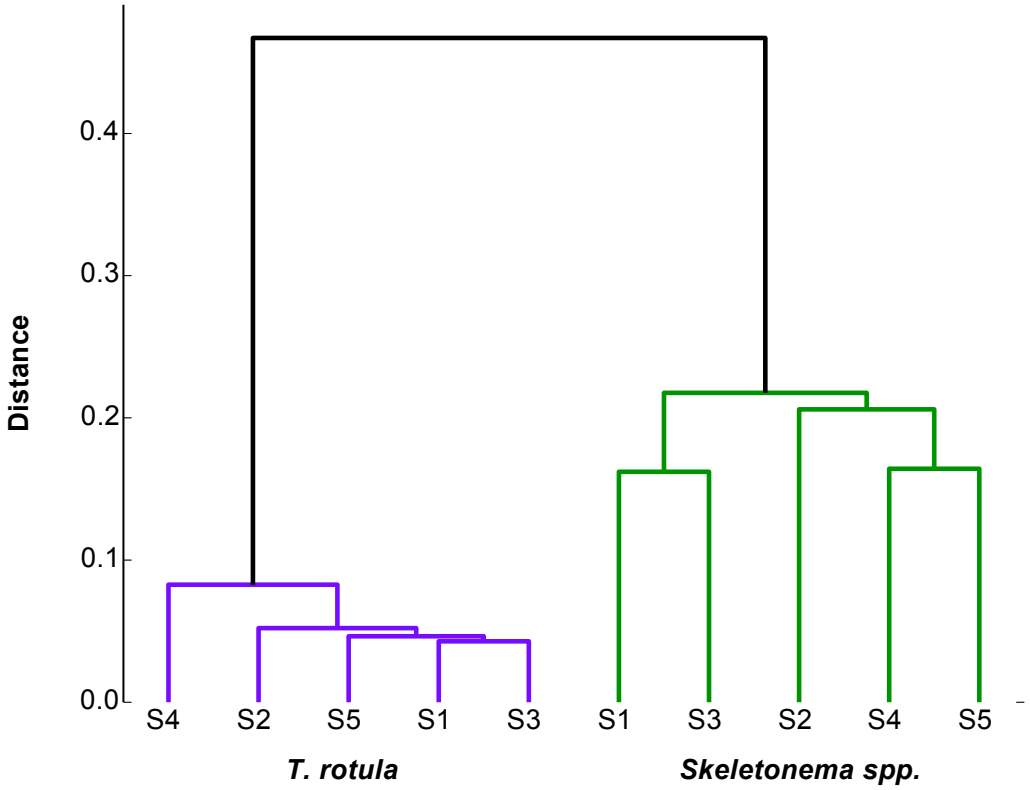


Figure B.3: Dendrogram depicting hierarchical clustering of samples based on relative expression of KEGG modules (Figure 2) across the five samples S1-S5 for *Skeletonema* spp. and *T. rotula*.

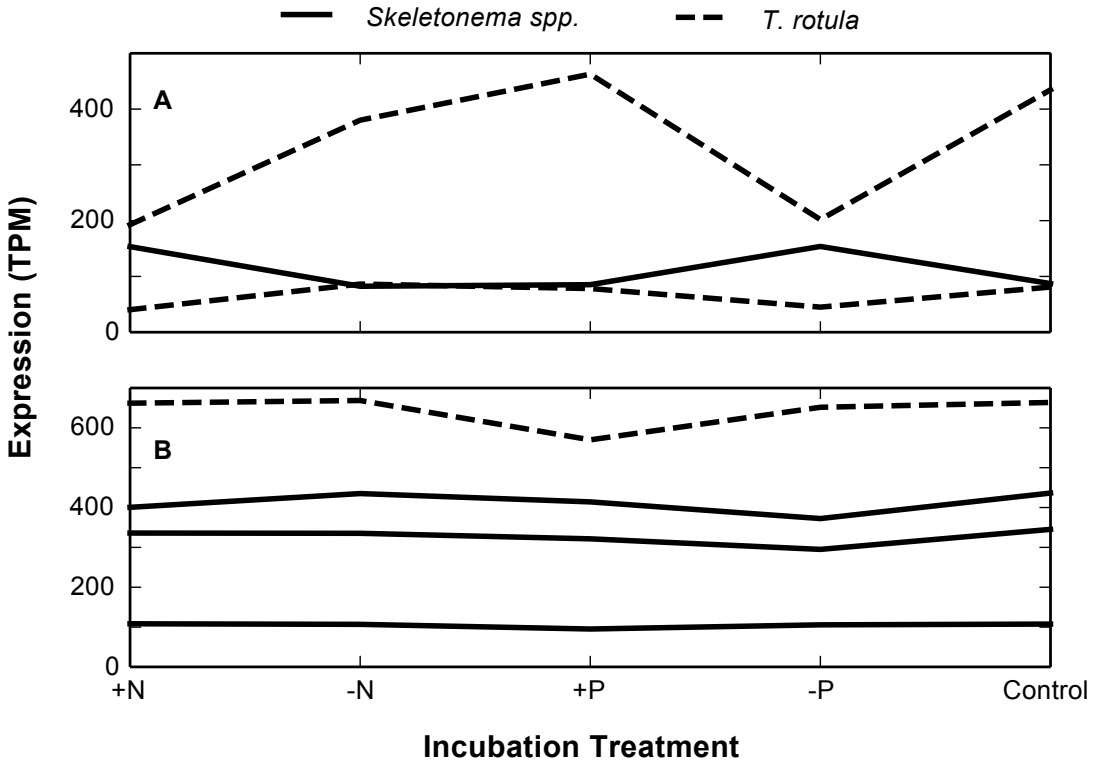


Figure B.4: Expression of stable reference genes identified based on literature and statistical parsing in nutrient amendment incubation. (A) The expression in tags per million (TPM) of stable reference genes identified in *T. rotula* (dashed line) and *Skeletonema* spp. (solid lines) based on homology (e-value $<$ 1e-5) to a known reference genes in *T. pseudonana*, ACT1 (Thaps_25772), in nutrient incubations. (B) Also shown are reference genes identified in the incubation experiments, using statistical analysis of sequence counts (Alexander et al., 2012; Wu et al., 2010), and nutrient incubations

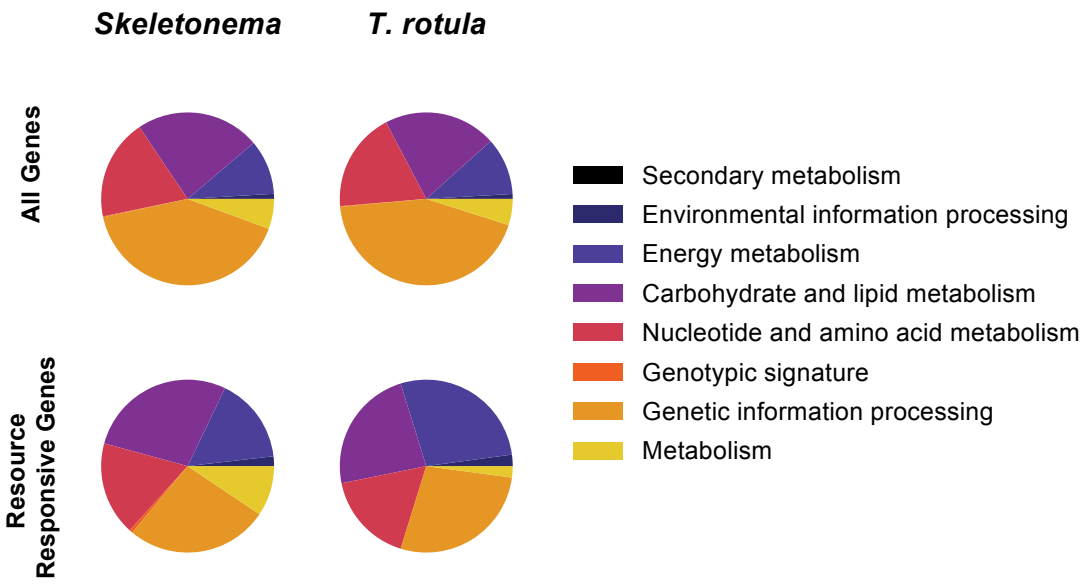


Figure B.5: Functional composition of the reference transcriptome and resource-responsive (RR) gene subset for *T. rotula* and *Skeletonema* spp. (A) RR gene sets were identified through cross comparison of like-nutrient incubations (i.e. +N vs. -N and +P vs. -P), using ASC (fold change = 2, post-$p > 0.95$). The relative functional categorization of the reference transcriptomes and RR gene set for *T. rotula* and *Skeletonema* spp. based on KEGG ontology as assigned by KAAS is depicted at the module-level.

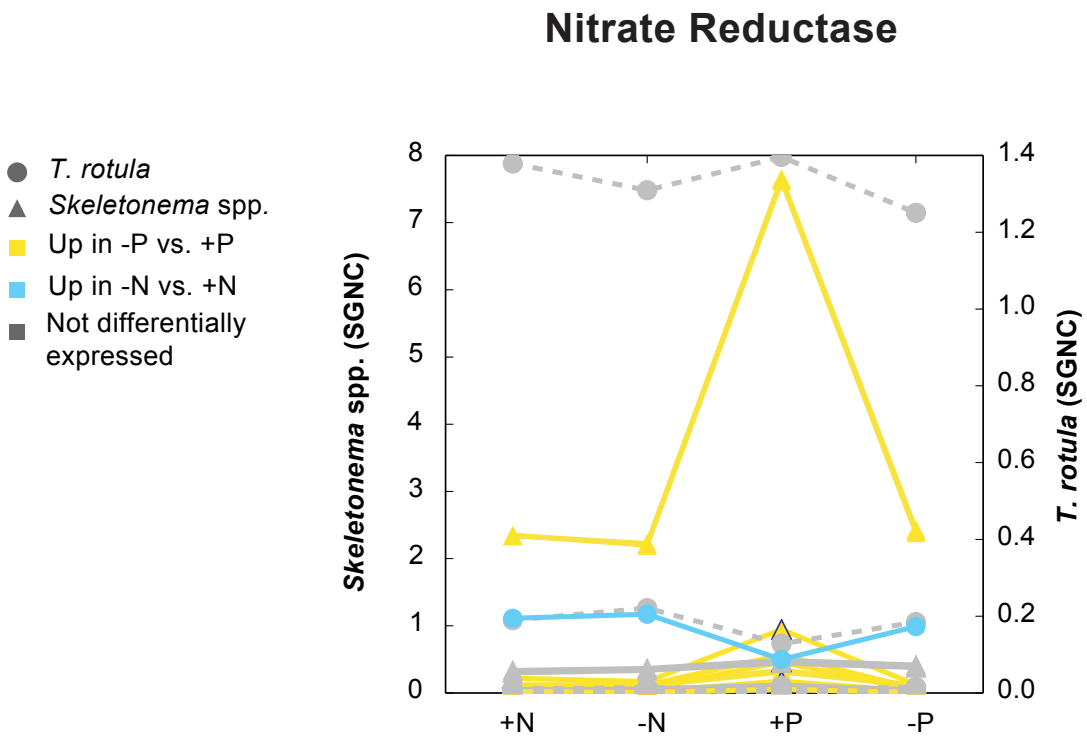


Figure B.6: The relative expression in stable gene normalized counts ($SGNC$) of the assimilatory nitrate reductase gene cluster across the incubation experiment treatments. Significance of regulation between the treatments is denoted by the color of the line; organisms are denoted by the shapes of the marker.

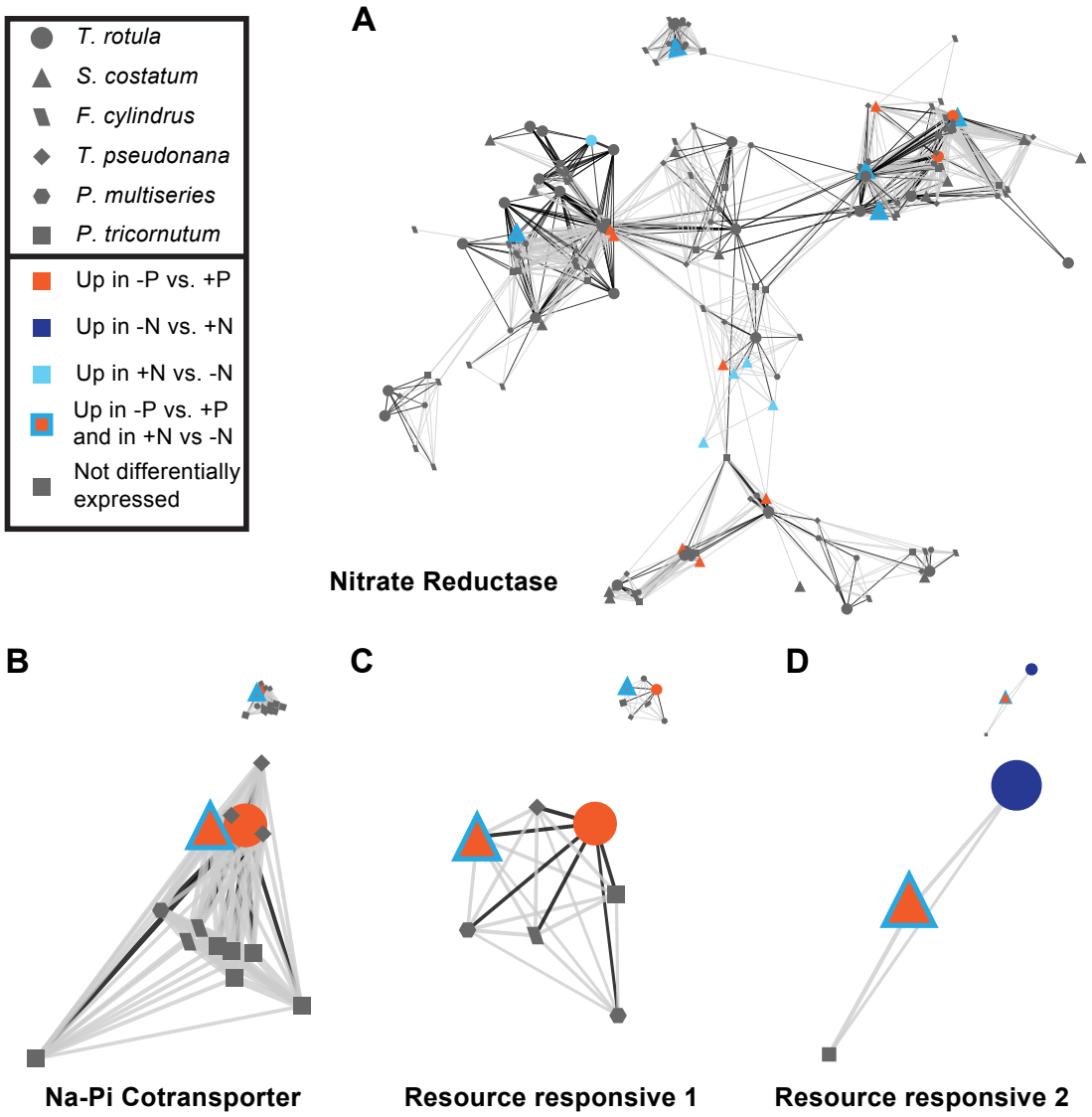


Figure B.7: Gene cluster known nutrient-responsive genes in *T. pseudonana*: (A) assimilatory nitrate reductase and (B) sodium-phosphate cotransporter and novel resource-responsive (RR) gene families: (C) RR1 and (D) RR2. Transcripts from the transcriptomes of *T. rotula* and *Skeletonema* spp. were clustered based upon relative homology with available diatom genomes: *F. cylindrus*, *P. tricornitum*, *P. multiseries*, and *T. pseudonana*. Symbols indicate different species, while color indicates regulation in the field incubation experiments. Two nodes within a gene cluster are connected by an edge if they share a homologous protein (reciprocal BLAST hit with a minimum of 1e-5 score and minimum 20% identity). Gene clusters are visualized using an edge-weighted spring-embedded model based on e-value, meaning that genes that are closer together are more similar. The width of the line correlates to the magnitude of the e-value, with lower e-values represented by thicker lines and higher e-values represented by thinner lines.

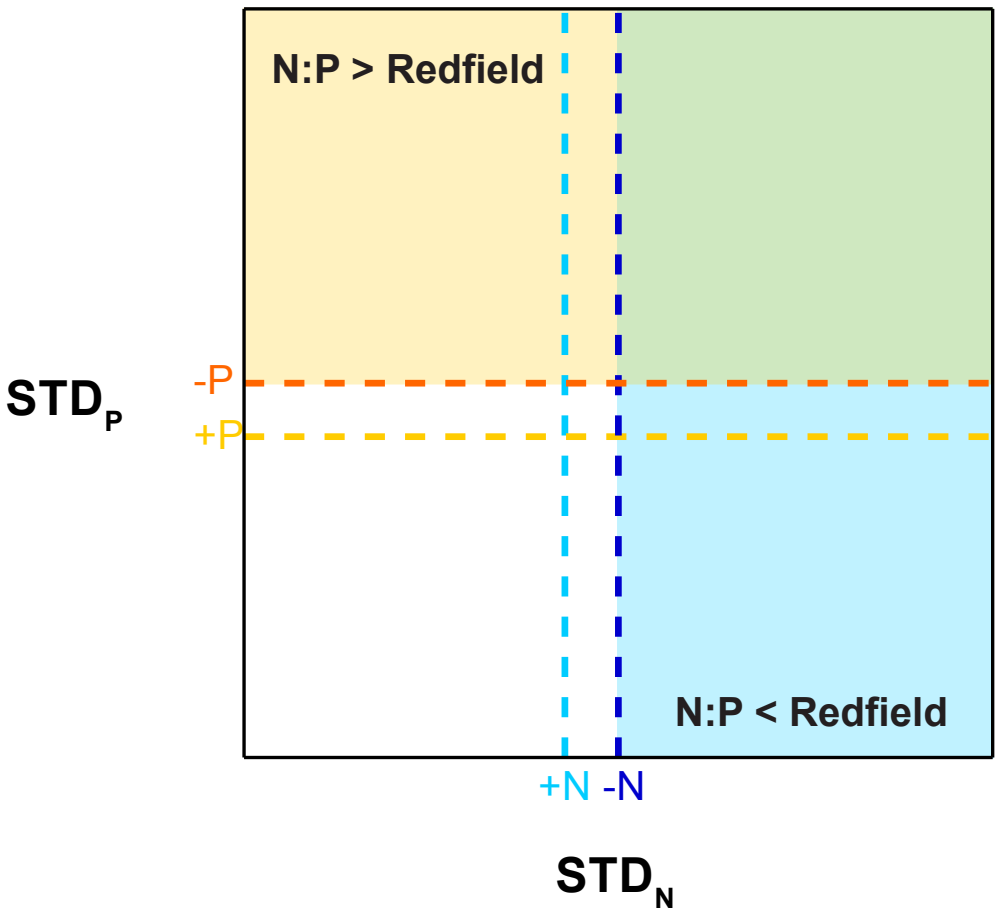


Figure B.8: A conceptual schematic of STD_N plotted against STD_P hypothesized regions of N:P > Redfield physiology and N:P < Redfield physiology highlighted.

