Phytoplankton Physiology Under Changing Nitrogen and Phosphorus Environments

by

Harriet Alexander

B.S., Wellesley College (2010)

Submitted to the MIT-WHOI Joint Program in Oceanography and Applied Ocean Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Author . . . .
MIT-WHOI Joint Program in Oceanography and Applied Ocean Engineering
December 18, 2015

Certified by . . . .
Sonya T. Dyhrman
Associate Professor, Columbia University
Thesis Supervisor

Accepted by . . . .
Samuel Laney
Chairman, Department Committee on Graduate Theses

Phytoplankton Physiology Under Changing Nitrogen and Phosphorus Environments

by

Harriet Alexander

Submitted to the MIT-WHOI Joint Program in Oceanography and Applied Ocean Engineering
on December 18, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, I designed and implemented a compiler which performs optimizations that reduce the number of low-level floating point operations necessary for a specific task; this involves the optimization of chains of floating point operations as well as the implementation of a "fixed" point data type that allows some floating point operations to simulated with integer arithmetic. The source language of the compiler is a subset of C, and the destination language is assembly language for a micro-floating point CPU. An instruction-level simulator of the CPU was written to allow testing of the code. A series of test pieces of codes was compiled, both with and without optimization, to determine how effective these optimizations were.

Thesis Supervisor: Sonya T. Dyhrman
Title: Associate Professor, Columbia University

Acknowledgments

Throughout graduate school I have been supported by the MIT Presdidential Fellowship, the National Defense Science and Engineering Graduate (NDSEG) Fellowship, the Ocean Life Institute Fellowship, and the WHOI Academic Progams Office. This thesis, like so many of my achievements and 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 want thank my advisor, Sonya Dyhrman. Over the last five years, Sonya has nurtured my scientific curiosity, while challenging me to delve deeper and work harder than I knew I could. [Something about how we travel and have fun?] It has been a great privalage to work with such a brilliant scientist whose integrity and devotion to science and her colleuges 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 all 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: Tatiana Rynearson, Mak Saito, Dave Karl, Sam Wilson. I am so happy to have intersected with such wonderful people during my time in the Dyhrman lab group. Sheean Haley, it goes without saying that you are the glue that holds us all together and the XX that keeps the lab running seamlessly. Your leavity and indefatigability both on land and at sea have kept me grounded and made all my science possible. To my fellow graduate students and postdocs, Louie Wurch, Abby Heithoff, Monica Rouco, Solange Duhmel, and Kyle Frishkorn, thank you for showing me the way and being awesome. 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. In particular, I wish to dedicate this thesis to the memory of my grandparents Betty and Sherwood Finley. I owe my work ethic, love of nature, and XX to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Photosynthetic microbes dominate the primary production of the ocean and account for roughly half of all primary production on earth [Field]; The diversity Chapter two describes the architecture of the μFPU unit, and the motivations for the design decisions made.

Chapter three describes the design of the compiler, as well as how the optimizations discussed in section 1.2 were implemented.

1.1

The idea of micro-optimization is motivated by the recent trends in computer architecture towards low-level parallelism and small, pipelineable instruction sets **??**. By getting rid of more complex instructions and concentrating on optimizing frequently used instructions, substantial increases in performance were realized.

Another important motivation was the trend towards placing more of the burden of performance on the compiler. Many of the new architectures depend on an intelligent, optimizing compiler in order to realize anywhere near their peak performance **???**. In these cases, the compiler not only is responsible for faithfully generating native code to match the source language, but also must be aware of instruction latencies, delayed branches, pipeline stages, and a multitude of other factors in order to generate fast code **?**.

Taking these ideas one step further, it seems that the floating point operations that are normally single, large instructions can be further broken down into smaller, simpler, faster instructions, with more control in the compiler and less in the hardware. This is the idea behind a micro-optimizing FPU; break the floating point instructions down into their basic components and use a small, fast implementation, with a large part of the burden of hardware allocation and optimization shifted towards compile-time.

Along with the hardware speedups possible by using a μFPU, there are also optimizations that the compiler can perform on the code that is generated. In a normal sequence of floating point operations, there are many hidden redundancies that can be eliminated by allowing the compiler to control the floating point operations down to their lowest level. These optimizations are described in detail in section 1.2.

1.2 Description of micro-optimization

In order to perform a sequence of floating point operations, a normal FPU performs many redundant internal shifts and normalizations in the process of performing a sequence of operations. However, if a compiler can decompose the floating point operations it needs down to the lowest level, it then can optimize away many of these redundant operations.

If there is some additional hardware support specifically for micro-optimization, there are additional optimizations that can be performed. This hardware support entails extra "guard bits" on the standard floating point formats, to allow several unnormalized operations to be performed in a row without the loss information[1]. A discussion of the mathematics behind unnormalized arithmetic is in appendix **??**.

The optimizations that the compiler can perform fall into several categories:

[1] A description of the floating point format used is shown in figures **??** and **??**.

1.2.1 Post Multiply Normalization

When more than two multiplications are performed in a row, the intermediate normalization of the results between multiplications can be eliminated. This is because with each multiplication, the mantissa can become denormalized by at most one bit. If there are guard bits on the mantissas to prevent bits from "falling off" the end during multiplications, the normalization can be postponed until after a sequence of several multiplies[2].

As you can see, the intermediate results can be multiplied together, with no need for intermediate normalizations due to the guard bit. It is only at the end of the operation that the normalization must be performed, in order to get it into a format suitable for storing in memory[3].

1.2.2 Block Exponent

In a unoptimized sequence of additions, the sequence of operations is as follows for each pair of numbers (m_1,e_1) and (m_2,e_2).

1. Compare e_1 and e_2.
2. Shift the mantissa associated with the smaller exponent $|e_1 - e_2|$ places to the right.
3. Add m_1 and m_2.
4. Find the first one in the resulting mantissa.
5. Shift the resulting mantissa so that normalized
6. Adjust the exponent accordingly.

[2] Using unnormalized numbers for math is not a new idea; a good example of it is the Control Data CDC 6600, designed by Seymour Cray. **?** The CDC 6600 had all of its instructions performing unnormalized arithmetic, with a separate `NORMALIZE` instruction.

[3] Note that for purposed of clarity, the pipeline delays were considered to be 0, and the branches were not delayed.

Out of 6 steps, only one is the actual addition, and the rest are involved in aligning the mantissas prior to the add, and then normalizing the result afterward. In the block exponent optimization, the largest mantissa is found to start with, and all the mantissa's shifted before any additions take place. Once the mantissas have been shifted, the additions can take place one after another[4]. An example of the Block Exponent optimization on the expression X = A + B + C is given in figure **??**.

1.3 Integer optimizations

As well as the floating point optimizations described above, there are also integer optimizations that can be used in the μFPU. In concert with the floating point optimizations, these can provide a significant speedup.

1.3.1 Conversion to fixed point

Integer operations are much faster than floating point operations; if it is possible to replace floating point operations with fixed point operations, this would provide a significant increase in speed.

This conversion can either take place automatically or or based on a specific request from the programmer. To do this automatically, the compiler must either be very smart, or play fast and loose with the accuracy and precision of the programmer's variables. To be "smart", the computer must track the ranges of all the floating point variables through the program, and then see if there are any potential candidates for conversion to floating point. This technique is discussed further in section **??**, where it was implemented.

The other way to do this is to rely on specific hints from the programmer that a certain value will only assume a specific range, and that only a specific precision is desired. This is somewhat more taxing on the programmer, in that he has to know the ranges that his values will take at declaration time (something normally abstracted

[4] This requires that for n consecutive additions, there are $\log_2 n$ high guard bits to prevent overflow. In the μFPU, there are 3 guard bits, making up to 8 consecutive additions possible.

away), but it does provide the opportunity for fine-tuning already working code.

Potential applications of this would be simulation programs, where the variable represents some physical quantity; the constraints of the physical system may provide bounds on the range the variable can take.

1.3.2 Small Constant Multiplications

One other class of optimizations that can be done is to replace multiplications by small integer constants into some combination of additions and shifts. Addition and shifting can be significantly faster than multiplication. This is done by using some combination of

$$\begin{aligned}
a_i &= a_j + a_k \\
a_i &= 2a_j + a_k \\
a_i &= 4a_j + a_k \\
a_i &= 8a_j + a_k \\
a_i &= a_j - a_k \\
a_i &= a_j \ll m\text{shift}
\end{aligned}$$

instead of the multiplication. For example, to multiply s by 10 and store the result in r, you could use:

$$\begin{aligned}
r &= 4s + s \\
r &= r + r
\end{aligned}$$

Or by 59:

$$\begin{aligned}
t &= 2s + s \\
r &= 2t + s \\
r &= 8r + t
\end{aligned}$$

Similar combinations can be found for almost all of the smaller integers[5]. **?**

1.4 Other optimizations

1.4.1 Low-level parallelism

The current trend is towards duplicating hardware at the lowest level to provide parallelism[6]

Conceptually, it is easy to take advantage to low-level parallelism in the instruction stream by simply adding more functional units to the μFPU, widening the instruction word to control them, and then scheduling as many operations to take place at one time as possible.

However, simply adding more functional units can only be done so many times; there is only a limited amount of parallelism directly available in the instruction stream, and without it, much of the extra resources will go to waste. One process used to make more instructions potentially schedulable at any given time is "trace scheduling". This technique originated in the Bulldog compiler for the original VLIW machine, the ELI-512. **??** In trace scheduling, code can be scheduled through many basic blocks at one time, following a single potential "trace" of program execution. In this way, instructions that *might* be executed depending on a conditional branch further down in the instruction stream are scheduled, allowing an increase in the potential parallelism. To account for the cases where the expected branch wasn't taken, correction code is inserted after the branches to undo the effects of any prematurely executed instructions.

[5] This optimization is only an "optimization", of course, when the amount of time spent on the shifts and adds is less than the time that would be spent doing the multiplication. Since the time costs of these operations are known to the compiler in order for it to do scheduling, it is easy for the compiler to determine when this optimization is worth using.

[6] This can been seen in the i860; floating point additions and multiplications can proceed at the same time, and the RISC core be moving data in and out of the floating point registers and providing flow control at the same time the floating point units are active. **?**

1.4.2 Pipeline optimizations

In addition to having operations going on in parallel across functional units, it is also typical to have several operations in various stages of completion in each unit. This pipelining allows the throughput of the functional units to be increased, with no increase in latency.

There are several ways pipelined operations can be optimized. On the hardware side, support can be added to allow data to be recirculated back into the beginning of the pipeline from the end, saving a trip through the registers. On the software side, the compiler can utilize several tricks to try to fill up as many of the pipeline delay slots as possible, as seendescribed by Gibbons. **?**

Chapter 2

Identifying reference genes with stable expression from high throughput sequence data

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., 2011; 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., 2011; 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* CCMP 1335 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 ($\sim$12 million), P-limited ($\sim$13 million), Fe-limited ($\sim$23 million), and co-limited ($\sim$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 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=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 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 (http://R-project.org).

2.4 Results

Transcript sequence data was generated from *T. pseudonana* CCMP 1335, 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., 2010), *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., 2011), *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 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 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 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 A-2). 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 A-2). 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), which stands in contrast to other clusters that had definite treatment driven expression patterns (Figure A-2). 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 /reffig:c2f2) 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).

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 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 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 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 of 14.18 TPM and 21.93 TPM, respectively.

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.

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; Data Sheet 3).

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 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 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 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., 2011; 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).

