UNIVERSITY OF CALIFORNIA, SAN DIEGO

Characterization of Small RNA Genes in the Marine Organisms *Ciona intestinalis* and *Thalassiosira pseudonana*

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor or Philosophy

in

Marine Biology

by

Trina M. Norden-Krichmar

Committee in charge:

Theresa Gaasterland, Co-Chair
Mark Hildebrand, Co-Chair
Andrew E. Allen
Eric E. Allen
Philip E. Bourne
Ronald S. Burton
Amy E. Pasquinelli
Victor D. Vacquier

2009

The dissertation of Trina M. Norden-Krichmar is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California, San Diego

2009

Table of Contents

List of Figures

List of Tables

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in BMC Genomics 2007. Norden-Krichmar, Trina M.; Holtz, Janette; Pasquinelli, Amy E.; and Gaasterland, Terry. "Computational prediction and experimental validation of Ciona intestinalis microRNA genes", BMC Genomics, 8:445, 2007. The dissertation author was the primary researcher and author of this paper.

Chapter 2, in full, has been submitted for publication. Norden-Krichmar, Trina M.; Allen, Andrew E.; Gaasterland, Terry; and Hildebrand, Mark. "Characterization of the small RNA transcriptome of the diatom, Thalassiosira pseudonana". The dissertation author was the primary researcher and author of this paper.

I would like to thank my committee members for their support and guidance throughout my academic career. A National Science Foundation Graduate Research Fellowship provided funding of my tuition and stipend for three years. Special thanks to my family and to my husband, Jeff, for their patience, advice and encouragement.

Vita

Bachelor of Science, Biochemistry
University of Maryland, College Park

Master of Science, Computer Science
George Washington University

Doctor of Philosophy, Marine Biology
Scripps Institution of Oceanography
University of California, San Diego

ABSTRACT OF THE DISSERTATION

Characterization of Small RNA Genes in the Marine Organisms *Ciona intestinalis* and *Thalassiosira pseudonana*

by

Trina M. Norden-Krichmar

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2009

Theresa Gaasterland, Co-Chair
Mark Hildebrand, Co-Chair

This dissertation describes an integration of computational and biological techniques to characterize small RNA genes in two key marine organisms, a sea squirt and a diatom. Eukaryotic small RNA genes, which are typically 19-31 nucleotides in length, regulate gene expression within cells in a temporal and state-dependent manner, controlling essential processes such as embryological development, cell differentiation, responses to environmental stress, and cellular defense mechanisms.

In the first chapter, computational methods were developed to predict evolutionarily conserved members of one class of small RNAs, known as microRNAs, in the sea squirt, *Ciona intestinalis*. The sea squirt is an important model organism due to its phylogenetic placement at the emergence of vertebrates. The microRNA prediction algorithm was designed to quickly screen the genome for the presence of conserved microRNAs, producing the first validated collection of microRNAs in the sea squirt. Additionally, a target prediction algorithm was implemented which identified potential target genes.

The second chapter built upon these techniques and expanded the search for other classes of endogenous small RNAs in the diatom, *Thalassiosira pseudonana*. Diatoms are unicellular phytoplankton, chosen as the model organism because of their global importance in processes such as carbon fixation and nutrient cycling. A small RNA cDNA library was constructed for exponentially growing *T. pseudonana*, and then pyrosequenced. Computational analysis of approximately 300,000 sequences yielded strong evidence of small RNA genes in *T. pseudonana*, including microRNAs, repeat-associated short interfering RNAs, and endogenous short interfering RNAs.

The third chapter focused on differential small RNA gene expression in *Thalassiosira pseudonana,* under various nutrient stress conditions. Small RNA cDNA libraries were constructed under conditions of exponential growth, silicon starvation, nitrogen starvation, and iron starvation. The libraries were then processed with high throughput SOLiD sequencing. A methodology was developed to computationally analyze the 150 million sequences, generating a profile of differential

small RNA expression between the conditions, as well as a core subset of small RNAs expressed across all conditions.

The novel computational techniques implemented in this dissertation can be applied to other organisms and aid in elucidating the roles of small RNAs in gene regulation.

Chapter 1. Computational prediction and experimental validation of *Ciona intestinalis* microRNA genes

BMC Genomics

BioMed Central

Research article

Open Access

Computational prediction and experimental validation of *Ciona intestinalis* microRNA genes

Trina M Norden-Krichmar*[1], Janette Holtz[2], Amy E Pasquinelli[2] and Terry Gaasterland[1]

Address: [1]Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, MSC 0202, La Jolla, CA 92093 USA and [2]Department of Biology, University of California, San Diego, La Jolla, CA 92093 USA

Email: Trina M Norden-Krichmar* - tnordenk@ucsd.edu; Janette Holtz - jholtz@ucsd.edu; Amy E Pasquinelli - apasquin@biomail.ucsd.edu; Terry Gaasterland - tgaasterland@ucsd.edu

* Corresponding author

Published: 29 November 2007

BMC Genomics 2007, **8**:445 doi:10.1186/1471-2164-8-445

Received: 30 July 2007
Accepted: 29 November 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/445

Abstract

Background: This study reports the first collection of validated microRNA genes in the sea squirt, *Ciona intestinalis*. MicroRNAs are processed from hairpin precursors to ~22 nucleotide RNAs that base pair to target mRNAs and inhibit expression. As a member of the subphylum Urochordata (Tunicata) whose larval form has a notochord, the sea squirt is situated at the emergence of vertebrates, and therefore may provide information about the evolution of molecular regulators of early development.

Results: In this study, computational methods were used to predict 14 microRNA gene families in *Ciona intestinalis*. The microRNA prediction algorithm utilizes configurable microRNA sequence conservation and stem-loop specificity parameters, grouping by miRNA family, and phylogenetic conservation to the related species, *Ciona savignyi*. The expression for 8, out of 9 attempted, of the putative microRNAs in the adult tissue of *Ciona intestinalis* was validated by Northern blot analyses. Additionally, a target prediction algorithm was implemented, which identified a high confidence list of 240 potential target genes. Over half of the predicted targets can be grouped into the gene ontology categories of metabolism, transport, regulation of transcription, and cell signaling.

Conclusion: The computational techniques implemented in this study can be applied to other organisms and serve to increase the understanding of the origins of non-coding RNAs, embryological and cellular developmental pathways, and the mechanisms for microRNA-controlled gene regulatory networks.

Background

Small non-coding RNA genes have emerged in recent years as regulators of transcription and translation in a time and cell-state dependent manner. In particular, mature microRNA (miRNA) molecules of approximately 22 nucleotides in length have been shown to inhibit gene expression by base pairing to target mRNA. MicroRNAs have been implicated in gene regulation during embryological development, disease, and cell differentiation [1]. Specifically, recent studies have linked miRNA to many diverse biological processes such as clearing of maternal mRNA during zebrafish embryogenesis [2], regulation of

the life span in *Caenorhabditis elegans* [3], and involvement in tumorigenesis in human cancer [4].

Although miRNAs were first discovered in the early 1990's, it has only been in the past five years that there has been an explosion in the literature on their formation and gene regulatory mechanisms, and on their abundance in the genomes of different organisms. MicroRNAs are now projected to occur at a frequency of approximately 0.5–1.5% of the total genes in the genome [5]. For the human genome, which contains approximately 30,000 protein coding and large non-coding RNA genes, there are currently 470 miRNAs reported in the miRNA database [6]. It is estimated that possibly 20 to 30% of human genes are targets of miRNA [7]. Since many miRNAs are evolutionarily conserved [8], discoveries in model organisms may contribute to understanding the role of specific miRNAs in gene regulatory networks for other organisms.

Recent studies have revealed many of the biological characteristics of miRNA formation and mechanism, which can be used for computational prediction of new miRNAs and their targets [9]. The microRNA primary transcripts (pri-miRNAs) contain one or multiple hairpin secondary structures. The ribonuclease Drosha complexes with a double-stranded RNA binding protein DGCR8, to process the pri-miRNA into a 70–100 nucleotide precursor miRNA (pre-miRNA). The pre-miRNA is exported from the nucleus to the cytoplasm via an Exportin-5 transport mechanism. Once in the cytoplasm, the RNase III enzyme Dicer cleaves the loop of the hairpin. The strands are separated, allowing the single-stranded mature microRNA whose 5'end has lower stability base pairings, to associate with the RNA-induced silencing complex (RISC), while the other strand is degraded. In RISC, the miRNA acts as a guide to recognize mRNA targets via base pairing. The mechanism of regulation is not entirely known, but the current model suggests that if the base pairing of the mature miRNA in the RISC complex shows sufficient complementarity to the mRNA target, then there will be cleavage and degradation of the mRNA target. With less complementarity, the miRNA/RISC represses translation of the mRNA target [10].

For animal miRNAs that partially base pair to target sequences, several general characteristics further constrain miRNA target prediction [10]. First, the miRNA binds to the 3' untranslated region (UTR) in most of the established mRNA targets in metazoans. Second, the strongest base-pairing between the miRNA and mRNA seems to occur at the 5' end of the miRNA, especially in the first 8 or 9 nucleotides. Third, sequence conservation in the UTRs of orthologous genes is a strong indicator of functional binding sites, and reduces false positive targets. Fourth, miRNA target regulation has been shown to occur when the same miRNA can recognize multiple sites in the 3'UTR of an mRNA target. Recently, multiple miRNAs in the same family, i.e., mir-48, mir-84, and mir-241 in *C.elegans*, were found to act in concert to control developmental genes [11]. Accurate target prediction must consider all of these properties of miRNA. Target prediction is difficult because of the very short length of the mature miRNA and the imperfect binding to the target. A variety of methods have been reported in the literature for the computational prediction of microRNA sequences and their targets [12].

This study involves the combination of computational and biological techniques to identify and validate microRNAs in the sea squirt, *Ciona intestinalis*. We performed these analyses in *C. intestinalis* for several reasons. First, as a member of the subphylum Urochordata (also known as Tunicata) whose larval form has a notochord, the sea squirt is situated phylogenetically at the emergence of the vertebrates [13], and therefore can provide information about the evolution of molecular regulators of early development. Second, the *C. intestinalis* genome has been completely sequenced [14], and another related urochordate genome, *Ciona savignyi*, has also been sequenced. Third, it has a compact genome of approximately 160 megabases (Mb), containing few repeated sequences. This small size and the consequent optimization of intergenic regions makes the genome tractable to study computationally. Fourth, *Ciona* is relatively easy to obtain and culture, and embryos can be manipulated experimentally with genetic techniques [15-17].

While another group has computationally predicted microRNAs for *Ciona intestinalis* [18], there are currently no experimentally validated microRNAs for *Ciona intestinalis* in the Sanger microRNA database, miRBase [6] (formerly known as "The microRNA Registry" [19]). To qualify for inclusion into the miRBase database, the expression of the microRNA must be detected using either the Northern blotting method, or from a library of cDNA made from size-fractionated RNA [20].

In the present study, we implemented microRNA gene and target prediction techniques to examine the genomes of the marine organisms *Ciona intestinalis* and *Ciona savignyi*. In our parameterized approach, we capitalized on the phylogenetic conservation of known miRNA gene families to screen quickly the genome of an organism for the possible presence of homologous miRNA genes. This technique greatly reduces the number of putative miRNAs that must be validated experimentally. Using the current approach, we predicted 14 miRNA sequences for *Ciona intestinalis*. We validated 8, out of 9 attempted, of the predicted miRNA sequences by Northern blot experiments. Following the prediction of the miRNA sequences, we computationally predicted and tabulated the most proba-

ble mRNA targets for these miRNAs. A subset of these targets exhibit conservation across phylogeny [21], and therefore may point to microRNA-controlled regulatory functions across evolutionary pathways.

Results and Discussion

Collection of known miRNA statistics

To tune the miRNA prediction algorithm parameters, miRNA and precursor sequences were analyzed to determine physical and sequence conservation characteristics, using the following three pairs of organisms: *Caenorhabditis elegans* vs. *Caenorhabditis briggsae*, *Drosophila melanogaster* vs. *Drosophila pseudoobscura*, and *Homo sapiens* vs. *Pan troglodytes*. Additional File 1 summarizes the statistics gathered for these three pairs of closely related organisms. For each species examined, the average percent identity of the two hairpin stem sequences was 78% or better, with a minimum of 65% identity. The average percent identity of the mature miRNA sequence between closely related species was 98%. The average hairpin loop length was 20 nucleotides, and ranged from 7 to 58 nucleotides. The percent GC content in the mature and precursor sequences was approximately 45%. These characteristics directed our microRNA prediction algorithm.

Computational prediction of the Ciona intestinalis *microRNAs*

Using the parameters from the statistical analysis (Figure 1), we predicted 14 miRNAs for *Ciona intestinalis*. Mapping the locations of these putative *Ciona* miRNAs to the *Ciona* genome yielded additional information for comparison to the available known miRNAs. Additional File 2 lists the putative *C. intestinalis* miRNAs, their mature miRNA sequence, oligonucleotide probe sequence, scaffold number, coordinates, strand, position, and nearest neighboring gene. Figures 2 and 3 show the *mfold* program's text and graphical output for the putative miR-72 and let-7 for *Ciona* and *C. elegans*. The *mfold* output demonstrates the conservation of the precursor hairpin structure with the related species, *Ciona savignyi*. Using the ClustalX program, the predicted mature miRNA sequences for *C. intestinalis* and *C. savignyi* were aligned with known mature miRNA sequences from other organisms. Figure 4 shows the mature miRNA multiple sequence alignments which in turn demonstrate the membership of the predicted *Ciona* miRNAs in existing miRNA gene families.

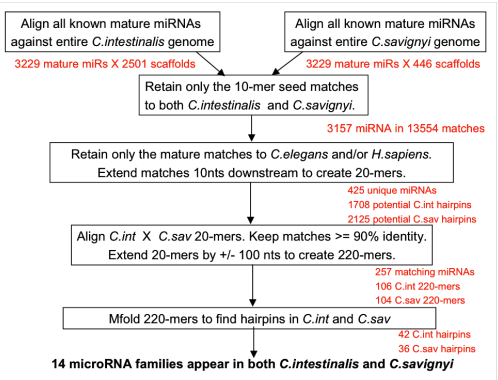

Figure 1 Flow diagram for microRNA prediction algorithm.

The number of miRNAs from this computational prediction is likely to be lower than the actual number of miRNA genes in C. intestinalis. Since miRNAs are believed to occur at a frequency of approximately 0.5–1.5% of the total genes in the genome [5], Ciona's 15,000 genes should have generated between 75 – 225 miRNAs. Of the 19 miRNA families that appear in both C. elegans and H. sapiens [22], our algorithm only predicted 5 of these families in C. intestinalis (Figure 5). We expected to find all of the miRNA families that were present in both C. elegans and H. sapiens. Therefore, this may be an indication that our algorithm is underpredicting the miRNAs. The parameters of our prediction algorithm could be relaxed in future studies, in order to produce additional miRNA sequence candidates.

Experimental validation of predicted miRNA sequences

We experimentally validated 8, out of 9 attempted, of the putative *C. intestinalis* miRNA sequences using Northern

Figure 2 RNA folding structure as calculated by the program *mfold* for **RNA folding structure as calculated by the program *mfold* for miR-72**. Text output of putative miR-72 for *Ciona intestinalis*, *Ciona savignyi* and *C.elegans*.

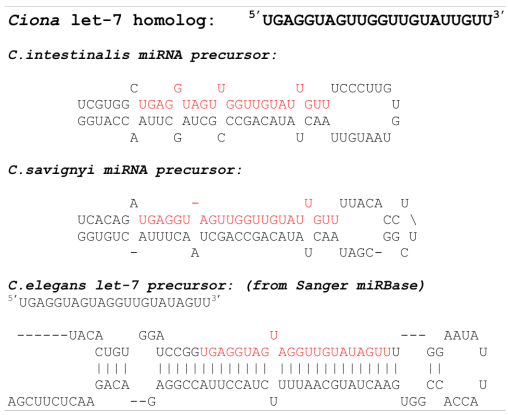

Figure 3
RNA folding structure as calculated by the program *mfold* for let-7. Text output of putative let-7 for *Ciona intestinalis*, *Ciona savignyi* and *C.elegans*.

blot analysis (Figure 6). To validate the strand polarity of the predicted mature miRNAs, we performed the Northern blot analysis with sense and anti-sense probes for the top and bottom strands of the let-7 and miR-72 *C. intestinalis* homolog predictions. In both the let-7 and miR-72 homologs, no hybridization to the anti-sense strand occurred. Therefore, the strand polarity of these two predicted miRNA sequences was confirmed.