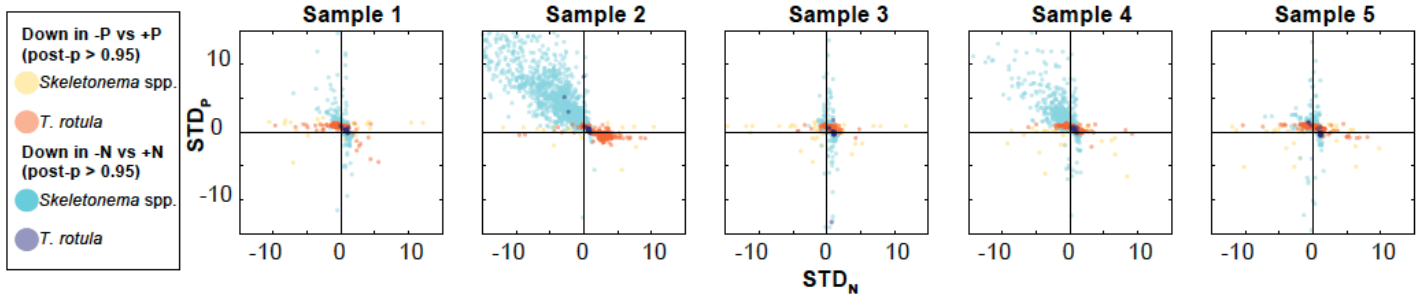


Figure B.9: Evolution of niche space indexing over time in Narragansett Bay for *T. rotula* and *Skeletonema* spp.. The stable gene normalized field signal from genes identified as significantly (2-fold change, post$-p > 0.95$) down-regulated in -P vs +P for Skeletonema spp. (yellow) and *T. rotula* (orange) and in -N vs +N for for *Skeletonema* spp. (cyan) and *T. rotula* (dark blue) was proportionalized relative to the expression for those genes in nutrient incubations, yielding the STD_N and STD_P. These data are plotted for Sample 1 through Sample 5.

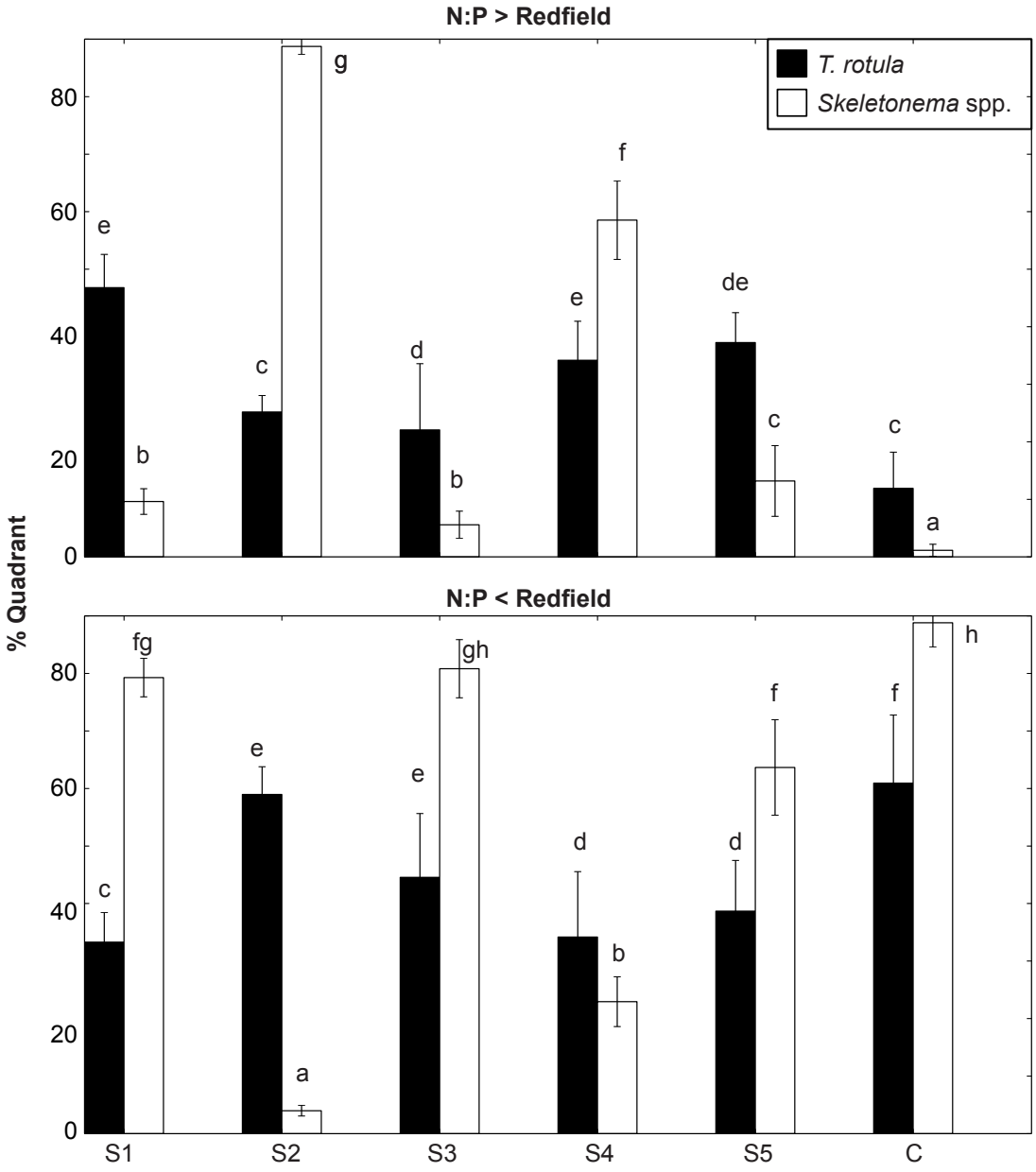


Figure B.10: The percentage of identified nutrient responsive genes falling into the N:P > Redfield and N:P < Redfield quadrants for *T. rotula* and *Skeletonema* spp.. The total number of genes falling into the N:P > Redfield quadrant ($STD_P > C$; $STD_N < C$, for $0.25 < C < 0.75$) and the N:P < Redfield quadrant ($STD_P <$ C; $STD_N > C$, for $0.25 < C < 0.75$). The value of C was varied over 10 different values and the average percentages of genes falling into each of the quadrants is depicted above based on the size of the circle at the median STD_N and STD_P for the genes in the quadrant. Similarity of data between species by quadrant was assessed using an analysis of variance (ANOVA) with a generalized linear model. The results from a post hoc Tukey test show the divergence of species across time ($p < 0.05$).

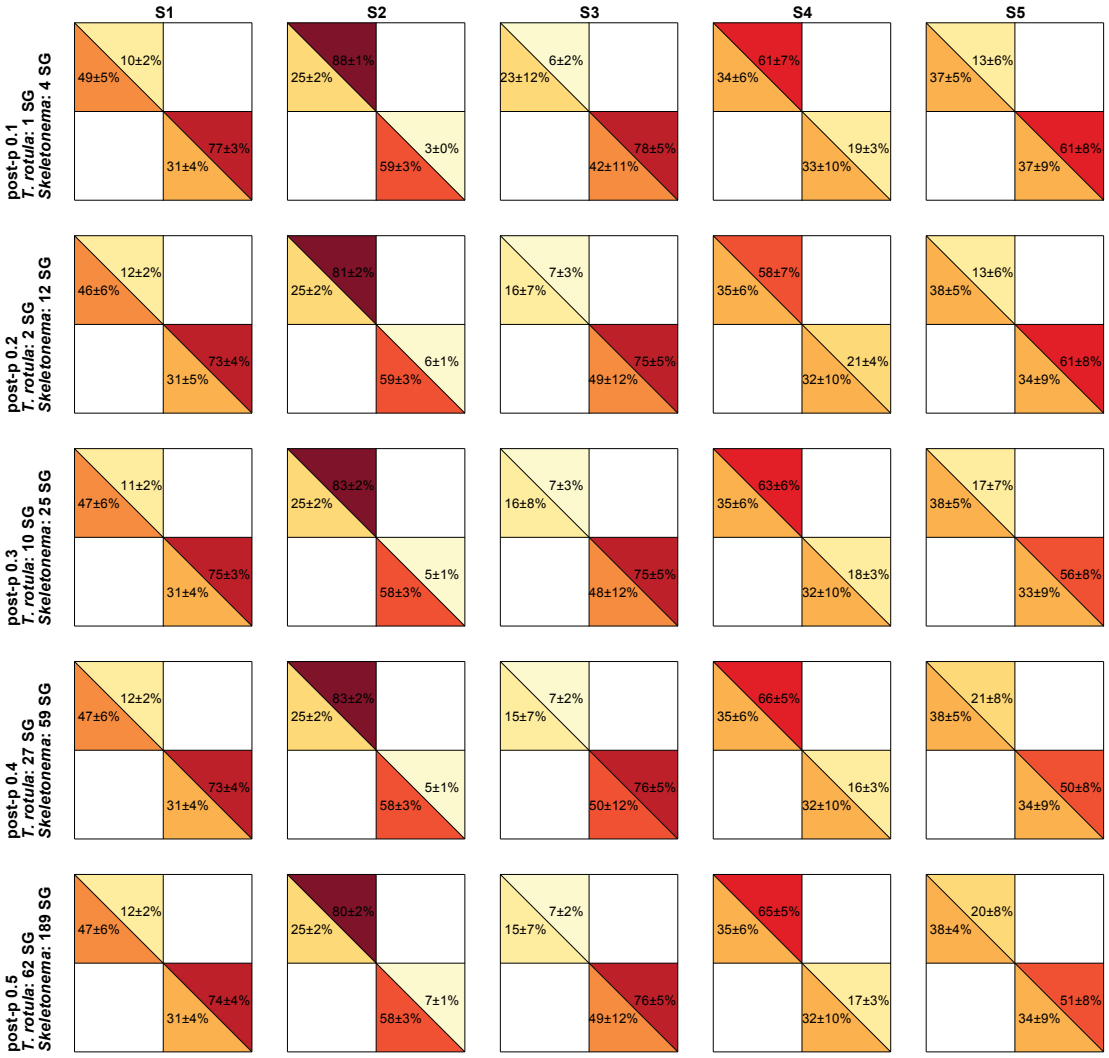


Figure B.11: The impact of stable gene selection on the quadrant localization of the resource responsive gene sets. The posterior probability cutoff used in the selection of stable genes was varied from 0.1 to 0.5 for a fold change of 1.25. The percentage of identified nutrient responsive genes falling into the N:P > Redfield and N:P < Redfield quadrants for *T. rotula* and *Skeletonema* spp. across the five sample points and five posterior probability values is depicted.

B.2 Supplemental Tables

Table B.1: The total number of paired end reads after quality control and trimming and the percentage of reads mapping to the *T. pseudonana* genome, *T. rotula* transcriptome, and *S. costatum* transcriptome.

Sample	Total library size (paired end reads)	Mapped representation in library		
		T. pseudonana	*T. rotula*	*S. costatum*
S1	89455034	2.98%	17.50%	33.50%
S2	64888267	0.41%	11.70%	54.90%
S3	103250243	0.39%	7.30%	9.00%
S4	45370867	0.68%	8.80%	8.30%
S5	55061692	0.88%	10.40%	11.20%
Ambient Control	51508197	0.27%	13.40%	8.00%
+N	58626239	0.43%	6.10%	5.30%
-N	44561851	0.41%	8.70%	8.30%
+P	51130364	0.29%	8.50%	8.00%
-P	58834022	0.40%	6.60%	6.50%

Table B.2: Nutrient concentrations used in nutrient amendment incubations.

Nutrient	**Treatment**				
	Ambient Control	+ P	+ N	- P	- N
Nitrate			10 μM	10 μM	
Phosphate		3 μM			3 μM
Silica				68 μM	68 μM
Iron				4.6 μM	4.6 μM
Vitamins				f/5	f/5

Table B.3: Total number of contigs in the *T. rotula* and *S. costatum* transcriptomes and the number of genes in each of the differentially regulated and stable groupings.

	T. rotula	***S. costatum***
Number of contigs in transcriptome	22362	27665
Pass 2 TPM cutoff	4318	20921
Up in -P vs +P	249	4754
Down in -P vs +P	335	52
Up in -N vs +N	196	9
Down in -N vs +N	49	1631
All differentially regulated (2 fold change, post-$p > 0.95$)	775	5136
Stable genes (1.25 fold change, post-$p < 0.1$)	1	4

B.3 Supplemental Data

Data Sheet 3-1 : Annotations based on KEGG Ontology for *Skeletonema* spp. and *T. rotula* transcriptomes. Data Sheet 3-1 can be downloaded from the online version of the manuscript of Alexander et al. (2015a) through *Proceedings of the National Academy of Sciences*.

Data Sheet 3-2 : Relative expression in tags per million (TPM) for genes identified as differentially or stably expressed in nutrient incubations. Data Sheet 3-2 can be downloaded from the online version of the manuscript of Alexander et al. (2015a) through *Proceedings of the National Academy of Sciences*.

Appendix C

Chapter 4 Supplemental Information

C.1 Supplemental Figures

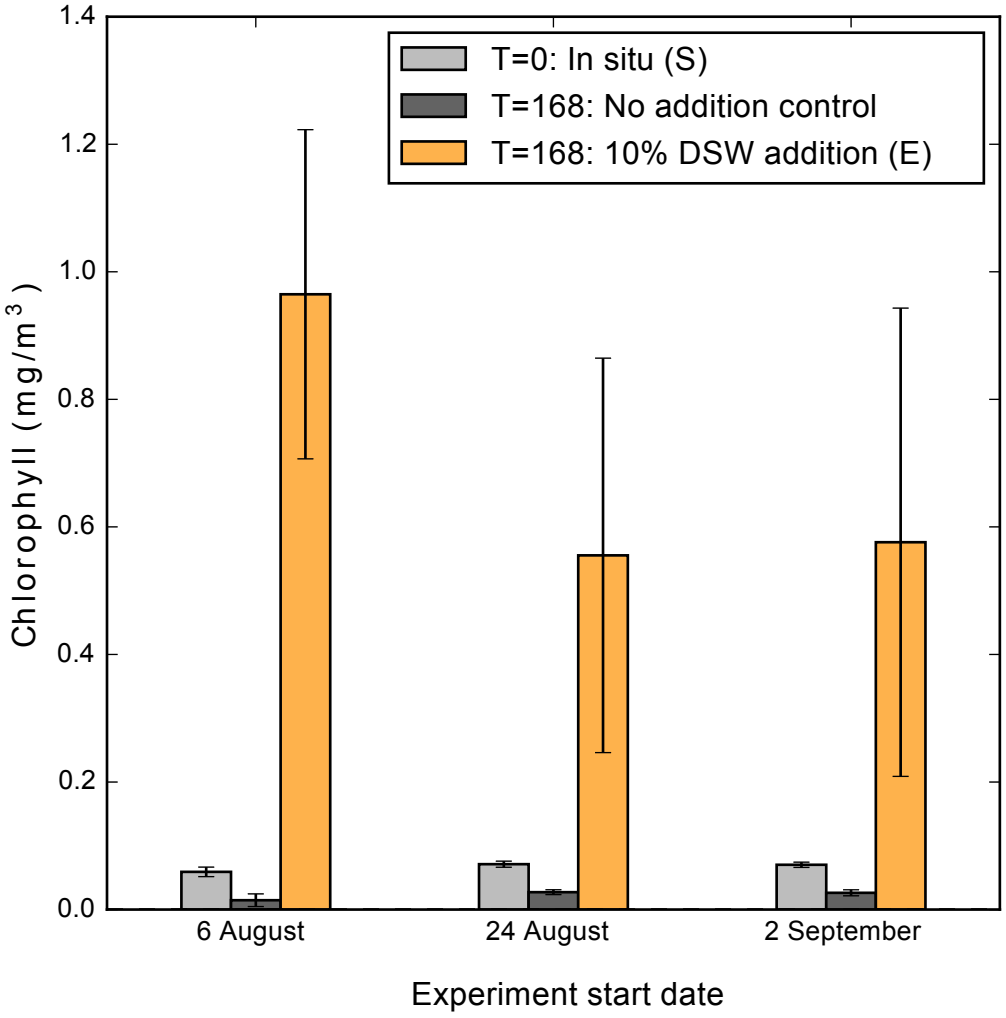


Figure C.1: Chlorophyll a of replicated experiments for *in situ* samples (S), a no addition control, and a 10% deep seawater (DSW) amendment (E). Incubation samples were harvested after 168 hours.

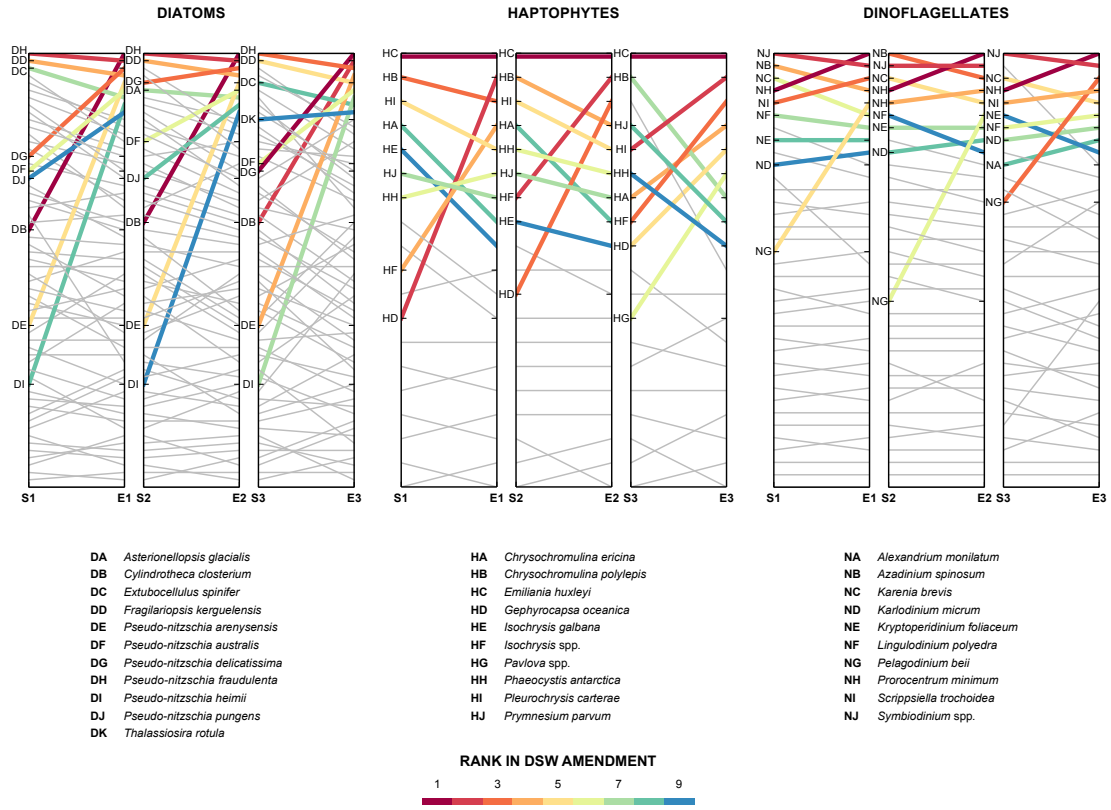


Figure C.2: Rank abundance shifts in the species composition of diatoms, haptophytes and dinoflagellates for the three experiments. The relative shift in rank abundance for each species is depicted for each incubation experiment (E1-E3) following deep seawater (DSW) addition. The nine most abundant taxa following DSW addition are highlighted for each of the functional groups. Although the species that recruited the reads are denoted here this is highly driven by the composition of the database and does not necessarily indicate the actual species present, but rather the closest species present in the database.