2.6 Figues

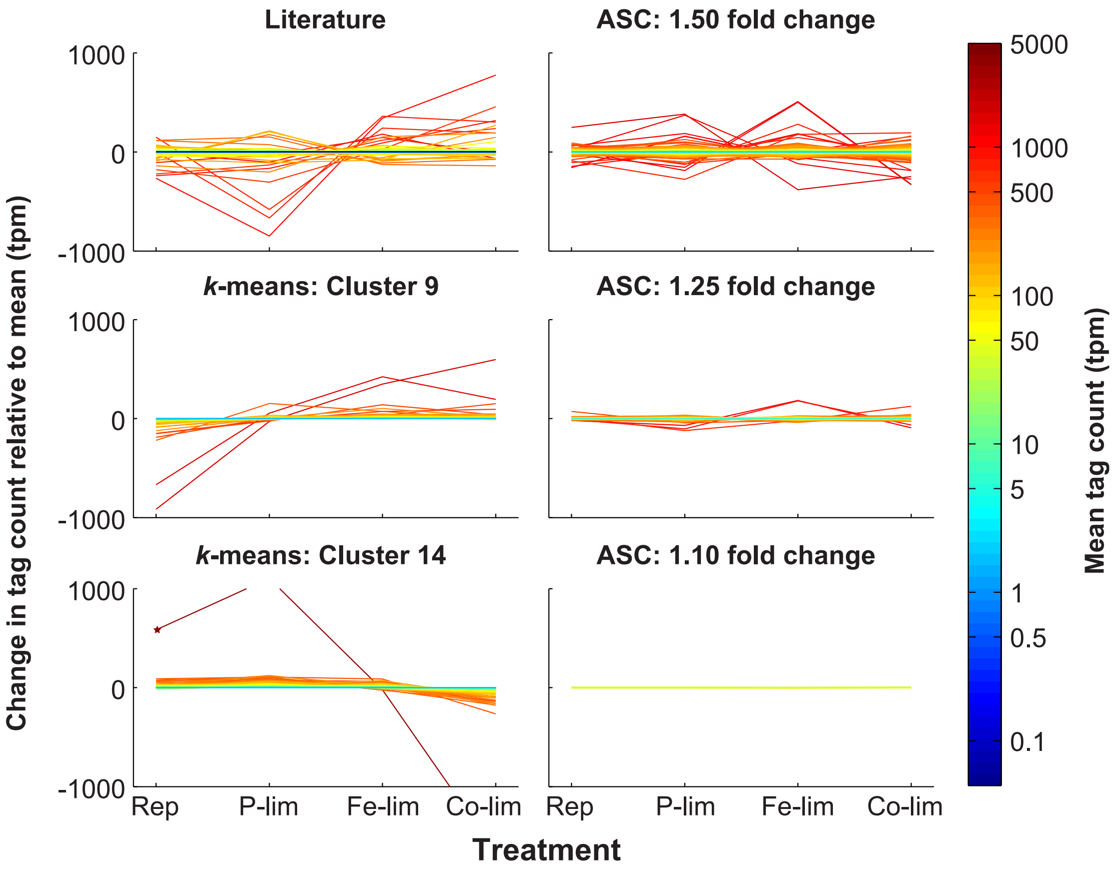


Figure 2-1: Expression patterns of putative reference genes identified through literature-based searches, k-means clustering, and ASC analysis. Through literature-based searches, a total of 101 genes homologous to reference genes from previous studies on plants and algae were identified in *T. pseudonana* 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 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.

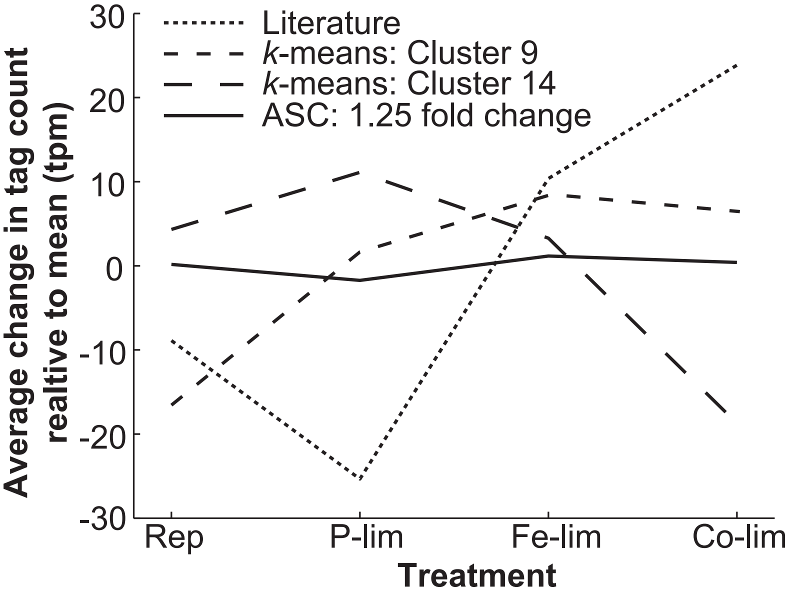


Figure 2-2: Average deviation from the mean level of expression for all genes found with literature-based searches, *k*-means clustering, and ASC analysis of 1.25 fold change. The average change in tag count from the mean expression (TPM) for all the genes identified through literature-based searches 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).

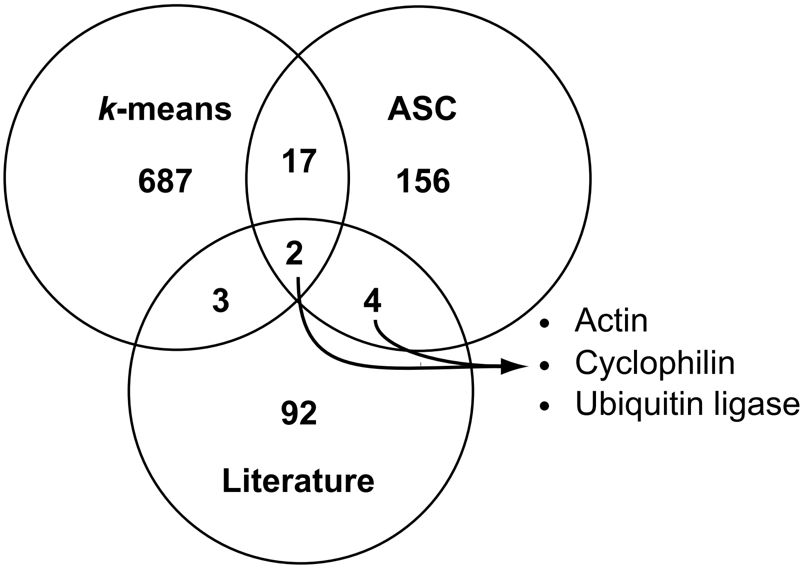


Figure 2-3: Comparison of possible reference genes found with literature-based searches, k-means clustering, and ASC analysis of 1.25 fold change. Venn diagram analysis was used to compare genes identified as candidate reference genes through literature-based homolog searches (totaling 101 genes), with 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).

Chapter 3

Metatranscriptome analyses indicate resource partitioning between diatoms in the field

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 âĂIJparadox of the plankton,âĂİ 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 resource 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 (Supplemental Table 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 (Supplemental Table 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 CCMP 3096 (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 âĂŞsensitive) (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 CCMP 1335 (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* CCMP 1102 v1.0, *Phaeodactylm tricornutum* CCMP 632 v2.0, *Pseudo-nitzschia multiseries* CLN-47 v1.0, and *Thalassiosira pseudonana* CCMP 1335 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 *Skele-*

tonema 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 A-13), 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 (Supplemental Table 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 A-10). 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) (Supplemental Table 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) (Supplemental Table 2). Eukaryotic metatranscriptomes from the five incubation treatments produced over 264 million 100 base pair, paired end cDNA reads (Supplemental Table 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 A-3), 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 A-3). *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 A-3). As such, subsequent analyses were focused on these two groups by remapping the data to representative transcriptomes: *T. rotula* and *S. costatum* (Supplemental Table 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 (Supplemental Table 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* (Supplemental Table 1), highlighting the specificity of the approach.

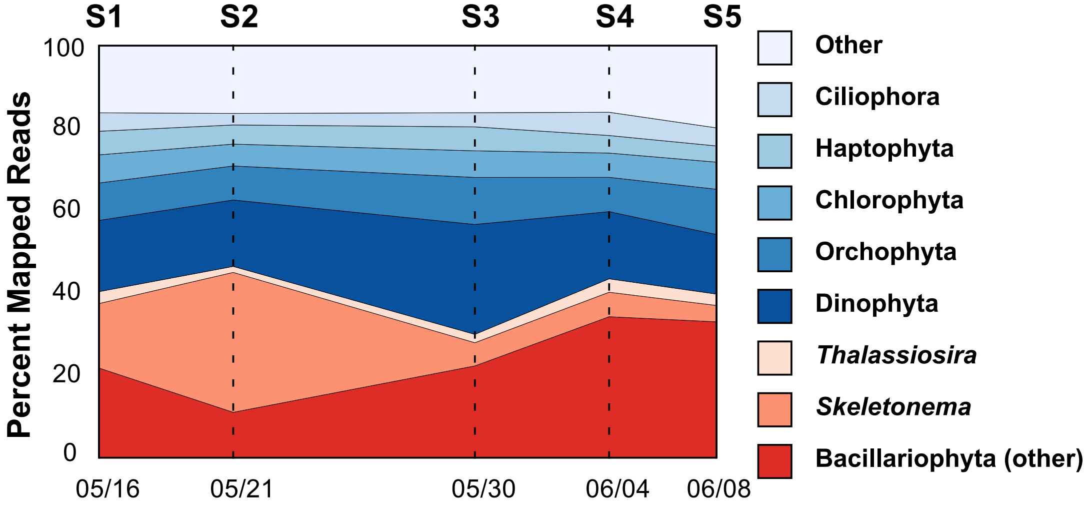


Figure 3-1: Taxonomic classification of RNA-seq paired end reads across the five field samples. Classification was determined by mapping to a database comprised of all publicly available transcriptomes through the Marine Microbial Eukaryotic Transcriptome Project (MMETSP) as of March 17, 2014.

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) (Supplemental Dataset 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 A-4), 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 A-5). 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 A-5), 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 expression 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