As a control, equal quantities of total RNA from *C. elegans* and *C. intestinalis* were run on the same Northern blot [results not shown]. When hybridizing against the probes for the miR-72 and let-7 *Ciona* homologs, the *C. intestinalis* lanes showed strong positive signals. The *C. elegans* lanes showed a weak response to the *Ciona* miR-72 homolog probe, and no response to the *Ciona* let-7 homolog probe. These results were as expected, since Northerns are extremely sensitive to the probe sequence. Since there were known mismatches to the probe at the ends of the *C. elegans* sequence, it did not hybridize as well to the miR-72 probe as *C. intestinalis*. Similarly, since the *Ciona* let-7 sequence contained 2 alignment gaps in the middle section the *C. elegans* sequence, we did not expect to see hybridization.

Hybridizations with 7 of the remaining predicted *C. intestinalis* miRNAs were attempted, yielding 6 additional positive Northern blot results. For these putative miRNA sequences, probes were made against the predicted mature miRNA sequences only, not against the bottom sequences of the hairpin structure.

MicroRNA expression data from other organisms confirmed the results that we obtained in the Northern analyses of our predicted *C. intestinalis* miRNA sequences [22-26]. For example, let-7 is highly expressed in adult tissue of *C. elegans*, *D. melanogaster*, and vertebrates [23]. Therefore, we were not surprised to find the let-7 homolog expressed in the adult tissue of *C. intestinalis*. Conversely, miR-47 has been experimentally validated in *C. elegans*, but has not yet been predicted or validated in vertebrates [22]. Although computationally there appears to be a miR-47 homolog in *C. intestinalis*, we were not able to detect its expression by Northern analysis. Since *C. intestinalis* is phylogenetically located at the emergence of vertebrates, the expression of miR-47 in *C. intestinalis* may be more like vertebrates in this case.

We chose not to validate our predictions of miR-302a, miR-33, miR-452*, miR-520d, and miR-7, based upon reports of the expression of these homologs in the literature. The miR-302a sequence was validated by others in mouse and human by cloning from embryonic stem cells. Likewise, the miR-7 sequence was found in early embryonic development of the Drosophila (embryo to 6 hours) [23]. Our validation was done with adult tissue. The miR-452* and miR-520d sequences were detected by the array-cloning technique [27]. The miR-33 sequence failed to validate by Northern in any of the tissues tested in a previous study, which included HeLa cells, mouse kidney, adult fish, frog ovary, and S2 cells [23]. Therefore, since it was unlikely to detect these homologs with Northern analyses in adult tissue, we have not attempted to validate them at this time.

Computational prediction of the Ciona intestinalis *mRNA targets*

Application of the target prediction algorithm allowed stepwise refinement of the 14,866 potential mRNA targets to a high confidence list for validation. The algorithm for mRNA target prediction was based upon the observed biological mechanism of miRNA in the regulation of gene expression. That is, the target prediction algorithm addresses the following miRNA properties: reverse complementary partial binding of the miRNA to the target; the most critical binding at the 5' end of the miRNA strand; and the binding to the 3'UTR of the target. Figure 7 contains a summary of the results of the computational target prediction. The 14,866 mRNA sequences were input to the prediction pipeline as potential targets for the *Ciona* miRNA. The Smith-Waterman alignment algorithm, highly sensitive to small sequences, reduced the number of potential targets to 572 unique mRNAs. These sequences were filtered to retain only reverse complementary matches in which the miRNA 5'end had the strongest affinity for the target.

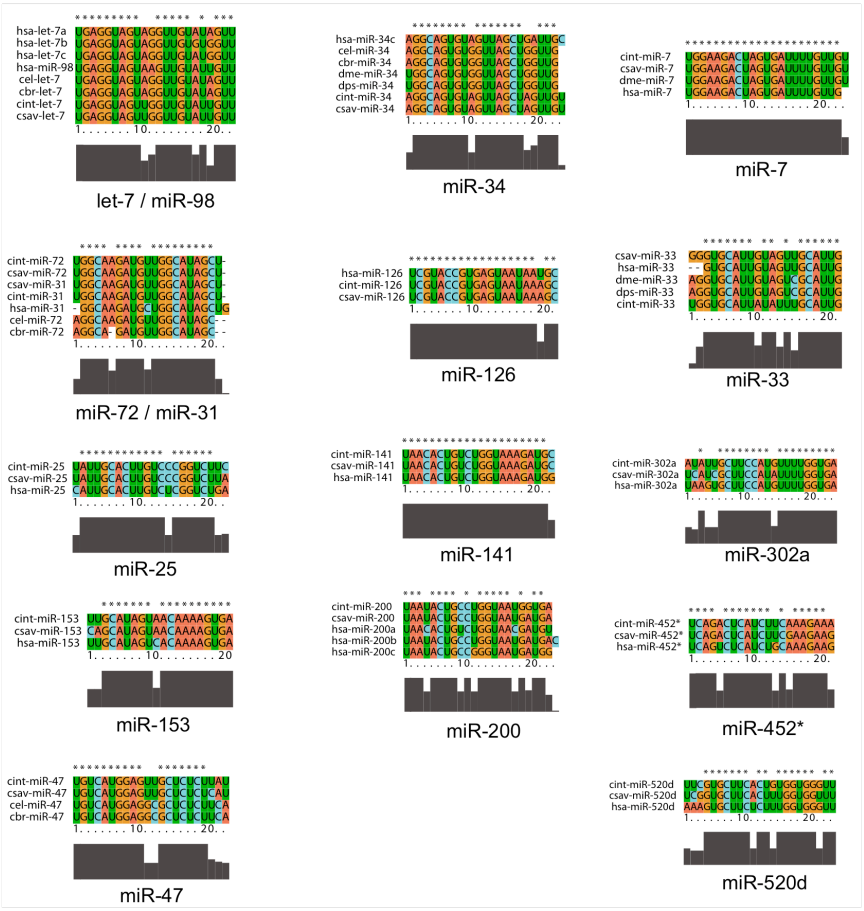

Figure 4
ClustalX alignments of the miRNA predictions. The first two columns contain the alignments for the miRNAs that we attempted to validate. The third column contains the alignments for the five miRNAs that we did not attempt to validate.

If 20 to 30% of the genes for an organism are regulated by miRNA, then *Ciona intestinalis* should have approximately 3000 to 4500 mRNA targets based upon its 15,000 gene size. However, since our algorithm underpredicted the number of *Ciona* miRNAs by approximately a factor of 10, this would explain a corresponding reduction of predicted targets. Therefore, 240 targets for 14 gene families is a reasonable number of possible targets. Preliminary experimentation with the program parameters demonstrated that varying the seed alignment lengths and sequence identity percentages reflected a direct correlation to the number of predicted targets. Configuring these parameters may then be used to adjust the system to produce more or less targets. Through computational analysis, we assigned functional descriptions to the target genes which had miRNA matches to the 3'UTR of the mRNAs. The target genes were also grouped according to their Gene Ontology (GO) terms (Figure 8). The majority of target gene functions are involved in metabolism, membrane transport, and cell signaling.

To enforce the relationship between the miRNA binding location and its potential to elicit a regulatory effect in the target gene, we computationally calculated and assigned

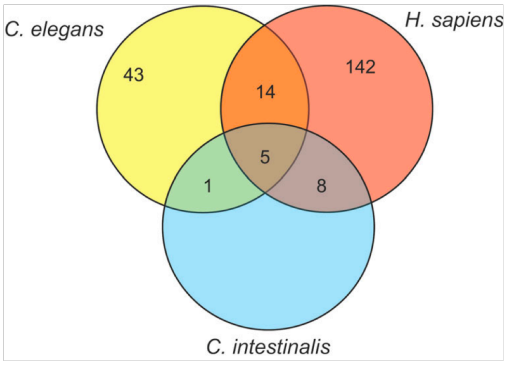

Figure 5 Venn diagram summarizing the distribution of the predicted *C. intestinalis* miRNAs into the *C. elegans* vs. *H. sapiens* families. Family counts for *C. elegans*, and for the intersection of C. elegans with *H. sapiens* were based on conserved 6-mer seeds [22]. The family count for *H. sapiens* was extracted from the miFam.dat microRNA family data file available from the miRNA database [6].

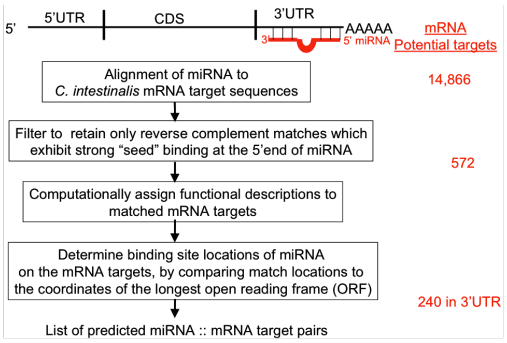

Figure 7 High level flow diagram of mRNA target prediction algorithm.

binding locations to each target. The results were tabulated and sorted according to the mRNA target match locations occurring in the 3' untranslated region (3'UTR), 5' untranslated region (5'UTR), and protein coding section (CDS). This procedure divided the 572 target matches of the *Ciona* mRNA sequences into the following subsets: 240 matches to the 3'UTR, 262 matches to the 5'UTR, and 70 matches to the CDS. As most established miRNA targets exhibit binding in the 3'UTR [28], this region was examined more closely. Further manual curation yielded three instances of multiple hits to the 3'UTR of the same target by different miRNAs. Additional File 3 contains a list of the predicted targets with single and multiple hits to the 3'UTR region. The predicted targets which contained multiple hits would be interesting to validate because there is evidence in other organisms that multiple miRNA binding sites in the 3'UTR contribute to more potent regulation [11,29].

miRNA homolog	Northern	5S rRNA	Result
let-7	(AS)		positive
miR-72	(AS)		positive
miR-34			positive
miR-126			positive
miR-25			positive
miR-47			not detected
miR-141			positive
miR-153			positive
miR-200			positive

Figure 6 Northern blot validation of miRNA predictions. **PAGE Northern blot validation of miRNA predictions**. PAGE Northern blot analyses using adult *C. intestinalis* total RNA were performed to determine the indicated miRNAs. Ethidium bromide staining of the 5S rRNA is shown as a control for RNA loading and quality. Anti-sense (AS) probes were tested for let-7 and miR-72.

Finally, we attempted to use phylogenetic homology with the *C. savignyi* genome to further prune the list of targets. There are no functional gene descriptions available for *C. savignyi*, and surprisingly, the *C. intestinalis* and *C. savignyi* genomes are more divergent than we had expected. Therefore, this step removed virtually all of the potential targets, and had to be abandoned from our algorithm. Other studies have also found that approximately 50% of the targets would have been missed if conservation had been used [2].

Because we could not use phylogenetic conservation with *C. savignyi* for the final step, we assessed whether other features supported the strength of the microRNA to mRNA target pairings. That is, we evaluated seed region match quality; we excluded G:U wobble pairs; and we examined whether some amount of compensatory binding at the 3' end of the miRNA was present. Figure 9 shows a sampling of the predicted mRNA targets. The RNAhybrid program [30] was applied to these miRNA:mRNA sequence pairs, to calculate the minimal free energy of the hybridization. The values obtained from RNAhybrid were consistent with energetically favorable hybridizations.

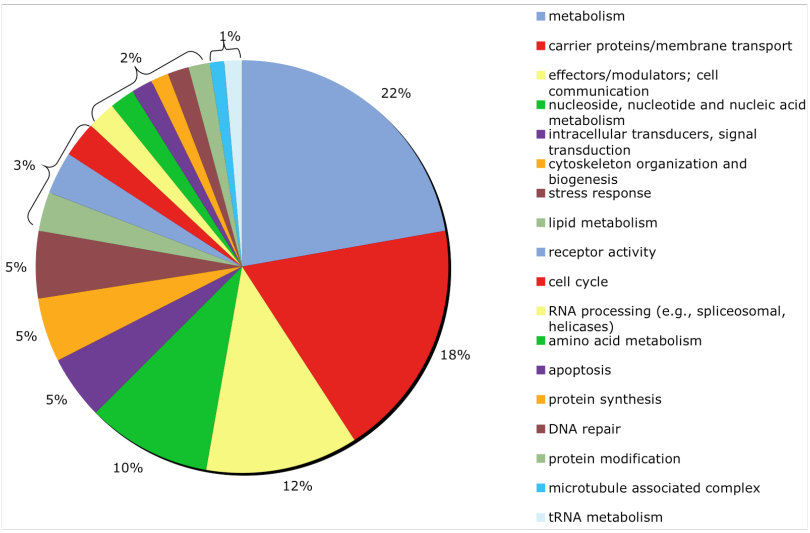

Figure 8 Gene ontology (GO) terms grouping of the mRNA targets.
Gene ontology (GO) terms grouping of the mRNA targets.

Notably, for approximately 25% of the predicted targets, these functional gene orthologs were also predicted for human/mouse/rat using the PicTar target prediction program [21].

Conclusion

In this study, we used computational methods to predict 14 miRNA sequences in the *Ciona intestinalis* genome. Of the 9 miRNA sequences tested experimentally using Northern blot analyses, we successfully validated the presence of 8 of our predicted *Ciona* miRNA homologs in the RNA of the adult tissue of *Ciona intestinalis*. Currently, no experimentally validated miRNAs are reported in the Sanger miRBase for *Ciona intestinalis*. The addition of our eight validated sequences to the database strengthens the evidence of phylogenetic miRNA sequence conservation. Additionally, a computational algorithm for mRNA target prediction was developed in this study, which takes advantage of the known biological properties of miRNA function and their target binding behavior.

Although previous studies have used phylogenetic homology for the prediction of miRNAs, our algorithm involved several novel concepts. First, since *Ciona* is located in an area of the phylogenetic tree for which miRNAs have not yet been validated, we chose to search for potential miRNA sequences across widely divergent phyla and subphyla. In other homology-driven approaches, the organism under study could be examined for known miRNAs from a member of the same Class or Order [31-33]. Second, in the initial step of our algorithm, the strongest constraint for alignment was placed upon the first 10 nucleotides, which includes the seed region of the miRNA. This differs from other methods that searched along the entire mature miRNA for perfect or near-perfect

C.intestinalis microRNA to mRNA target binding	Functional gene category
ci0100146217 5′AAC-AUAC-ACCA ACUACCUCA3′ let-7/miR-98 3′UUG UAUG UGGU--UGAUGGAGU5′ mfe: -26.8 kcal/mol	Galactosyl-transferase
ci0100143185 5′AGC-AUG--AAC AUCUUGCCA3′ miR-72/miR-31 3′UCG UAC UUG-UAGAACGGU5′ mfe: -23.3 kcal/mol	ATP binding cassette (ABC) transporter
ci0100142874 5′UCACUUU UGU UA CUAUGCA3′ miR-153 3′AGUGAAA-ACA-AU---GAUACGUU5′ mfe: -20.8 kcal/mol	Uroporphyrinogen decarboxylase
ci0100133471 5′AC AUU-CUCAGGCAGUAUUA3′ miR-200 3′GUG UAA G-GUCCGUCAUAAU5′ mfe: -25.5 kcal/mol	Kinesin motor domain

Figure 9 Sampling of predicted mRNA targets
Sampling of predicted mRNA targets. ("mfe" is the minimal free energy of the duplex, as calculated by RNAhybrid.)

matches [12,34-36]. This constraint was chosen because the seed region is important in target binding [7], and has been used in other studies as the basis for determining miRNA family membership [22]. Third, our algorithm used a second homology filter to compare the miRNA candidates between two species for which miRNAs have not yet been validated. Because the two genomes of *Ciona* are so divergent, these conserved ~22 nucleotide sequences are even more compelling as potentially valid miRNA genes.