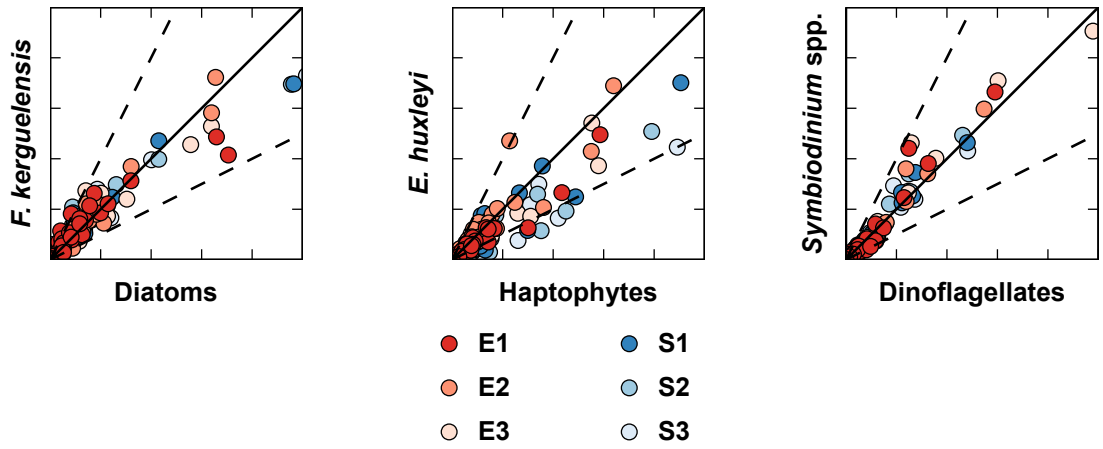


Figure C.3: Comparison of the quantitative metabolic fingerprint (QMF) between the whole functional group and representative taxa. The proportion of reads falling into each of the modules depicted in Figure 2 is plotted for S1-S3 and E1-E3, comparing the summed functional group signal and that of a representative taxon. Color of the marker indicates the sample; solid and dashed lines mark the 1:1 and 1:2 lines, respectively.

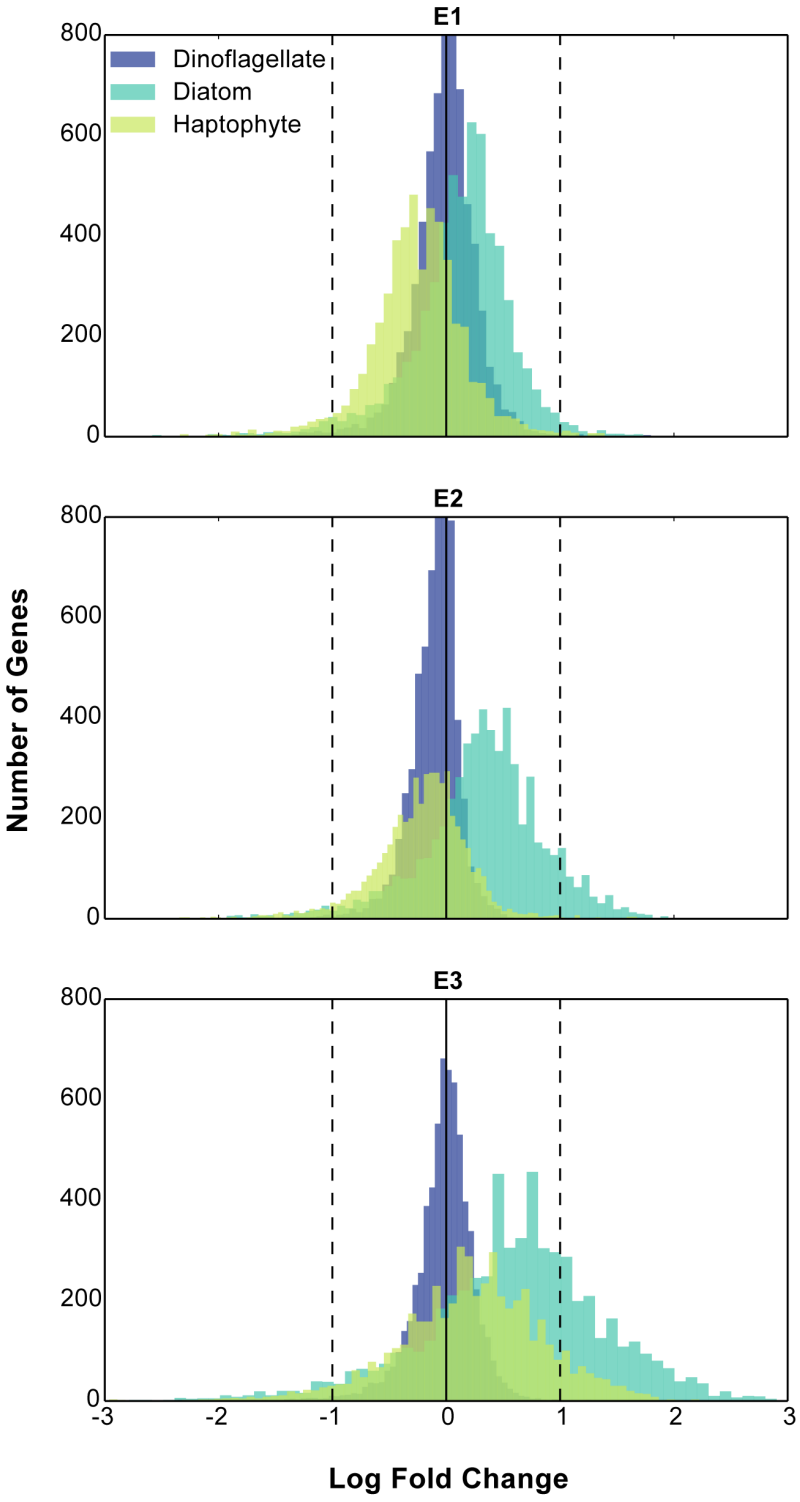


Figure C.4: Distribution of log fold change following deep seawater (DSW) addition. Histogram of the number of genes falling within each of the log fold change bins for diatoms, haptophytes and dinoflagellates. Solid line indicates no fold change; dashed lines indicate 2 fold-change both up and down.

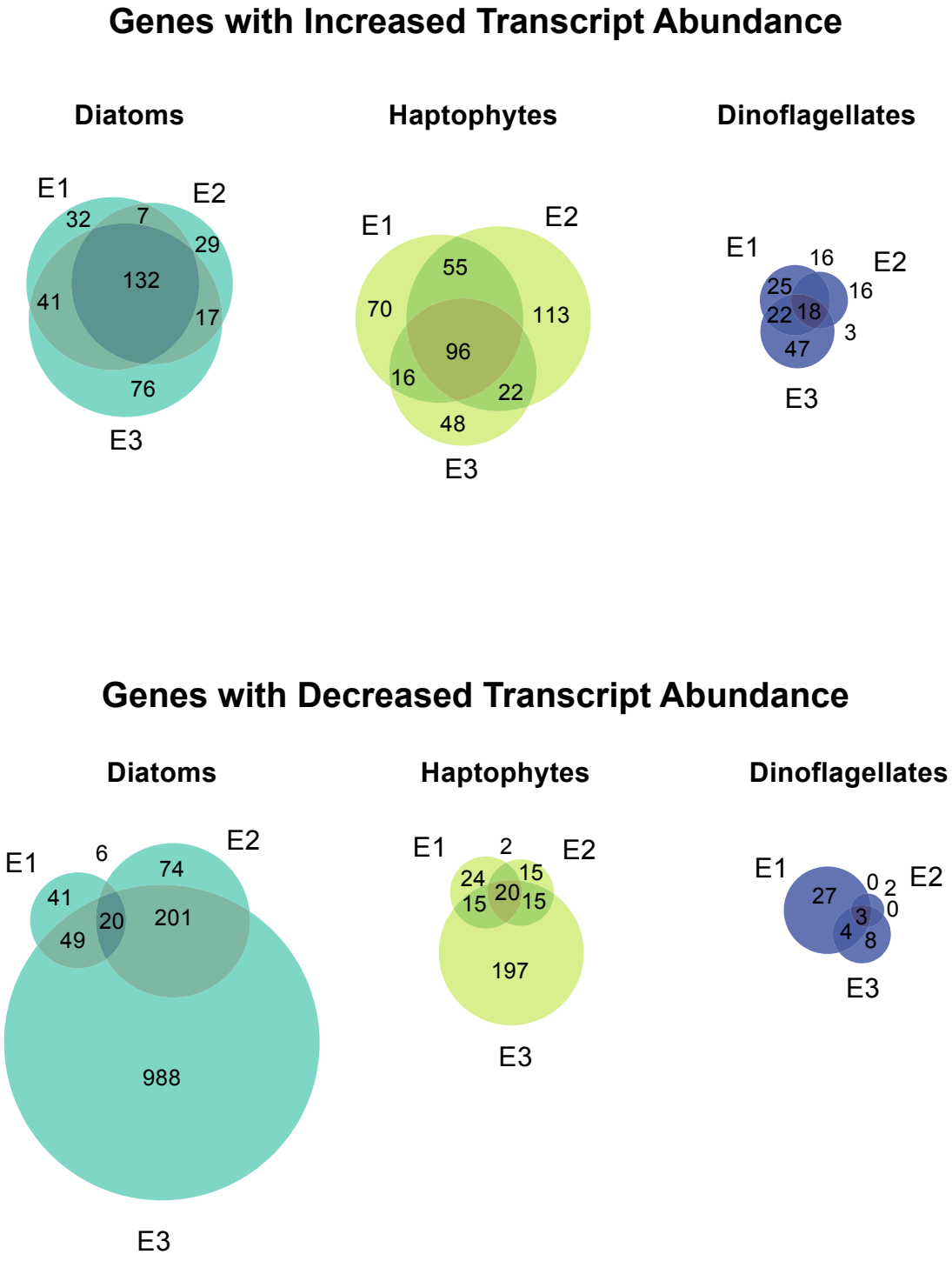


Figure C.5: Weighted Venn diagrams of genes with significantly different abundances following deep seawater (DSW) addition by functional group. The uniqueness of KEGG orthologs with increased or decreased abundances as determined by ASC (2 fold-change, post-p > 0.95) across experiments was assessed for diatoms, haptophytes, and dinoflagellates.

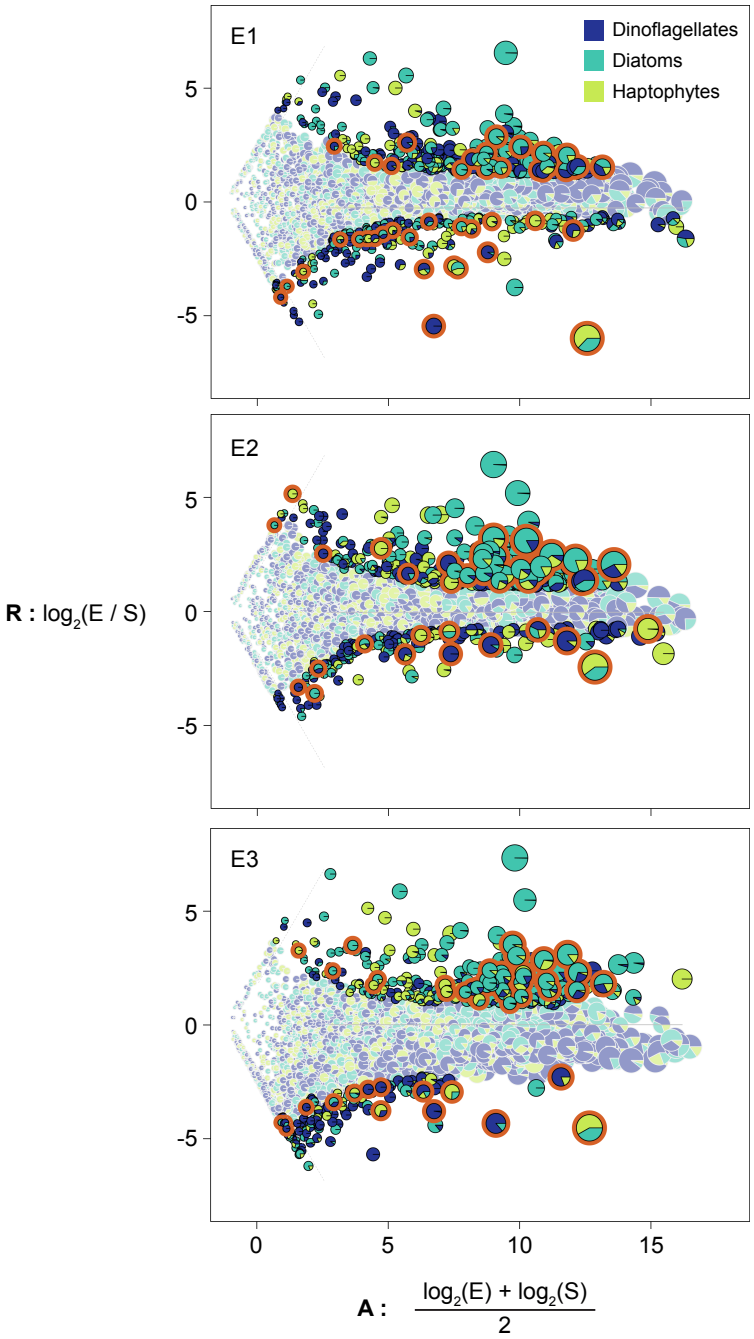


Figure C.6: Microbial Assemblage Normalized Transcript Analysis (MANTA) ratio-averaged plots for global shifts in expression of KEGG orthologs. Fold change ratio (R) and average read count (A) are plotted for read counts in the *in situ* (S) and deep seawater (DSW) amendment (E) samples across the three sample pairs (S1:E1, S2:E2, S3:E3). The trimmed mean of fold-change values is noted as a gray solid line; orthologs unique to one library are separated by gray dashed lines. Pies indicate the taxonomic distribution of orthologous reads across the three functional groups. KEGG orthologs that were significantly differentially expressed (DE) (adjusted $P > 0.05$) are outlined in black and those not significantly DE are outlined in gray. DE KEGG orthologs that fall in the Energy Metabolism KEGG module are outlined in orange.

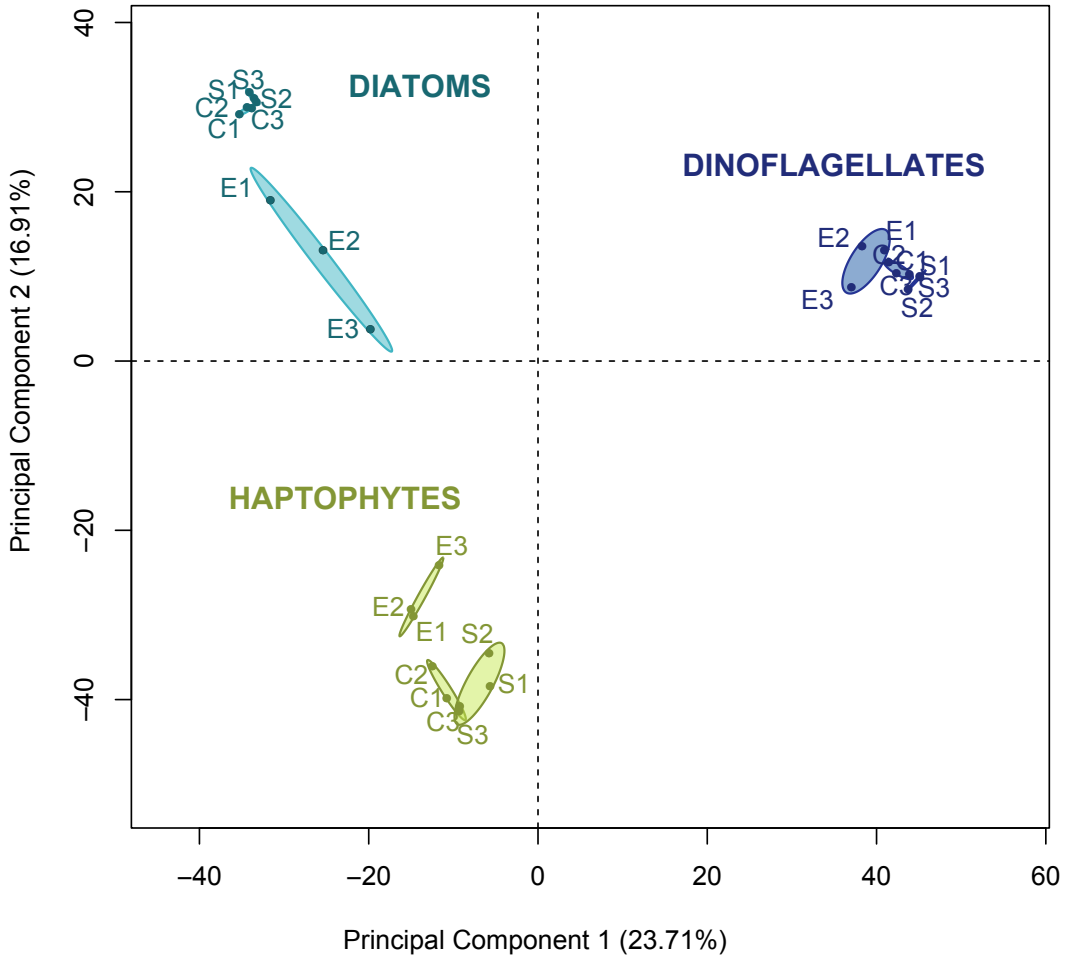


Figure C.7: Principal component analysis of the quantitative metabolic fingerprint (QMF) signals across *in situ*, no addition control, and deep seawater (DSW) amended samples. Principal component analysis of the QMF signals for each of the functional groups across *in situ* (S1-S3), control no addition (C1-C3) and DSW amendment (E1-E3); 95% confidence ellipses are indicated for each of the sample types by functional group.