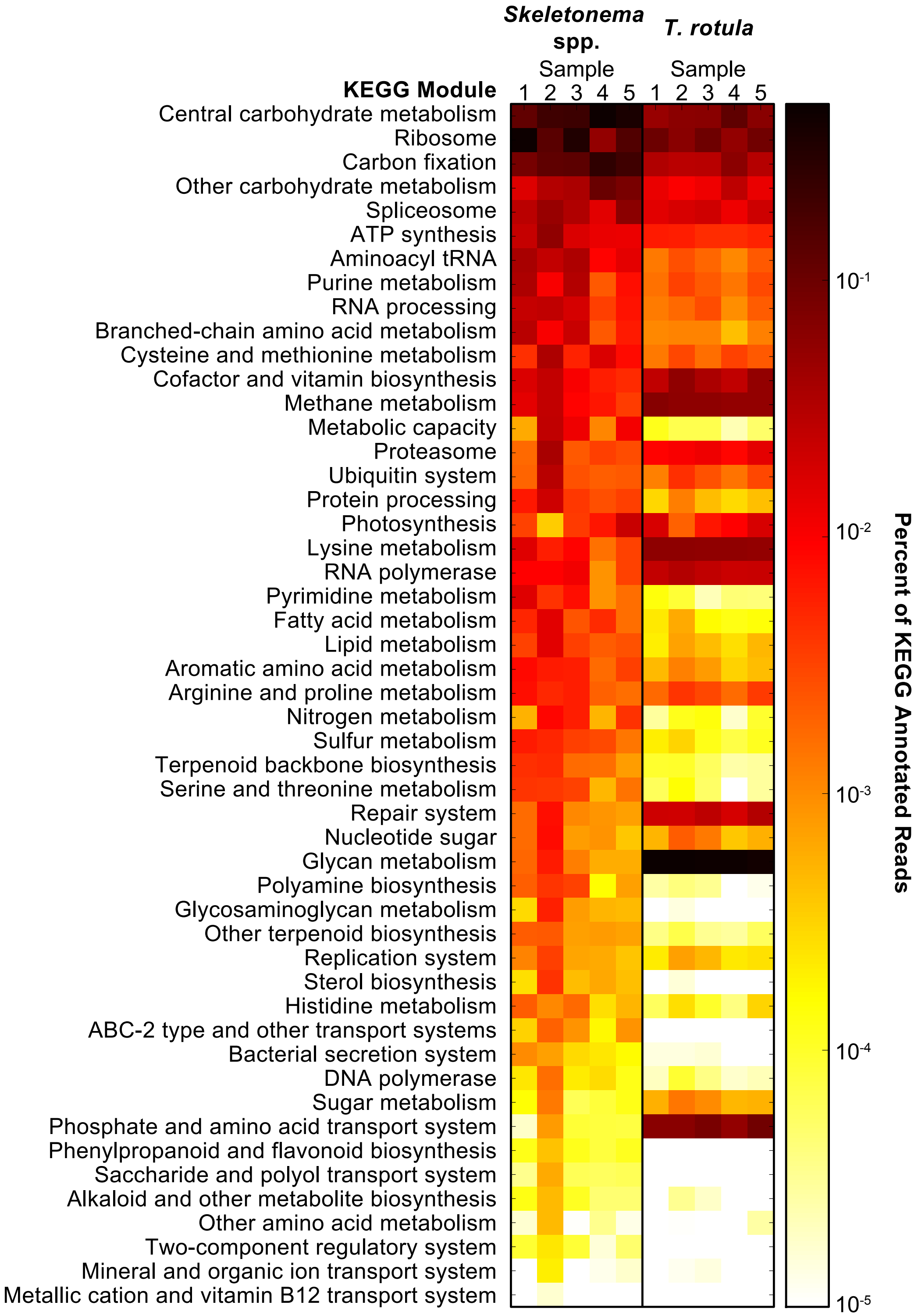


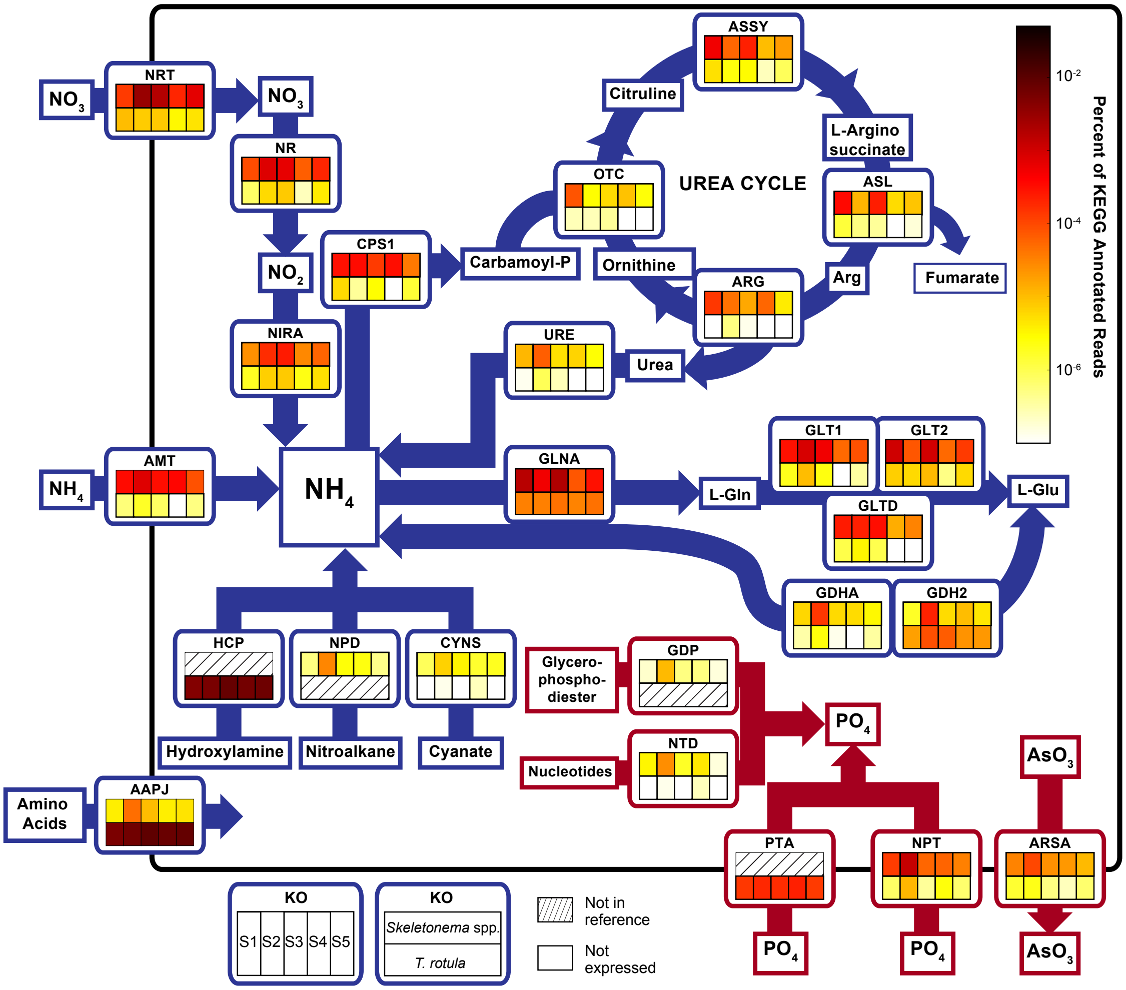
Figure 3-2: Quantitative metabolic fingerprint (QMF) depicting the relative expression of KEGG modules for *Skeletonema* spp. and *T. rotula* in Narragansett Bay across the five sampling time points (S1-S5). Color indicates the proportion of total reads mapping to each KEGG module relative to all KEGG annotated reads.

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 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



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

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.

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 A-6) 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 (Supplemental Dataset 2, Supplemental Table 3). As is common with phytoplankton studies (Marchetti et al., 2012), the majority of the RR genes do not have a KEGG annotation (Figure A-6A). 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 A-7). 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 A-7). 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) (Supplemental Dataset 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., 2011; Dyhrman et al., 2006; Bruhn et al., 2010). Overall, these genes demonstrated patterns of regulation *in situ* (Figure 3-4B, Figure A-8) 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 **??**). In fact, members of this large gene family (Figure A-9) showed disparate regulation in both species (Figure A-8). 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

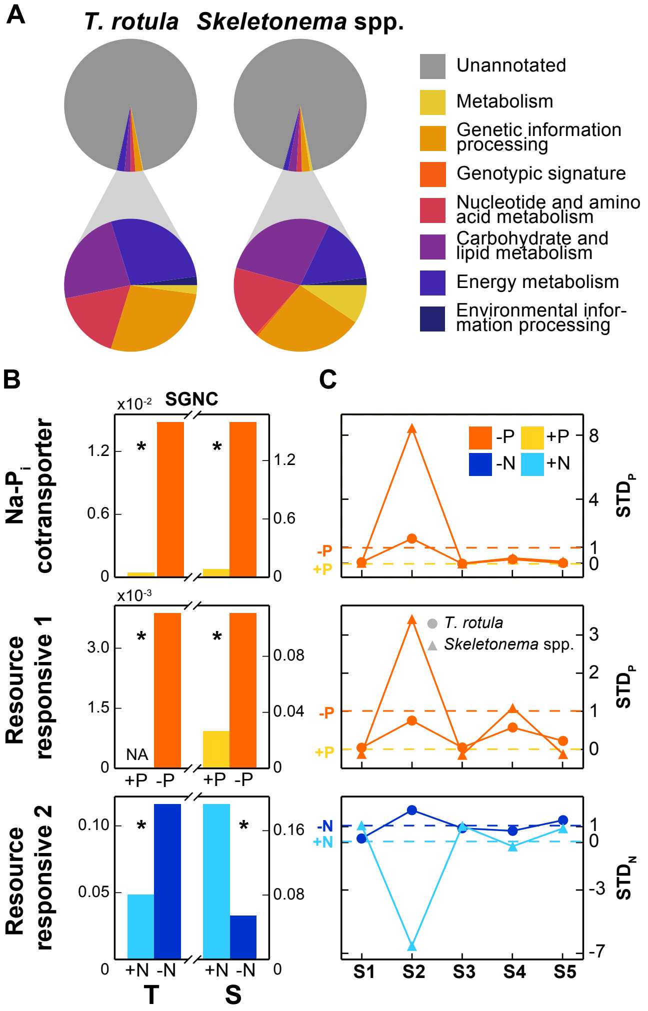


Figure 3-4: Functional composition of the resource-responsive (RR) gene sets for *T. rotula* and *Skeletonema* spp. (A), the relative expression in the incubation samples (B) and standardized transcriptional differentiation (STD) scores (C) for a known P-responsive gene, sodium-phosphate cotransporter, 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.

Skeletonema spp. and *T. rotula* (Figure 3-4B, Figure A-9). 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 A-9). 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 A-9) 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, 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 (Supplemental Dataset 2) to expand upon the single gene analyses above. The RR genes were plotted based on the $STD_N : STD_P$ (Figure A-10) to examine how similar the pattern was to the incubation N:P ratio (Figure 3-5A, Figure A-11). 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

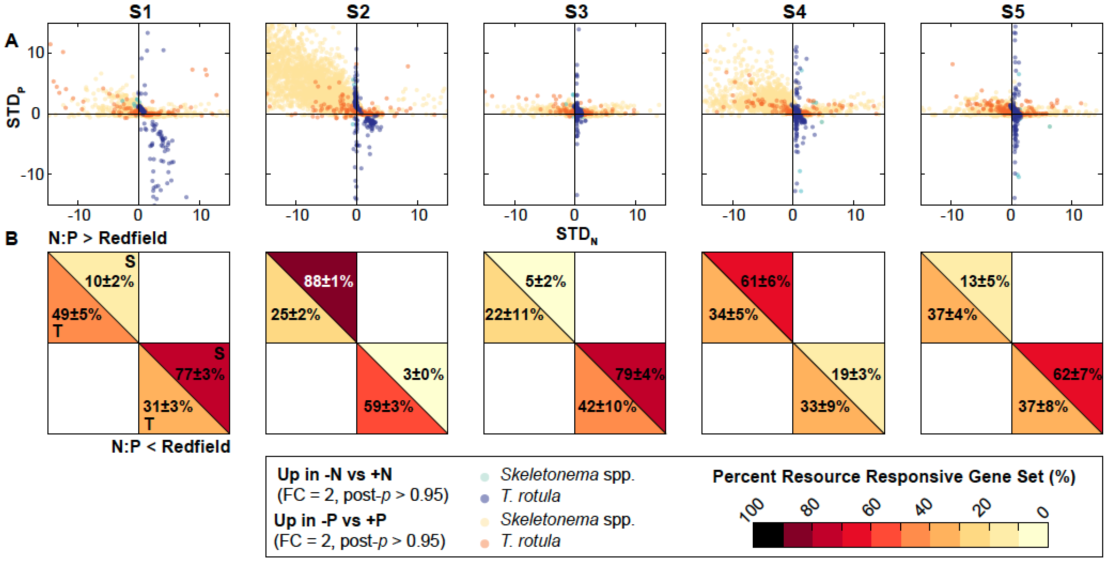


Figure 3-5: Evolution of resource-responsive (RR) gene partitioning over time in Narragansett Bay for *T. rotula* and *Skeletonema* spp.. (A) The stable gene normalized field signal for each gene identified as significantly (2-fold change, post-p $>$ 0.95) up-regulated in -P vs. +P for *Skeletonema* 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).

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 A-12). 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-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.

