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 (∼12 million), P-limited (∼13 million), Fe-limited (∼23 million), and co-limited (∼26 million). Tags were mapped to gene models (predicted protein coding regions) with a pipeline designed by Genesifter Inc., requiring 100% identity and covering 9759 genes. Tag counts within a gene were pooled and normalized to the size of the library, with resulting data expressed in tags per million (tpm). Genes with normalized tag counts less than 2.5 tpm for each of the four treatments were excluded (Figure S1), leaving 7380 genes in the analysis. The data discussed in this publication have been deposited in NCBIâĂŹs 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 em-

ployed 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 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 S2). 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 S2). 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 1; Data Sheet 2), which stands in contrast to other clusters that had definite treatment driven expression patterns (Figure S2). Despite the relatively small deviations from the mean value, genes in

Clusters 9 and 14 displayed both clear patterns of regulation, as demonstrated by the average change in tag count relative to the mean (Figure 2) 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 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 1; Data Sheet 3).

The bin with the 1.10 cutoff had two genes (NCBI: 7446346 and 7452192), which are both hypothetical proteins (Figure 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 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). 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 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 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 stud-

ies 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 Acknowledgements

The authors acknowledge Genesifter Inc. for their assistance with pipeline development, and Z. Wu for helpful discussions. This research was funded by the National Science Foundation grant #OCE-0723667 (to STD, MAS, BDJ, and TAR). HA is funded under a National Defense Science and Engineering Graduate (NDSEG) Fellowship.

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

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 (Adib et al., 2004; 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 extremely 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., 2010a, 2010b), 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; DâĂŹhaeseleer, 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.

3.3 Materials and Methods

3.3.1 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 S1), leaving 7380 genes in the analysis. The data discussed in this publication have been deposited in NCBIâĂŹs 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).

3.3.2 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, 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).

3.4 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, 2008b; 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., 2010a, 2010b; Leggat et al., 2011), *Prorocentrum minimum* (Guo and Ki, 2011), *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, Îś-, Îš-, Îş-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 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 S2). 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 S2). 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 1; Data Sheet 2), which stands in contrast to other clusters that had definite treatment driven expression patterns (Figure S2). Despite the relatively small deviations from the mean value, genes in Clusters 9 and 14 displayed both clear patterns of regulation, as demonstrated by the average change in tag count relative to the mean (Figure 2) and the presence of âĂIJoutlierâĂİ genes with differential expression such as NCBI: 7451632, which was downregulated in the Co-limited treatment for Cluster 14 (Figure 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 1; Data Sheet 3).

The bin with the 1.10 cutoff had two genes (NCBI: 7446346 and 7452192), which are both hypothetical proteins (Figure 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 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). If all genes in a group were perfectly con-

stitutively 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 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 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).

3.5 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 et al., 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 (DâĂŹhaeseleer, 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 anophag-*

efferens, *Phaeodactylum tricornutum*, and *Prorocentrum minimum* (Siaut et al., 2007; McGinn and Morel, 2008a; Guo and Ki, 2011; 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).

3.6 Acknowledgements

The authors acknowledge Genesifter Inc. for their assistance with pipeline development, and Z. Wu for helpful discussions. This research was funded by the National Science Foundation grant #OCE-0723667 (to STD, MAS, BDJ, and TAR). HA is

funded under a National Defense Science and Engineering Graduate (NDSEG) Fellowship.

Bibliography

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. ISSN 1091-6490. doi: 10.1073/pnas.0711370105. URL http://www.pnas.org/content/105/30/10438 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2492447&tool=pmcentrez&re

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. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/25.17.3389. URL http://nar.oxfordjournals.org/content/25/17/3389.

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. ISSN 0023-6837. URL http://dx.doi.org/10.1038/labinvest.3700303.

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. ISSN 1095-9203. doi: 10.1126/science.1101156. URL http://www.sciencemag.org/content/306/5693/79.abstract http://www.ncbi.nlm.nih.gov/pubmed/15459382.