C.2 Supplemental Tables

Table C.1: Macronutrient concentrations in control no addition (C), DSW-amended incubations (E), and 700 m water used in DSW amendment incubations

Treatment	**Time post inoculation (hours)**	**NO_2 + NO_3** (μM)	**PO_4** (μM)	**Si** (μM)
C (control no addition) *	168	0.12 ± 0.03	0.12 ± 0.02	1.91 ± 0.2
E (+ 10% DSW) *	168	$1.9 \pm .93$	0.23 ± 0.05	8.46 ± 3.11
DSW (700 m water) *	N/A	37.5 ± 1.68	3.14 ± 0.03	83.4 ± 9.33

** Nutrient data averaged for E1 and E2, nutrients were not assayed on E3.*

Appendix D

Chapter 5 Supplemental Information

D.1 Supplemental Figures

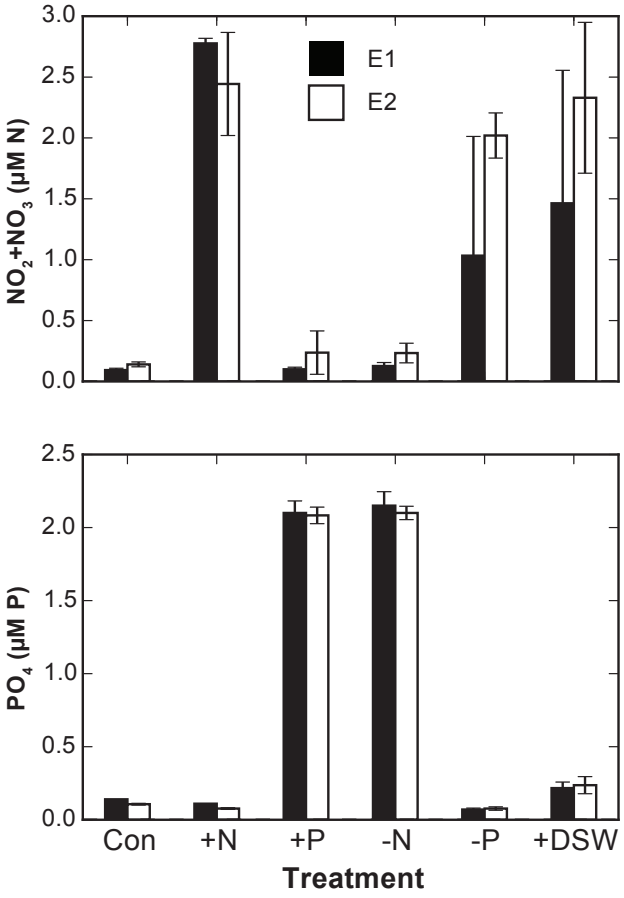


Figure D.1: Inorganic nitrogen and phosphorus concentrations at the point of RNA sampling (7 days post-inoculation) for each of the six treatments in E1 and E2, averaged across triplicate bottles (n=3).

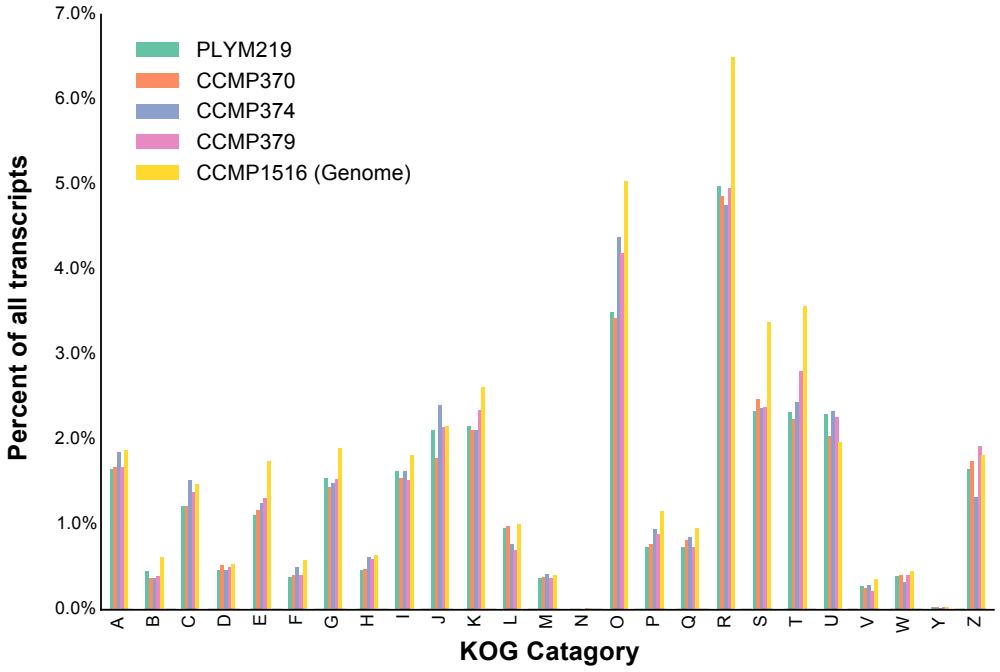


Figure D.2: The percent of genes falling into each of the KOG classes for each of the five strains.

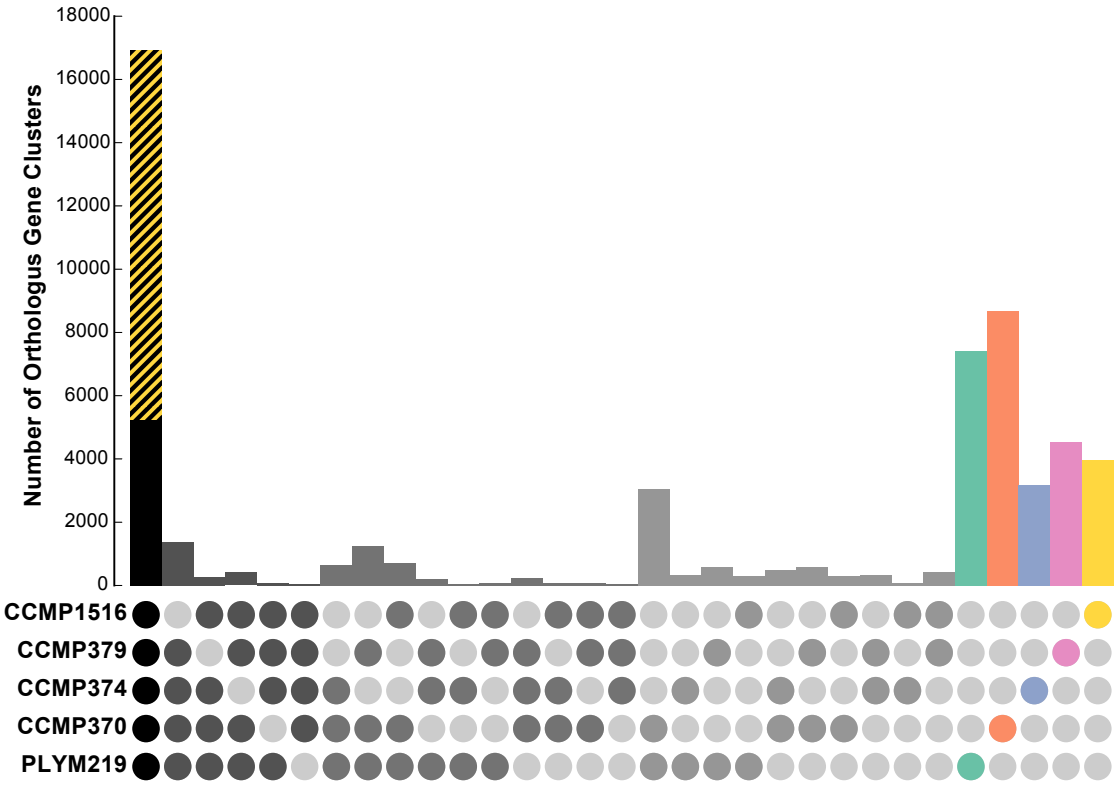


Figure D.3: The number of orthologous groups falling into each of the possible strain sets across the five strains surveyed. The relative strain membership is depicted in a scatter plot along the x-axis ranging from the first row of 'shared' or 'core' genes, common to all strains (black), through variable memberships across some but not all strains, to sets comprised of only one strain (colored). Genes common to all strains in this study are shown in black. Genes identified as 'core' in CCMP1516, the genome strain, by Read et al. (2013), but that were not identified in some or all of the other strains were added to the 'shared' set and are indicated in yellow hatching.

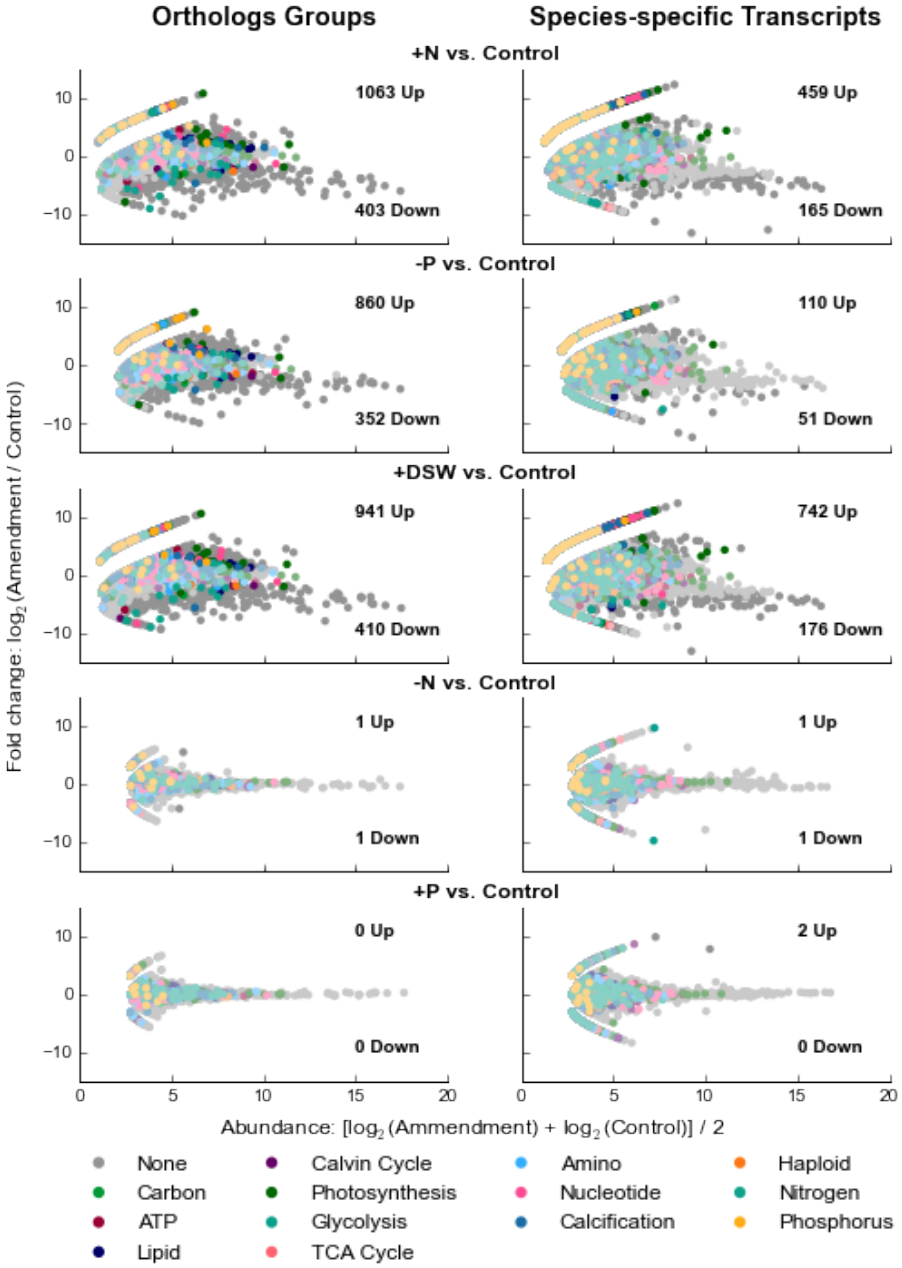


Figure D.4: Log normalized fold change plotted against log normalized average abundance for each of the five amended treatments compared to the no-addition control. edgeR was used to assess the average abundance and log fold change for each of the orthologous groups (left column) and strain-specific transcripts (right column). Genes are colored by generalized metabolic function. The intensity of the color indicates significance, with opaque indicating significance (FDR < 0.05).

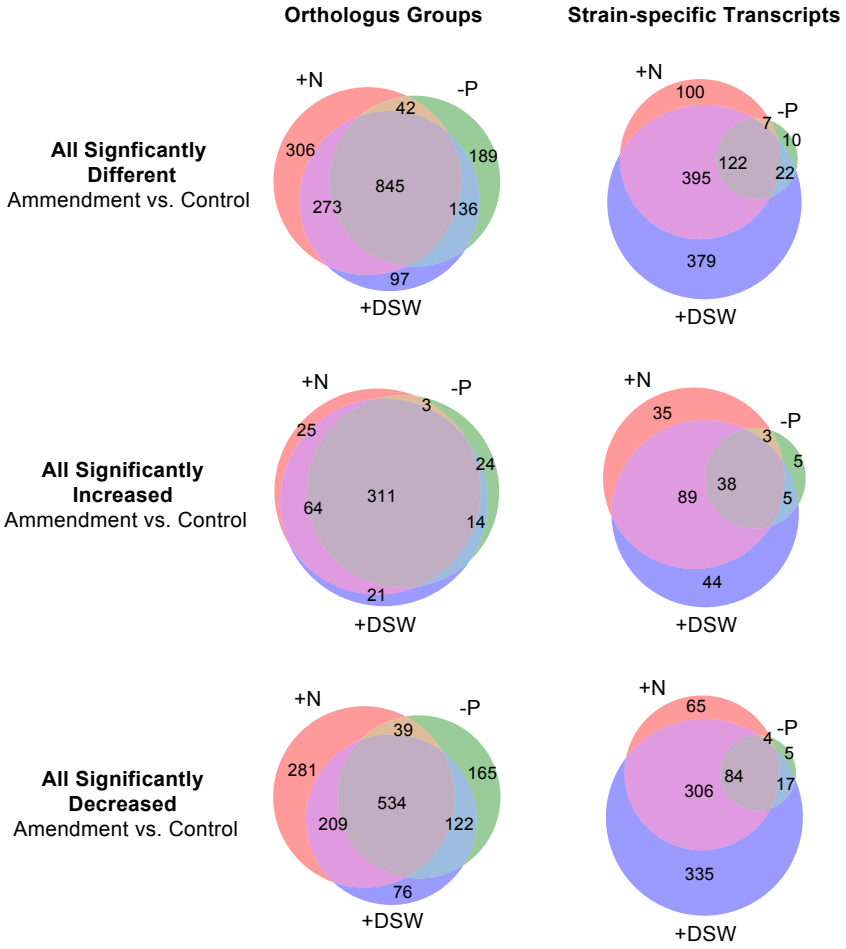


Figure D.5: Weighted Venn diagrams of significantly (FDR < 0.05) different, increased, and decreased orthologus groups and species-specific transcripts across each of the amendments to which N was added (+N, -P, +DSW).

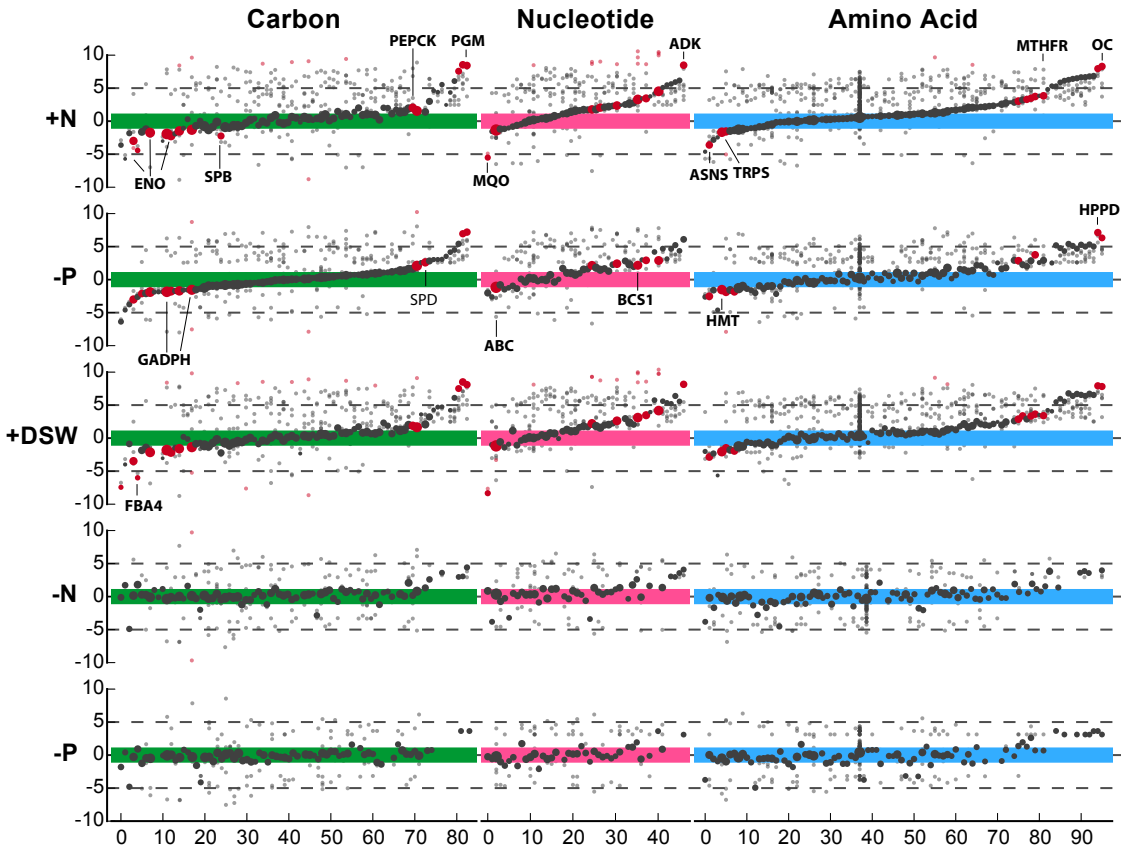


Figure D.6: Fold change of genes associated with carbon, nucleotide, and amino acid metabolism across each of the incubation amendments compared to the no addition control. The log fold change of orthologous groups associated with carbon, nucleotide, and amino acid metabolism was assessed with edgeR across the five amended incubations compared to the no addition control are plotted in opaque grey. The size of the orthologous group marker is proportionate to the log of the mean abundance across the two treatments. Orthologous groups are that are significantly differentially abundant (FDR < 0.05) are plotted highlighted in red. Individual transcripts within an orthologous group are plotted in light grey or red to indicate significance of fold change. Genes of interest are labeled with abbreviations as follows, labels in bold indicate significant regulation in two or more conditions.