Appendix A

Supplemental Information

A.1 Appendix for Chapter 2

A.1.1 Supplemental Figures

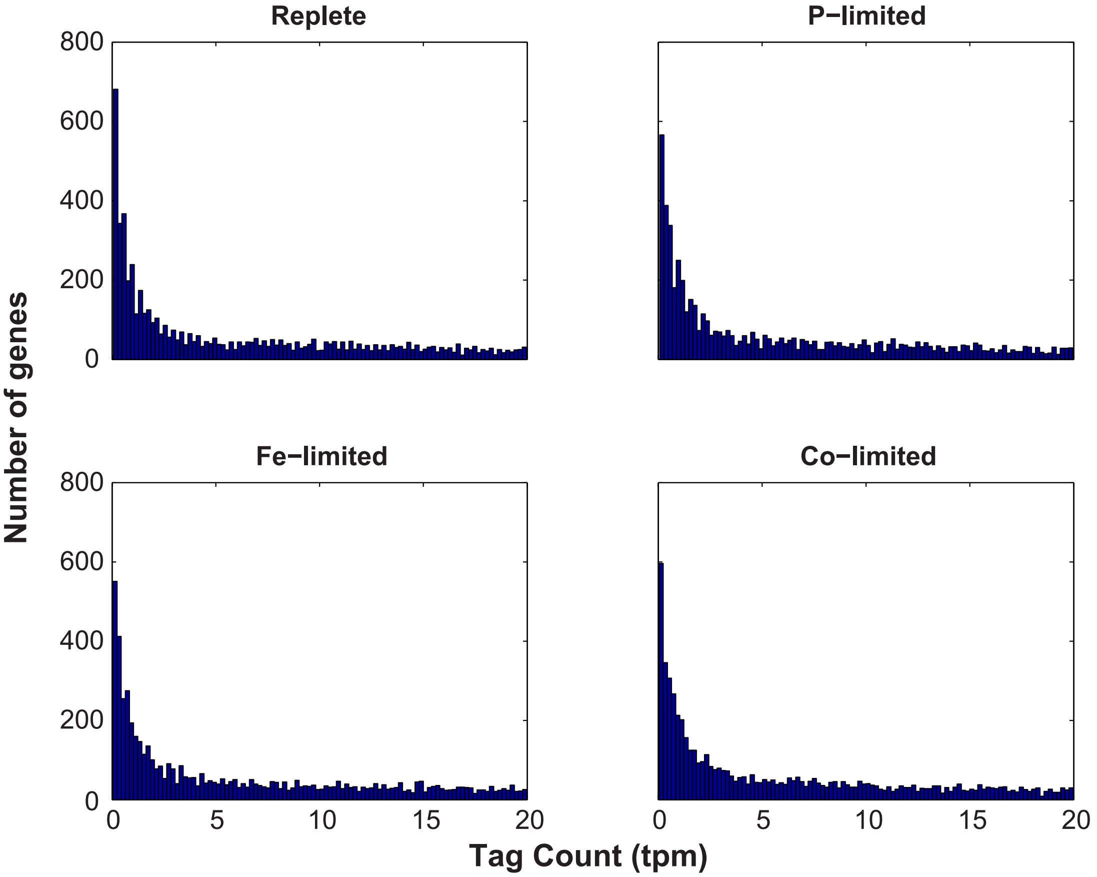


Figure A-1: Histogram analysis of the distribution of normalized tag counts (TPM) for each gene across each of the four treatments (Replete, P-limited, Fe-limited, and co-limited). The abundance of normalized tag counts (TPM) was assessed, tallying the total number of genes with a given tag count. Only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

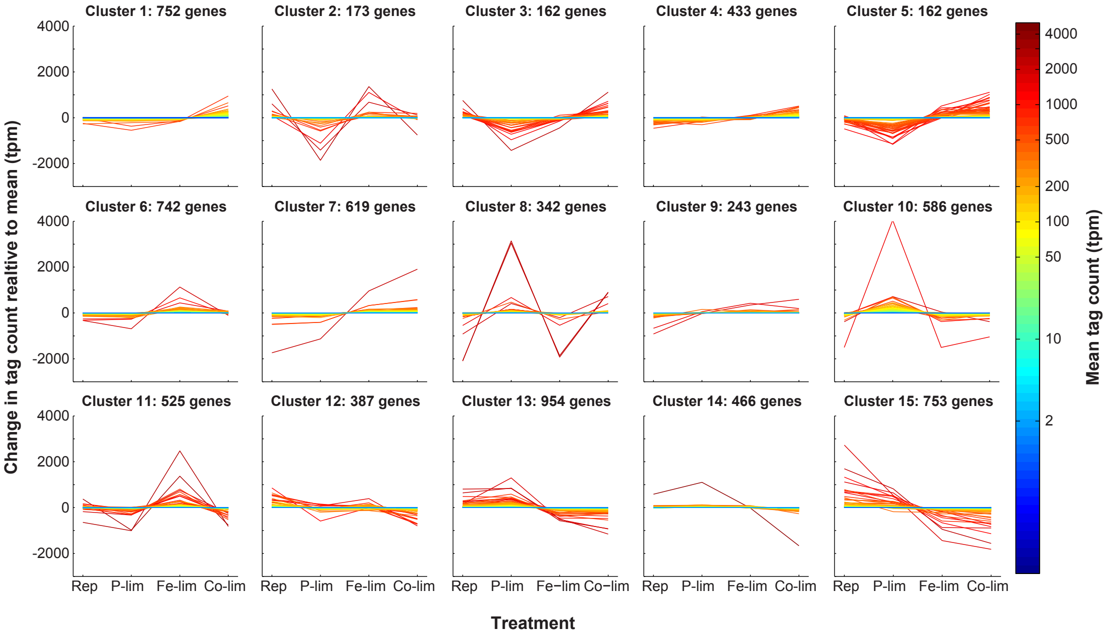


Figure A-2: *K*-means clustering of normalized genes. The 7380 genes that passed the 2.5 TPM cutoff were clustered into 15 clusters using the *k*-means algorithm under the Pearson correlation coefficient. Tag counts normalized to total library size (in TPM) for each gene are plotted relative to the mean (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 Appendix for Chapter 3

A.2.1 Supplemental Figures

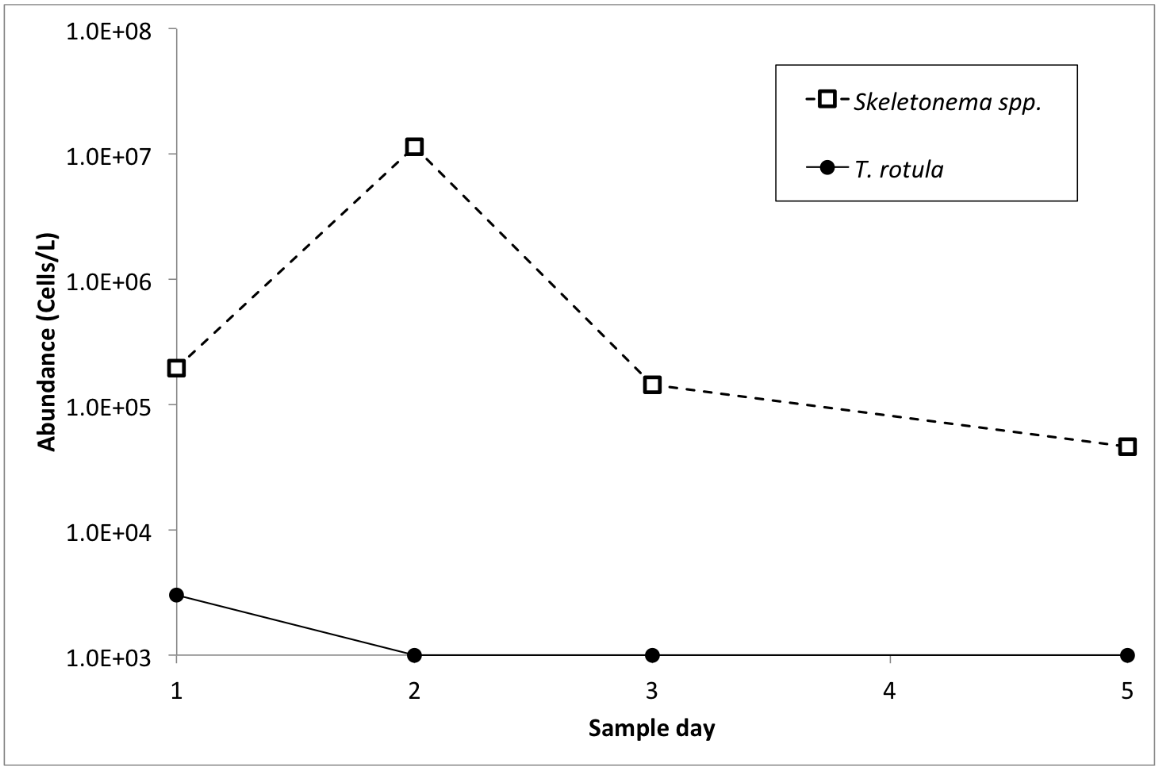


Figure A-3: Abundance estimation from cell counts of *Skeletonema* spp. and *T. rotula* across the five sample points during the spring of 2012.

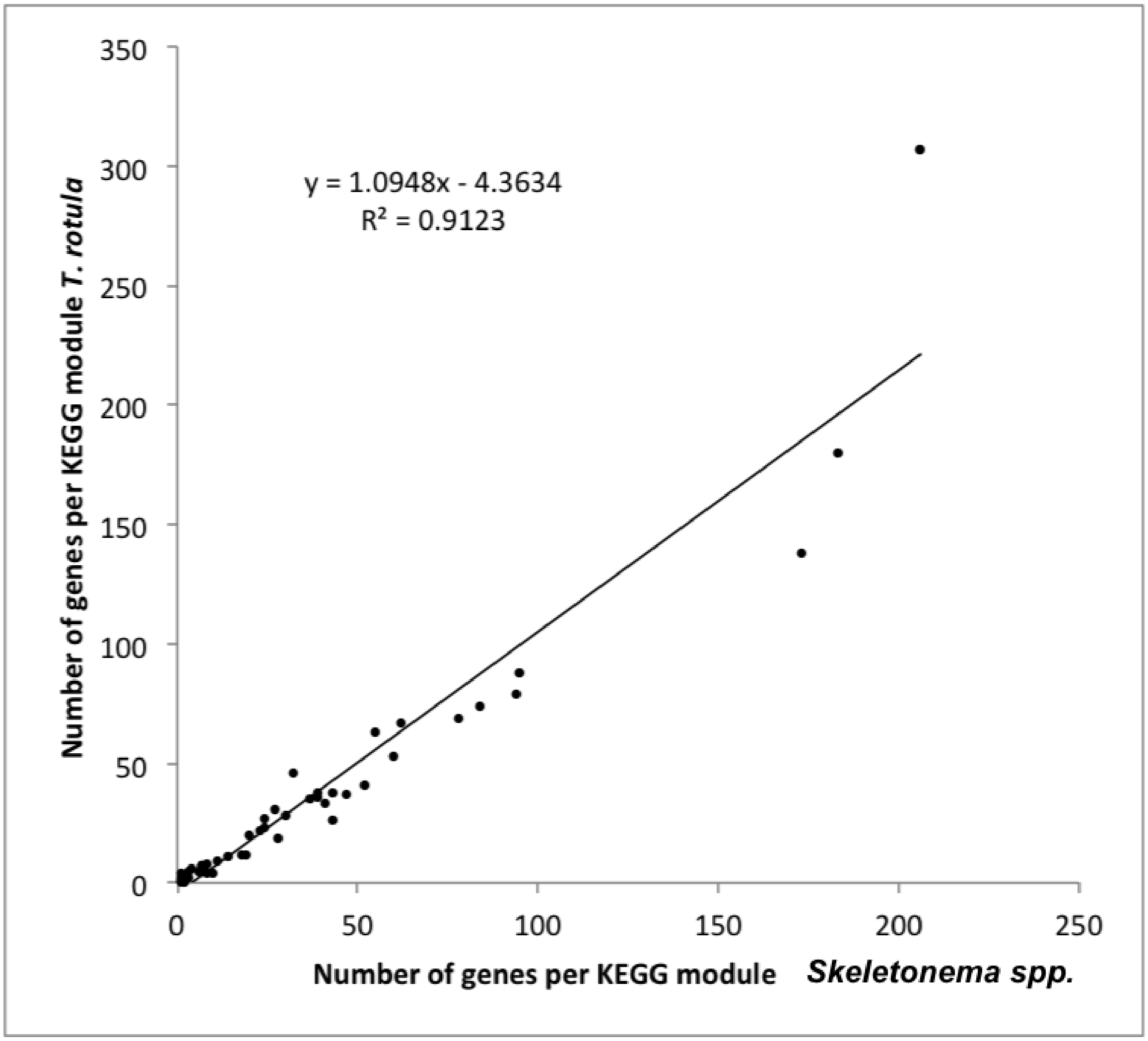


Figure A-4: Total number of genes assigned to each KEGG module for *Skeletonema* spp. and *T. rotula*.