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. URL http://www.plantphysiol.org/content/136/3/3649.abstract.

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. ISSN 1434-4610. doi: 10.1016/j.protis.2011.07.008. URL http://www.sciencedirect.com/science/article/pii/S1434461011000630.

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. ISSN 1529-8817. doi: 10.1111/j.1529-8817.2008.00571.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1529-8817.2008.00571.x/abstract.

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. ISSN 1529-8817. doi: 10.1111/j.1529-8817.2010.00809.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1529-8817.2010.00809.x/abstract.

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. ISSN 0960-9822. doi: 10.1016/j.cub.2005.09.011. URL http://www.sciencedirect.com/science/article/pii/S096098220501033X.

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. ISSN 1471-2164. doi: 10.1186/1471-2164-12-159. URL http://www.biomedcentral.com/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. ISSN 0736-6205. URL http://www.ncbi.nlm.nih.gov/pubmed/14579739.

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. URL http://www.plantphysiol.org/content/139/1/5.abstract.

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. ISSN 1434-4610. 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. URL http://ec.asm.org/content/5/7/1157.abstract.

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. ISSN 1932-6203. doi: 10.1371/journal.pone.0000898. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1976390&tool=pmcentrez&re
UR - http://dx.plos.org/10.1371/journal.pone.0000898,http://dx.plos.org/10.1371/jou

Patrik D'haeseleer. How does gene expression clustering work? *Nature biotechnology*, 23(12):1499–501, December 2005. ISSN 1087-0156. doi: 10.1038/nbt1205-1499. URL http://dx.doi.org/10.1038/nbt1205-1499 http://www.ncbi.nlm.nih.gov/pubmed/16333293.

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. ISSN 0003-2697. doi: 10.1016/j.ab.2005.05.022. URL http://www.ncbi.nlm.nih.gov/pubmed/16054107.

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. ISSN 1932-6203. doi: 10.1371/journal.pone.0033768. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3315573&tool=pmcentrez&re

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. URL http://www.pnas.org/content/105/17/6368.abstract.

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. ISSN 0959-440X. 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. ISSN 1088-9051. doi: 10.1101/gr.397002.

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. ISSN 1618-0429. doi: 10.1016/j.ejop.2011.11.001. URL http://dx.doi.org/10.1016/j.ejop.2011.11.001.

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. ISSN 0035-9254. doi: 10.2307/2346830. URL http://www.jstor.org/stable/2346830.

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. ISSN 1529-8817. doi: 10.1111/j.1529-8817.2007.00359.x. URL http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1529-8817.2007.00359.

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. ISSN 1559-0305. doi: 10.1007/s12033-012-9539-2. URL http://www.ncbi.nlm.nih.gov/pubmed/22528553.

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. ISSN 1535-3893. doi: 10.1021/pr900475f. URL http://dx.doi.org/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. ISSN 0022-0981. doi: 10.1016/j.jembe.2011.07.032.

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. ISSN 0022-0981. doi: 10.1016/j.jembe.2009.01.010. URL http://www.sciencedirect.com/science/article/pii/S0022098109000380.

Kenneth J. Livak and Thomas D. Schmittgen. Analysis of relative gene expression using real-time quantitative PCR and the 2âĹŠÎĂCT method. *Methods*, 25(4):

402–408, December 2001. ISSN 1046-2023. doi: 10.1006/meth.2001.1262. URL http://www.sciencedirect.com/science/article/pii/S1046202301912629.

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. ISSN 00243590. doi: 10.4319/lo.2006.51.4.1729. URL http://www.www.aslo.org/lo/toc/vol_51/issue_4/1729.html.

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. ISSN 1469-7793. doi: 10.1113/jphysiol.2004.074047. URL http://onlinelibrary.wiley.com/doi/10.1113/jphysiol.2004.074047/abstract.

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. URL http://mbe.oxfordjournals.org/content/27/10/2268.abstract.

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. ISSN 1399-3054. doi: 10.1111/j.1399-3054.2007.01039.x. URL http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1399-3054.2007.01039.