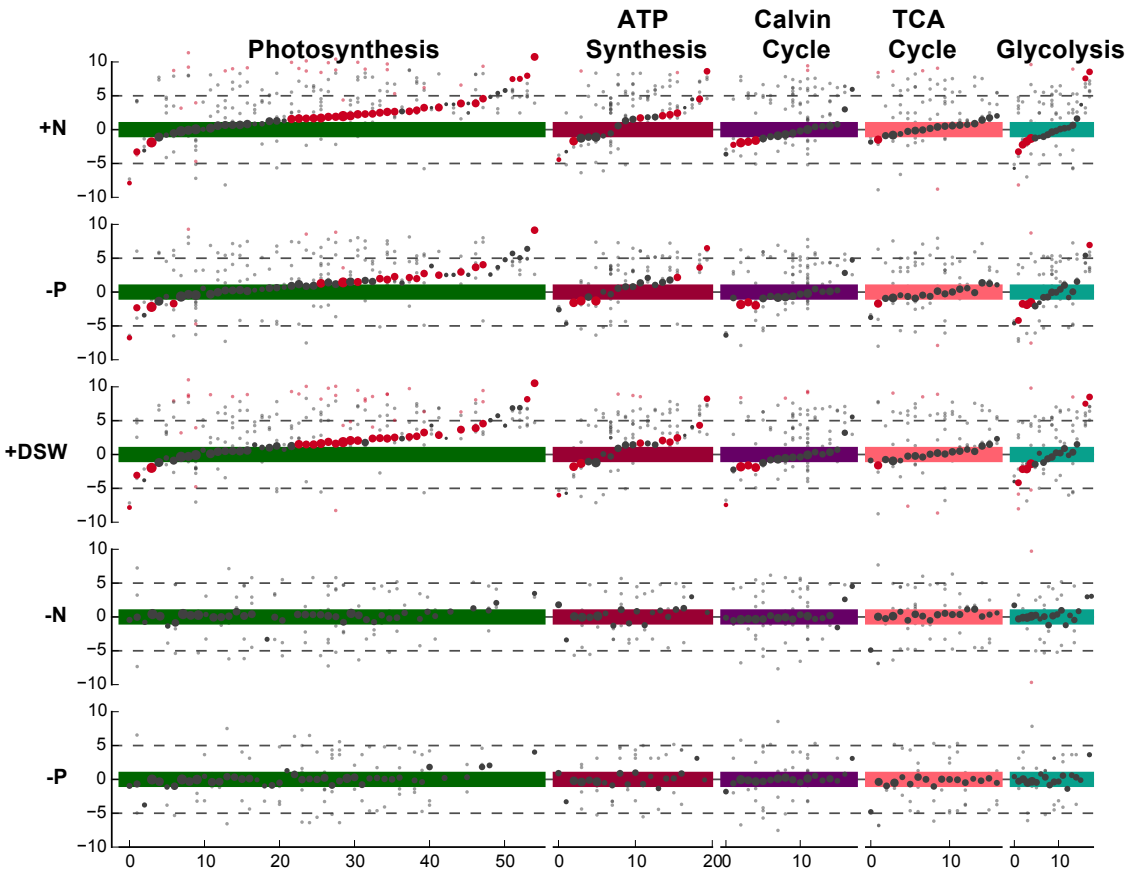


Figure D.7: Fold change of genes associated with photosynthesis, ATP synthesis, Calvin cycle, TCA cycle, and glycolysis across each of the incubation amendments compared to the no addition control. The log fold change of orthologous groups associated with photosynthesis, ATP synthesis, Calvin cycle, TCA cycle, and glycolysis was assessed with edgeR across the five amended incubations compared to the no addition control are plotted in opaque grey. The size of the orthologous group marker is proportionate to the log of the mean abundance across the two treatments. Orthologous groups are that are significantly differentially abundant (FDR < 0.05) are plotted highlighted in red. Individual transcripts within an orthologous group are plotted in light grey or red to indicate significance of fold change. Genes of interest are labeled with abbreviations as follows, labels in bold indicate significant regulation in two or more conditions.

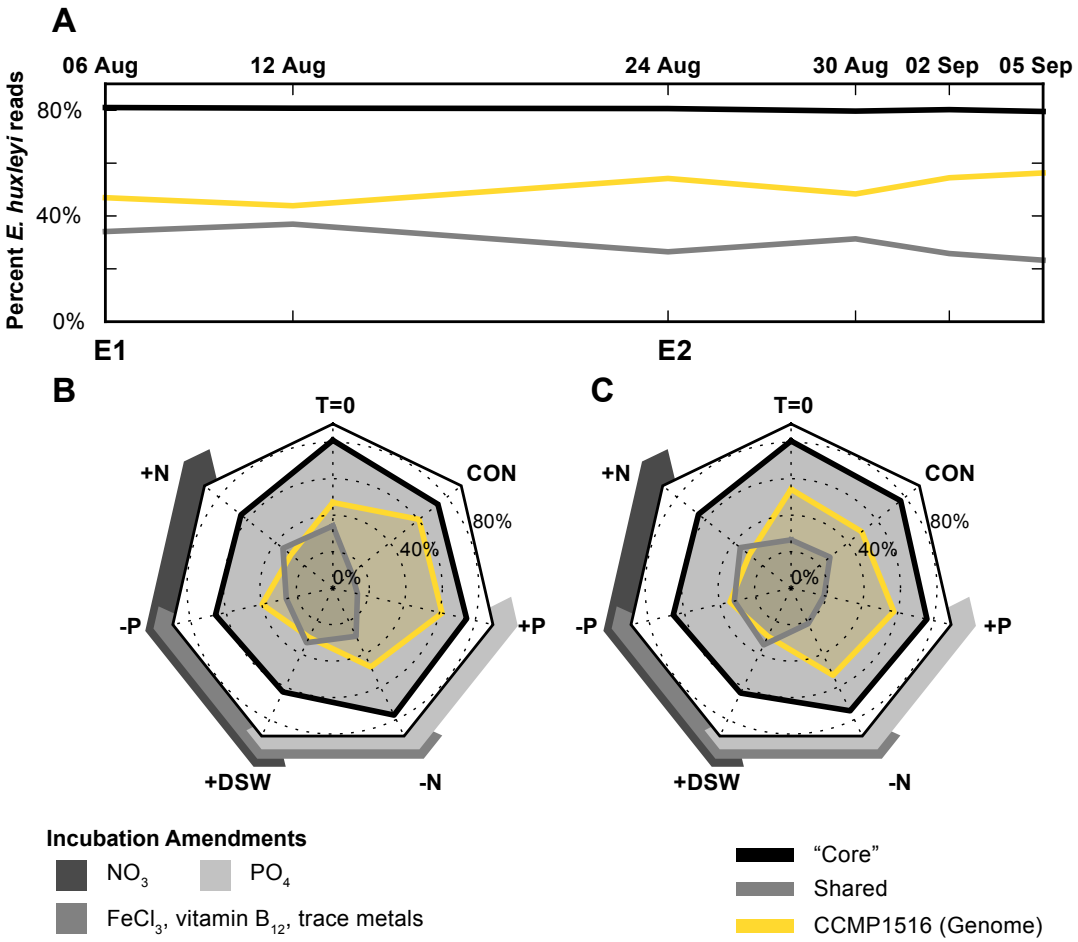


Figure D.8: The relative expression of 'core', shared, and CCMP1516-specific transcripts across time and in incubation experiments. The percentage of all mapped reads corresponding to the genes considered to be âĂŸcoreâĂŹ by Read et al. (2013) (black), found to be shared across the five strains used in this study (grey), or originally considered to be unique to CCMP1516 are plotted for each in situ sample (A) and in each of the two replicated incubation experiments, E1 (B) and E2 (C). Nutrients added to incubation experiments are indicated on the exterior of the radar plots, indicating the addition of nitrate, phosphate, trace metals, and vitamins.

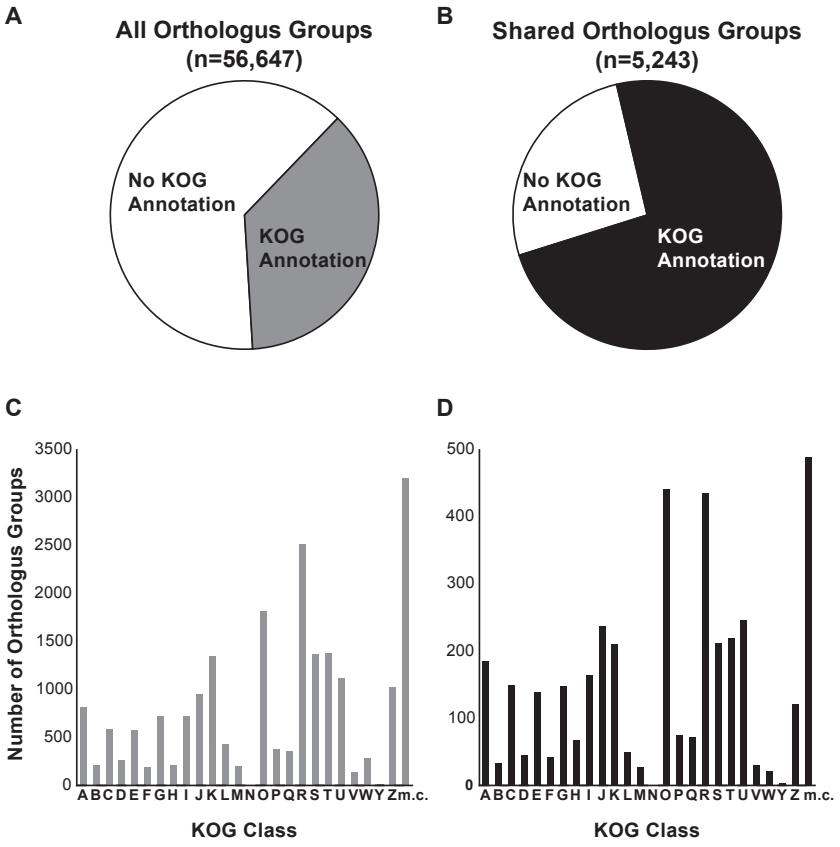


Figure D.9: Annotation of orthologous groups using KOG orthology for all *E. huxleyi* orthologous groups and for shared orthologous groups. The relative percentage of orthologous groups able to be annotated for all orthologous groups (A) and orthologous group shared amongst the five studied strains (B) are shown. The number of orthologous groups falling into each KOG class or multiple classes (m.c.) is shown for both all orthologous groups (C) and shared groups (D).

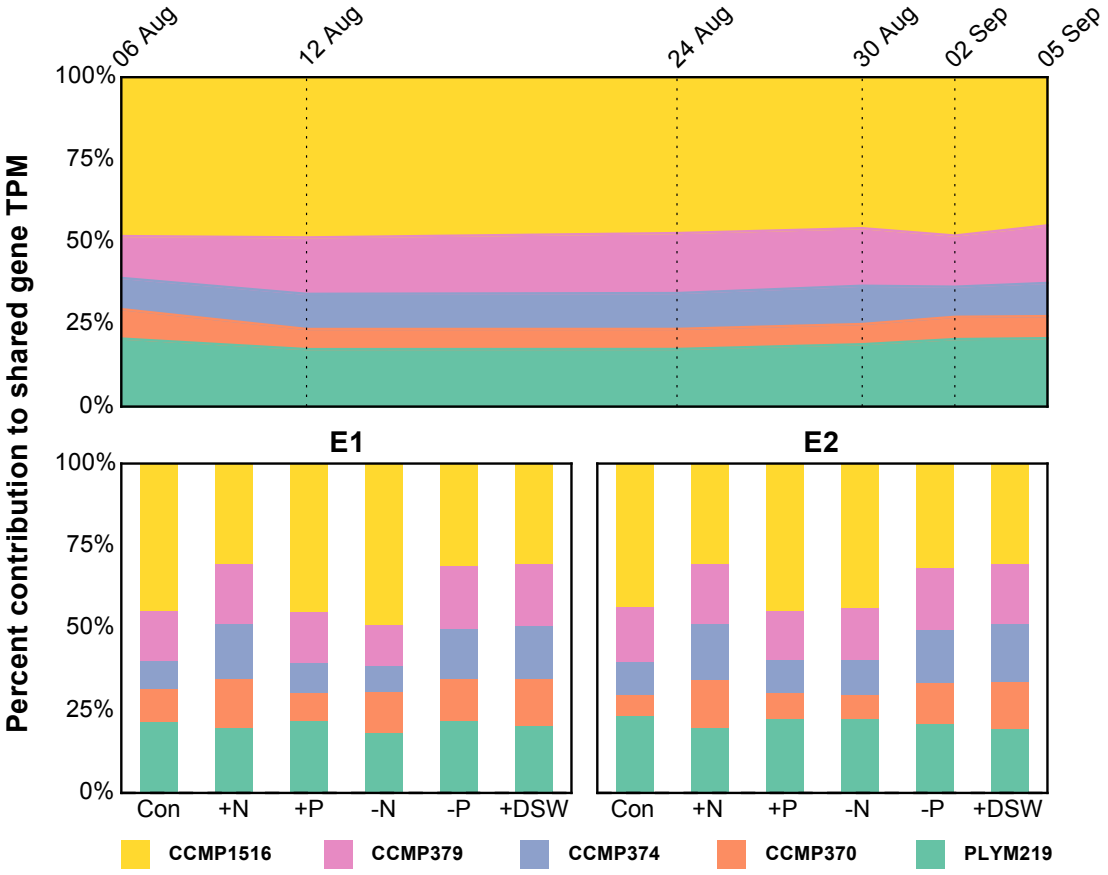


Figure D.10: ****The RSEM estimated contribution of each strain to the abundance of the shared set of genes in the field and incubation experiments.

Bibliography

Alcolombri, U., Ben-Dor, S., Feldmesser, E., Levin, Y., Tawfik, D. S., and Vardi, a. Identification of the algal dimethyl sulfide-releasing enzyme: A missing link in the marine sulfur cycle. *Science*, 348(6242):1466–1469, 2015. doi: 10.1126/science.aab1586.

Alexander, H., Jenkins, B. D., Rynearson, T. a., Saito, M. A., Mercier, M. L., and Dyhrman, S. T. Identifying reference genes with stable expression from high throughput sequence data. *Frontiers in Microbiology*, 3(November):385, January 2012. doi: 10.3389/fmicb.2012.00385.

Alexander, H., Jenkins, B. D., Rynearson, T. A., and Dyhrman, S. T. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proceedings of the National Academy of Sciences of the United States of America*, 112(17):E2182–E2190, April 2015a. doi: 10.1073/pnas.1421993112.

Alexander, H., Rouco, M., Haley, S. T., Wilson, S. T., Karl, D. M., and Dyhrman, S. T. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proceedings of the National Academy of Sciences*, 112(44):E5972–E5979, October 2015b. doi: 10.1073/pnas.1518165112.

Allen, A. E., Laroche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P. J., Finazzi, G., Fernie, A. R., and Bowler, C. Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(30):10438–43, July 2008. doi: 10.1073/pnas.0711370105.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997. doi: 10.1093/nar/25.17.3389.

Amin, S. A., Hmelo, L. R., van Tol, H. M., Durham, B. P., Carlson, L. T., Heal, K. R., Morales, R. L., Berthiaume, C. T., Parker, M. S., Djunaedi, B., Ingalls, A. E., Parsek, M. R., Moran, M. A., and Armbrust, E. V. Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature*, 522:98–101, May 2015. doi: 10.1038/nature14488.

Anders, S., Pyl, P. T., and Huber, W. HTSeq - A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 31(2):166–9, September 2014. doi: 10.1093/bioinformatics/btu638.

Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data.

Antonov, J., Goldstein, D. R., Oberli, A., Baltzer, A., Pirotta, M., Fleischmann, A., Altermatt, H. J., and Jaggi, R. Reliable gene expression measurements from degraded RNA by quantitative real-time PCR depend on short amplicons and a proper normalization. *Laboratory Investigation*, 85(8):1040–1050, August 2005. doi: 10.1038/labinvest.3700303.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov, V. V., Kröger, N., Lau, W. W. Y., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P., and Rokhsar, D. S. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science (New York, N.Y.)*, 306(5693): 79–86, October 2004. doi: 10.1126/science.1101156.

Avonce, N., Leyman, B., Mascorro-Gallardo, J. O., Van Dijck, P., Thevelein, J. M., and Iturriaga, G. The *Arabidopsis* trehalose-6-P synthase AtTPS1 gene is a regulator of glucose, abscisic acid, and stress signaling. *Plant Physiology*, 136(3):3649 –3659, November 2004. doi: 10.1104/pp.104.052084.

Baas-Becking, L. Geobiologie; of inleiding tot de milieukunde. 1934.

Balvanera, P., Pfisterer, A. B., Buchmann, N., He, J.-S., Nakashizuka, T., Raffaelli, D., and Schmid, B. Quantifying the evidence for biodiversity effects on ecosystem functioning and services. *Ecology letters*, 9(10):1146–56, October 2006. doi: 10.1111/j.1461-0248.2006.00963.x.

Beaufort, L., Probert, I., de Garidel-Thoron, T., Bendif, E. M., Ruiz-Pino, D., Metzl, N., Goyet, C., Buchet, N., Coupel, P., Grelaud, M., Rost, B., Rickaby, R. E. M., and de Vargas, C. Sensitivity of coccolithophores to carbonate chemistry and ocean acidification. *Nature*, 476(7358):80–3, August 2011. doi: 10.1038/nature10295.

Behl, S., Donval, A., and Stibor, H. The relative importance of species diversity and functional group diversity on carbon uptake in phytoplankton communities. *Limnology and Oceanography*, 56(2):683–694, 2011. doi: 10.4319/lo.2011.56.2.0683.

Bender, S. J., Parker, M. S., and Armbrust, E. V. Coupled effects of light and nitrogen source on the urea cycle and nitrogen metabolism over a diel cycle in the marine diatom *Thalassiosira pseudonana*. *Protist*, 163(2):232–251, March 2012. doi: 10.1016/j.protis.2011.07.008.

Bender, S. J., Durkin, C. A., Berthiaume, C. T., Morales, R. L., and Armbrust, E. V. Transcriptional responses of three model diatoms to nitrate limitation of growth. *Frontiers in Marine Science*, 1, March 2014. doi: 10.3389/fmars.2014.00003.

Benitez-Nelson, C. R., Bidigare, R. R., Dickey, T. D., Landry, M. R., Leonard, C. L., Brown, S. L., Nencioli, F., Rii, Y. M., Maiti, K., Becker, J. W., Bibby, T. S., Black, W., Cai, W.-J., Carlson, C. A., Chen, F., Kuwahara, V. S., Mahaffey, C., McAndrew, P. M., Quay, P. D., Rappé, M. S., Selph, K. E., Simmons, M. P., and Yang, E. J. Mesoscale eddies drive increased silica export in the subtropical Pacific Ocean. *Science (New York, N.Y.)*, 316(5827):1017–21, May 2007. doi: 10.1126/science.1136221.

Berg, G. M., Shrager, J., Glöckner, G., Arrigo, K. R., and Grossman, A. R. Understanding nitrogen limitation in *Aureococcus anophagefferens* (Pelagophyceae) through cDNA and qRT-PCR analysis. *Journal of Phycology*, 44(5):1235–1249, October 2008. doi: 10.1111/j.1529-8817.2008.00571.x.