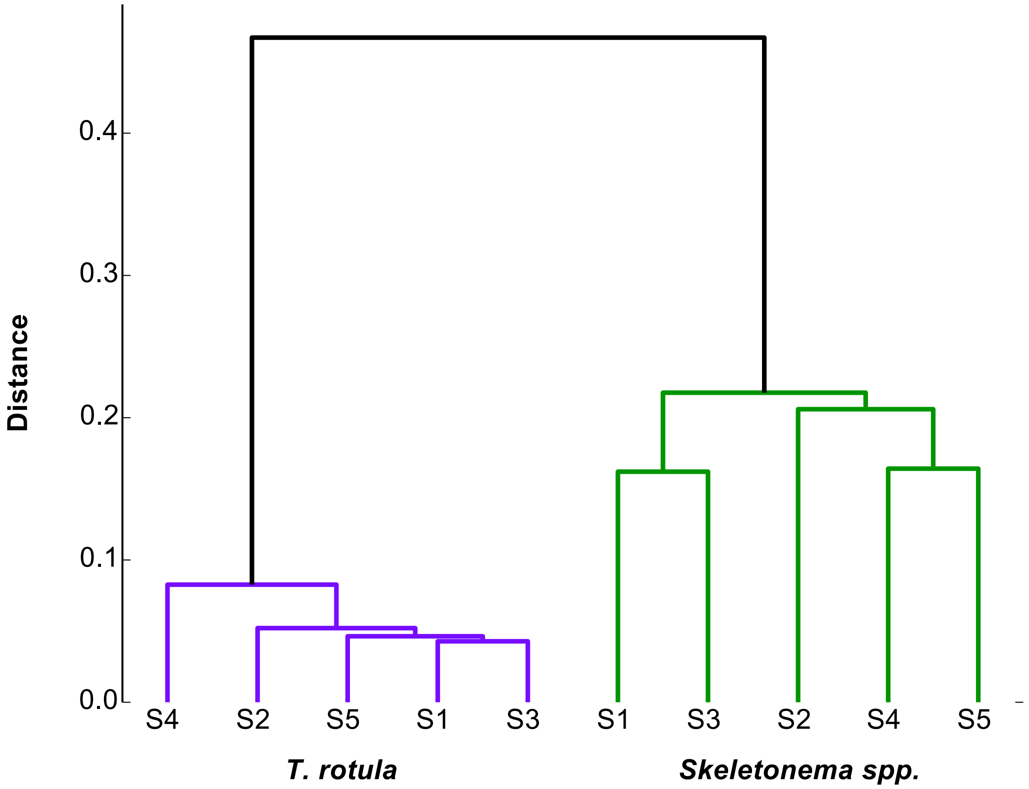


Figure A-5: Dendrogram depicting hierarchical clustering of samples based on relative expression of KEGG modules (Figure 2) across the five samples S1-S5 for *Skeletonema* spp. and *T. rotula*.

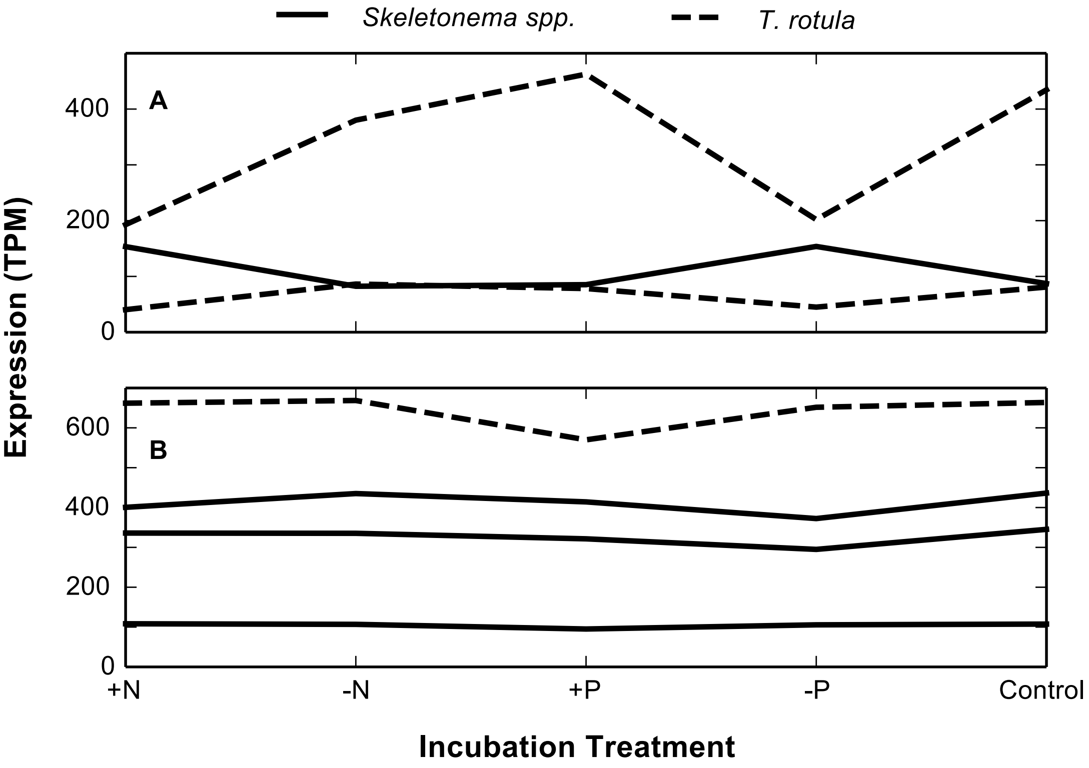


Figure A-6: Expression of stable reference genes identified based on literature and statistical parsing in nutrient amendment incubation. (A) The expression in tags per million (TPM) of stable reference genes identified in *T. rotula* (dashed line) and *Skeletonema* spp. (solid lines) based on homology (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

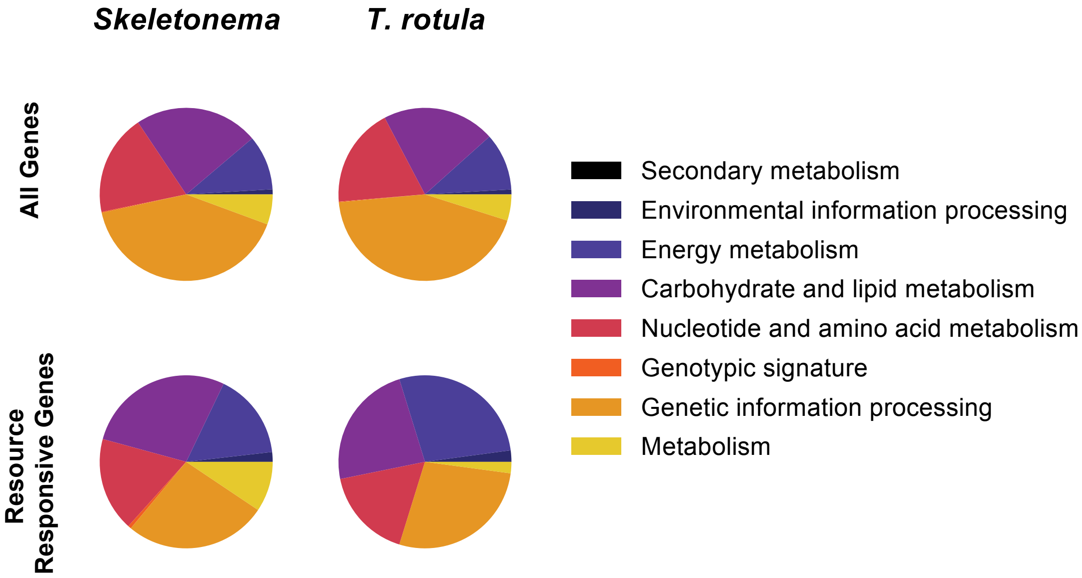


Figure A-7: Functional composition of the reference transcriptome and resource-responsive (RR) gene subset for *T. rotula* and *Skeletonema* spp. (A) RR gene sets were identified through cross comparison of like-nutrient incubations (i.e. +N vs. -N and +P vs. -P), using ASC (fold change = 2, post-$p > 0.95$). 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.

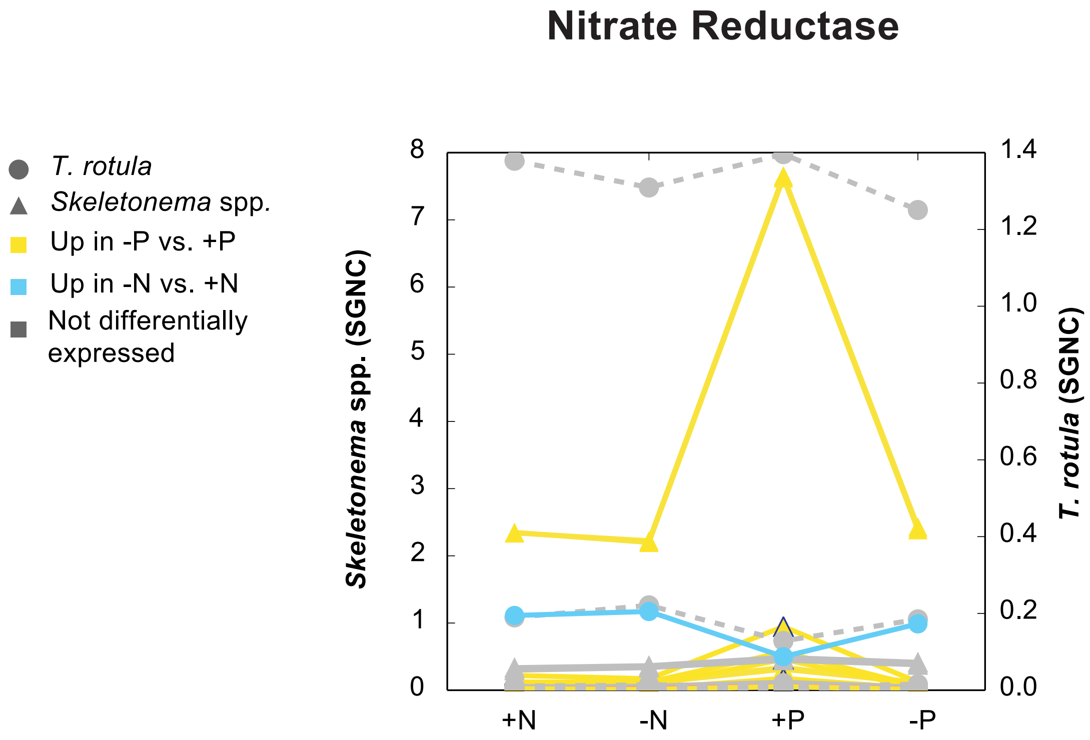


Figure A-8: The relative expression in stable gene normalized counts ($SGNC$) of the assimilatory nitrate reductase gene cluster across the incubation experiment treatments. Significance of regulation between the treatments is denoted by the color of the line; organisms are denoted by the shapes of the marker.