In the course of this study, another laboratory published an initial computational prediction of miRNAs for *Ciona intestinalis* [18], based upon homology between predicted precursor hairpin structures in *Ciona intestinalis*, *Ciona savignyi*, and *Oikopleura dioica*. Their published prediction of 41 miRNA precursor sequences overlapped with ours by just one microRNA, let-7. In a subsequent study [37], a clustering approach was used with the *Ciona* non-coding RNA candidate sequences. By this method, the authors predicted 58 miRNAs, of which only let-7, miR-7, miR-124, and miR-126 coincided with known miRNAs. Our method predicted let-7, miR-7, and miR-126, as well as 11 other conserved miRNA families. While their approach exhibits great potential for predicting conserved and novel miRNA candidates, they report no experimental validation in either study, so we cannot evaluate the accuracy of their prediction method's results.

Our miRNA prediction code was written to be highly configurable, allowing for different binding strengths, as well as step-wise refinement to prune the candidates by phylogenetic conservation with homologous miRNA gene families. The entire prediction pipeline was designed so that intermediate results could be extracted at each step and examined for interesting patterns. As more characteristics of miRNA are discovered, code modules can be added and removed to reflect the changing paradigm. Establishing statistics for the characteristics of known miRNA genes in other organisms enabled the creation of a parameterized prediction approach with a high success rate. Genomic sequence analysis for conserved miRNA genes, coupled with secondary structure folding, is not necessarily sufficient to find miRNA genes in an organism unless parameters are applied to aid in the selection process. In a recent study involving the prediction of porcine miRNA genes through homology, only 7 out of 20 attempted predictions were validated by Northern blot [31]. The details of the prediction algorithm in this study were not shown, so it is not clear if parameterization was applied during the process. Our code was also designed so that parameters could be configured to tighten or relax the constraints on the matches. The current parameters may be too restrictive, if the percentages reported in the literature [5] are extendable from human miRNAs to *Ciona* miRNAs. That is, we should have obtained approximately 75 to 225 miRNAs based upon the 15,000 gene size of the *Ciona intestinalis* genome. In future studies, several constraints could be relaxed in order to potentially predict additional miRNAs in *Ciona intestinalis*. In particular, the length of alignment and percent identity between the *C. intestinalis* and *C. savignyi* mature miRNA candidates could be reduced. Further, the requirement of structural similarity, involving a reasonable conservation of bulge size and location in the hairpin stem region of the predicted miRNA precursors, could be decreased between these two species of *Ciona*.

Additionally, some miRNA genes and their targets may have been missed by the phylogenetic conservation constraint to the *C. savignyi* genome. Our inability to find orthologous targets across *C. intestinalis* and *C. savignyi* may be due to several factors. First, the *C. savignyi* genome has ambiguous and incomplete regions, which prevented sufficient alignment with the *C. intestinalis* genome. Second, the two species of *Ciona* have recently been found to be far more divergent than expected for such morphologically similar species. Using an analysis of 18S rRNA sequences, it has been estimated that "the divergence between the two species of *Ciona* is slightly greater than that between human and chick [38]." Additionally, the *Ciona* species exhibit an extremely high allelic polymorphism. The allelic polymorphism for *C. intestinalis* has been estimated to be an average of 1.2% [13]. Sequencing of two haplotypes of *C. savignyi* had an extremely high heterozygosity rate of approximately 4.6% [39]. Because of the high polymorphic rate, the *C. savignyi* genome could not be assembled using the classical whole-genome assembly method. Instead, two haplotypes were sequenced separately, and then merged to produce the reference sequence. To confirm our predictions in *C. savignyi*, it would be necessary to examine the pre-assembled alleles for the sequence. Therefore, it was necessary to reevaluate our target prediction algorithm and remove the constraint of phylogenetic conservation with *C. savignyi*, in order to produce a larger sample set for validation.

Despite these bioinformatic challenges, our miRNA prediction success rate demonstrates that this algorithm may be used for other organisms, including those with poorly or unannotated genomes. Further, our study highlights the benefit of the selection of *C. intestinalis* as a model organism for miRNA experimentation. In particular, since miRNA expression appears to be tissue-specific, many miRNAs of higher vertebrates might not be detected due to the specific tissue or organ that is sampled. In *Ciona*, the entire adult organism, including all organs, can be homogenized and prepared for RNA extraction. Therefore, all tissue types of *Ciona* are represented in the

homogenate that is used for miRNA detection, thereby increasing the probability of successful validation.

A disadvantage to our miRNA prediction method is that it will not find novel miRNAs in a genome. However, the information of the existence of known ones in different species may be used to refine the method of finding novel ones. In the genome of an organism, one can find many hairpin loops, and short conserved sequences as potential miRNA candidates. However, not all of these candidates will produce a mature miRNA. Therefore, as more miRNAs are experimentally confirmed, refinements can be applied to the prediction algorithms.

Our computational predictions could also be used to construct a gene regulatory network model, using hierarchical clustering of the miRNA and mRNA target binding strengths. Initial experimental gene regulatory networks are being constructed at several laboratories for *Ciona intestinalis* [40,41]. However, there are currently no published miRNA-based regulatory networks for *Ciona intestinalis*. The development of a computational gene regulatory network model of miRNA to target mRNA mappings for *Ciona* would facilitate the process of experimental validation and gene regulatory mechanism discovery.

The outcome of this study involves the implementation of computational techniques which can be applied to the study of other organisms. In particular, the technique can be used to quickly screen the genome of any organism for the presence of existing miRNA homologs. The prediction and validation of these factors may increase the understanding of evolution of microRNA-controlled regulatory relationships and give insight into the origins of microRNA networks in diverse animal species.

Methods

Collection of known miRNA statistics

The analysis of characteristics of known miRNAs for several species yielded parameters used to predict miRNAs and their mRNA targets in the two *Ciona* genomes. Mature miRNA and precursor sequences were obtained from the Sanger Institute's miRBase database ([6], July 2006). The statistics collected for these known miRNA molecules included loop length, percent identity in the stem region of the hairpin structure of each miRNA, percent identity of the mature miRNA sequences between the two closely related species, and percent GC content of the sequences.

Computational prediction of the Ciona intestinalis *microRNAs*

Computational prediction of *Ciona intestinalis* and *Ciona savignyi* microRNAs was performed with genome data available from the Department of Energy Joint Genome Institute (JGI) and the Broad Institute. The *Ciona intestinalis* genome, Assembly v1.0 (April 2002) and Annotation (V1.0), were obtained from the JGI website [14]. The *Ciona savignyi* genome, Assembly v1.0 (April 2003), was obtained from the Broad Institute website [42]. Using statistics for characteristics of known microRNAs gathered in the first phase of this study, a miRNA prediction algorithm was implemented to search the *Ciona intestinalis* genome for conserved mature microRNA genes (Figure 1).

The entire *C. intestinalis* and *C. savignyi* genomes were searched for conservation with the seed region of the known mature miRNA sequences from the Sanger miRBase. Although it has been estimated that 60% of miRNA primary transcripts are found in the intergenic regions [43], we chose to search the entire genome to avoid the omission of any miRNA genes that might occur in introns or coding regions. The list of mature miRNA sequences for all organisms was obtained and used as query sequences. Queries were aligned locally with the target genomes using the FASTA/ssearch34 program [44]. Software was written to examine the results and extract matches of high similarity. Matches of 90% identity or better to the first 10 nucleotides of the known miRNAs were retained for further processing. Conservation of 90% identity or better in this seed region, between *C. intestinalis* and *C. savignyi*, was enforced. To remove repeated matches to miRNA gene family members, only the matches to *C. elegans* and/or *H. sapiens* were retained. The rationale for this choice was the following: if the sequences existed in *C.elegans* and/or *H.sapiens*, they were likely to exist in *Ciona* as well. To prune the number of potential hairpin structures, the list of seed matches between *C. intestinalis* and *C. savignyi* were extended by 10 nucleotides downstream from the match. These 20-mers were searched for 90% identity between the two species. For all such conserved matches, 100 nucleotides upstream and downstream from the match boundaries were extracted from the genomes and examined for a hairpin structure.

The RNA folding software *mfold* [45] was used to confirm the hairpin structure as the lowest energy folding form. The low energy state is indicative of the secondary structure that the RNA sequence is most likely to adopt. The structures were manually curated for the presence of hairpins with the mature miRNA sequence in the stem region, loop lengths between 10 and 50 nucleotides, and reasonable conservation of bulge size and location in the hairpin stem region between the two species. The procedure resulted in the prediction of 18 miRNA molecules that appear in both *C. intestinalis* and *C. savignyi*.

The ClustalX version 1.83 program [46] was used to confirm the miRNA gene family membership, and to verify

homology to *C. elegans* and/or *H. sapiens*. ClustalX analysis yielded 14 miRNA gene families for *Ciona intestinalis*.

The genome browsing capability of the Satoh laboratory web site was utilized to examine the JGI Gene V1 and the Kyoto Grailexp Gene 2005 tracks of the *Ciona intestinalis* genome [42]. The coordinates for the putative *Ciona* miRNAs were mapped to the gene locations to determine the position and nearest neighboring genes.

Experimental validation of predicted miRNA sequences

We chose nine sequences from our list of 14 predicted miRNAs for experimental validation. The validation process involved the collection, culturing, dissection, and RNA extraction from adult *Ciona intestinalis*. The total RNA was analyzed via the Northern blot protocol with end-labeled DNA oligonucleotide probes, specific to our predicted miRNA sequences.

Adult *Ciona intestinalis* were collected in Mission Bay, San Diego, CA. The adults were cultured at Scripps Institute of Oceanography in an aquarium of constantly flowing filtered seawater, which was pumped from approximately 1000 feet offshore. The *C. intestinalis* were kept under constant light conditions to suppress the release of gametes.

Prior to RNA extraction, the tunic of the *C. intestinalis* was removed to avoid the inclusion of contaminating organisms and material that may be growing on the surface. The fresh tissue of the entire organism, without the tunic, was immediately homogenized in the Trizol reagent (Invitrogen), and a total RNA extraction protocol was performed.

We used polyacrylamide gel electrophoresis (PAGE) Northern blot methods to detect the presence of miRNA in the RNA samples [8,47]. For a positive control on some of the Northern blots, an equal amount of total RNA from the nematode *Caenorhabditis elegans* was loaded in adjacent lanes to the *C. intestinalis* RNA. The *C. elegans* RNA was extracted from wild-type N2 grown at 20C, and collected at 53 hours in young adult stage. This stage of *C.elegans* may contain some fertilized oocytes.

DNA oligonucleotides (Allele Biotechnology, Inc. and Integrated DNA Technologies, Inc.) were designed with reverse complementary sequence to the putative mature miRNA sequences. For the predicted let-7 and miR-72 *Ciona* miRNA homologs, oligonucleotides were designed as probes for both the top and bottom strands of the hairpin structure. For all predicted miRNAs, the probe was designed by extending the mature miRNA sequence by 3 nucleotides on each end. Additional File 2 lists mature predicted miRNA sequences and their corresponding probes.

Computational prediction of the Ciona intestinalis *mRNA targets*

Following the prediction of the miRNA sequences, we computationally predicted and tabulated the most probable mRNA targets for these miRNAs. Figure 7 contains a flow diagram for the mRNA target prediction algorithm. The process aligns miRNA sequences against potential target sequences, filters the output data according to observed miRNA target binding characteristics, assigns gene functional descriptions, determines the binding locations, and prunes by phylogenetic homology with *C. savignyi*. Each individual step will be discussed in more detail in subsequent sections.

The Smith-Waterman algorithm performs a local alignment, as opposed to a global alignment. Although, it is more time-consuming, local alignment is more sensitive to finding smaller query sequences in the target sequences. A series of filtering and ranking algorithms were implemented in the Perl programming language. The procedure was written such that parameters could easily be modified and was based upon slight variations in a previously reported target prediction procedure [7] which capitalizes on the seed region binding tendencies. In particular, the output matches from the Smith-Waterman alignment were passed through a filter to confirm that the miRNA exhibited reverse complementary binding to the target. The matches were then classified according to the length of their seed binding as follows: *6-mer* matches with perfect complementary for nucleotides 2 – 7 of the 5' end of the miRNA; *7-mer* matches at nucleotides 2 – 8; *7-mer* matches at nucleotides 1 – 7; *8-mer* matches at nucleotides 2 – 9; and *8-mer* matches located at nucleotides 1 – 8. Our algorithm did not enforce the presence of an adenosine (A) on either side of the *6-mer* seed match.

The filtered matches were then ranked in two ways: by number of matches of miRNA for each mRNA, and by number of matches of mRNA per miRNA. The locations on the mRNA target were retained for processing of the relative hit location and for conservation with *C. savignyi*. These steps are will be elaborated in the next four paragraphs.

The potential target gene data set used for the target prediction was constructed as follows. Predicted gene transcripts for *Ciona intestinalis* (Annotation, V1.0) were obtained from the Satoh laboratory web site [48]. These predicted mRNA sequences did not necessarily contain the 5'UTR and 3'UTR regions of the mRNA. Thus, 1500 nucleotides upstream and downstream from the transcript boundaries were extracted from the *C. intestinalis* genome. We focused on matches to the 3'UTR of the target genes as the most likely genes regulated by miRNA.

Because the *Ciona intestinalis* genome is incompletely annotated, we computationally assigned functional descriptions to the sequences. The functional descriptions for most transcription factor genes and some signaling molecules, based on cDNA libraries and *in situ* hybridization studies, are available from the *Ciona intestinalis* online database [49]. However, many of the predicted transcripts across the genome were not functionally annotated. Therefore, the available data files were augmented with functional descriptions to the mRNAs based on blast alignments to proteins from the non-redundant database, "nr" [50], using the MESH component of MAGPIE [51]. The top match for each mRNA target was chosen as a surrogate functional description. The mRNA functional description list generated by MESH was joined to the InterPro (IPR) functional descriptions from the *Ciona intestinalis* online database, to produce a list of mRNA identifiers and their functional descriptions.

The Gene Ontology (GO) terms for the mRNA targets were assigned via custom software following alignment of the mRNA sequences with the GO database. The GO identifiers were then input into the Gene Ontology (GO) Terms Classification Counter [52], using the EGAD2GO classification filter for higher level grouping.

To further restrict the number of potential targets, the location of each binding site of a miRNA to a target gene was classified by location: 5'UTR, CDS, or 3'UTR. Matches to the 3'UTR of the mRNA were retained as the most probable targets of miRNA regulation. The EMBOSS *transeq* program [53] was used to translate the *C. intestinalis* mRNAs into 6 possible reading frames. A Perl script was implemented to select the longest open reading frame (ORF) and tabulate its coordinates in the mRNA sequence. The region of the sequence upstream of the ORF was assumed to be the 5'UTR, and the region downstream of the stop codon of the ORF was assumed to be the 3'UTR. The ORF itself was classified as the coding sequence (CDS). Next, another Perl script was written to compare the location of the miRNA match in the mRNA sequence, to the locations of the 5'UTR, CDS, and 3'UTR. Each miRNA match was classified as a 5'UTR, CDS, or 3'UTR match. This information was applied to the list of potential targets, which could then be sorted and ranked by the hit location.

The final step in the mRNA target prediction process examined the mRNA targets for conservation in *C. savignyi*. Currently, predicted transcripts of *C. savignyi* do not exist. *C.savignyi* "pseudo" mRNAs were generated as follows. The tera-tblastn program on the TimeLogic board aligned the *C.intestinalis* mRNA gene sequences against the entire *C. savignyi* genome. The top match of the *C. savignyi* genome to each of the mRNA sequences was classified as an orthologous *C. savignyi* mRNA. The target prediction code was applied to the *C. savignyi* subsequences, to check that miRNAs matched the 3'UTR of orthologous genes to *C. intestinalis*, according to the seed match criteria described earlier.

Authors' contributions

TMNK designed and implemented the algorithms, and participated in the experimental validation. JH participated in the experimental validation. AEP guided development of the algorithm, and oversaw the experimental work. TG contributed to the development of the algorithm, and oversaw the project. All authors read and approved the final manuscript.