Bertrand, E. M., Saito, M. A., Rose, J. M., Riesselman, C. R., Lohan, M. C., Noble, A. E., Lee, P. A., DiTullio, G. R., Maeve, C., Saitol, M. A., and Sedwick, P. Vitamin B12 and iron colimitation of phytoplankton growth in the Ross Sea. *Limnology and Oceanography*, 52(3):1079–1093, May 2007.

Bertrand, E. M., Allen, A. E., Dupont, C. L., Norden-Krichmar, T. M., Bai, J., Valas, R. E., and Saito, M. A. Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. *Proceedings of the National Academy of Sciences of the United States of America*, 109(26):E1762–71, June 2012. doi: 10.1073/pnas.1201731109.

Biller, S. J., Berube, P. M., Lindell, D., and Chisholm, S. W. Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology*, 13(1):13–27, December 2014. doi: 10.1038/nrmicro3378.

Bollmann, J. and Herrle, J. O. Morphological variation of *Emiliania huxleyi* and sea surface salinity. *Earth and Planetary Science Letters*, 255(3-4):273–288, March 2007. doi: 10.1016/j.epsl.2006.12.029.

Borkman, D. G. and Smayda, T. Multidecadal (1959–1997) changes in Skeletonema abundance and seasonal bloom patterns in Narragansett Bay, Rhode Island, USA. *Journal of Sea Research*, 61(1-2):84–94, January 2009. doi: 10.1016/j.seares.2008.10.004.

Bowler, C., Allen, A. E., Badger, J. H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R. P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J. A., Brownlee, C., Cadoret, J.-P., Chiovitti, A., Choi, C. J., Coesel, S., De Martino, A., Detter, J. C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M. J. J., Jenkins, B. D., Jiroutova, K., Jorgensen, R. E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P. G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P. J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L. K., Montsant, A., Oudot-Le Secq, M.-P., Napoli, C., Obornik, M., Parker, M. S., Petit, J.-L., Porcel, B. M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T. A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M. R., Taylor, A. R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L. S., Rokhsar, D. S., Weissenbach, J., Armbrust, E. V., Green, B. R., Van de Peer, Y., and Grigoriev, I. V. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239–44, November 2008. doi: 10.1038/nature07410.

Brown, C. W. and Yoder, J. A. Coccolithophorid blooms in the global ocean. *Journal of Geophysical Research*, 99(C4):7467, 1994. doi: 10.1029/93JC02156.

Bruhn, A., LaRoche, J., and Richardson, K. *Emiliania huxleyi* (prymnesiophyceae): nitrogen-metabolism genes and their expression in response to external nitrogen sources. *Journal of Phycology*, 46(2):266–277, April 2010. doi: 10.1111/j.1529-8817.2010.00809.x.

Bustin, S. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, 25(2):169–193, October 2000. doi: 10.1677/jme.0.0250169.

Cadotte, M. W. Experimental evidence that evolutionarily diverse assemblages result in higher productivity. *Proceedings of the National Academy of Sciences of the United States of America*, 110(22):8996–9000, May 2013. doi: 10.1073/pnas.1301685110.

Caldwell, P. E., Walkiewicz, M., and Stern, M. Ras activity in the *Drosophila* prothoracic gland regulates body size and developmental rate via ecdysone release. *Current Biology*, 15(20):1785–1795, October 2005. doi: 10.1016/j.cub.2005.09.011.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., Narwani, A., Mace, G. M., Tilman, D., Wardle, D. A., Kinzig, A. P., Daily, G. C., Loreau, M., Grace, J. B., Larigauderie, A., Srivastava, D. S., and Naeem, S. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59–67, June 2012. doi: 10.1038/nature11148.

Carvalho, R. N. and Lettieri, T. Proteomic analysis of the marine diatom *Thalassiosira pseudonana* upon exposure to benzo(a)pyrene. *BMC Genomics*, 12(1):159, March 2011. doi: 10.1186/1471-2164-12-159.

Casciotti, K., Trull, T., Glover, D., and Davies, D. Constraints on nitrogen cycling at the subtropical North Pacific Station ALOHA from isotopic measurements of nitrate and particulate nitrogen. *Deep Sea Research Part II: Topical Studies in Oceanography*, 55 (14-15):1661–1672, July 2008. doi: 10.1016/j.dsr2.2008.04.017.

Cavender-Bares, K. K., Karl, D. M., and Chisholm, S. W. Nutrient gradients in the western North Atlantic Ocean: relationship to microbial community structure and comparison to patterns in the Pacific Ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, 48(11):2373–2395, November 2001. doi: 10.1016/S0967-0637(01)00027-9.

Chung, C.-C., Hwang, S.-P. L., and Chang, J. Identification of a high-affinity phosphate transporter gene in a prasinophyte alga, Tetraselmis chui, and its expression under nutrient limitation. *Applied and environmental microbiology*, 69(2):754–9, February 2003.

Coker, J. S. and Davies, E. Selection of candidate housekeeping controls in tomato plants using EST data. *BioTechniques*, 35(4):740–742, 744, 746 passim, October 2003.

Connel, J. and Connell, J. Diversity and the coevolution of competitors, or the ghost of competition past. *Oikos*, 35(2):131–138, October 1980.

Cortés, M. Y. M., Bollmann, J., and Thierstein, H. H. R. Coccolithophore ecology at the HOT station ALOHA, Hawaii. *Deep Sea Research Part II: Topical Studies in Oceanography*, 48(8-9):1957–1981, January 2001. doi: 10.1016/S0967-0645(00)00165-X.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., and Scheible, W.-R. Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiology*, 139(1):5 –17, 2005. doi: 10.1104/pp.105.063743.

Davis, A. K. and Palenik, B. Characterization of a modular, cell-surface, protein and identification of a new gene family in the diatom *Thalassiosira pseudonana*. *Protist*, 159 (2):195–207, April 2008. doi: 10.1016/j.protis.2007.09.006.

Davis, A. K., Hildebrand, M., and Palenik, B. Gene expression induced by copper stress in diatom *Thalassiosira pseudonana*. *Eukaryotic Cell*, 5(7):1157 –1168, July 2006. doi: 10.1128/EC.00042-06.

de Jonge, H. J. M., Fehrmann, R. S. N., de Bont, E. S. J. M., Hofstra, R. M. W., Gerbens, F., Kamps, W. A., de Vries, E. G. E., van der Zee, A. G. J., te Meerman, G. J., and ter Elst, A. Evidence based selection of housekeeping genes. *PloS one*, 2(9):e898, January 2007. doi: 10.1371/journal.pone.0000898.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Luke, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sullivan, M. B., and Velayoudon, D. Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237):1261605–1261605, May 2015. doi: 10.1126/science.1261605.

de Wit, R., Bouvier, T., Wit, R. D., and Bouvier, T. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental microbiology*, 8(4):755–8, April 2006. doi: 10.1111/j.1462-2920.2006.01017.x.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A. Z., Robbens, S., Partensky, F., Degroeve, S., Echeynié, S., Cooke, R., Saeys, Y., Wuyts, J., Jabbari, K., Bowler, C., Panaud, O., Piégu, B., Ball, S. G., Ral, J.-P., Bouget, F.-Y., Piganeau, G., De Baets, B., Picard, A., Delseny, M., Demaille, J., Van de Peer, Y., and Moreau, H. Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. *Proceedings of the National Academy of Sciences*, 103(31):11647–11652, 2006. doi: 10.1073/pnas.0604795103.

D'haeseleer, P. How does gene expression clustering work? *Nature biotechnology*, 23(12): 1499–501, December 2005. doi: 10.1038/nbt1205-1499.

Dheda, K., Huggett, J. F., Chang, J. S., Kim, L. U., Bustin, S. A., Johnson, M. A., Rook, G. A. W., and Zumla, A. The implications of using an inappropriate reference gene for real-time reverse transcription PCR data normalization. *Analytical Biochemistry*, 344(1): 141–143, September 2005. doi: 10.1016/j.ab.2005.05.022.

Doney, S. C., Ruckelshaus, M., Emmett Duffy, J., Barry, J. P., Chan, F., English, C. A., Galindo, H. M., Grebmeier, J. M., Hollowed, A. B., Knowlton, N., Polovina, J., Rabalais, N. N., Sydeman, W. J., and Talley, L. D. Climate change impacts on marine ecosystems. *Annual Review of Marine Science*, 4(1):11–37, January 2012. doi: 10.1146/annurev-marine-041911-111611.

Dupont, C. L., McCrow, J. P., Valas, R., Moustafa, A., Walworth, N., Goodenough, U., Roth, R., Hogle, S. L., Bai, J., Johnson, Z. I., Mann, E., Palenik, B., Barbeau, K. A., Craig Venter, J., and Allen, A. E. Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *The ISME journal*, 9(5): 1076–92, May 2015. doi: 10.1038/ismej.2014.198.

Dyhrman, S. T. and Anderson, D. M. Urease activity in cultures and field populations of the toxic dinoflagellate Alexandrium. *Limnology and Oceanography*, 48(2):647–655, March 2003. doi: 10.4319/lo.2003.48.2.0647.

Dyhrman, S. T. and Palenik, B. Characterization of ectoenzyme activity and phosphate-regulated proteins in the coccolithophorid *Emiliania huxleyi*. *Journal of Plankton Research*, 25(10):1215–1225, October 2003. doi: 10.1093/plankt/fbg086.

Dyhrman, S. T. and Palenik, B. P. The identification and purification of a cell-surface alkaline phosphatase from the dinoflagellate *Prorocentrum minimum* (Dinophyceae). *Journal of Phycology*, 33(4):602–612, August 1997. doi: 10.1111/j.0022-3646.1997.00602.x.

Dyhrman, S. T., Haley, S. T., Birkeland, S. R., Wurch, L. L., Cipriano, M. J., and McArthur, A. G. Long serial analysis of gene expression for gene discovery and transcriptome profiling in the widespread marine coccolithophore *Emiliania huxleyi*. *Applied and Environmental Microbiology*, 72(1):252–260, January 2006. doi: 10.1128/AEM.72.1.252-260.2006.

Dyhrman, S. T., Jenkins, B. D., Rynearson, T. A., Saito, M. A., Mercier, M. L., Alexander, H., Whitney, L. P., Drzewianowski, A., Bulygin, V. V., Bertrand, E. M., Wu, Z., Benitez-Nelson, C., and Heithoff, A. The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response. *PloS one*, 7(3):e33768, January 2012. doi: 10.1371/journal.pone.0033768.

Elton, C. S. *The Ecology of Invasions by Animals and Plants*. Springer US, Boston, MA, 1958. ISBN 978-0-412-11430-4. doi: 10.1007/978-1-4899-7214-9.

Feng, Y., Warner, M. E., Zhang, Y., Sun, J., Fu, F.-X., Rose, J. M., and Hutchins, D. A. Interactive effects of increased pCO_2, temperature and irradiance on the marine coccolithophore Emiliania huxleyi (Prymnesiophyceae). *European Journal of Phycology*, 43(1): 87–98, February 2008. doi: 10.1080/09670260701664674.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. G. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science (New York, N.Y.)*, 281(5374):237–240, July 1998. doi: 10.1126/science.281.5374.237.

Flatt, T., Min, K.-J., D'Alterio, C., Villa-Cuesta, E., Cumbers, J., Lehmann, R., Jones, D. L., and Tatar, M. *Drosophila* germ-line modulation of insulin signaling and lifespan. *Proceedings of the National Academy of Sciences*, 105(17):6368 –6373, April 2008. doi: 10.1073/pnas.0709128105.

Follows, M. J. and Dutkiewicz, S. Modeling diverse communities of marine microbes. *Annual review of marine science*, 3(1):427–51, January 2011. doi: 10.1146/annurev-marine-120709-142848.

Follows, M. J., Dutkiewicz, S., Grant, S., and Chisholm, S. W. Emergent biogeography of microbial communities in a model ocean. *Science (New York, N.Y.)*, 315(5820):1843–6, March 2007. doi: 10.1126/science.1138544.

Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M., and DeLong, E. F. Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Frontiers in Microbiology*, 6:469, May 2015. doi: 10.3389/fmicb.2015.00469.

Frada, M., Probert, I., Allen, M. J., Wilson, W. H., and de Vargas, C. The "Cheshire Cat" escape strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (41):15944–9, October 2008. doi: 10.1073/pnas.0807707105.

Frada, M. J., Bidle, K. D., Probert, I., and de Vargas, C. In situ survey of life cycle phases of the coccolithophore *Emiliania huxleyi* (Haptophyta). *Environmental microbiology*, 14 (6):1558–69, June 2012. doi: 10.1111/j.1462-2920.2012.02745.x.

Frischkorn, K. R., Harke, M. J., Gobler, C. J., and Dyhrman, S. T. De novo assembly of *Aureococcus anophagefferens* transcriptomes reveals diverse responses to the low nutrient and low light conditions present during blooms. *Frontiers in Microbiology*, 5:375, July 2014. doi: 10.3389/fmicb.2014.00375.

Furnas, M. J. Growth rates of summer nanoplankton (<10 micrometers) populations in lower Narragansett Bay, Rhode Island, USA. *Marine Biology*, 70(1):105–115, 1982. doi: 10.1007/BF00397301.

Furnas, M. J. Community structure, biomass and productivity of size-fractionated summer phytoplankton populations in lower Narragansett Bay, Rhode Island. *Journal of Plankton Research*, 5(5):637–655, 1983. doi: 10.1093/plankt/5.5.637.

Gerstein, M. and Jansen, R. The current excitement in bioinformatics– analysis of whole-genome expression data: how does it relate to protein structure and function? *Current Opinion in Structural Biology*, 10(5):574–584, October 2000. doi: 10.1016/S0959-440X(00)00134-2.

Gibbons, F. D. and Roth, F. P. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, 12(10):1574–1581, October 2002. doi: 10.1101/gr.397002.

Gifford, S. M., Sharma, S., Rinta-Kanto, J. M., and Moran, M. A. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *The ISME Journal*, 5(3):461–472, March 2011. doi: 10.1038/ismej.2010.141.

Gifford, S. M., Sharma, S., Booth, M., and Moran, M. A. Expression patterns reveal niche diversification in a marine microbial assemblage. *The ISME Journal*, 7(2):281–298, February 2013. doi: 10.1038/ismej.2012.96.

Green, J., Course, P., and Tarran, G. The life-cycle of *Emiliania huxleyi*: A brief review and a study of relative ploidy levels analysed by flow cytometry. *Journal of Marine Systems*, 9(1-2):33–44, October 1996. doi: 10.1016/0924-7963(96)00014-0.

Guillard, R. R. L. *Culture of marine invertebrate animals*. Springer US, Boston, MA, 1975. ISBN 978-1-4615-8716-3. doi: 10.1007/978-1-4615-8714-9.

Guo, R. and Ki, J.-S. Evaluation and validation of internal control genes for studying gene expression in the dinoflagellate Prorocentrum minimum using real-time PCR. *European journal of protistology*, 48(3):199–206, August 2012. doi: 10.1016/j.ejop.2011.11.001.

Hackett, J. D., Anderson, D. M., Erdner, D. L., and Bhattacharya, D. Dinoflagellates: a remarkable evolutionary experiment. *American Journal of Botany*, 91(10):1523–1534, 2004.

Halary, S., Leigh, J. W., Cheaib, B., Lopez, P., and Bapteste, E. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1):127–32, January 2010. doi: 10.1073/pnas.0908978107.

Halary, S., McInerney, J. O., Lopez, P., and Bapteste, E. EGN: a wizard for construction of gene and genome similarity networks. *BMC evolutionary biology*, 13(1):146, January 2013. doi: 10.1186/1471-2148-13-146.

Hartigan, J. A. and Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, January 1979. doi: 10.2307/2346830.

Hennon, G. M. M., Ashworth, J., Groussman, R. D., Berthiaume, C., Morales, R. L., Baliga, N. S., Orellana, M. V., and Armbrust, E. V. Diatom acclimation to elevated CO_2 via cAMP signalling and coordinated gene expression. *Nature Climate Change*, 5(8):761–765, June 2015. doi: 10.1038/nclimate2683.

Holligan, P. M., Fernández, E., Aiken, J., Balch, W. M., Boyd, P., Burkill, P. H., Finch, M., Groom, S. B., Malin, G., Muller, K., Purdie, D. A., Robinson, C., Trees, C. C., Turner, S. M., and van der Wal, P. A biogeochemical study of the coccolithophore, *Emiliania huxleyi*, in the North Atlantic. *Global Biogeochemical Cycles*, 7(4):879–900, December 1993. doi: 10.1029/93GB01731.

Hothorn, M., Neumann, H., Lenherr, E. D., Wehner, M., Rybin, V., Hassa, P. O., Uttenweiler, A., Reinhardt, M., Schmidt, A., Seiler, J., Ladurner, A. G., Herrmann, C., Scheffzek, K., and Mayer, A. Catalytic core of a membrane-associated eukaryotic polyphosphate polymerase. *Science (New York, N.Y.)*, 324(5926):513–6, April 2009. doi: 10.1126/science.1168120.

Huisman, J. and Weissing, F. Biodiversity of plankton by species oscillations and chaos. *Nature*, 402(6760):407–410, November 1999. doi: 10.1038/46540.

Huisman, J., Johansson, A. M., Folmer, E. O., and Weissing, F. J. Towards a solution of the plankton paradox: the importance of physiology and life history. *Ecology Letters*, 4 (5):408–411, September 2001. doi: 10.1046/j.1461-0248.2001.00256.x.

Hutchins, D. A., Witter, A. E., Butler, A., and Luther, G. W. Competition among marine phytoplankton for different chelated iron species. *Nature*, 400(6747):858–861, August 1999. doi: 10.1038/23680.

Hutchinson, G. E. The Paradox of the Plankton. *The American Naturalist*, 95(882):137, January 1961. doi: 10.1086/282171.

Iglesias-Rodriguez, D. M., Schofield, O. M., Batley, J., Medlin, L. K., and Hayes, P. K. Intraspecific genetic diversity in the marine coccolithophore *Emiliania huxleyi* (Prymnesiophyceae): the use of microsatellite analysis in marine phytoplankton population studies. *Journal of Phycology*, 42(3):526–536, June 2006. doi: 10.1111/j.1529-8817.2006.00231.x.

Iglesias-Rodriguez, M. D., Halloran, P. R., Rickaby, R. E. M., Hall, I. R., Colmenero-Hidalgo, E., Gittins, J. R., Green, D. R. H., Tyrrell, T., Gibbs, S. J., von Dassow, P., Rehm, E., Armbrust, E. V., and Boessenkool, K. P. Phytoplankton calcification in a high-CO_2 world. *Science (New York, N.Y.)*, 320(5874):336–40, April 2008. doi: 10.1126/science.1154122.

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., and Chisholm, S. W. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science (New York, N.Y.)*, 311(5768):1737–40, March 2006. doi: 10.1126/science.1118052.