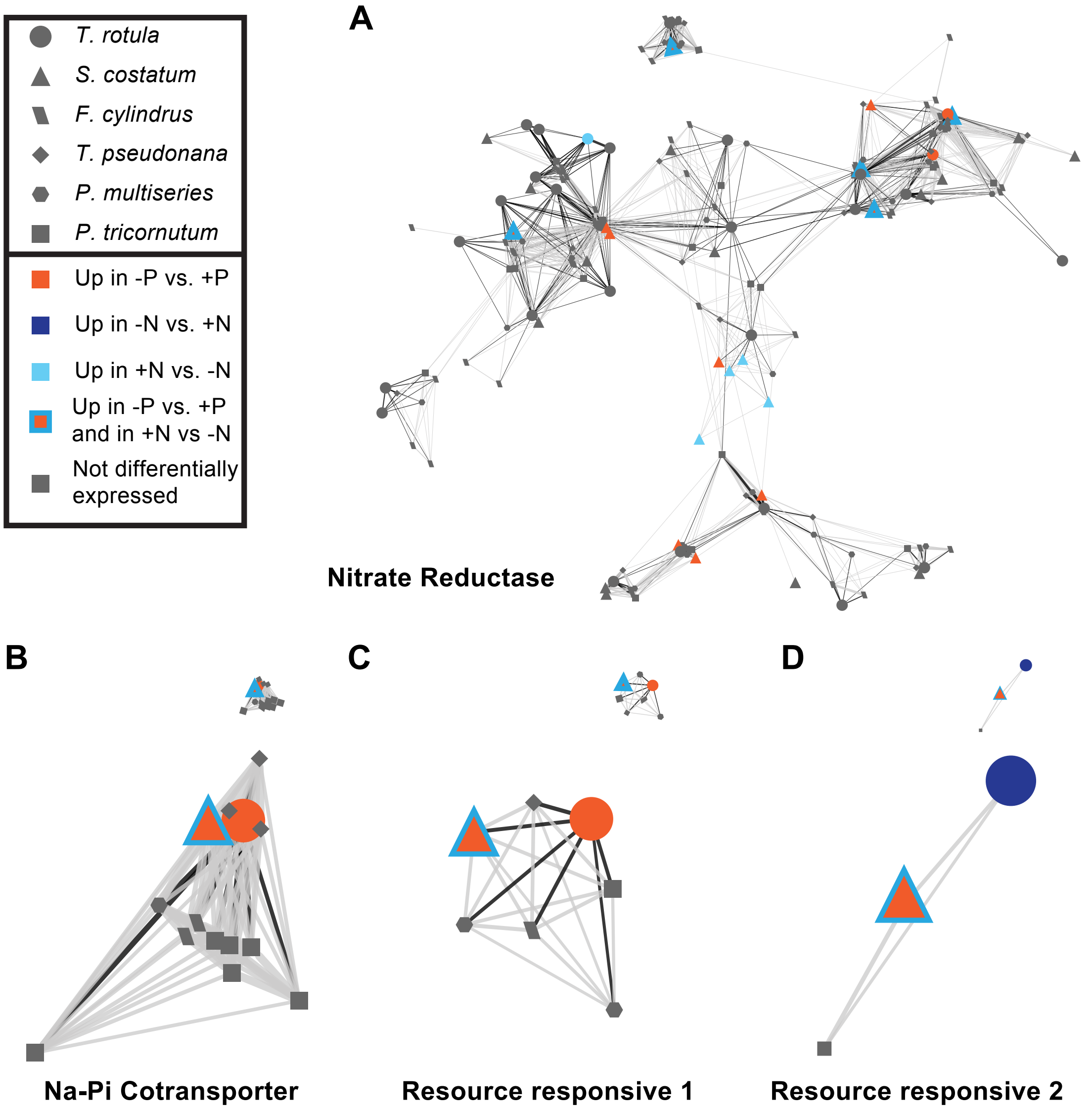


Figure A-9: Gene cluster known nutrient-responsive genes in *T. pseudonana*: (A) assimilatory nitrate reductase and (B) sodium-phosphate cotransporter and novel resource-responsive (RR) gene families: (C) RR1 and (D) RR2. Transcripts from the transcriptomes of *T. rotula* and *Skeletonema* spp. were 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.

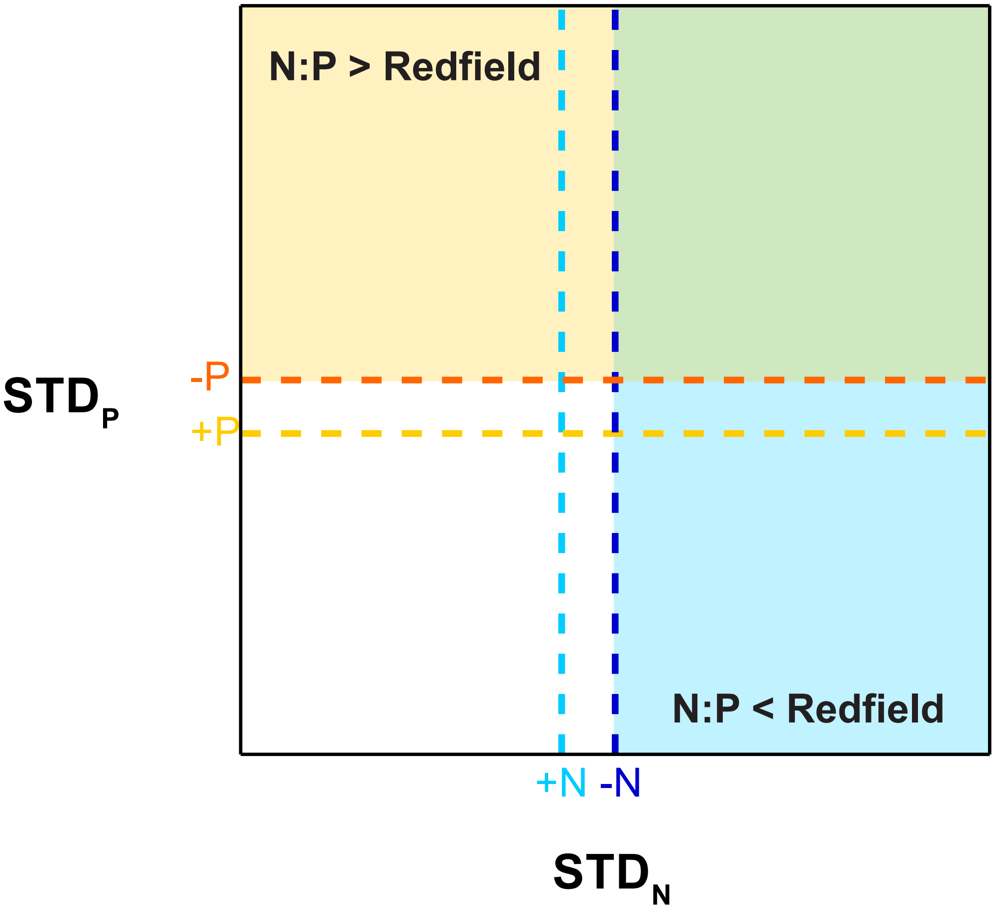


Figure A-10: A conceptual schematic of STD_N plotted against STD_P hypothesized regions of N:P > Redfield physiology and N:P < Redfield physiology highlighted.

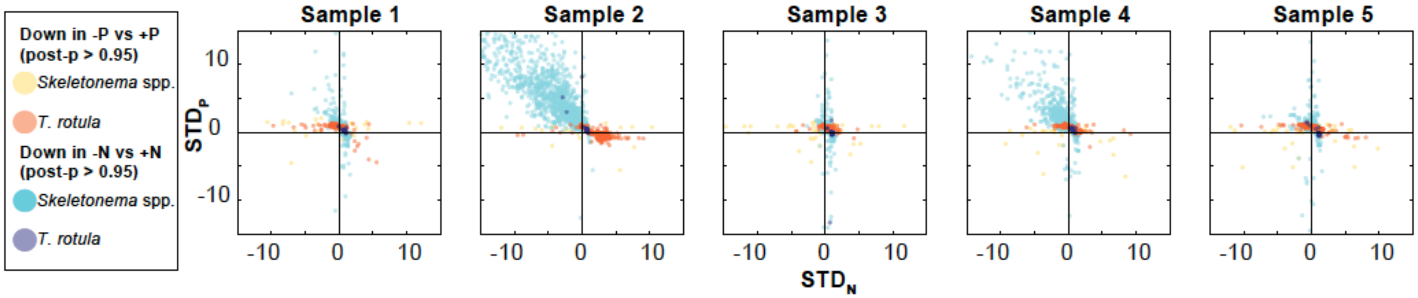


Figure A-11: Evolution of niche space indexing over time in Narragansett Bay for *T. rotula* and *Skeletonema* spp.. The stable gene normalized field signal from genes identified as significantly (2-fold change, post$-p > 0.95$) down-regulated in -P vs +P for Skeletonema spp. (yellow) and *T. rotula* (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.

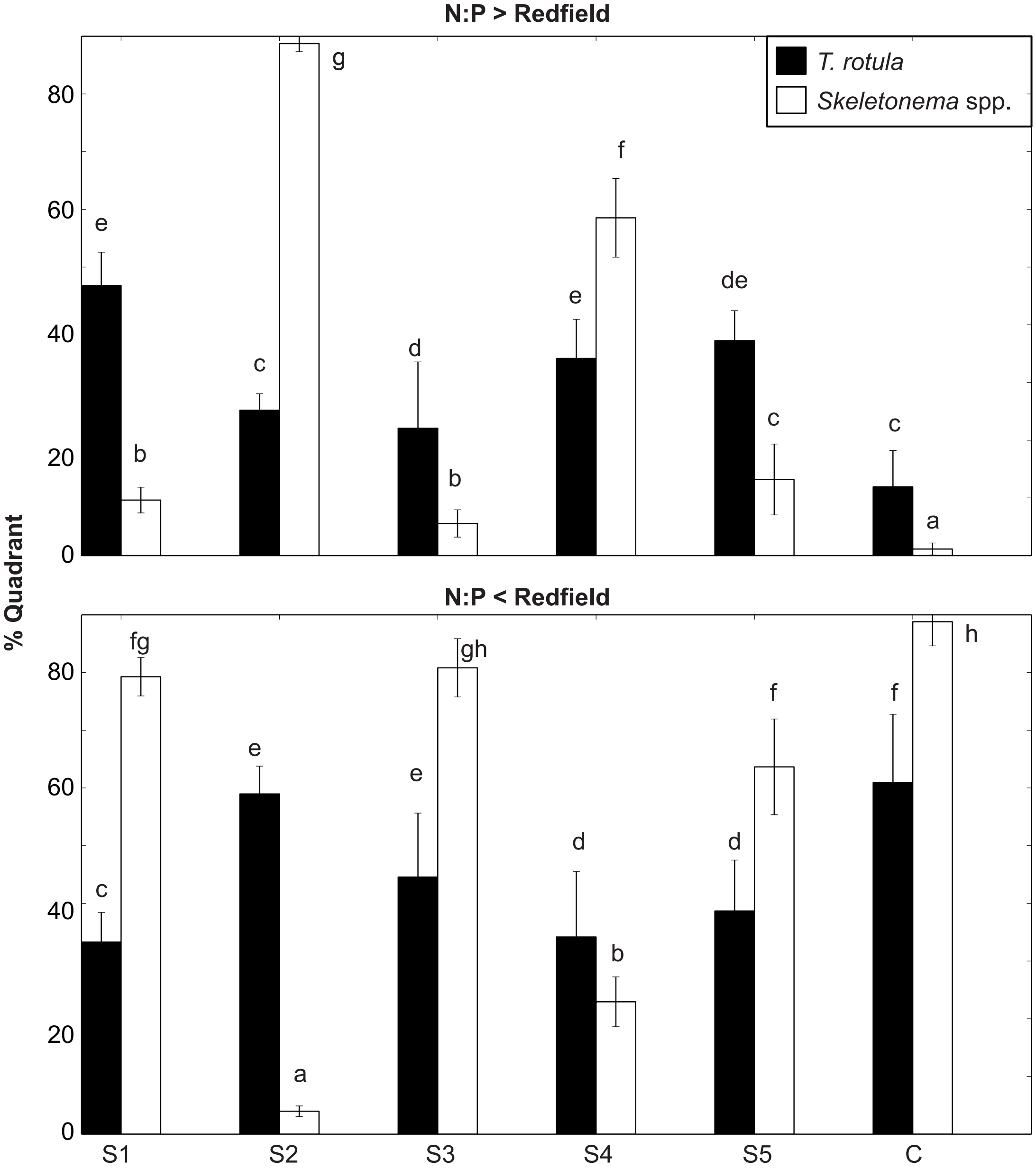


Figure A-12: The percentage of identified nutrient responsive genes falling into the N:P $>$ Redfield and N:P $<$ Redfield quadrants for *T. rotula* and *Skeletonema* spp.. The total number of genes falling into the N:P $>$ Redfield quadrant ($STD_P > C$; $STD_N < C$, for $0.25 < C < 0.75$) and the N:P $<$ 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 ($p < 0.05$).