Additional material

Additional File 1

Analysis of the characteristics of known miRNAs. This file summarizes the miRNA statistics gathered for three pairs of closely related organisms: Caenorhabditis elegans *vs.* Caenorhabditis briggsae, Drosophila melanogaster *vs.* Drosophila pseudoobscura, *and* Homo sapiens *vs.* Pan troglodytes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-445-S1.pdf]

Additional File 2

List of predicted miRNAs for C. intestinalis *and their corresponding probe sequences. This file lists the putative* C. intestinalis *miRNAs, their mature miRNA sequence, oligonucleotide probe sequence, scaffold number, coordinates, strand, position, and nearest neighboring gene.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-445-S2.xls]

Additional File 3

List of predicted miRNA targets for C. intestinalis. *This file contains a list of the predicted targets with single and multiple hits to the 3'UTR region.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-445-S3.xls]

Acknowledgements

Funding for Trina Norden-Krichmar was provided by a National Science Foundation Graduate Research Fellowship. Funding for Terry Gaasterland was provided by a grant to the Scripps Genome Center from the Rancho Santa Fe Foundation, created by Louis Simpson. Funding for Amy Pasquinelli and Janette Holtz was provided by grants from NIH and the Searle Scholars Program. Special thanks to Sheila Podell for her advice and help throughout the project, and to Alexey Novoradovsky for his code contribution for the GO terms mapping.

References

1. Kloosterman WP, Plasterk RH: **The diverse functions of microRNAs in animal development and disease.** *Dev Cell* 2006, **11(4):**441-450.

2. Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF: **Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs.** *Science* 2006, **312(5770):**75-79.
3. Boehm M, Slack F: **A developmental timing microRNA and its target regulate life span in C. elegans.** *Science* 2005, **310(5756):**1954-1957.
4. Calin GA, Sevignani C, Dan Dumitru C, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM: **Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers.** *P Natl Acad Sci USA* 2004, **101(9):**2999-3004.
5. Carthew RW: **Gene regulation by microRNAs.** *Curr Opin Genet Dev* 2006, **16(2):**203-208.
6. **The Sanger miRBase of microRNA data** [http://microrna.sanger.ac.uk/]
7. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell* 2005, **120(1):**15-20.
8. Pasquinelli AE, McCoy A, Jimenez E, Salo E, Ruvkun G, Martindale MQ, Baguna J: **Expression of the 22 nucleotide let-7 heterochronic RNA throughout the Metazoa: a role in life history evolution?** *Evol Dev* 2003, **5(4):**372-378.
9. Kim VN: **MicroRNA biogenesis: Coordinated cropping and dicing.** *Nat Rev Mol Cell Bio* 2005, **6(5):**376-385.
10. Nilsen TW: **Mechanisms of microRNA-mediated gene regulation in animal cells.** *Trends Genet* 2007, **23(5):**243-249.
11. Abbott AL, Alvarez-Saavedra E, Miska EA, Lau NC, Bartel DP, Horvitz HR, Ambros V: **The let-7 microRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans.** *Dev Cell* 2005, **9(3):**403-414.
12. Yoon S, De Micheli G: **Computational identification of microRNAs and their targets.** *Birth Defects Res C Embryo Today* 2006, **78(2):**118-128.
13. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, Harafuji N, Hastings KEM, Ho I, Hotta K, Huang W, Kawashima T, Lemaire P, Martinez D, Meinertzhagen IA, Necula S, Nonaka M, Putnam N, Rash S, Saiga H, Satake M, Terry A, Yamada L, Wang HG, Awazu S, Azumi K, Boore J, Branno M, Chin-bow S, DeSantis R, Doyle S, Francino P, Keys DN, Haga S, Hayashi H, Hino K, Imai KS, Inaba K, Kano S, Kobayashi K, Kobayashi M, Lee BI, Makabe KW, Manohar C, Matassi G, Medina M, Mochizuki Y, Mount S, Morishita T, Miura S, Nakayama A, Nishizaka S, Nomoto H, Ohta F, Oishi K, Rigoutsos I, Sano M, Sasaki A, Sasakura Y, Shoguchi E, Shin-i T, Spagnuolo A, Stainier D, Suzuki MM, Tassy O, Takatori N, Tokuoka M, Yagi K, Yoshizaki F, Wada S, Zhang C, Hyatt PD, Larimer F, Detter C, Doggett N, Glavina T, Hawkins T, Richardson P, Lucas S, Kohara Y, Levine M, Satoh N, Rokhsar DS: **The draft genome of Ciona intestinalis: Insights into chordate and vertebrate origins.** *Science* 2002, **298(5601):**2157-2167.
14. **DOE Joint Genome Institute, Ciona intestinalis genome** [http://genome.jgi-psf.org/ciona4/ciona4.home.html]
15. Zeller RW: **Generation and use of transgenic ascidian embryos.** *Methods Cell Biol* 2004, **74:**713-730.
16. Moody R, Davis SW, Cubas F, Smith WC: **Isolation of developmental mutants of the ascidian Ciona savignyi.** *Molecular and General Genetics* 1999, **262(1):**199-206.
17. Cirino P, Toscano A, Caramiello D, Macina A, Miraglia V, Monte A: **Laboratory culture of the ascidian Ciona intestinalis (L.): a model system for molecular developmental biology research.** *Mar Mod Elec Rec* 2002 [http://www.mbl.edu/BiologicalBulletin/MMER/cirino/CirCon.html].
18. Missal K, Rose D, Stadler PF: **Non-coding RNAs in Ciona intestinalis.** *Bioinformatics* 2005, **21:**77-78.
19. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32:**D109-D111.
20. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen XM, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T: **A uniform system for microRNA annotation.** *RNA* 2003, **9(3):**277-279.
21. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37(5):**495-500.
22. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: **Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans.** *Cell* 2006, **127(6):**1193-1207.
23. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs.** *Science* 2001, **294(5543):**853-858.
24. Palakodeti D, Smielewska M, Graveley BR: **MicroRNAs from the Planarian Schmidtea mediterranea: A model system for stem cell biology.** *RNA* 2006, **12(9):**1640-1649.
25. Sempere LF, Cole CN, Mcpeek MA, Peterson KJ: **The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint.** *J Exp Zool (Mol Dev Evol)* 2006, **306B:**1-14.
26. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of Caenorhabditis elegans.** *Gene Dev* 2003, **17(8):**991-1008.
27. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z: **Identification of hundreds of conserved and nonconserved human microRNAs.** *Nat Genet* 2005, **37(7):**766-770.
28. Ambros V: **The functions of animal microRNAs.** *Nature* 2004, **431(7006):**350-355.
29. Wightman B, Ha I, Ruvkun G: **Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans.** *Cell* 1993, **75(5):**855-862.
30. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes.** *RNA* 2004, **10(10):**1507-1517.
31. Kim HJ, Cui XS, Kim EJ, Kim WJ, Kim NH: **New porcine microRNA genes found by homology search.** *Genome* 2006, **49(10):**1283-1286.
32. Weber MJ: **New human and mouse microRNA genes found by homology search.** *Febs J* 2005, **272(1):**59-73.
33. Chatterjee R, Chaudhuri K: **An approach for the identification of microRNA with an application to Anopheles gambiae.** *Acta Biochim Pol* 2006, **53(2):**303-309.
34. Chaudhuri K, Chatterjee R: **MicroRNA detection and target prediction: integration of computational and experimental approaches.** *DNA Cell Biol* 2007, **26(5):**321-337.
35. Doran J, Strauss WM: **Bio-informatic trends for the determination of miRNA-target interactions in mammals.** *DNA Cell Biol* 2007, **26(5):**353-360.
36. Lindow M, Gorodkin J: **Principles and limitations of computational microRNA gene and target finding.** *DNA Cell Biol* 2007, **26(5):**339-351.
37. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R: **Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering.** *Plos Comput Biol* 2007, **3(4):**e65.
38. Johnson DS, Davidson B, Brown CD, Smith WC, Sidow A: **Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance.** *Genome Res* 2004, **14(12):**2448-2456.
39. Vinson JP, Jaffe DB, O'Neill K, Karlsson EK, Stange-Thomann N, Anderson S, Mesirov JP, Satoh N, Satou Y, Nusbaum C, Birren B, Galagan JE, Lander ES: **Assembly of polymorphic genomes: algorithms and application to Ciona savignyi.** *Genome Res* 2005, **15(8):**1127-1135.
40. Cone AC, Zeller RW: **Using ascidian embryos to study the evolution of developmental gene regulatory networks.** *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 2005, **83(1):**75-89.
41. Shi WY, Levine M, Davidson B: **Unraveling genomic regulatory networks in the simple chordate, Ciona intestinalis.** *Genome Res* 2005, **15(12):**1668-1674.
42. **The Broad Institute Ciona savignyi Database** [http://www.broad.mit.edu/annotation/ciona/]
43. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T: **New microRNAs from mouse and human.** *RNA* 2003, **9(2):**175-179.
44. Hudson D: **FASTA software.** 1995 [http://fasta.bioch.virginia.edu/].
45. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31(13):**3406-3415.
46. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ: **Multiple sequence alignment with Clustal x.** *Trends Biochem Sci* 1998, **23(10):**403-405.

47. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.** *Cell* 1993, **75(5):**843-854.
48. **Laboratory for Developmental Biology and Genome Biology, Kyoto University, Ascidian website** [http://ghost.zool.kyoto-u.ac.jp/]
49. Satou Y, Kawashima T, Shoguchi E, Nakayama A, Satoh N: **An integrated database of the ascidian, Ciona intestinalis: towards functional genomics.** *Zoolog Sci* 2005, **22(8):**837-843.
50. **The National Center for Biotechnology Information website** [http://www.ncbi.nlm.nih.gov/]
51. Gaasterland T, Sensen CW: **Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture.** *Biochimie* 1996, **78(5):**302-310.
52. Hu Z Bao, J., and Reecy, JM: **A Gene Ontology (GO) Terms Classification Counter.** *Plant and Animal Genome XV Conference, San Diego, CA* 2007 [http://www.animalgenome.org/bioinfo/tools/countgo].
53. Rice P, Longden I, Bleasby A: **EMBOSS: The European molecular biology open software suite.** *Trends Genet* 2000, **16(6):**276-277.

1.1 Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in BMC Genomics 2007. Norden-Krichmar, Trina M.; Holtz, Janette; Pasquinelli, Amy E.; and Gaasterland, Terry. "Computational prediction and experimental validation of Ciona intestinalis microRNA genes", BMC Genomics, 8:445, 2007. The dissertation author was the primary researcher and author of this paper.

Chapter 2. Characterization of the small RNA transcriptome of the diatom, *Thalassiosira pseudonana*

2.1 Abstract

Background

This study presents the first characterization of endogenous small RNAs in a diatom, *Thalassiosira pseudonana*. Small RNAs act as transcriptional and translational regulators, controlling specific target genes over a variety of biological functions. Diatoms are unicellular photosynthetic organisms that play major environmental roles in food webs and in global carbon fixation.

Results

A small RNA cDNA library was constructed for exponentially growing *T. pseudonana*, and then subjected to highly parallel pyrosequencing. From the computational analysis of approximately 300,000 sequences in the library, there exists evidence of small RNA genes, including microRNAs, repeat-associated short interfering RNAs, and endogenous short interfering RNAs. The diatom small RNA system apparently has unique features, including a novel putative Dicer, a differential bias in the first nucleotide of the small RNAs, putative microRNA sequences that are not evolutionarily conserved, and a high percentage of short interfering RNAs originating from protein-coding and repetitive areas of the genome. The diatom genome contains elements similar to plant small RNA systems, such as the RNAi machinery, the transcription from similar genomic regions, and binding sites of the small RNAs occurring primarily in the coding section of the predicted targets.

Conclusions

The characterization of the small RNA repertoire present in the transcriptome of *T. pseudonana* opens the door to a wide range of possible gene regulatory mechanisms in diatoms. Furthermore, because of the unique origin of diatoms from a double endosymbiotic event, the small RNAs found in diatoms may provide insight into the evolutionary history of regulatory small RNAs.

2.2 Background

The ocean environment is subject to dynamic changes which frequently occur on short time scales, and organisms in this environment must have mechanisms to adapt rapidly to these changes. Phytoplankton are key players in oceanic processes, and because many of these are not motile, they are fully subject to changes in environmental conditions. As a consequence, gene expression control mechanisms in phytoplankton must be tailored to enable rapid changes in cellular metabolism.

Diatoms are unicellular photosynthetic phytoplankton. They are photosynthetic heterokonts, and have a distinct evolutionary history relative to land plants and red, green, and glaucophyte algae in that they are the result of two endosymbiotic events, one of which was a eukaryote/eukaryote endosymbiosis [1]. This leads to a unique genetic complement in diatoms [2], and by inference, potentially unique gene expression control mechanisms. Diatoms play globally important environmental roles. They are extremely ecologically successful and adaptable, and are found in marine, freshwater, terrestrial, and frozen environments. They are responsible for 20% of global carbon fixation [3, 4], and as such not only provide a major source of carbon for food webs, but are key players in atmospheric carbon cycling and its attendant environmental issues. Diatoms are also abundant producers of neutral lipids, and are model organisms for algal-based biofuels

development [5]. Additionally, diatom cell walls are made of intricately structured silica [6], which is of interest both from a pure cell biology standpoint, and for applications in nanotechnology [7].

Several features of diatoms make them amenable to bioinformatics and molecular biological study. The genome sequences of two diatom species, *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, are complete, compact (34.5 Mbp and 27.4 Mbp, respectively), and contain approximately 11,800 and 10,400 predicted genes, respectively [1, 8, 9]. There are also extensive EST databases available ([10][http://www.biologie.ens.fr/diatomics/EST]). Genetic experimental tools used with diatoms include transformation with selectable markers using the biolistic gene gun method [11], gene tagging with green fluorescent protein [12, 13], and mRNA expression knockdown approaches [14]. Gene expression control mechanisms are not well characterized in diatoms, although they are expected to be as complex as those in other eukaryotes. For example, post-transcriptional regulation has been demonstrated for the nitrate reductase gene [15] and the silicon transporter gene family [16], which suggests the involvement of small RNA regulation of gene expression.

Small non-coding RNA genes have been found in numerous organisms. For example, the microRNA class of small RNAs have been found in multicellular animals, plants, viruses, and the unicellular green algae *Chlamydomonas reinhardtii*

[17, 18]. Small RNAs act as transcriptional and translational regulators of gene expression. Their ability to silence specific genes provides a wide range of biological functions, ranging from gene regulation during embryological developmental and cell differentiation, to genome rearrangement [19]. In eukaryotic organisms, the different types of small RNAs range in size from 19-31 nucleotides. Besides their similarity in length, small RNAs share the common feature of also associating with the Argonaute family proteins. Once they are bound in an Argonaute protein complex called the RNA-induced silencing complex (RISC), the small RNAs are able to guide these protein effector complexes by base-pairing perfectly or partially to a target mRNA, which regulates gene expression by either degradation by endonucleolytic cleavage or mRNA target sequestration (translational regulation). Some small RNAs can also regulate transcription of the target mRNA by forming a RITS (RNA induced transcriptional silencing) complex that affects chromatin structure. Although the small RNA genes are all derived from double-stranded RNA (dsRNA), the various types of small RNA genes may be distinguished from one another by differences in their biogenesis.

Of the classes of small RNAs, the microRNA (miRNA) family is perhaps the most extensively characterized. MicroRNAs are projected to occur at a frequency of approximately 0.5-1.5% of the total genes in the genome of an organism [20], but it is estimated that 20 to 30% of human genes are targets of miRNA [21], thus their regulatory role is substantial. Many of the characteristics of miRNA formation and

mechanism are now known and can be used to predict and identify miRNAs and their target genes [22]. One of the defining characteristics of miRNA biogenesis involves an intermediate step containing a double stranded hairpin secondary structure. The loop of the hairpin structure is cleaved by the RNase III enzyme Dicer. The strands of the dsRNA are unwound by a helicase, and one of the strands associates with Argonaute to form the RISC.