Jones, B. M., Iglesias-Rodriguez, M. D., Skipp, P. J., Edwards, R. J., Greaves, M. J., Young, J. R., Elderfield, H., and O'Connor, C. D. Responses of the *Emiliania huxleyi* proteome to ocean acidification. *PloS one*, 8(4):e61868, January 2013. doi: 10.1371/journal.pone.0061868.

Kang, L.-K., Hwang, S.-P. L., Lin, H.-J., Chen, P.-C., and Chang, J. Establisment of minimal and maximal transcript levels for nitrate transporter genes genes for detecting nitrogen deficiency in the marine phytoplankton Isochrysis glabana (Prymnesiophycae) and *Thalassiosira pseudonana* (Bacillariophycae). *Journal of Phycology*, 45(4):864–872, August 2009. doi: 10.1111/j.1529-8817.2009.00698.x.

Karl, D., Letelier, R., Tupas, L., Dore, J., Christian, J., and Hebel, D. The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature*, 388 (6642):533–538, August 1997.

Karl, D. M. Nutrient dynamics in the deep blue sea. *Trends in Microbiology*, 10(9):410–418, September 2002. doi: 10.1016/S0966-842X(02)02430-7.

Karl, D. M., Björkman, K. M., Dore, J. E., Fujieki, L., Hebel, D. V., Houlihan, T., Letelier, R. M., and Tupas, L. M. Ecological nitrogen-to-phosphorus stoichiometry at station ALOHA. *Deep Sea Research Part II: Topical Studies in Oceanography*, 48(8-9):1529–1566, 2001. doi: 10.1016/S0967-0645(00)00152-1.

Karl, D. M., Laws, E. A., Morris, P., Williams, P. J. L., and Emerson, S. Global carbon cycle: metabolic balance of the open sea. *Nature*, 426(6962):32, November 2003. doi: 10.1038/426032a.

Karl, D., Church, M., Dore, J. E., Letelier, R. M., and Mahaffey, C. Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6): 1842–1849, February 2012. doi: 10.1073/pnas.1120312109.

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., and Chisholm, S. W. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science (New York, N.Y.)*, 344(6182):416–20, May 2014. doi: 10.1126/science.1248575.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S.,

Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. B., Schroeder, D. C., Simpson, A. G. B., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., and Worden, A. Z. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology*, 12(6):e1001889, June 2014. doi: 10.1371/journal.pbio.1001889.

Kim, I.-N., Lee, K., Gruber, N., Karl, D. M., Bullister, J. L., Yang, S., and Kim, T.-W. Increasing anthropogenic nitrogen in the North Pacific Ocean. *Science (New York, N.Y.)*, 346(6213):1102–6, November 2014. doi: 10.1126/science.1258396.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–45, September 2009. doi: 10.1101/gr.092759.109.

Kustka, A. B., Allen, A. E., and Morel, F. M. M. Sequence analysis and transcriptional regulation of iron acquisition genes in two marine diatoms. *Journal of Phycology*, 43(4): 715–729, August 2007. doi: 10.1111/j.1529-8817.2007.00359.x.

Langer, G., Nehrke, G., Probert, I., Ly, J., and Ziveri, P. Strain-specific responses of *Emiliania huxleyi* to changing seawater carbonate chemistry. *Biogeosciences*, 6(11):2637–2646, November 2009. doi: 10.5194/bg-6-2637-2009.

Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, April 2012. doi: 10.1038/nmeth.1923.

Lanoix, D., Lacasse, A.-A., St-Pierre, J., Taylor, S. C., Ethier-Chiasson, M., Lafond, J., and Vaillancourt, C. Quantitative PCR pitfalls: the case of the human placenta. *Molecular biotechnology*, 52(3):234–43, November 2012. doi: 10.1007/s12033-012-9539-2.

Lawerence, C. and Menden-Deuer, S. Drivers of protistan grazing pressure: seasonal signals of plankton community composition and environmental conditions. *Marine Ecology Progress Series*, 459:39–52, July 2012. doi: 10.3354/meps09771.

Lee, F. W.-F., Morse, D., and Lo, S. C.-L. Identification of two plastid proteins in the dinoflagellate *Alexandrium affine* that are substantially down-regulated by nitrogen-depletion. *Journal of Proteom Research*, 8(11):5080–5092, 2009. doi: 10.1021/pr900475f.

Leggat, W., Yellowlees, D., and Medina, M. Recent progress in Symbiodinium transcriptomics. *Journal of Experimental Marine Biology and Ecology*, 408(1-2):120–125, November 2011. doi: 10.1016/j.jembe.2011.07.032.

Leonardos, N. and Geider, R. J. Elevated atmospheric carbon dioxide increases organic carbon fixation by *Emiliania huxleyi* (Haptophyta), under nutrient-limited high-light conditions. *Journal of Phycology*, 41(6):1196–1203, December 2005. doi: 10.1111/j.1529-8817.2005.00152.x.

Lessard, E. J., Merico, A., and Tyrrell, T. Nitrate : phosphate ratios and *Emiliania huxleyi* blooms. *Limnology and Oceanography*, 50(3):1020–1024, 2005. doi: 10.4319/lo.2005.50.3.1020.

Letelier, R. M., Bidigare, R. R., Hebel, D. V., Ondrusek, M., Winn, C. D., and Karl, D. M. Temporal variability of phytoplankton community structure based on pigment analysis. *Limnology and Oceanography*, 38(7):1420–1437, 1993. doi: 10.4319/lo.1993.38.7.1420.

Li, B. and Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, January 2011. doi: 10.1186/1471-2105-12-323.

Li, H. and Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5):589–95, March 2010. doi: 10.1093/bioinformatics/btp698.

Li, L., Stoeckert, C. J., and Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–89, September 2003. doi: 10.1101/gr.1224503.

Li, W. and Godzik, A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–9, July 2006. doi: 10.1093/bioinformatics/btl158.

Li, Y. and Smayda, T. J. Temporal variability of chlorophyll in Narragansett Bay, 1973–1990. *ICES Journal of Marine Science: Journal du Conseil*, 55(4):661–667, August 1998. doi: 10.1006/jmsc.1998.0383.

Lin, S., Sandh, G., Zhang, H., Cheng, J., Perkins, K., Carpenter, E. J., and Bergman, B. Two flavodoxin genes in *Trichodesmium* (Oscillatoriales, Cyanophyceae): Remarkable sequence divergence and possible functional diversification. *Journal of Experimental Marine Biology and Ecology*, 371(1):93–101, March 2009. doi: 10.1016/j.jembe.2009.01.010.

Litchman, E., Klausmeier, C. A., Miller, J. R., Schofield, O. M., and Falkowski, P. G. Multi-nutrient, multi-group model of present and future oceanic phytoplankton communities. *Biogeosciences*, 3(4):585–606, November 2006. doi: 10.5194/bg-3-585-2006.

Litchman, E. and Klausmeier, C. A. Trait-based community ecology of phytoplankton. *Annual Review of Ecology, Evolution, and Systematics*, 39(1):615–639, 2008. doi: 10.1146/annurev.ecolsys.39.110707.173549.

Livak, K. J. and Schmittgen, T. D. Analysis of relative gene expression using real-time quantitative PCR and the 2–$\Delta\Delta$CT method. *Methods*, 25(4):402–408, December 2001. doi: 10.1006/meth.2001.1262.

Lohse, M., Bolger, A. M., Nagel, A., Fernie, A. R., Lunn, J. E., Stitt, M., and Usadel, B. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40(W1):W622–W627, June 2012. doi: 10.1093/nar/gks540.

López-Urrutia, A., San Martin, E., Harris, R. P., and Irigoien, X. Scaling the metabolic balance of the oceans. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8739–44, June 2006. doi: 10.1073/pnas.0601137103.

Mackinder, L., Wheeler, G., Schroeder, D., Riebesell, U., and Brownlee, C. Molecular Mechanisms Underlying Calcification in Coccolithophores. *Geomicrobiology Journal*, September 2010.

Mackinder, L., Wheeler, G., Schroeder, D., von Dassow, P., Riebesell, U., and Brownlee, C. Expression of biomineralization-related ion transport genes in *Emiliania huxleyi*. *Environmental microbiology*, 13(12):3250–65, December 2011. doi: 10.1111/j.1462-2920.2011.02561.x.

Maldonado, M. T., Allen, A. E., Chong, J. S., Lin, K., Leus, D., Karpenko, N., and Harris, S. L. Copper-dependent iron transport in coastal and oceanic diatoms. *Limnology and Oceanography*, 51(4):1729–1743, 2006. doi: 10.4319/lo.2006.51.4.1729.

Mann, D. G. and Droop, S. J. M. 3. Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia*, 336(1-3):19–32, October 1996. doi: 10.1007/BF00010816.

Marchetti, A., Varela, D. E., Lance, V. P., Johnson, Z., Palmucci, M., Giordan, M., and Armbrust, E. V. Iron and silicic acid effects on phytoplankton productivity, diversity, and chemical composition in the central equatorial Pacific Ocean. *Limnology and Oceanography*, 55(1):11–29, 2005.

Marchetti, A., Schruth, D. M., Durkin, C. a., Parker, M. S., Kodner, R. B., Berthiaume, C. T., Morales, R., Allen, A. E., and Armbrust, E. V. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6):E317–25, February 2012. doi: 10.1073/pnas.1118408109.

Margalef, R. On certain unifying principles in ecology. *American Naturalist*, 1963.

Margalef, R. Life-forms of phytoplankton as survival alternatives in an unstable environment. *Oceanologica Acta*, 1(4):493–509, 1978.

Marionneau, C., Couette, B., Liu, J., Li, H., Mangoni, M. E., Nargeot, J., Lei, M., Escande, D., and Demolombe, S. Specific pattern of ionic channel gene expression associated with pacemaker activity in the mouse heart. *Journal of Physiology*, 562(1):223–234, January 2005. doi: 10.1113/jphysiol.2004.074047.

Martin, J. H. Phytoplankton-zooplankton relationships in NarragansettBay. IV. The seasonal importance of grazing. *Limnology and Oceanography*, 15(3):413–418, 1970. doi: 10.4319/lo.1970.15.3.0413.

Martiny, A. C., Treseder, K., and Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *The ISME Journal*, 7(4):830–838, 2013. doi: 10.1038/ismej.2012.160.

McAndrew, P., Björkman, K., Church, M., Morris, P., Jachowski, N., leB Williams PJ, and Karl, D. Metabolic response of oligotrophic plankton communities to deep water nutrient enrichment. *Marine Ecology Progress Series*, 332:63–75, March 2007. doi: 10.3354/meps332063.

McCarren, J., Becker, J. W., Repeta, D. J., Shi, Y., Young, C. R., Malmstrom, R. R., Chisholm, S. W., and DeLong, E. F. Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proceedings of the National Academy of Sciences of the United States of America*, 107(38):16420–7, September 2010. doi: 10.1073/pnas.1010732107.

McDonald, S. M., Plant, J. N., and Worden, A. Z. The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: a case study of *Micromonas*. *Molecular Biology and Evolution*, 27(10):2268 –2283, October 2010. doi: 10.1093/molbev/msq113.

McGinn, P. J. and Morel, F. M. M. Expression and regulation of carbonic anhydrases in the marine diatom *Thalassiosira pseudonana* and in natural phytoplankton assemblages from Great Bay, New Jersey. *Physiologia Plantarum*, 133(1):78–91, May 2008a. doi: 10.1111/j.1399-3054.2007.01039.x.

McGinn, P. J. and Morel, F. M. Expression and inhibition of the carboxylating and decarboxylating enzymes in photosynthetic C4 pathway of marine diatoms. *Plant Physiology*, 146(1):300 –309, January 2008b. doi: 10.1104/pp.107.110569.

McKew, B. A., Metodieva, G., Raines, C. A., Metodiev, M. V., and Geider, R. J. Acclimation of *Emiliania huxleyi* (1516) to nutrient limitation involves precise modification of the proteome to scavenge alternative sources of N and P. *Environmental microbiology*, 17(10): 4050–62, October 2015. doi: 10.1111/1462-2920.12957.

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., Matsumoto, K., Montzka, S. A., Raper, S. C. B., Riahi, K., Thomson, A., Velders, G. J. M., and Vuuren, D. P. The RCP greenhouse gas concentrations and their extensions from 1765 to 2300. *Climatic Change*, 109(1-2):213–241, August 2011. doi: 10.1007/s10584-011-0156-z.

Menden-Deuer, S. and Rowlett, J. Many ways to stay in the game: individual variability maintains high biodiversity in planktonic microorganisms. *Journal of the Royal Society, Interface / the Royal Society*, 11(95):20140031, June 2014. doi: 10.1098/rsif.2014.0031.

Min, X. J., Butler, G., Storms, R., and Tsang, A. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic acids research*, 33(Web Server issue):W677–80, July 2005. doi: 10.1093/nar/gki394.

Mock, T., Samanta, M. P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., Bondurant, S. S., Richmond, K., Rodesch, M., Kallas, T., Huttlin, E. L., Cerrina, F., Sussman, M. R., and Armbrust, E. V. Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences of the United States of America*, 105(5): 1579–84, February 2008. doi: 10.1073/pnas.0707946105.

Moore, J. K. Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model. *Global Biogeochemical Cycles*, 18(4):1–21, 2004. doi: 10.1029/2004GB002220.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35(Web Server):W182–W185, May 2007. doi: 10.1093/nar/gkm321.

Moseley, J. L., Chang, C.-W., and Grossman, A. R. Genome-based approaches to understanding phosphorus deprivation responses and PSR1 control in *Chlamydomonas reinhardtii*. *Eukaryotic Cell*, 5(1):26 –44, January 2006. doi: 10.1128/EC.5.1.26-44.2006.

Moustafa, A., Evans, A. N., Kulis, D. M., Hackett, J. D., Erdner, D. L., Anderson, D. M., and Bhattacharya, D. Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PloS one*, 5 (3):e9688, January 2010. doi: 10.1371/journal.pone.0009688.

Naeem, S., Thompson, L. J., Lawler, S. P., Lawton, J. H., and Woodfin, R. M. Declining biodiversity can alter the performance of ecosystems. *Nature*, 368(6473):734–737, April 1994. doi: 10.1038/368734a0.

Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A., and Quéguiner, B. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles*, 9(3):359–372, September 1995. doi: 10.1029/95GB01070.

Nicot, N., Hausman, J.-F., Hoffmann, L., and Evers, D. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. *Journal of Experimental Botany*, 56(421):2907 –2914, November 2005. doi: 10.1093/jxb/eri285.

Nielsen, E. S. Productivity of the Oceans. *Annual Review of Plant Physiology*, 11(1):341–362, June 1960. doi: 10.1146/annurev.pp.11.060160.002013.

Nixon, S., Granger, S., and Nowicki, B. An assessment of the annual mass balance of carbon, nitrogen, and phosphorus in Narragansett Bay. *Biogeochemistry*, 31(1):15–61, October 1995. doi: 10.1007/BF00000805.

North, B. B. and Stephens, G. C. Amino acid transport in *Nitzschia ovalis arnott*. *Journal of Phycology*, 8(1):64–68, March 1972. doi: 10.1111/j.1529-8817.1972.tb04003.x.

Ogawa, N., DeRisi, J., and Brown, P. O. New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Molecular Biology of the Cell*, 11(12):4309–4321, December 2000. doi: 10.1091/mbc.11.12.4309.

Ottesen, E. A., Marin, R., Preston, C. M., Young, C. R., Ryan, J. P., Scholin, C. A., and DeLong, E. F. Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *The ISME journal*, 5(12):1881–95, December 2011. doi: 10.1038/ismej.2011.70.

Ottesen, E. A., Young, C. R., Gifford, S. M., Eppley, J. M., Marin, R., Schuster, S. C., Scholin, C. A., and DeLong, E. F. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science (New York, N.Y.)*, 345(6193):207–12, July 2014. doi: 10.1126/science.1252476.

Oviatt, C., Buckley, B., and Nixon, S. Annual phytoplankton metabolism in Narragansett Bay calculated from survey field measurements and microcosm observations. *Estuaries*, 4 (3):167, September 1981. doi: 10.2307/1351472.

Paasche, E. Roles of nitrogen and phosphorus in coccolith formation in *Emiliania huxleyi* (Prymnesiophyceae). *European Journal of Phycology*, 33(1):33–42, February 1998. doi: 10.1080/09670269810001736513.

Paasche, E. A review of the coccolithophorid Emiliania huxleyi (Prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions. *Phycologia*, 40(6):503–529, November 2001. doi: 10.2216/i0031-8884-40-6-503.1.

Paasche, E. and Brubak, S. Enhanced calcification in the coccolithophorid *Emiliania huxleyi* (Haptophyceae) under phosphorus limitation. *Phycologia*, 33(5):324–330, September 1994. doi: 10.2216/i0031-8884-33-5-324.1.

Palenik, B. P. and Henson, S. E. The use of amides and other organic nitrogen sources by the phytoplankton *Emiliania huxleyi*. *Limnology and Oceanography*, 42(7):1544–1551, 1997. doi: 10.4319/lo.1997.42.7.1544.

Park, H., McGinn, P. J., and Morel, F. M. M. Expression of cadmium carbonic anhydrase of diatoms in seawater. *Aquatic Microbial Ecology*, 51(2):183–193, May 2008. doi: 10.3354/ame01192.

Parsons, T. R., Yoshiaki, M., and Lalli, C. M. *A Manual of Chemical & Biological Methods for Seawater Analysis*. Pergamon Press, Oxford, 1984. ISBN 0-08-030288-2.

Pearson, G. A., Lago-Leston, A., Cánovas, F., Cox, C. J., Verret, F., Lasternas, S., Duarte, C. M., Agusti, S., and Serrão, E. A. Metatranscriptomes reveal functional variation in diatom communities from the Antarctic Peninsula. *The ISME Journal*, 9(10):2275–2289, October 2015. doi: 10.1038/ismej.2015.40.

Perry, A. M. J., Limnology, S., Jan, N., and Perry, M. J. Phosphate utilization by an oceanic diatom in phosphorus-limited chemostat culture and in the oligotrophic waters of the Central North Pacific. 21(1):88–107, 1976.

Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*, 29(9):e45, May 2001.

Pfaffl, M. W., Tichopad, A., Prgomet, C., and Neuvians, T. P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – Excel-based tool using pair-wise correlations, March 2004.

Poretsky, R. S., Sun, S., Mou, X., and Moran, M. A. Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environmental Microbiology*, 12(3):616–627, March 2010. doi: 10.1111/j.1462-2920.2009.02102.x.

Poulton, A. J., Adey, T. R., Balch, W. M., and Holligan, P. M. Relating coccolithophore calcification rates to phytoplankton community dynamics: Regional differences and implications for carbon export. *Deep Sea Research Part II: Topical Studies in Oceanography*, 54(5-7):538–557, March 2007. doi: 10.1016/j.dsr2.2006.12.003.

Ptacnik, R., Solimini, A. G., Andersen, T., Tamminen, T., Brettum, P. l., Lepistö, L., Willén, E., and Rekolainen, S. Diversity predicts stability and resource use efficiency in natural phytoplankton communities. *Proceedings of the National Academy of Sciences of the United States of America*, 105(13):5134–8, April 2008. doi: 10.1073/pnas.0708328105.