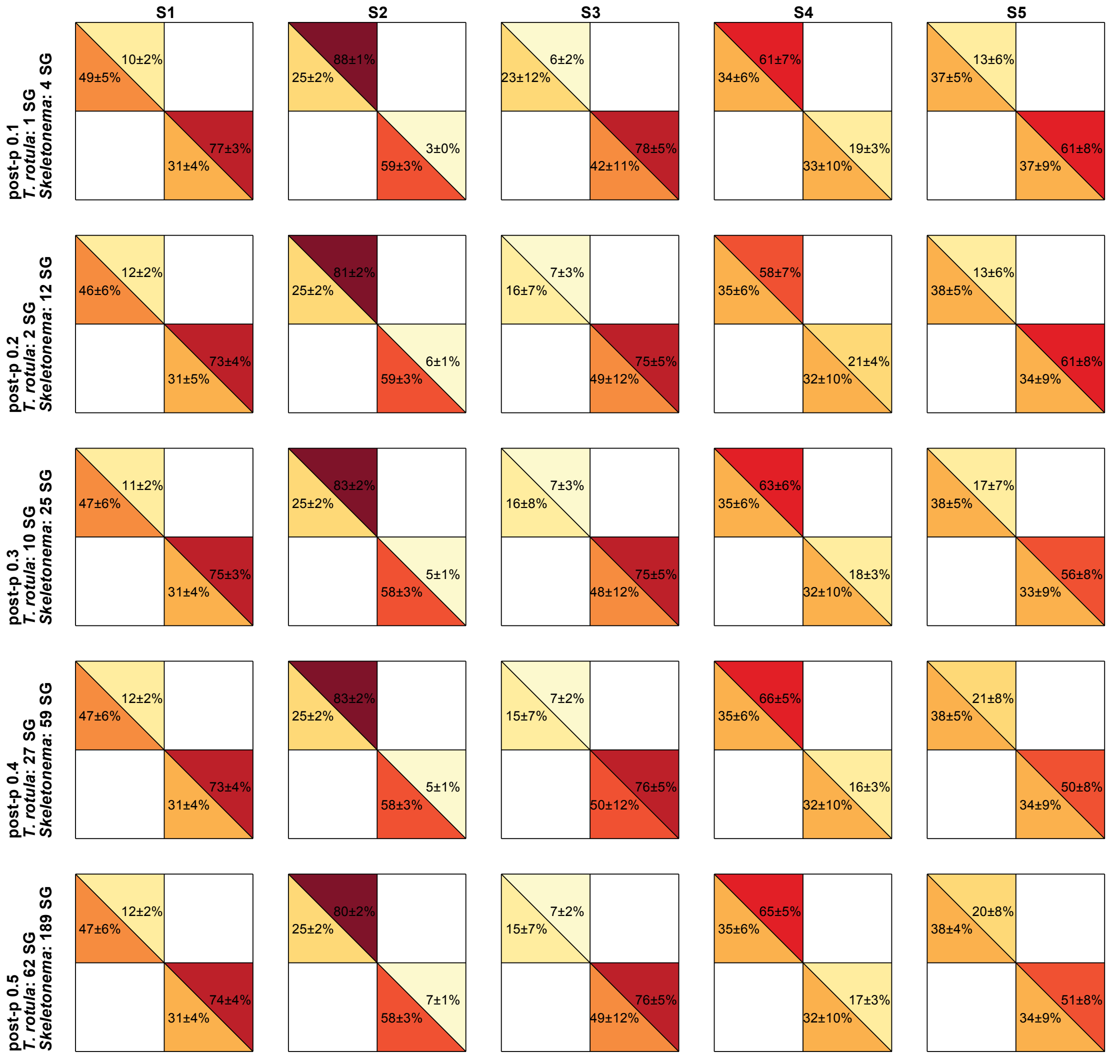


Figure A-13: The impact of stable gene selection on the quadrant localization of the resource responsive gene sets. The posterior probability cutoff used in the selection of stable genes was varied from 0.1 to 0.5 for a fold change of 1.25. The percentage of identified nutrient responsive genes falling 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.

Bibliography

Harriet Alexander, Bethany D Jenkins, Tatiana a Rynearson, Mak A Saito, Melissa L Mercier, and Sonya T Dyhrman. 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.

Andrew E Allen, Julie Laroche, Uma Maheswari, Markus Lommer, Nicolas Schauer, Pascal J Lopez, Giovanni Finazzi, Alisdair R Fernie, and Chris Bowler. 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.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. 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.

Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. 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.

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

Janine Antonov, Darlene R Goldstein, Andrea Oberli, Anna Baltzer, Marco Pirotta, Achim Fleischmann, Hans J Altermatt, and Rolf Jaggi. 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, June 2005.

E. Virginia Armbrust, John A. Berges, Chris Bowler, Beverley R. Green, Diego Martinez, Nicholas H. Putnam, Shiguo Zhou, Andrew E. Allen, Kirk E. Apt, Michael Bechner, Mark A. Brzezinski, Balbir K. Chaal, Anthony Chiovitti, Aubrey K. Davis, Mark S. Demarest, J. Chris Detter, Tijana Glavina, David Goodstein, Masood Z. Hadi, Uffe Hellsten, Mark Hildebrand, Bethany D. Jenkins, Jerzy Jurka, Vladimir V. Kapitonov, Nils Kröger, Winnie W. Y. Lau, Todd W. Lane, Frank W. Larimer, J. Casey Lippmeier, Susan Lucas, Mónica Medina, Anton Montsant, Miroslav Obornik, Micaela Schnitzler Parker, Brian Palenik, Gregory J. Pazour, Paul M. Richardson, Tatiana A. Rynearson, Mak A. Saito, David C. Schwartz, Kimberlee Thamatrakoln, Klaus Valentin, Assaf Vardi, Frances P. Wilkerson, and Daniel S. Rokhsar. 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.

Nelson Avonce, Barbara Leyman, José O. Mascorro-Gallardo, Patrick Van Dijck, Johan M. Thevelein, and Gabriel Iturriaga. 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.

Patricia Balvanera, Andrea B Pfisterer, Nina Buchmann, Jing-Shen He, Tohru Nakashizuka, David Raffaelli, and Bernhard Schmid. 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.

Stephan Behl, Anne Donval, and Herwig Stibor. 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.

Sara J. Bender, Micaela S. Parker, and E. Virginia Armbrust. 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.

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

Gry Mine Berg, Jeff Shrager, Gernot Glöckner, Kevin R Arrigo, and Arthur R Grossman. 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.

David G. Borkman and Theodore Smayda. Multidecadal (1959âĂŞ1997) 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.

Chris Bowler, Andrew E Allen, Jonathan H Badger, Jane Grimwood, Kamel Jabbari, Alan Kuo, Uma Maheswari, Cindy Martens, Florian Maumus, Robert P Otillar, Edda Rayko, Asaf Salamov, Klaas Vandepoele, Bank Beszteri, Ansgar Gruber, Marc Heijde, Michael Katinka, Thomas Mock, Klaus Valentin, Fréderic Verret, John A Berges, Colin Brownlee, Jean-Paul Cadoret, Anthony Chiovitti, Chang Jae Choi, Sacha Coesel, Alessandra De Martino, J Chris Detter, Colleen Durkin, Angela Falciatore, Jérome Fournet, Miyoshi Haruta, Marie J J Huysman, Bethany D

Jenkins, Katerina Jiroutova, Richard E Jorgensen, Yolaine Joubert, Aaron Kaplan, Nils Kröger, Peter G Kroth, Julie La Roche, Erica Lindquist, Markus Lommer, Véronique Martin-Jézéquel, Pascal J Lopez, Susan Lucas, Manuela Mangogna, Karen McGinnis, Linda K Medlin, Anton Montsant, Marie-Pierre Oudot-Le Secq, Carolyn Napoli, Miroslav Obornik, Micaela Schnitzler Parker, Jean-Louis Petit, Betina M Porcel, Nicole Poulsen, Matthew Robison, Leszek Rychlewski, Tatiana A Rynearson, Jeremy Schmutz, Harris Shapiro, Magali Siaut, Michele Stanley, Michael R Sussman, Alison R Taylor, Assaf Vardi, Peter von Dassow, Wim Vyverman, Anusuya Willis, Lucjan S Wyrwicz, Daniel S Rokhsar, Jean Weissenbach, E Virginia Armbrust, Beverley R Green, Yves Van de Peer, and Igor V Grigoriev. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239–44, November 2008. doi: 10.1038/nature07410.

Annette Bruhn, Julie LaRoche, and Katherine Richardson. 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.

S. Bustin. 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.

Marc W Cadotte. 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.

Philip E. Caldwell, Magdalena Walkiewicz, and Michael Stern. 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.

Bradley J Cardinale, J Emmett Duffy, Andrew Gonzalez, David U Hooper, Charles Perrings, Patrick Venail, Anita Narwani, Georgina M Mace, David Tilman, David A Wardle, Ann P Kinzig, Gretchen C Daily, Michel Loreau, James B Grace, Anne Larigauderie, Diane S Srivastava, and Shahid Naeem. Biodiversity loss and its impact on humanity. *Nature*, 486(7401):59–67, June 2012. doi: 10.1038/nature11148.

Raquel N Carvalho and Teresa Lettieri. 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.

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

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

Tomasz Czechowski, Mark Stitt, Thomas Altmann, Michael K. Udvardi, and Wolf-Rüdiger Scheible. 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.

Aubrey K. Davis and Brian Palenik. 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.

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

Hendrik J. M. de Jonge, Rudolf S. N. Fehrmann, Eveline S. J. M. de Bont, Robert M. W. Hofstra, Frans Gerbens, Willem A. Kamps, Elisabeth G. E. de Vries, Ate G. J. van der Zee, Gerard J. te Meerman, and Arja ter Elst. Evidence based selection of housekeeping genes. *PloS one*, 2(9):e898, January 2007. doi: 10.1371/journal.pone.0000898.

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

K Dheda, J F Huggett, J S Chang, L U Kim, S A Bustin, M A Johnson, G A W Rook, and A Zumla. 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.

Chris L Dupont, John P McCrow, Ruben Valas, Ahmed Moustafa, Nathan Walworth, Ursula Goodenough, Robyn Roth, Shane L Hogle, Jing Bai, Zackary I Johnson, Elizabeth Mann, Brian Palenik, Katherine A Barbeau, J Craig Venter, and Andrew E Allen. 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.

Sonya T Dyhrman, Sheean T Haley, Shanda R Birkeland, Louie L Wurch, Michael J Cipriano, and Andrew G McArthur. 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.

Sonya T Dyhrman, Bethany D Jenkins, Tatiana A Rynearson, Mak A Saito, Melissa L Mercier, Harriet Alexander, Leann P Whitney, Andrea Drzewianowski, Vladimir V Bulygin, Erin M Bertrand, Zhijin Wu, Claudia Benitez-Nelson, and Abigail Heithoff. 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.

Charles S. Elton. *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.

Christopher B. Field, Michael J. Behrenfeld, James T Randerson, and Paul G. Falkowski. 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.

Thomas Flatt, Kyung-Jin Min, Cecilia D'Alterio, Eugenia Villa-Cuesta, John Cumbers, Ruth Lehmann, D. Leanne Jones, and Marc Tatar. *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.

Michael J Follows, Stephanie Dutkiewicz, Scott Grant, and Sallie W Chisholm. 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.

Kyle R Frischkorn, Matthew J Harke, Christopher J Gobler, and Sonya T Dyhrman. 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, January 2014. doi: 10.3389/fmicb.2014.00375.

M. J. Furnas. 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.

Miles J. Furnas. 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.

Mark Gerstein and Ronald Jansen. The current excitement in bioinformaticsâĂŤanalysis 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.

Francis D. Gibbons and Frederick P. Roth. 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.