Plant miRNAs possess many of the same characteristics as animal miRNAs, such as their mature miRNA length of approximately 22 nucleotides, and their formation from a hairpin precursor. However, their miRNA precursor length is more variable, and ranges from 50 to more than 350 nucleotides [23]. Generally, animal miRNAs repress gene expression by mediating translational attenuation through (multiple) miRNA-binding sites located within the 3′ untranslated region of the target gene. In contrast, almost all plant miRNAs regulate their targets by directing mRNA cleavage at single sites in the coding regions [24]. In plants, base pairing between the target and the miRNA is nearly perfect, but a few mismatches may be allowed. In animals, base pairing is usually imperfect, with the exception of a “seed” region, which enables more diversity in targets for a given miRNA [25]. Interestingly, no miRNA family has been found which is conserved between plant and animals [26], and it is suggested that miRNAs may have different evolutionary origins (Ambrose 2004).

Besides microRNAs, the remaining types of small RNAs may be grouped together collectively as endogenous short interfering RNAs (siRNA). Endogenous siRNAs have been found in multicellular and unicellular plants and animals [27]. The common characteristic of siRNA genes is that their biogenesis involves double-stranded RNA, without a hairpin precursor. The siRNA types of relevance to this study include repeated-associated siRNAs, trans-acting siRNA, and natural antisense siRNA.

Repeat-associated siRNA (rasiRNA) are 24-27 nucleotides in length, and are produced by the cleavage of long dsRNA which is derived from multiple transposons and repetitive sequences. Their association with the Argonaute family of proteins includes the Piwi subfamily: Piwi, Aubergine, and Argonaute. They silence homologous retrotransposons and repetitive sequences in the genome in sense and antisense orientations. RasiRNAs have been found in plants, *Drosophila*, *Trypanosoma brucei*, and fission yeast (*S. pombe*). The biogenesis of rasiRNAs seems to occur without the involvement of the protein Dicer, and involves an amplification process [18, 27, 28].

Trans-acting siRNA (tasiRNA) are 21-22 nucleotides in length, and have currently only been found in plants and moss. Their biogenesis involves a miRNA-mediated cleavage of a non-coding precursor transcript. These cleaved transcripts are converted into dsRNA by RNA-dependent RNA polymerase, and then processed by a

Dicer-like protein. The tasiRNAs require two miRNA complementary sites on the precursor transcript to enable cleavage for production. Their mature form guides cleavage of their target mRNAs [29].

Natural antisense transcript-derived siRNAs (nat-siRNA) are produced by transcription in opposite directions of overlapping gene pairs. Their biogenesis is Dicer dependent, and has been found in the plant *Arabidopsis* [18], and in mammalian genomes [30]. RNA is normally transcribed in the 5' to 3' direction from the antisense strand of DNA. As a result, the mRNA reads in the sense direction. In natural antisense transcription, RNA is also transcribed from the opposite strand of DNA, allowing it to interact with sense RNA and possibly regulate it. *Cis*-encoded natural antisense transcripts, or *cis*-NAT siRNA, are generated from the opposite strand of the same locus in the DNA as the sense RNA. *Trans*-encoded natural antisense transcripts, or *trans*-NAT siRNA, can be generated from another area of the genome, and may not have perfect complementarity to the opposite strand [31]. Natural antisense transcripts may occur in lengths of several thousand nucleotides [30].

While most eukaryotic small RNA research to date has focused on multicellular plants and animals, there have recently been studies in unicellular eukaryotes. In particular, several types of small RNAs were reported in the unicellular green algae *Chlamydomonas reinhardtii*, following 454 sequencing of a small cDNA library [17, 18]. Since *Chlamydomonas* has 3 Dicer-like proteins and 2 Argonaute

proteins, it is not surprising that it has the capability for RNAi. Overall, these studies identified miRNAs, phased siRNAs, tasiRNAs, and nat-siRNAs. Interestingly, 47 out of the 68 miRNAs identified in *Chlamydomonas* had longer miRNA hairpin precursors (150-729 nucleotides) as compared with the reported length for plants and animals (i.e., usually less than 150 nucleotides). All of the miRNAs found were novel and did not computationally match known plant and animal miRNAs. The genome of the green algae *Ostreococcus* was also searched, and no matches with the *Chlamydomonas* miRNAs were found. These results are significant, since they imply that deep sequencing might be the key to discovering miRNAs in organisms that have not yet been studied extensively for small RNAs. Highly expressed plant and animal small RNAs were initially characterized using traditional cloning approaches [32-34]. Re-examination of these organisms by deep sequencing approaches has revealed a larger population that includes miRNAs expressed at lower levels [35]. miRNAs with lower expression levels are generally not conserved between organisms, suggesting that they play specialized roles [35].

Since small regulatory RNAs have been found in unicellular and multicellular eukaryotes as well as in prokaryotes [36, 37], it is very likely that the diatom will possess small RNA regulatory mechanisms. The demonstration of post-transcriptional regulation of gene expression in nitrate reductase [15] and silicon transporters [16], coupled with a need for rapid responses to environmental changes in these organisms support this hypothesis. In this study, we applied a 454 sequencing approach to clone

and characterize classes of small RNAs from the diatom *Thalassiosira pseudonana*, determine their similarity to plant and animal small RNAs, examine their genomic distribution, and identify potential target mRNAs for regulation.

2.3 Results

Computational search for RNA interference (RNAi) machinery

The first step in this study required determining if the necessary RNA interference (RNAi) machinery was present in the diatom, using a rigorous bioinformatics search into the genome of *Thalassiosira pseudonana*, version 3.0. The *T. pseudonana* genome is complete and readily available for download from the Department of Energy Joint Genome Institute (JGI) website ([http://genome.jgi-psf.org/Thaps3/Thaps3.home.html]) [1, 38]. However, the predicted transcripts do not contain the functional descriptions of small RNA-related proteins such as Dicer and Argonaute in the gene models. Therefore, the annotation described here involved sequence similarity, pattern similarity, and keyword similarity against the list of known RNAi proteins and motifs.

Combining the results of these analyses provided evidence of several key components of small RNA processing machinery in the *T. pseudonana* genome. From the sequence similarity search, it was determined that the genome possesses DEAD components, the methylation components (HEN1, HDAC6, MET1, DDM1), RNA helicases, and RNA-dependent RNA polymerases. Although *T. pseudonana* did not have full sequence alignments for Argonaute or Dicer, it had good partial alignments. This may be sufficient for the production of small RNAs, since the pattern/motif search via HMM analysis showed that *T. pseudonana* has many of the necessary domains present, including the PIWI, PAZ, Helicase_C, DEAD, double-stranded RNA

binding motif (DSRM), Ribonuclease_3 (RNase III), Chromo,WW, Hist_deacetyl, and zf_UBP domains. *T. pseudonana* does not have evidence of the protein Drosha, which is utilized in miRNA biogenesis in animals [22]. Additionally, *T. pseudonana* does not have the DUF283 domain, so it apparently does not have an animal-like Dicer present. However, this data suggests that diatoms possess one or more of the homologous plant Dicer-like proteins, as well as other important RNAi-related proteins and motifs (Figure 2.1).

Mapping the motifs to chromosomes for *T. pseudonana*, it was found that the matches to the PAZ and PIWI domains fell on the same transcript in the proper order for Argonaute proteins. This finding was confirmed by the subsequent mapping of one Argonaute homolog for *T. pseudonana* on the Superfamily website [39]. There was also evidence of three homologs of RNA dependent RNA polymerase (RdRP), and one protein encoding two Ribonuclease III (RNaseIII) domains. The gene encoding the two RNaseIII domains additionally encodes a PAZ domain, thus showing homology to the *Giardia intestinalis* Dicer gene [40]. A second Dicer-like homolog present in *T. pseudonana* appears to contain the domains for DEAD/DEAH box helicases (DEAD), Helicase conserved C-terminal domain (HELICc), a weak match to the PAZ domain, and a double-stranded RNA binding motif (DSRM). This is unlike most other organisms with RNAi pathways, which additionally contain two RNaseIII domains in the Dicer protein. Since *T. pseudonana* contains a Dicer-like protein that has two RNaseIII domains and a PAZ domain, these two Dicer-like

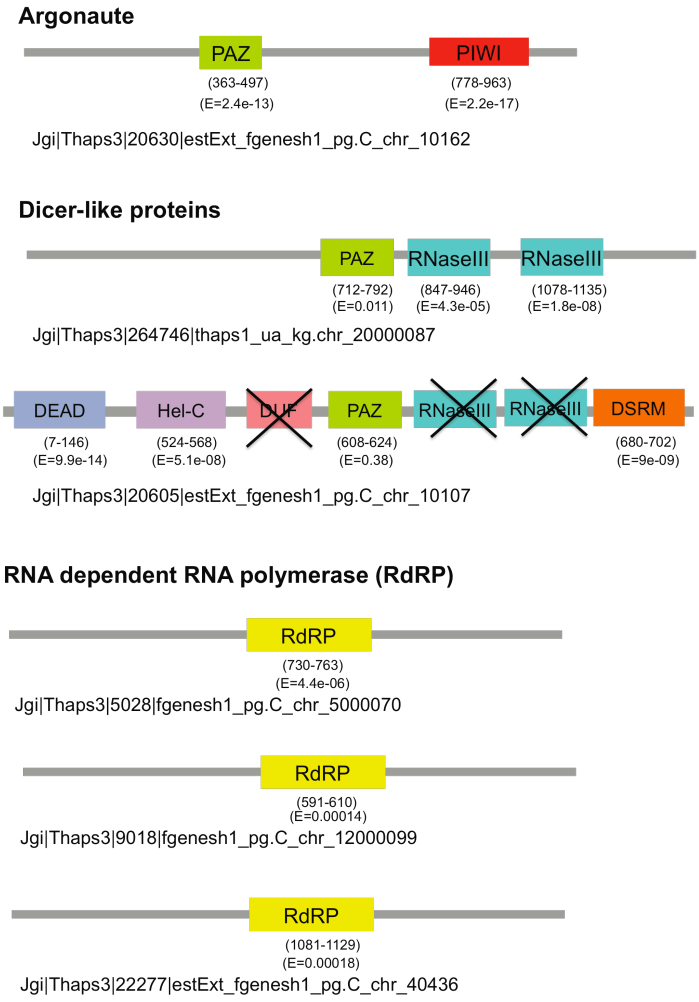

Figure 2.1. Evidence of RNAi machinery in the *T. pseudonana* genome

Schematic diagram of the *T. pseudonana* genes demonstrating homology to the Argonaute, Dicer, and RNA dependent RNA polymerase (RdRp) families of proteins. The gene names refer to the filtered gene models from the *T. pseudonana* JGI website (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html). In parentheses below each motif are the residue coordinates and HMM E-value for the motif in the gene. The typical Dicer motifs, DUF283 and RNaseIII, which were not find in transcript 20605 are denoted with an 'X' through the motif. Abbreviations used in this diagram: DEAD - DEAD-like helicase, DSRM - Double-stranded RNA binding domain, DUF - DUF283 domain, Hel-C - Helicase C-terminal domain, PAZ - PAZ domain, PIWI - PIWI domain, RdRp - RNA dependent RNA polymerase, RNaseIII - Ribonuclease III domain.

homologs may act in conjunction. Figure 2.1 contains a schematic diagram of the key eukaryotic RNAi-related proteins present in the *T. pseudonana* genome. Homologs for bacterial small RNA processing, such as RNA polymerase sigma factor (rpoS) and the LSM domain of the Hfq chaperone protein [36, 37], were also found in the *T. pseudonana* genome. This suggests that bacterial, as well as eukaryotic, small RNA mechanisms may be possible in *T. pseudonana*, which is consistent with the presence of relatively high numbers of bacterial gene homologs found in the diatom genomes [8].

General characteristics of the small RNA library

A small RNA cDNA library was constructed from RNA isolated from exponentially-growing *T. pseudonana* using established procedures (Methods and [41]). The final gel purification step of the small RNA cDNA library construction revealed the presence of the expected size small RNAs that amplified with the known linkers (Figure 2.2). There were 3 predominant bands in the gel, which could represent different subclasses of small RNAs. After confirming the presence of cloned small RNAs in a subsample of the library by TOPO cloning and Sanger sequencing, the purified products were sequenced with parallel pyrosequencing on a 454 platform [42], resulting in 305,454 reads.

Figure 2.3 contains a flow chart of the results of the computational analysis steps. Computational removal of the adapter sequences from the reads determined that

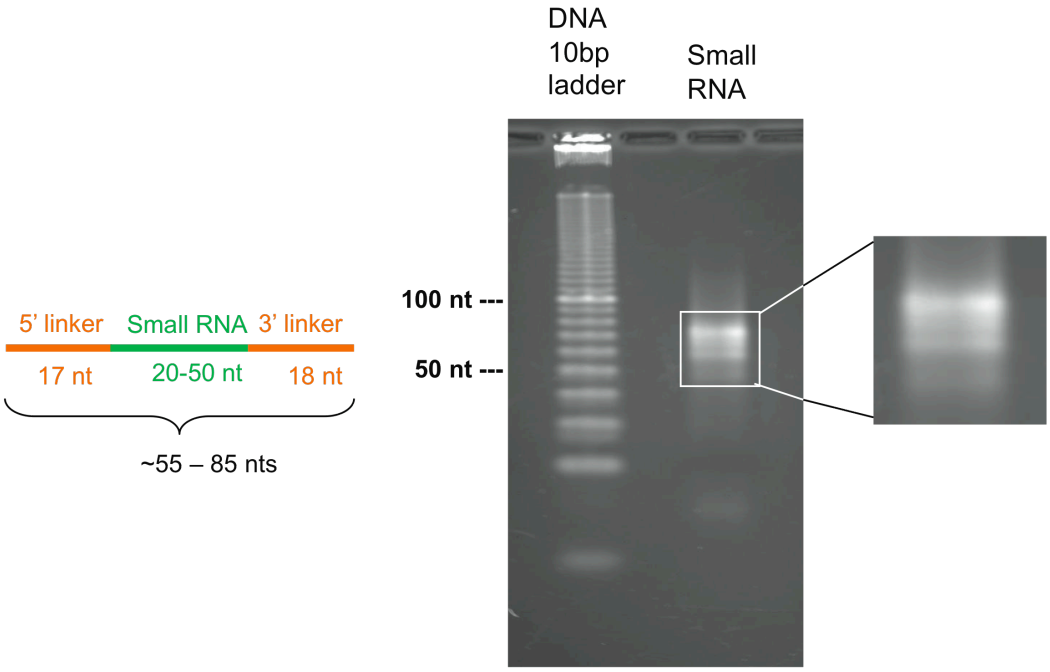

Figure 2.2. *T. pseudonana* preparatory gel showing presence of small RNA bands after library construction and amplification

Left, schematic diagram of small RNA cloning strategy, in which 5' and 3' linkers were added to size-selected small RNAs. Center, agarose gel electrophoretic separation of final amplified small RNA products. Right, zoomed gel detail highlights the presence of several predominant bands and size classes.