Quackenbush, J. Computational analysis of microarray data. *Nature Reviews Genetics*, 2 (6):418–427, June 2001. doi: 10.1038/35076576.

Radonic, A., Thulke, S., Mackay, I. M., Landt, O., Siegert, W., and Nitsche, A. Guideline to reference gene selection for quantitative real-time PCR. *Biochemical and Biophysical Research Communications*, 313(4):856–862, January 2004. doi: 10.1016/j.bbrc.2003.11.177.

Raven, J., Caldeira, K., and Elderfield, H. *Ocean acidification due to increasing atmospheric carbon dioxide*. 2005.

Read, B. A., Kegel, J., Klute, M. J., Kuo, A., Lefebvre, S. C., Maumus, F., Mayer, C., Miller, J., Monier, A., Salamov, A., Young, J., Aguilar, M., Claverie, J.-M., Frickenhaus, S., Gonzalez, K., Herman, E. K., Lin, Y.-C., Napier, J., Ogata, H., Sarno, A. F., Shmutz, J., Schroeder, D., de Vargas, C., Verret, F., von Dassow, P., Valentin, K., Van de Peer, Y., Wheeler, G., Dacks, J. B., Delwiche, C. F., Dyhrman, S. T., Glöckner, G., John, U., Richards, T., Worden, A. Z., Zhang, X., and Grigoriev, I. V. Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature*, 499(7457):209–13, July 2013. doi: 10.1038/nature12221.

Redfield, A. C. The biological control of chemical factors in the environment. *American Scientist*, 46:205–221, 1958.

Richerson, P., Armstrong, R., and Goldman, C. R. Contemporaneous Disequilibrium, a New Hypothesis to Explain the "Paradox of the Plankton". *Proceedings of the National Academy of Sciences of the United States of America*, 67(4):1710–1714, December 1970. doi: 10.1073/pnas.67.4.1710.

Richier, S., Fiorini, S., Kerros, M.-E., von Dassow, P., and Gattuso, J.-P. Response of the calcifying coccolithophore Emiliania huxleyi to low pH/high pCO_2: from physiology to molecular level. *Marine biology*, 158(3):551–560, January 2011. doi: 10.1007/s00227-010-1580-8.

Riebesell, U., Zondervan, I., Rost, B., Tortell, P. D., Zeebe, R. E., and Morel, F. M. Reduced calcification of marine plankton in response to increased atmospheric CO_2. *Nature*, 407 (6802):364–7, September 2000. doi: 10.1038/35030078.

Riegman, R. and Stolte, W. Nutrient uptake and alkaline phosphatase (EC 3: 1: 3: 1) activity of *Emiliania huxleyi* (Prymnesiophyceae) during growth under N and P limitation in continuous. *Journal of Phycology*, 36:87–96, 2000.

Rinta-Kanto, J. M., Sun, S., Sharma, S., Kiene, R. P., Moran, M. A., and RintaâĂŘKanto, J. M. Bacterial community transcription patterns during a marine phytoplankton bloom. *Environmental microbiology*, 14(1):228–39, January 2012. doi: 10.1111/j.1462-2920.2011.02602.x.

Robinson, M. and Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):1–9, March 2010. doi: 10.1186/gb-2010-11-3-r25.

Rokitta, S. D., Von Dassow, P., Rost, B., and John, U. *Emiliania huxleyi* endures N-limitation with an efficient metabolic budgeting and effective ATP synthesis. *BMC genomics*, 15(1):1051, January 2014. doi: 10.1186/1471-2164-15-1051.

Rosic, N. N., Pernice, M., Dove, S., Dunn, S., and Hoegh-Guldberg, O. Gene expression profiles of cytosolic heat shock proteins Hsp70 and Hsp90 from symbiotic dinoflagellates

in response to thermal stress: possible implications for coral bleaching. *Cell Stress and Chaperones*, 16(1):69–80, September 2010a. doi: 10.1007/s12192-010-0222-x.

Rosic, N. N., Pernice, M., Rodriguez-Lanetty, M., and Hoegh-Guldberg, O. Validation of housekeeping genes for gene expression studies in *Symbiodinium* exposed to thermal and light stress. *Marine Biotechnology*, 13(3):355–365, July 2010b. doi: 10.1007/s10126-010-9308-9.

Rouco, M., Branson, O., Lebrato, M., and Iglesias-Rodríguez, M. D. The effect of nitrate and phosphate availability on Emiliania huxleyi (NZEH) physiology under different CO_2 scenarios. *Frontiers in microbiology*, 4:155, January 2013. doi: 10.3389/fmicb.2013.00155.

Roy, S. and Chattopadhyay, J. Towards a resolution of âĂŸthe paradox of the planktonâĂŹ: A brief overview of the proposed mechanisms. *Ecological Complexity*, 4(1-2):26–33, March 2007. doi: 10.1016/j.ecocom.2007.02.016.

Saeed, A. I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. TM4: a free, open-source system for microarray data management and analysis. *BioTechniques*, 34(2):374–378, February 2003.

Saeed, A. I., Bhagabati, N. K., Braisted, J. C., Liang, W., Sharov, V., Howe, E. A., Li, J., Thiagarajan, M., White, J. A., and Quackenbush, J. TM4 microarray software suite. *Methods in Enzymology*, 411:134–193, 2006. doi: 10.1016/S0076-6879(06)11009-5.

Saito, M. A., Goepfert, T. J., and Ritt, J. T. Some thoughts on the concept of colimitation: Three definitions and the importance of bioavailability. *Limnology and Oceanography*, 53 (1):276–290, 2008. doi: 10.4319/lo.2008.53.1.0276.

Saito, M. A., McIlvin, M. R., Moran, D. M., Goepfert, T. J., DiTullio, G. R., Post, A. F., and Lamborg, C. H. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science (New York, N.Y.)*, 345(6201):1173–7, September 2014. doi: 10.1126/science.1256450.

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., Pico, A. R., Bader, G. D., and Ideker, T. A travel guide to Cytoscape plugins. *Nature methods*, 9(11): 1069–76, November 2012. doi: 10.1038/nmeth.2212.

Sakshaug, E. Limiting nutrients and maximum growth rates for diatoms in Narragansett Bay. *Journal of Experimental Marine Biology and Ecology*, 28(2):109–123, July 1977. doi: 10.1016/0022-0981(77)90110-1.

Sakshaug, E., Granéli, E., Elbrächter, M., and Kayser, H. Chemical composition and alkaline phosphatase activity of nutrient-saturated and P-deficient cells of four marine dinoflagellates. *Journal of Experimental Marine Biology and Ecology*, 77(3):241–254, May 1984. doi: 10.1016/0022-0981(84)90122-9.

Sarno, D., Kooistra, W. H. C. F., Medlin, L. K., Percopo, I., and 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. *Journal of Phycology*, 41 (1):151–176, February 2005. doi: 10.1111/j.1529-8817.2005.04067.x.

Schmittgen, T. D. and Livak, K. J. Analyzing real-time PCR data by the comparative CT method. *Nature Protocols*, 3(6):1101–1108, June 2008. doi: 10.1038/nprot.2008.73.

Sciandra, A., Harlay, J., Lefèvre, D., Lemée, R., Rimmelin, P., Denis, M., and Gattuso, J. Response of coccolithophorid Emiliania huxleyi to elevated partial pressure of CO_2 under nitrogen limitation. *Marine Ecology Progress Series*, 261:111–122, 2003. doi: 10.3354/meps261111.

Serra, J. L., Llama, M. J., and Cadenas, E. Nitrate utilization by the diatom *Skeletonema costatum*: II. Regulation of nitrate uptake. *Plant physiology*, 62(6):991–994, December 1978. doi: 10.1104/pp.62.6.991.

Shi, Y., Tyson, G. W., Eppley, J. M., and DeLong, E. F. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *The ISME journal*, 5(6):999–1013, July 2011. doi: 10.1038/ismej.2010.189.

Shi, Y., McCarren, J., and DeLong, E. F. Transcriptional responses of surface water marine microbial assemblages to deep-sea water amendment. *Environmental microbiology*, 14(1):191–206, January 2012. doi: 10.1111/j.1462-2920.2011.02598.x.

Shrestha, R. P., Tesson, B., Norden-Krichmar, T., Federowicz, S., Hildebrand, M., and Allen, A. E. Whole transcriptome analysis of the silicon response of the diatom *Thalassiosira pseudonana*. *BMC genomics*, 13(1):499, January 2012. doi: 10.1186/1471-2164-13-499.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A., and Bowler, C. Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. *Gene*, 406(1-2):23–35, December 2007. doi: 10.1016/j.gene.2007.05.022.

Silver, M. W., Bargu, S., Coale, S. L., Benitez-Nelson, C. R., Garcia, A. C., Roberts, K. J., Sekula-Wood, E., Bruland, K. W., and Coale, K. H. Toxic diatoms and domoic acid in natural and iron enriched waters of the oceanic Pacific. *Proceedings of the National Academy of Sciences of the United States of America*, 107(48):20762–7, November 2010. doi: 10.1073/pnas.1006968107.

Smayda, T. J. Bioassay of the growth potential of the surface water of lower Narragansett Bay over an annual cycle using the diatom *Thalassiosira pseudonana*. *Limnology and Oceanography*, 19(6):889–901, November 1974.

Smayda, T. J. Cryptic planktonic diatom challenges phytoplankton ecologists. *Proceedings of the National Academy of Sciences of the United States of America*, 108(11):4269–70, March 2011. doi: 10.1073/pnas.1100997108.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–2, February 2011. doi: 10.1093/bioinformatics/btq675.

Sommer, U. Nutrient competition between phytoplankton species in multispecies chemostat experiments. *Archiv für hydrobiologie*, 96(4):399–416, 1983.

Sommer, U. Comparison between steady state and non-steady state competition: Experiments with natural phytoplankton. *Limnology and Oceanography*, 30(2):335–346, March 1985. doi: 10.4319/lo.1985.30.2.0335.

Sournia, A., Chrdtiennot-Dinet, M.-J., and Ricard, M. Marine phytoplankton: how many species in the world ocean? *Journal of Plankton Research*, 13(5):1093–1099, January 1991. doi: 10.1093/plankt/13.5.1093.

Steiner, C. F., Long, Z. T., Krumins, J. A., and Morin, P. J. Temporal stability of aquatic food webs: partitioning the effects of species diversity, species composition and enrichment. *Ecology Letters*, 8(8):819–828, August 2005. doi: 10.1111/j.1461-0248.2005.00785.x.

Steinke, M., Wolfe, G. V., and Kirst, G. O. Partial characterisation of dimethylsulfoniopropionate (DMSP) lyase isozymes in 6 strains of *Emiliania huxleyi*, March 1998.

Strickland, J. D. Phytoplankton and marine primary production. *Annual review of microbiology*, 19:127–62, January 1965. doi: 10.1146/annurev.mi.19.100165.001015.

Striebel, M., Behl, S., and Stibor, H. The coupling of biodiversity and productivity in phytoplankton communities: consequences for biomass stoichiometry. *Ecology*, 90(8): 2025–2031, August 2009. doi: 10.1890/08-1409.1.

Strom, S. and Bright, K. Inter-strain differences in nitrogen use by the coccolithophore *Emiliania huxleyi*, and consequences for predation by a planktonic ciliate. *Harmful Algae*, 8:811–816, 2009.

Stuart, R. K., Dupont, C. L., Johnson, D. A., Paulsen, I. T., and Palenik, B. Coastal strains of marine *Synechococcus* species exhibit increased tolerance to copper shock and a distinctive transcriptional response relative to those of open-ocean strains. *Applied and Environmental Microbiology*, 75(15):5047 –5057, August 2009. doi: 10.1128/AEM.00271-09.

Sverdrup, H. On conditions for the vernal blooming of phytoplankton. *Journal du Conseil*, 18(3):287–295, 1953.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–285, July 1999. doi: 10.1038/10343.

Tett, P. and Barton, E. Why are there about 5000 species of phytoplankton in the sea? *Journal of Plankton Research*, 17(8):1693–1704, 1995. doi: 10.1093/plankt/17.8.1693.

Tilman, D., Reich, P. B., Knops, J., Wedin, D., Mielke, T., and Lehman, C. Diversity and productivity in a long-term grassland experiment. *Science (New York, N.Y.)*, 294(5543): 843–5, October 2001. doi: 10.1126/science.1060391.

Tilman, D. Resource competition between plankton algae: an experimental and theoretical approach. *Ecology*, 58(2):338–348, March 1977. doi: 10.2307/1935608.

Tilman, D., Wedin, D., and Knops, J. Productivity and sustainability influenced by biodiversity in grassland ecosystems. *Nature*, 379(6567):718–720, February 1996. doi: 10.1038/379718a0.

Van Mooy, B. A. S., Fredricks, H. F., Pedler, B. E., Dyhrman, S. T., Karl, D. M., Koblízek, M., Lomas, M. W., Mincer, T. J., Moore, L. R., Moutin, T., Rappé, M. S., and Webb, E. a. Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. *Nature*, 458(7234):69–72, March 2009. doi: 10.1038/nature07659.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., and Speleman, F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology*, 3(7):research0034, June 2002. doi: 10.1186/gb-2002-3-7-research0034.

Villar, E., Farrant, G. K., Follows, M., Garczarek, L., Speich, S., Audic, S., Bittner, L., Blanke, B., Brum, J. R., Brunet, C., Casotti, R., Chase, A., Dolan, J. R., D'Ortenzio, F., Gattuso, J.-P., Grima, N., Guidi, L., Hill, C. N., Jahn, O., Jamet, J.-L., Le Goff, H., Lepoivre, C., Malviya, S., Pelletier, E., Romagnan, J.-b., Roux, S., Santini, S., Scalco, E., Schwenck, S. M., Tanaka, A., Testor, P., Vannier, T., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Boss, E., de Vargas, C., Gorsky, G., Ogata, H., Pesant, S., Sullivan, M. B., Sunagawa, S., Wincker, P., Karsenti, E., Bowler, C., Not, F., Hingamp, P., and Iudicone, D. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science (New York, N.Y.)*, 348(6237):1261447, May 2015. doi: 10.1126/science.1261447.

Villareal, T. A., Adornato, L., Wilson, C., and Schoenbaechler, C. A. Summer blooms of diatom-diazotroph assemblages and surface chlorophyll in the North Pacific Gyre: a disconnect. *Journal of Geophysical Research*, 116(C3):C03001, March 2011. doi: 10.1029/2010JC006268.

Villareal, T. A., Brown, C. G., Brzezinski, M. A., Krause, J. W., and Wilson, C. Summer diatom blooms in the North Pacific Subtropical Gyre: 2008-2009. *PloS one*, 7(4):e33109, January 2012. doi: 10.1371/journal.pone.0033109.

von Dassow, P., Ogata, H., Probert, I., Wincker, P., Da Silva, C., Audic, S., Claverie, J.-M., and de Vargas, C. Transcriptome analysis of functional differentiation between haploid and diploid cells of *Emiliania huxleyi*, a globally significant photosynthetic calcifying cell. *Genome biology*, 10(10):R114, January 2009. doi: 10.1186/gb-2009-10-10-r114.

Whitney, L. P., Lins, J. J., Hughes, M. P., Wells, M. L., Chappell, P. D., and Jenkins, B. D. Characterization of putative iron responsive genes as species-specific indicators of iron stress in thalassiosiroid diatoms. *Frontiers in microbiology*, 2(November):234, January 2011. doi: 10.3389/fmicb.2011.00234.

Winn, C. D. and Karl, D. M. Laboratory calibrations of the [3H]adenine technique for measuring rates of RNA and DNA synthesis in marine microorganisms. *Applied and Environmental Microbiology*, 47(4):835–842, 1984.

Wu, Z., Jenkins, B. D., Rynearson, T. A., Dyhrman, S. T., Saito, M. A., Mercier, M., and Whitney, L. P. Empirical bayes analysis of sequencing-based transcriptional profiling without replicates. *BMC bioinformatics*, 11(1):564, January 2010. doi: 10.1186/1471-2105-11-564.

Wurch, L. L., Bertrand, E. M., Saito, M. A., Van Mooy, B. A. S., and Dyhrman, S. T. Proteome changes driven by phosphorus deficiency and recovery in the brown tide-forming alga *Aureococcus anophagefferens*. *PloS one*, 6(12):e28949, January 2011a. doi: 10.1371/journal.pone.0028949.

Wurch, L. L., Haley, S. T., Orchard, E. D., Gobler, C. J., and Dyhrman, S. T. Nutrient-regulated transcriptional responses in the brown tide-forming alga *Aureococcus anophag-*

efferens. *Environmental microbiology*, 13(2):468–81, February 2011b. doi: 10.1111/j.1462-2920.2010.02351.x.

Xu, Y., Wahlund, T. M., Feng, L., Shaked, Y., and Morel, F. M. M. A novel alkaline phosphatase in the cocolithophore *Emiliania huxleyi* (Prymnesiophyceae) and its regulation by phosphorus. *Journal of Phycology*, 42(4):835–844, August 2006. doi: 10.1111/j.1529-8817.2006.00243.x.

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309 –318, April 2001. doi: 10.1093/bioinformatics/17.4.309.

Zehr, J. and Turner, P. Nitrogen fixation: Nitrogenase genes and gene expression. volume 30, pages 271–286. Elsevier, 2001. ISBN 978-0-12-521530-5.

Zhao, T., Wang, W., Bai, X., and Qi, Y. Gene silencing by artificial microRNAs in *Chlamydomonas*. *Plant Journal*, 58(1):157–164, April 2009. doi: 10.1111/j.1365-313X.2008.03758.x.

Zingone, A., Percopo, I., Sims, P. A., and Sarno, D. Diversity in the genus *Skeletonema* (Bacillariophyceae). I. A reexamination of the type material of *S. costatum* with the description of *S. grevillei* sp. nov. *Journal of Phycology*, 41(1):140–150, February 2005. doi: 10.1111/j.1529-8817.2005.04066.x.

Zondervan, I., Rost, B., and Riebesell, U. Effect of CO_2 concentration on the PIC/POC ratio in the coccolithophore Emiliania huxleyi grown under light-limiting conditions and different daylengths. *Journal of Experimental Marine Biology ...*, 2002.

Zondervan, I. The effects of light, macronutrients, trace metals and CO_2 on the production of calcium carbonate and organic carbon in coccolithophoresâĂŤA review. *Deep Sea Research Part II: Topical Studies in Oceanography*, 54(5-7):521–537, March 2007. doi: 10.1016/j.dsr2.2006.12.004.

Zubkov, M. V., Fuchs, B. M., Tarran, G. A., Burkill, P. H., and Amann, R. High rate of uptake of organic nitrogen compounds by *Prochlorococcus* cyanobacteria as a key to their dominance in oligotrophic oceanic waters. *Applied and Environmental Microbiology*, 69 (2):1299–1304, February 2003. doi: 10.1128/AEM.69.2.1299-1304.2003.