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

Scott M. Gifford, Shalabh Sharma, Melissa Booth, and Mary Ann Moran. Expression patterns reveal niche diversification in a marine microbial assemblage. *The ISME journal*, 7(2):281–98, February 2013. doi: 10.1038/ismej.2012.96.

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

Ruoyu Guo and Jang-Seu Ki. 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.

Sébastien Halary, Jessica W Leigh, Bachar Cheaib, Philippe Lopez, and Eric Bapteste. 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.

Sébastien Halary, James O McInerney, Philippe Lopez, and Eric Bapteste. 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.

J. A. Hartigan and M. A. Wong. 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.

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

Jef Huisman, Anna M. Johansson, Eelke O. Folmer, and Franz J. Weissing. 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.

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

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

Zackary I Johnson, Erik R Zinser, Allison Coe, Nathan P McNulty, E. Malcolm S Woodward, and Sallie W Chisholm. 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.

Lee-Kuo Kang, Sheng-Ping L. Hwang, Hsing-Juh Lin, Pei-Chung Chen, and Jeng Chang. 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.

Patrick J. Keeling, Fabien Burki, Heather M. Wilcox, Bassem Allam, Eric E. Allen, Linda A. Amaral-Zettler, E. Virginia Armbrust, John M. Archibald, Arvind K. Bharti, Callum J. Bell, Bank Beszteri, Kay D. Bidle, Connor T. Cameron, Lisa

Campbell, David A. Caron, Rose Ann Cattolico, Jackie L. Collier, Kathryn Coyne, Simon K. Davy, Phillipe Deschamps, Sonya T. Dyhrman, Bente Edvardsen, Ruth D. Gates, Christopher J. Gobler, Spencer J. Greenwood, Stephanie M. Guida, Jennifer L. Jacobi, Kjetill S. Jakobsen, Erick R. James, Bethany Jenkins, Uwe John, Matthew D. Johnson, Andrew R. Juhl, Anja Kamp, Laura A. Katz, Ronald Kiene, Alexander Kudryavtsev, Brian S. Leander, Senjie Lin, Connie Lovejoy, Denis Lynn, Adrian Marchetti, George McManus, Aurora M. Nedelcu, Susanne Menden-Deuer, Cristina Miceli, Thomas Mock, Marina Montresor, Mary Ann Moran, Shauna Murray, Govind Nadathur, Satoshi Nagai, Peter B. Ngam, Brian Palenik, Jan Pawlowski, Giulio Petroni, Gwenael Piganeau, Matthew C. Posewitz, Karin Rengefors, Giovanna Romano, Mary E. Rumpho, Tatiana Rynearson, Kelly B. Schilling, Declan C. Schroeder, Alastair G B Simpson, Claudio H. Slamovits, David R. Smith, G Jason Smith, Sarah R. Smith, Heidi M. Sosik, Peter Stief, Edward Theriot, Scott N. Twary, Pooja E. Umale, Daniel Vaulot, Boris Wawrik, Glen L. Wheeler, William H. Wilson, Yan Xu, Adriana Zingone, and Alexandra Z. Worden. 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.

Adam B Kustka, Andrew E Allen, and François M. M Morel. 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.

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

Dave Lanoix, Andrée-Anne Lacasse, Joey St-Pierre, Sean C. Taylor, Maude Ethier-Chiasson, Julie Lafond, and Cathy Vaillancourt. Quantitative PCR pitfalls: the case of the human placenta. *Molecular biotechnology*, 52(3):234–43, November 2012. doi: 10.1007/s12033-012-9539-2.

C Lawerence and S Menden-Deuer. 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.

Fred Wang-Fat Lee, David Morse, and Samuel Chun-Lap Lo. 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.

William Leggat, David Yellowlees, and Monica Medina. Recent progress in *Symbiodinium* transcriptomics. *Journal of Experimental Marine Biology and Ecology*, 408 (1âĂŞ2):120–125, November 2011. doi: 10.1016/j.jembe.2011.07.032.

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

Weizhong Li and Adam Godzik. 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.

Yaqin Li and Theodore J Smayda. Temporal variability of chlorophyll in Narragansett Bay, 1973âĂŞ1990. *ICES Journal of Marine Science: Journal du Conseil*, 55(4): 661–667, August 1998. doi: 10.1006/jmsc.1998.0383.

Senjie Lin, Gustaf Sandh, Huan Zhang, Jiujun Cheng, Keri Perkins, Edward J. Carpenter, and Birgitta Bergman. 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.

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

M. Lohse, A. M. Bolger, A. Nagel, A. R. Fernie, J. E. Lunn, M. Stitt, and B. Usadel. 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.

Maria T. Maldonado, Andrew E. Allen, Joanne S. Chong, Kevin Lin, Dan Leus, Nataliya Karpenko, and Shannon L. Harris. 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.

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

Adrian Marchetti, David M Schruth, Colleen a Durkin, Micaela S Parker, Robin B Kodner, Chris T Berthiaume, Rhonda Morales, Andrew E Allen, and E Virginia Armbrust. 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.

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

Céline Marionneau, Brigitte Couette, Jie Liu, Huiyu Li, Matteo E Mangoni, Joël Nargeot, Ming Lei, Denis Escande, and Sophie Demolombe. 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.

John H. Martin. 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.

Jay McCarren, Jamie W. Becker, Daniel J. Repeta, Yanmei Shi, Curtis R. Young, Rex R. Malmstrom, Sallie W. Chisholm, and Edward F. DeLong. 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.

Sarah M. McDonald, Joshua N. Plant, and Alexandra Z. Worden. 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.

Patrick J McGinn and François M. M Morel. 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.

Patrick J. McGinn and François M.M. Morel. 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.

Susanne Menden-deuer and Julie Rowlett. Many ways to stay in the game : individual variability maintains high biodiversity in planktonic microorganisms Many ways to stay in the game : individual variability maintains high biodiversity in planktonic microorganisms. (March), 2014.

Xiang Jia Min, Gregory Butler, Reginald Storms, and Adrian Tsang. 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.

Thomas Mock, Manoj Pratim Samanta, Vaughn Iverson, Chris Berthiaume, Matthew Robison, Karie Holtermann, Colleen Durkin, Sandra Splinter Bondurant, Kathryn Richmond, Matthew Rodesch, Toivo Kallas, Edward L. Huttlin, Francesco Cerrina, Michael R. Sussman, and E. Virginia Armbrust. 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.

J Keith Moore. 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.

Jeffrey L. Moseley, Chiung-Wen Chang, and Arthur R. Grossman. 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.

Ahmed Moustafa, Andrew N. Evans, David M. Kulis, Jeremiah D. Hackett, Deana L. Erdner, Donald M. Anderson, and Debashish Bhattacharya. 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.

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

David M. Nelson, Paul Tréguer, Mark A. Brzezinski, Aude Leynaert, and Bernard Quéguiner. 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.

Nathalie Nicot, Jean-François Hausman, Lucien Hoffmann, and Danièle Evers. 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.

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

S.W. Nixon, S.L. Granger, and B.L. Nowicki. 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.

Barbara B. North and Grover C. Stephens. 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.

N. Ogawa, J. DeRisi, and P. O. Brown. 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.

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

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

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

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

Michael W. Pfaffl, Ales Tichopad, Christian Prgomet, and Tanja P. Neuvians. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper âĂŞ Excel-based tool using pair-wise correlations. *Biotechnology Letters*, 26(6):509–515, March 2004. doi: 10.1023/B:BILE.0000019559.84305.47.

Rachel S Poretsky, Shulei Sun, Xiaozhen Mou, and Mary Ann Moran. 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.

Robert Ptacnik, Angelo G Solimini, Tom Andersen, Timo Tamminen, På l Brettum, Liisa Lepistö, Eva Willén, and Seppo Rekolainen. 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.

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

Aleksandar Radonic, Stefanie Thulke, Ian M Mackay, Olfert Landt, Wolfgang Siegert, and Andreas Nitsche. 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.

P. Richerson, R. Armstrong, and C. R. Goldman. 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.

Sophie Richier, Sarah Fiorini, Marie-Emmanuelle Kerros, Peter von Dassow, and Jean-Pierre Gattuso. Response of the calcifying coccolithophore Emiliania huxleyi to low pH/high pCO2: from physiology to molecular level. *Marine Biology*, 158(3):551–560, November 2010. doi: 10.1007/s00227-010-1580-8.

Johanna M Rinta-Kanto, Shulei Sun, Shalabh Sharma, Ronald P Kiene, Mary Ann Moran, and Johanna M RintaâĂŘKanto. 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.

Nedeljka N. Rosic, Mathieu Pernice, Sophie Dove, Simon Dunn, and Ove Hoegh-Guldberg. 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.

Nedeljka N. Rosic, Mathieu Pernice, Mauricio Rodriguez-Lanetty, and Ove Hoegh-Guldberg. 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.

Shovonlal Roy and J. Chattopadhyay. 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.

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

Alexander I Saeed, Nirmal K Bhagabati, John C Braisted, Wei Liang, Vasily Sharov, Eleanor A Howe, Jianwei Li, Mathangi Thiagarajan, Joseph A White, and John Quackenbush. TM4 microarray software suite. *Methods in Enzymology*, 411:134–193, 2006. doi: 10.1016/S0076-6879(06)11009-5.

Mak A. Saito, Tyler J. Goepfert, and Jason T. Ritt. 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.

Rintaro Saito, Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, Samad Lotia, Alexander R Pico, Gary D Bader, and Trey Ideker. A travel guide to Cytoscape plugins. *Nature methods*, 9(11):1069–76, November 2012. doi: 10.1038/nmeth.2212.

Egil Sakshaug. 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.

Diana Sarno, Wiebe H. C. F. Kooistra, Linda K. Medlin, Isabella Percopo, and Adriana Zingone. 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.

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

J. L. Serra, M. J. Llama, and E. Cadenas. 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.

Roshan Prakash Shrestha, Benoit Tesson, Trina Norden-Krichmar, Stephen Federowicz, Mark Hildebrand, and Andrew E Allen. 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.

Magali Siaut, Marc Heijde, Manuela Mangogna, Anton Montsant, Sacha Coesel, Andrew Allen, Alessandro Manfredonia, Angela Falciatore, and Chris Bowler. 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.

Theodore J. Smayda. 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.

Theodore J Smayda. 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.

Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. 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.

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

ULRICH Sommer. 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.

A. Sournia, M.-J. Chrdtiennot-Dinet, and M. Ricard. 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.

Christopher F. Steiner, Zachary T. Long, Jennifer A. Krumins, and Peter J. Morin. 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.

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

Maren Striebel, Stephan Behl, and Herwig Stibor. 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.

Rhona K. Stuart, Chris L. Dupont, D. Aaron Johnson, Ian T. Paulsen, and Brian Palenik. 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.

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

P Tett and ED Barton. 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.

D Tilman, P B Reich, J Knops, D Wedin, T Mielke, and C Lehman. 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.

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

Benjamin A S Van Mooy, Helen F Fredricks, Byron E Pedler, Sonya T Dyhrman, David M Karl, Michal Koblízek, Michael W Lomas, Tracy J Mincer, Lisa R Moore, Thierry Moutin, Michael S Rappé, and Eric a Webb. Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. *Nature*, 458(7234): 69–72, March 2009. doi: 10.1038/nature07659.

Jo Vandesompele, Katleen De Preter, Filip Pattyn, Bruce Poppe, Nadine Van Roy, Anne De Paepe, and Frank Speleman. 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.

LeAnn P. Whitney, Jeremy J. Lins, Margaret P Hughes, Mark L Wells, P. Dreux Chappell, and Bethany D. Jenkins. 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.

Zhijin Wu, Bethany D Jenkins, Tatiana A Rynearson, Sonya T Dyhrman, Mak A Saito, Melissa Mercier, and LeAnn P Whitney. Empirical bayes analysis of sequencing-based transcriptional profiling without replicates. *BMC bioinformatics*, 11(1):564, January 2010. doi: 10.1186/1471-2105-11-564.

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

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

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

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

Adriana Zingone, Isabella Percopo, Pat A. Sims, and Diana Sarno. 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.

M. V. Zubkov, B. M. Fuchs, G. A. Tarran, P. H. Burkill, and R. Amann. 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.