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. URL http://www.plantphysiol.org/content/146/1/300.abstract.

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. ISSN 1091-6490. doi: 10.1073/pnas.0707946105. URL http://www.pnas.org/content/105/5/1579.abstract http://www.pnas.org/content/105/5/1579.long http://www.pubmedcentral.nih.gov/articl

J Keith Moore. Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model. *Global Biogeochemical Cycles*, 18(4): 1–21, 2004. ISSN 0886-6236. doi: 10.1029/2004GB002220. URL http://www.agu.org/pubs/crossref/2004/2004GB002220.shtml.

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. URL http://ec.asm.org/content/5/1/26.abstract.

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. ISSN 1932-6203. doi: 10.1371/journal.pone.0009688. URL UR - http://dx.doi.org/10.1371/journal.pone.0009688,http://dx.doi.org/10.1371/journal.po http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2837391&tool=pmcentrez&re

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. URL http://jxb.oxfordjournals.org/content/56/421/2907.abstract.

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. URL http://www.int-res.com/abstracts/ame/v51/n2/p183-193/.

Michael W. Pfaffl. A new mathematical model for relative quantification in real-time RTâĂŞPCR. *Nucleic Acids Research*, 29(9):e45, May 2001. ISSN 0305-1048.

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. ISSN 0141-5492. doi: 10.1023/B:BILE.0000019559.84305.47. URL http://www.springerlink.com.libproxy.mit.edu/content/g6732750704887n8/.

John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, June 2001. ISSN 1471-0056. doi: 10.1038/35076576. URL http://dx.doi.org/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. ISSN 0006291X. doi: 10.1016/j.bbrc.2003.11.177. URL http://linkinghub.elsevier.com/retrieve/pii/S0006291X03025646.

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. ISSN 0025-3162, 1432-1793. doi: 10.1007/s00227-010-1580-8. URL http://www.springerlink.com/content/l183m3673k745632/.

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. ISSN 1355-8145, 1466-1268. doi: 10.1007/s12192-010-0222-x. URL http://www.springerlink.com/content/x00x7r24l8396386/.

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. ISSN 1436-2228, 1436-2236. doi: 10.1007/s10126-010-9308-9. URL http://www.springerlink.com/content/g8171496u72523x3/.

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. ISSN 0736-6205. URL http://www.ncbi.nlm.nih.gov/pubmed/12613259.

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. ISSN 0076-6879. doi: 10.1016/S0076-6879(06)11009-5. URL http://www.ncbi.nlm.nih.gov/pubmed/16939790.

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. ISSN 00243590. doi: 10.4319/lo.2008.53.1.0276. URL http://ww.w.aslo.org/lo/toc/vol_53/issue_1/0276.html.

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. ISSN 1754-2189. doi: 10.1038/nprot.2008.73. URL http://dx.doi.org/10.1038/nprot.2008.73.

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. ISSN 03781119. doi: 10.1016/j.gene.2007.05.022.

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. URL http://aem.asm.org/content/75/15/5047.abstract.

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. ISSN 1061-4036. doi: 10.1038/10343. URL http://dx.doi.org/10.1038/10343.

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. ISSN 1465-6914. doi: 10.1186/gb-2002-3-7-research0034. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=126239&tool=pmcentrez&ren

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. ISSN 1664-302X. doi: 10.3389/fmicb.2011.00234. URL http://www.frontiersin.org/Aquatic_Microbiology/10.3389/fmicb.2011.00234/pdf/abstra http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3223615&tool=pmcentrez&re

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. ISSN 1471-2105. doi: 10.1186/1471-2105-11-564. URL http://www.biomedcentral.com/1471-2105/11/564 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3098101&tool=pmcentrez&re

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. ISSN 1932-6203. doi: 10.1371/journal.pone.0028949. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3237563&tool=pmcentrez&re

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. URL http://bioinformatics.oxfordjournals.org/content/17/4/309.abstract.

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. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2008.03758.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-313X.2008.03758.x/abstract.