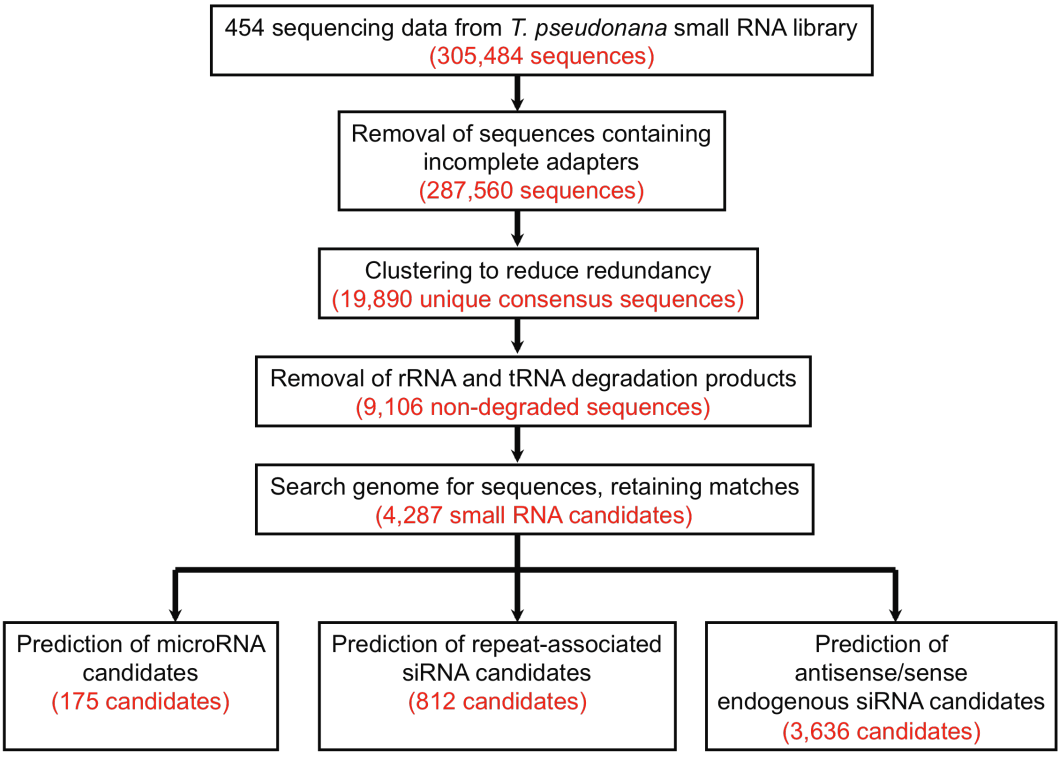

Figure 2.3. Flow chart of the computational analysis steps performed on the *T. pseudonana* 454 data sequences

17,924 sequences did not have the full adapters on both ends of the read, leaving 287,560 sequences in the candidate pool. Alignment of the candidate pool with the *T. pseudonana* assembled genome and allowing one mismatch resulted in 175,420 sequences that met this criterion. Additionally, 2140 sequences matched the *T. pseudonana* chloroplast, 382 matched the *T. pseudonana* mitochondria, and 1702 matched the bottom_drawer sequences. Under further analysis, the organellar sequences did not exhibit the characteristics of functional small RNAs. The inability of the remaining sequences to align with the *T. pseudonana* assembled genome at the one mismatch cutoff may be due to sequencing errors in the data set, sequencing errors in the *T. pseudonana* assembled genome itself, or the sequence data might have been derived from unsequenced regions of the assembled genome. Additionally, the *T. pseudonana* assembled genome is a consensus sequence between two haplotypes [1], so it is possible that there will be mismatches due to different alleles. The percentage of sequences that did not match the assembled genome is similar to the results found in other studies [18, 43].

To eliminate redundancy, the total set of sequences was clustered using a parameter of 100% identity over a minimum of 80% of their length, which reduced the candidate pool from 287,560 to a set of 19,890 unique consensus sequences. Removal of sequences that did not match the *T. pseudonana* genome, and sequences that showed sequence similarity to degraded rRNA or tRNA, resulted in 4,287 unique

sequences as small RNA candidates. Table 2.1 shows the list of sequences found in different genomic categories.

The length distribution of the small RNA library candidate sequences was tabulated after removal of rRNA degradation products and alignment with the *T. pseudonana* genome, and is shown graphically in Figure 2.4. The lengths are consistent with the library construction protocol, as we had size-selected a gel region comprising a range between 20 – 50 nucleotides. The average length is approximately 30 nucleotides, with a representation at 22 - 24 nucleotides and the majority of sequence lengths centered from 28 – 32 nucleotides, which could correspond to the enrichment in bands in Figure 2.2. In some other organisms, small RNA lengths are tightly constrained [27]; these results suggested that several types of small RNAs might be present in the diatom.

Small RNA studies in other organisms [44, 45] indicate that the 5' nucleotide is most commonly U, however in *T. pseudonana* the largest number of sequences began with G (Figure 2.5).

Table 2.1. *T. pseudonana* small RNA library sequences matching different categories of genome

	Reads	Redundant sequences	Nonredundant sequences	Total sequences	Percentage of total
Nuclear genome	**total**	171,899	3,521	175,420	61.0%
	unique	3,644	3,521	7,165	36.0%
Repetitive regions	**total**	2,220	441	2,661	0.9%
	unique	371	441	812	4.1%
Noncoding RNA	**total**	240,194	6,292	246,486	85.7%
	unique	4,492	6,292	10,784	54.2%
Chloroplast	**total**	1,996	144	2,140	0.7%
	unique	147	144	291	1.5%
Mitochondria	**total**	361	21	382	0.1%
	unique	47	21	68	0.3%
Bottom_drawer	**total**	1,517	185	1,702	0.6%
	unique	178	185	363	1.8%
Aligns with nuclear genome but is not degraded rRNA	**total**	12,844	2,453	15,297	5.3%
	unique	1,834	2,453	4,287	21.6%
Aligns with nuclear genome but is degraded rRNA	**total**	159,129	1,072	160,201	55.7%
	unique	1,818	1,072	2,890	14.5%

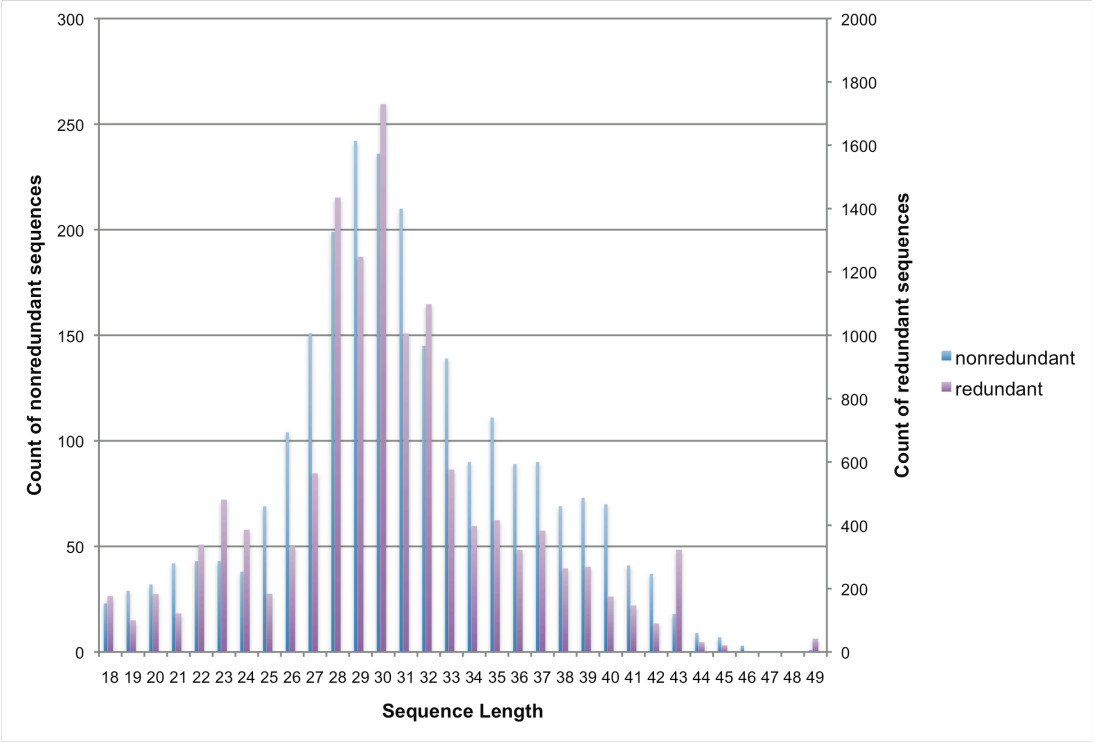

Figure 2.4. Length distribution of *T. pseudonana* small RNA candidate sequences

Length distribution was calculated after removal of RNA degradation products and alignment with the *T. pseudonana* genome.

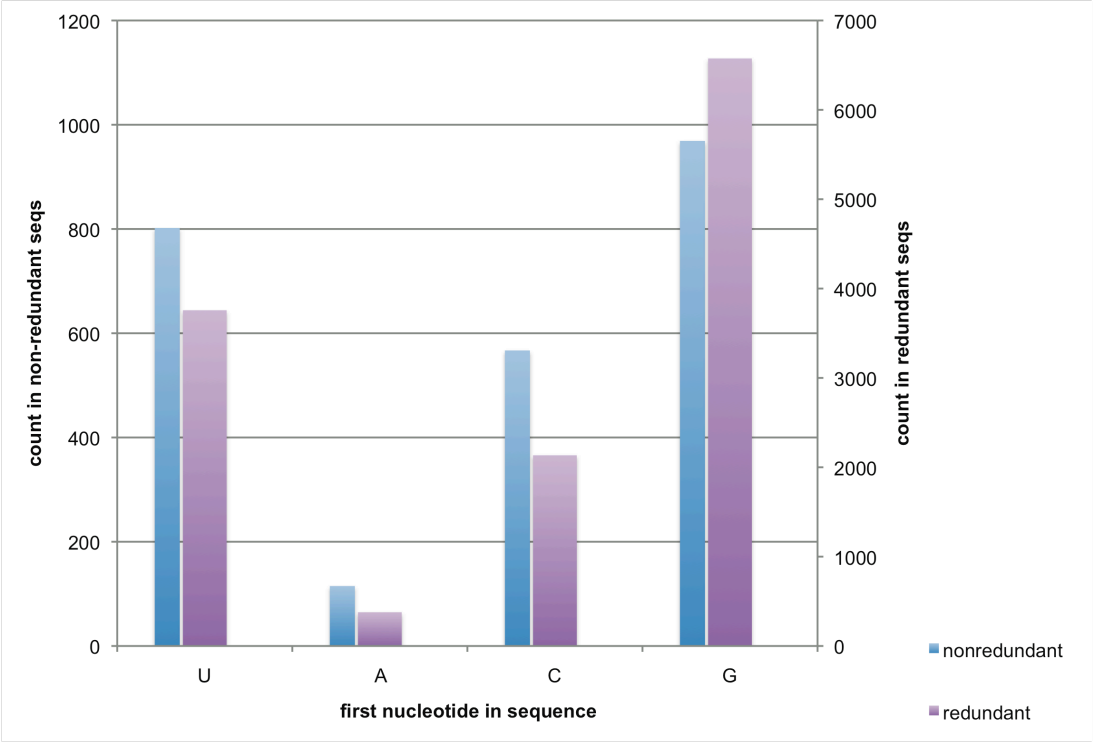

Figure 2.5. Nucleotide frequency at the 5' end of the small RNA candidate sequences

Nucleotide frequency was tabulated after removal of RNA degradation products and alignment with the *T. pseudonana* genome.

To investigate the origin and distribution of the candidate small RNA sequences in the genome, the alignment coordinates of the sequences were binned into histograms along each chromosome. As an example, Figure 2.6 shows the location and frequency of generation of small RNAs along chromosomes 6 and 7. Values above the x-axis signify that the small RNA was transcribed along the plus strand of DNA, and below the x-axis for the minus strand. From these graphs, it can be observed that the small RNAs are not evenly distributed on the chromosomes, but instead are grouped into clusters or hotspots. These results are consistent with other studies of small RNAs [46].

Prediction of miRNA candidates

Our initial small RNA characterization analysis focused on searching for intramolecular hairpins and miRNA candidates. In the entire size range of the small RNA sequences (i.e., 18 to 49 nucleotides), the folding structures of the precursor sequences produced by the mfold software was examined for intramolecular hairpin candidates. In particular, the hairpins were considered good candidates if they contained greater than 65% base pairing in the stem region, and a loop length greater than or equal to 10 nucleotides. The base pairing requirement is based upon the accepted criteria for annotating microRNA [47]. The loop length parameter is based on statistics gathered from known miRNAs in the Sanger miRBase from a previous

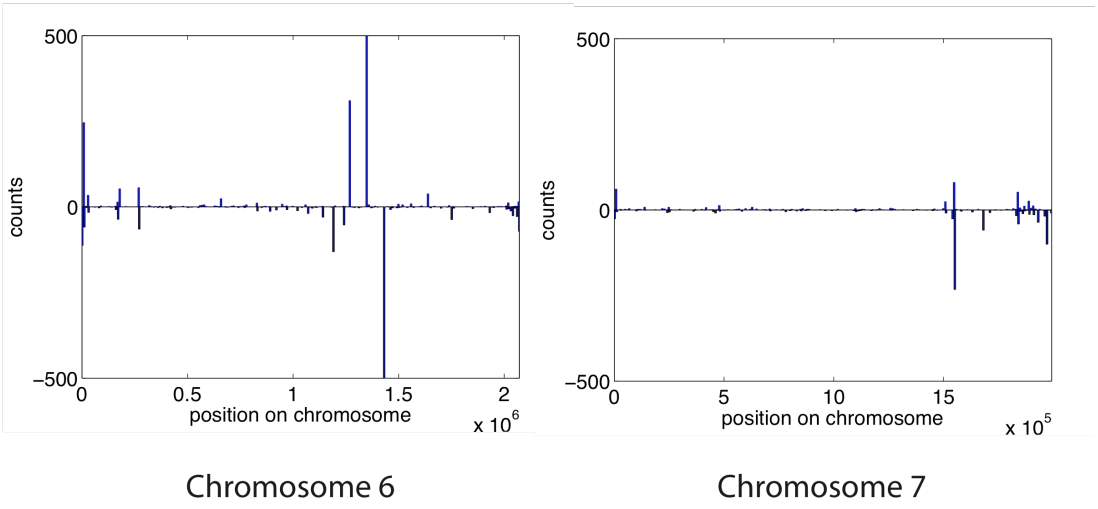

Figure 2.6. Histograms of small RNA candidate abundance mapped along the *T. pseudonana* chromosomes 6 and 7

A binsize of 10000 was used for both plots. Bars above the line represent the plus strand of DNA, and below the line represent the complimentary strand.

study [48], and by experimental evidence that Drosha RNase III processing of animal miRNAs requires a loop length greater than 10 nucleotides [49].

Of the structures that met these criteria, 175 intramolecular hairpins were found over the entire size range. In the 18-24 nucleotide size range that is characteristic for miRNAs, 70 miRNA candidates were predicted. Figure 2.7 contains the hairpin structure of three of the miRNA candidates. The precursor sequences of the miRNA candidate hairpin structures ranged in length from 52 to 291 nucleotides, with an average length of 108 nucleotides and a median length of 87 nucleotides. These lengths are similar to those of plant miRNA precursors, which range from 50 to more than 350 nucleotides [23]. This is different from the miRNAs present in *Chlamydomonas*, for which the majority of the miRNA precursors ranged from 150 to 729 nucleotides [17].

Figure 2.8 illustrates the percentages of the predicted miRNA candidates mapped relative to the *T. pseudonana* annotated genes. Intergenic regions produced 45% of the miRNA candidates, while an approximately equal percentage was found in exons (sense and antisense), and only 9% in introns. These results suggest similarity to the characteristics of plant miRNAs, for which the majority of the candidates occur within intergenic regions of the genome [26].

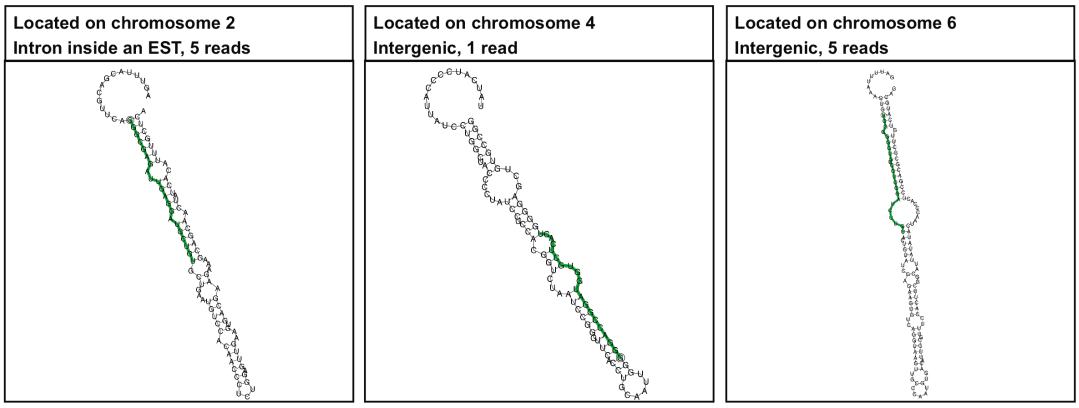

Figure 2.7. Examples of predicted secondary structures for miRNA candidates found in the *T. pseudonana* small RNA library

The mature miRNA portions are highlighted in green.

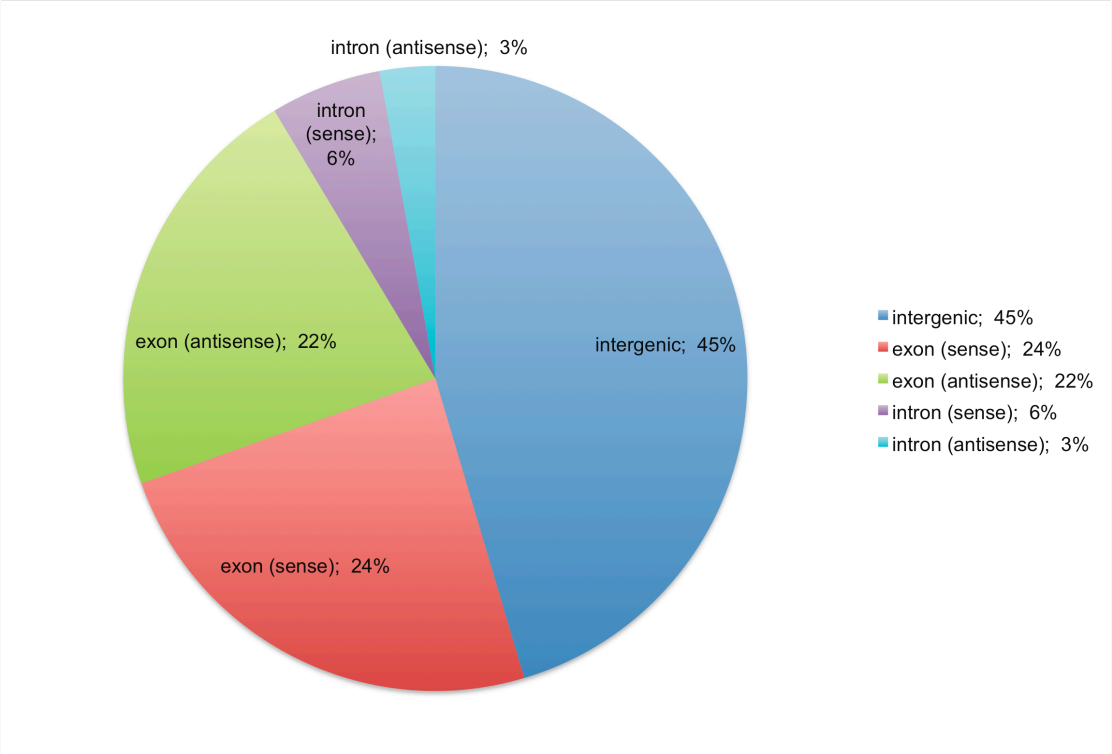

Figure 2.8. Percentages of the miRNA candidates mapped relative to *T. pseudonana* annotated gene locations

The search for miRNA:miRNA* duplexes in the data set only produced one such pair in the proper polarity. Both arms of the hairpin were represented by a very low number of reads. The absense of miRNA:miRNA* duplexes may be due to the overall low abundance of reads which folded into hairpin structures. That is, when the mature miRNA candidate is detected as only a few reads, it is unlikely that the typically lower abundance miRNA* [45] will be found in the data. A more likely scenario involves the immediate degradation of the miRNA* during cellular processing, and thereby underrepresentation in the sequenced data set.

In order to investigate if the miRNA candidates were produced via a mirtron mechanism [50], the precursor sequence of the candidates that were located inside an intron were examined in more detail. Mirtrons utilize intron splicing for producing the precursor miRNA, instead of Drosha and Pasha, and are characteristically found occupying the entire intron. None of the miRNA candidates identified through the hairpin structure analysis fit this criterion, suggesting that mirtrons may not be present in diatoms.

We applied Northern blot analysis to three of our miRNA candidates, shown schematically in Figure 2.7, to determine whether their expression levels were detectable and confirm their estimated sizes. Dilutions of control oligonucleotides were included, and although hybridization was observed to these, we did not detect a compelling positive hybridization to the total RNA (data not shown). This result is

not surprising, as the candidates were present at a very low level in the small RNA library, that is, at between 1 and 5 reads in the total library of ~300,000 sequences. Therefore, it is highly possible that the miRNA candidates were expressed at level lower than the detection sensitivity of a Northern blot.

Another way to confirm the authenticity of the miRNA candidates was by comparison with another sequence data set. In a separate study from the one described here, total RNA samples from *T. pseudonana* cell cultures were processed according to the SOLiD Small RNA Expression Kit protocol (Applied Biosystems, #4399443), and subjected to Applied Biosystems SOLiD next generation high throughput sequencing. Both biological and technical replicates were examined. The SOLiD sequences were examined for the presence of the putative miRNA and miRNA* candidates from the 454 sequencing data. Of the pool of 175 miRNA candidates in the current study, 86% of the predicted mature miRNA sequences were found in the SOLiD data set. Additionally, 64% of the miRNA* arms were found in the SOLiD data set. Since these candidates were detected from different cell preparations, using different small RNA extraction protocols and next generation sequencing techniques, this reinforces the evidence that these miRNAs and miRNA:miRNA* duplexes are valid and are endogenously present in the cell (Norden-Krichmar, T.M., Allen, A.E., and Hildebrand, M., in preparation).

To determine if known conserved miRNAs were present, the 4,287 *T. pseudonana* small RNA library sequences were compared to the database of all known miRNAs. For sequences that were in the normal length range of miRNAs (18-24nt), the seed region was examined for homology to the Sanger miRBase. When enforcing a 90% identity constraint for the first 10 nucleotides, only 12 unique sequences matched the Sanger miRBase. Of these sequences, only 1 unique sequence folded into a reasonable hairpin loop, for a precursor length of +/- 100nts surrounding the putative mature miRNA sequence. This sequence matched a miRNA found in *Arabidopsis thaliana*, ath-miR-823. The putative *T. pseudonana* miRNA demonstrated excellent seed region sequence and structural homology, since the mature miRNA was found on the 3'arm of the hairpin for both the ath-miR-823 and the putative *T. pseudonana* miRNA. Figure 2.9 compares the secondary structures of ath-miR-823 and the putative *T. pseudonana* miRNA.

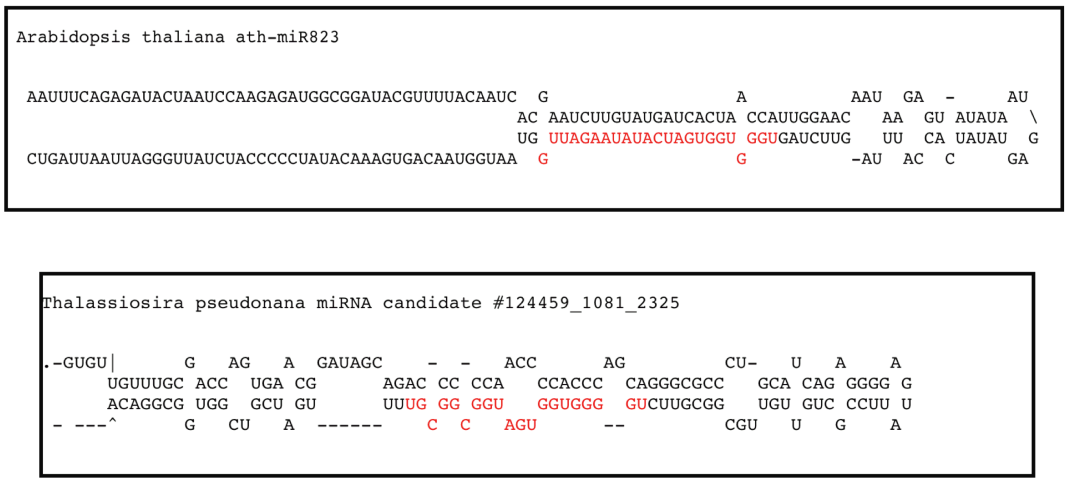

Figure 2.9. Conserved *T. pseudonana* candidate demonstrating sequence and structural homology to *Arabidopsis thaliana* miRNA, ath-miR-823

Red lettering denotes the sequence of the mature miRNA.

Prediction of endogenous siRNA candidates

Short interfering small RNAs (siRNAs) are generated from double stranded RNA in the transcriptome by several mechanisms [51]. The three types of siRNAs that were explored in this study were repeat-associated siRNAs, natural antisense transcribed siRNAs, and trans-acting siRNAs.

Repeat-associated siRNAs, or rasiRNAs, have been found to silence homologous retrotransposons and other repetitive sequences in the genome of an organism in sense and antisense orientations. In this study, the small RNA sequence data was aligned with the RepeatMasker repetitive elements in the *T. pseudonana* genome. A total of 812 unique small RNA sequences from the small RNA candidate pool mapped to the repetitive elements. Figure 2.10 contains two views of the results of the mapping, one with the general types of repetitive elements and one in finer detail. From these charts, an almost equal split can be observed in the amount of small RNAs derived from DNA transposons and long-terminal repeat retrotransposons (LTR). Figure 2.10 also contains a chart depicting the relative percentage of repeat sequence types, as mapped by repeat mapper, in the *T. pseudonana* genome. As reported [1, 8, 52], the most common transposable element in the *T. pseudonana* genome are the long-terminal repeat retrotransposons, *Copia* and *gypsy*. However, the small RNA sequences showed a slightly higher percentage occurrence in the DNA transposons, such as *Harbinger* and *MuDR*. The relatively high occurrence of small

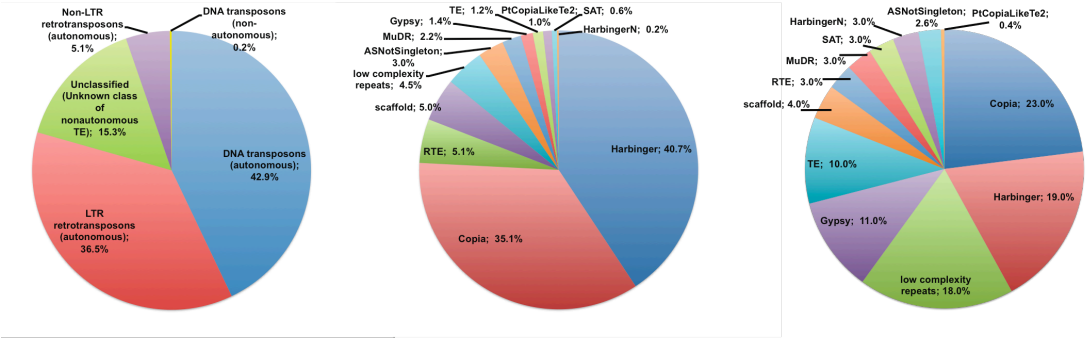

Figure 2.10. Percentage of small RNA sequences in each repetitive element class

Left, percentage of small RNA sequences relative to general class of transposons. Center, percentage of small RNA sequences relative to specific subclasses of transposons. Right, percentage of repetitive sequence types in the *T. pseudonana* genome.

RNA library sequences derived from repetitive elements suggests an important role in regulation.

Natural antisense transcript-derived siRNA (nat-siRNA) are produced from double stranded RNA formed from by transcription of overlapping gene regions. These endogenous short interfering RNAs may act as cis or trans regulatory elements. From the pool of *T. pseudonana* small RNA sequences, after removing repeat-associated and intramolecular hairpin candidates, the sequences were examined for their orientation in relation to the genomic DNA and the predicted gene transcripts. Figure 2.11 contains a count of the endogenous siRNA candidates that may be acting in *cis* or *trans* regulation. Some of the antisense transcription candidates were highly represented in the data set.

The last type of endogenous siRNA that was investigated in this study was trans-acting siRNA (tasiRNA). These siRNAs are 21-22 nucleotides in length, and their biogenesis involves a miRNA-mediated cleavage of their precursor transcripts [53]. Predictive software has been written for this type of small RNA [54], and was utilized to process the *T. pseudonana* small RNA data set. None of the *T. pseudonana* small RNA sequences were flagged as tasiRNA, since the program did not predict the occurrence with a pvalue less than 0.001. This may be due to the software's requirement of a high number of 21-nt sequences occurring at phased intervals. Since the *T. pseudonana* data set possessed a relatively low number of 21-nt

	Sequences In Exons	Sequences In Introns
sense	1673	248
antisense	548	92
Intergenic only	1127	
Intergenic and protein-coding region	136	

Figure 2.11. Number of endogenous siRNA candidates which may be acting in *cis* or *trans* regulation

sequences, it is unlikely that there would be enough occurrences to detect a phased relationship.

Search for homology of sequences to *Phaeodactylum tricornutum* genome

In an effort to determine if the *T. pseudonana* small RNA sequences were conserved in the phylogenetically closest species for which there exists a complete genome, the consensus sequences were aligned with the genome of the diatom, *Phaeodactylum tricornutum.* No perfect full-length matches to the *P. tricornutum* genome occurred when the *T. pseudonana* small RNA library consensus sequences from which ribosomal degradation products were removed was aligned to it. Perfect alignment only occurred with rRNA degradation products in the *T. pseudonana* data set. When one mismatch in the alignment was permitted, only 6 sequences were found, but all of these sequences mapped to ribosomal RNA degradation products in the *P. tricornutum* genome. Allowing two mismatches, which corresponds to approximately 90% identity between the two genomes, resulted in four additional matches. These four sequences were not generated from similar genomic areas in the two genomes. For example, one sequence matched a repetitive area of the *T. pseudonana* genome, but matched the coding section of an mRNA in *P. tricornutum*. Therefore, at the depth of coverage in this study, the small RNA candidates identified in *T. pseudonana* are not conserved in *P. tricornutum*.

Prediction of mRNA targets for the putative small RNAs

Target prediction was performed for all of the putative small RNAs characterized in this study. Using binding characteristics of both plant and animal

targets, a substantial list of possible targets was created. For the animal miRNA binding characteristics, reverse complement matches of the small RNA candidate to the target gene were classified according to the length of their seed binding, retaining *6-mer* matches with perfect complementary for nucleotides 2 – 7 of the 5' end of the small RNA. For the more stringent plant-like binding, the entire length of the small RNA sequence was examined for a minimum of 80% similarity. The percentage of predicted targets for the putative miRNAs relative to the total number of genes in *T. pseudonana* using the plant criteria was 20%, and using the animal criteria was 22%. For both plant and animal criterion, c.a. 96% of the predicted binding sites in targets were within the coding section (CDS), 2-3% were within the 3'-UTR, and 1.3% were in the 5'UTR. A majority of the targets binding in the CDS is a characteristic of plant miRNA target binding [55].

The target genes were also grouped according to their Gene Ontology (GO) terms. The majority of target gene functions are involved in metabolism, membrane transport, and nucleic acid metabolism. However, these gene functions also comprise the majority of gene functions in the entire set of annotated *T. pseudonana* filtered gene models. Therefore, in an attempt to further tease out functional bias, the ratio of the fraction of the target gene functional representation to the overall gene set was compared. Table 2.2 contains the list of the top 12 gene functions, in descending percentage of occurrence, for the targets that were predicted for the miRNA candidates and for the antisense transcription candidates. In both cases, the target binding was

enforced at 80% identity for the entire length of the small RNA to target mRNA. For the miRNA candidates, the predicted target genes had a high proportion of functions related to transcription factors and post-translational modification, in addition to metabolic processing genes. The predicted targets for the antisense transcription small RNAs, on the other hand, were more uniformly involved in metabolic processes. Because of the stringency of the target matches, and in some cases, the presence of multiple binding sites, many of the target genes demonstrate potential to control key cellular processes. While the lower stringency, animal-like binding settings produced a greater number of potential targets, the functional groupings remained similar to those targets predicted with plant-like binding characteristics.

Additionally, the list of targets for the putative *T. pseudonana* miRNA that demonstrated seed sequence and secondary structure similarity to ath-miR-823 were closely examined for conserved target matches. In Arabidopsis, the target of ath-miR823 was validated by 5'RACE to be CMT3, a CpNpG DNA cytosine methyltransferase [35]. Therefore, we searched for predicted targets in the 80% identity pool for the conserved miRNA that were similar in function. We found three

Table 2.2. Predicted target gene functional categories for miRNA and antisense transcription candidates

Predicted target gene functional categories	
microRNA candidates	**antisense transcription candidates**
RNA polymerases, DNA-directed RNA polymerase activity	amino acid metabolism
tricarboxylic acid cycle	RNA polymerases, DNA-directed RNA polymerase activity
post-translational modification/targeting	tRNA metabolism
amino acid metabolism	tricarboxylic acid cycle
proteolysis and peptidolysis	protein synthesis
energy/TCA cycle	glycolysis
RNA synthesis, transcription, DNA-dependent	RNA processing (e.g., spliceosomal, helicases)
transcription factors/activity	ribosomal proteins
RNA processing (e.g., spliceosomal, helicases)	energy/TCA cycle
tRNA metabolism	cell cycle
lipid metabolism	nucleoside, nucleotide and nucleic acid metabolism
carrier proteins/membrane transport	RNA synthesis, transcription, DNA-dependent

Functional categories are in an ordered list as a ratio to the frequency of the functional groups in the genome.

predicted targets for this miRNA candidate with functional descriptions matching “methyltransferase” or “methylase”. These results suggest that this miRNA:mRNA target pair may be evolutionarily conserved.

2.4 Discussion

Combining the results from all of the sequencing data analyses, this study demonstrates strong evidence of small RNA expression in the transcriptome of the diatom, *Thalassiosira pseudonana*. Disregarding the sequences that matched degraded ribosomal or transfer RNA, there exists a large percentage of sequences that may be acting as endogenous siRNAs, repeat-associated RNAs, or miRNAs. Figure 2.12 shows the percentage of small RNA sequences in each category covered in this study.

An essential component for determining the feasibility of the existence of small RNAs in an organism is the presence of proteins known to be involved in the biogenesis and action of the small RNAs. A recent study on the origins of RNA-mediated silencing [40], outlines the required RNAi machinery that an organism must possess as the following: one Dicer, one Argonaute-like protein, and one RNA-dependent RNA polymerase (RdRP). Dicer is involved in the biogenesis of miRNAs, natural antisense siRNAs, trans-acting siRNAs, and some small RNA produced from direct repeat sequences [56]. RNA-dependent RNA polymerases are also involved in the generation of repeat-associated siRNAs, trans-acting siRNAs, and antisense siRNAs. The presence of Argonaute, which is an integral part of the RNA-induced silencing complex (RISC), is necessary to guide the small RNA to its target. The diatom, *T. pseudonana*, contains these components, so the production of small RNAs and the mechanisms of RNAi can occur in this organism. In combination with reports

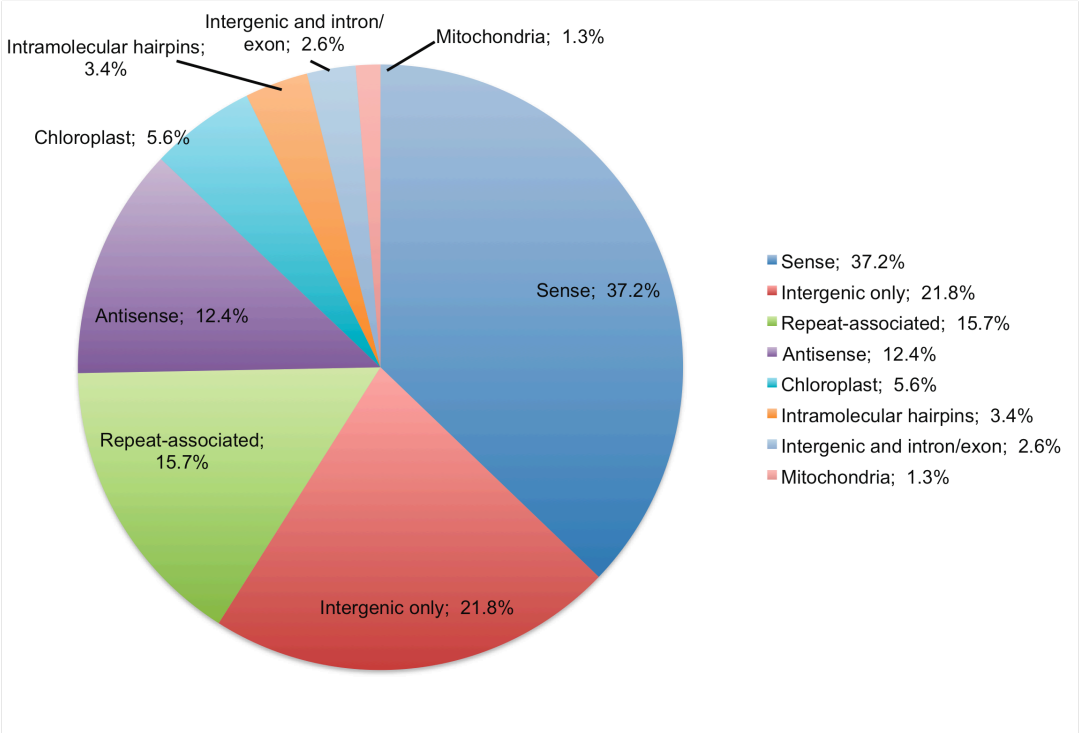

Figure 2.12. Summary of the small RNA sequence distribution in the *T. pseudonana* genome

of RNAi in a diatom [14], experimental evidence of post-transcriptional gene regulation [15], and the data produced from the sequencing of the small RNA library in this study, there is a strong case for a functional small RNA system in diatoms.

A recent study, using different computational analysis techniques, found similar evidence of the RNAi protein machinery present in *T. pseudonana* [14]. In particular, they found evidence of Argonaute, two types of Dicer-like proteins, and several RNA dependent RNA polymerases (RdRp). The main difference between our analyses involves their omission of the PAZ domains in the Dicer-like protein candidates. We found that one putative Dicer-like protein contains a match to the PAZ domain upstream from the two RNaseIII domains. However, since this *T. pseudonana* Dicer-like homolog is missing the DEAD-like helicase, HELIc, and DUF domains, this motif subset is most similar to that of *G. intestinalis*, which also contains only the PAZ and two RNaseIII domains [40].

The second putative *T. pseudonana* Dicer-like protein contains a weak match to the PAZ domain, but it was included in our annotation since the domain is found in the characteristic domain ordering of the Dicer protein in other organisms; DEAD-like helicase, HELIc, PAZ, and DSRM. The absence of the RNaseIII domains in Dicer has not been reported in other organisms. However, there are examples in *T. brucei* and *Entamoeba histolytica* where the Dicer proteins have single RNaseIII domains, which may act as dimers [40]. Thus, interaction between the two *T.*

pseudonana Dicer-like proteins is a possible mechanism for reconstituting full Dicer activity.

The function of the PAZ domain in Dicer has been linked to the interaction with the 3' overhang of small RNAs in the silencing pathway [57]. Therefore, the presence of the PAZ domain in these Dicer-like homologs may facilitate their functionality in the RNAi machinery of *T. pseudonana*.

Although homologs for some of the bacterial components of small RNA processing were found in the *T. pseudonana* genome, the presence of bacterial small RNAs was not pursued in this study. The protocol used for the sequencing in this study involved a size-selection step, which excluded sequences greater than ~50 nucleotides in length. Since bacterial small RNAs are generally 80-100 nucleotides in length, we would not see this type of RNA in our sequencing data.

Several pieces of evidence suggest that diatom small RNAs are more similar to plant than animal small RNAs. The *T. pseudonana* genome lacks the protein Drosha, which is involved in animal miRNA biogenesis. It also lacks a DUF283 domain, which is characteristic of animal Dicer proteins. Other plant-like characteristics include 1) the predominance of siRNAs in the small RNA candidate pool [58], 2) the location of most predicted miRNAs in the intergenic regions (Figure 2.8), 3) similar lengths of miRNA precursors, and 4) most predicted targets in the coding regions.

The number of miRNA candidates and intramolecular hairpins falls within the expected range as identified in other organisms. It has been estimated that miRNAs represent 0.5 to 1.5% of the total genes of an organism [20]. Since there are 11,390 genes annotated in *T. pseudonana*, then a range of 55 to 165 miRNA genes should be present. In this study, we found a total of 175 intramolecular hairpins of all length classes, and 70 hairpins of the miRNA length class, i.e., 18-24 nucleotides.

Approximately half of these miRNA candidates were transcribed from the intergenic region of the genome. These results are also similar to those found in a small RNA study of the moss, *Physcomitrella patens*, which reported 43% of the miRNAs were generated from intergenic regions, and the remainder of miRNA loci originated from exons, introns, and exon-intron boundaries [59]. In the green algae, *Chlamydomonas*, the majority of miRNAs were generated from intronic, rather than intergenic regions as they are in plants [18]. In animals, the majority of miRNA are transcribed from intronic regions [60]. In comparison to these studies, the number of miRNA candidates originating from introns in the *T. pseudonana* data set is lower than expected. This may be due to the small size and amount of intronic sequence in the *T. pseudonana* genome. It is estimated that in the *T. pseudonana* genome, there are 1.52 introns/gene, with an average size of 125-135 bp, and a median length of 90-92 bp [8]. Additionally, it is estimated that 39% of the genes are single-exon gene models. Therefore, it is possible that there is reduced ability to produce miRNA

precursors due to either insufficient availability of introns, or the short average length of intronic DNA.

Although mirtron candidates were not identified in our examination, the search did provide some impetus for future studies. The *T. pseudonana* miRNA candidates that were found inside introns were located in the genes for a copper transporter, an RNA helicase, calmodulin, vegetative cell wall protein gp1, and glycine-rich RNA-binding protein 2. One miRNA candidate was located partially inside the intron and exon of DNA topoisomerase II.

From Figure 2.12, it can also be observed that 15.7% of the small RNA candidates were derived from repetitive regions of the genome. The percentage of 15.7% is notable, considering that only approximately 2% of the *T. pseudonana* genome is composed of interspersed repeats and transposable elements [25]. In contrast, the *Chlamydomonas* small RNA study [18] found a lower percentage of repeat-associated small RNAs in their library than in the genome. That is, repeats account for 6.11% of the *Chlamydomonas* genome, but less than 4% of the small RNAs in their study were derived from repeats. The amoeba *Dictyostelium discoideum*, on the other hand, was found to have approximately 68% of small RNAs derived from the DIRS-1 retrotransposon [43]. DIRS-1 is the most abundant retrotransposon in the amoeba, and is believed to function as centromeres at mitosis. Small RNAs found in repetitive regions near centromeres are also a hallmark of gene

regulation in yeast, *S. pombe* [61]. Currently, no evidence for centromeric sequences has been found, based on G+C content, transposable element distribution, or gene poor regions in the *T. pseudonana* genome [8]. Therefore, no statement can be made concerning the relation of the locations of the repetitive regions or the repeat-associated small RNAs to the centromeres. However, the relatively high level of expression of small RNA sequences found in the repetitive regions of the *T. pseudonana* genome suggests that they may play an important role in regulation and silencing of the transposable elements. The transcriptional activity of transposable elements in diatoms is known to be elevated under stress conditions, such as nitrogen starvation [52].

As shown in Figure 2.12, the majority of the endogenous siRNA candidates were transcribed from sense, antisense, or intergenic RNA, suggesting a possible natural antisense regulatory role in the cell. Antisense transcription is prevalent in the mammalian genome, with original estimates that 20% of the human transcripts may form sense-antisense pairs [62], to later estimates of twice that value [30]. Regulation by antisense transcription has also been reported in the plant, *Arabidopsis thaliana* [63]. The *Chlamydomonas reinhardtii* small RNA study [18] classified over half of their total small RNA reads as originating in protein-coding genes and intergenic regions. Therefore, the abundance of putative natural antisense transcript-derived siRNAs found in the *T. pseudonana* small RNA candidate pool is in line with these other studies.

Figure 2.11 contains a count of the endogenous sense-antisense siRNA candidates that may be acting in *cis* or *trans* regulation. The results of the analysis determined that there were a large number of sequences that were transcribed from intergenic regions, which is consistent with studies of plant small RNAs [64]. However, the most interesting candidates were the sequences that are transcribed in the antisense direction to the introns and exons, or that are mapped to intergenic and protein coding regions, since these characteristics suggests that the small RNAs could form double-stranded RNA with the protein coding genes, generating endogenous siRNAs that have regulatory properties.

The length distribution of the small RNA candidates corresponds to the percentages of the different small RNA classes found in *T. pseudonana*, as shown in Figure 2.12. Since intramolecular hairpins and miRNA candidates , which are typically ~22 nucleotides in length, only comprised 3.4% of the small RNA candidate pool, this size class had a low representation in the data. The repeat-associated candidates, which are typically ~24 nucleotides in length, were predicted to represent approximately 15.7% of the data. Therefore, these repeat-associated candidates may be producing the peak in the ~23-24 nucleotide range in the data. Sense-antisense siRNAs are typically ~21 nucleotides, but sense-antisense transcript pairs may be as large as several thousand nucleotides (Beiter et al., 2009). The sense-antisense siRNAs, therefore, may also be contributing to the peaks at the greater lengths. There

are other categories of small RNAs, such as scanRNAs [65, 66], piwi-interacting or piRNAs [67, 68], and heterochromatic or hc-siRNAs [69], which may also produce size classes up to 31 nucleotides in length. These types of small RNA, if present in *T. pseudonana*, would require determination by their functionality or biogenesis characteristics.

The largest number of small RNA candidates in *T. pseudonana* possessed a G as the 5' terminal nucleotide. The second most prevalent 5' nucleotide in the data set was U, which is the typical 5' nucleotide for miRNAs in other organisms [44, 45]. The difference in the diatom may be due to differences in the small RNA composition and characteristics relative to other organisms. Recently, it was determined that the 5' terminal nucleotide is involved in sorting the different types of small RNAs to particular Argonaute proteins in *Arabidopsis thaliana* [70-72]. The studies found that U was most often associated with AGO1, A with AGO2 and AGO4, and C with AGO5. Since there is only one Argonaute protein annotated in the *T. pseudonana* genome, which must process the different types and lengths of small RNAs, it may be adapted to work optimally with a 5' U, but also can act on a 5' G. Another possibility would be that other currently underdetermined mechanisms for cleavage and processing are present in *T. pseudonana*.

Target prediction for the small RNA candidates using both animal and plant criteria yielded a large number of potential targets, with the majority falling within the

coding regions of the genes, rather than in the UTRs. The characteristics of intergenic miRNAs and potential targets in the coding region of genes are consistent with the biogenesis and function of plant miRNAs [55]. Nevertheless, because we do not have functional proof, we included the animal criteria in the analysis. We found that the highest percentage of targets had functions relating to transcription (Table 2.2). This result corresponds to previous studies of miRNA target gene validations in land plants, which demonstrates that a majority of conserved miRNA target transcription factor mRNAs [45, 55]. Additionally, the percentage of predicted targets for the putative miRNAs was approximately 20% of the total genes in the *T. pseudonana* genome. This percentage agrees with the estimate that 20-30% of human genes are the targets of miRNAs [21]. The presence of conserved target genes for the ath-miR-823 homolog was encouraging, since conserved target functionality and regulation by miRNAs is often evolutionarily linked [55]. Additionally, a large number of the predicted mRNA target genes contained multiple binding sites to the small RNA candidates. Since multiple binding is linked with stronger regulation in other organisms [73], these targets would have higher potential for future validation.

The lack of phylogenetic conservation of our small RNA candidates to those found in other organisms was not surprising. Recently, it was determined that although the unicellular green algae *Chlamydomonas reinhardtii* possess miRNA genes, they do not share miRNA sequence conservation with plants and animals [17, 18]. Since the diatom is a unicellular brown algae, which is not represented in the Sanger miRBase,

it is very likely that diatom miRNAs will provide a new set of miRNAs to the database. Furthermore, there was no apparent sequence homology of the centric diatom *T. pseudonana* small RNA candidates to the pennate diatom, *Phaeodactylum tricornutum*. One explanation relates to the molecular divergence between the pennate and centric diatoms approximately 90 million years ago, which is estimated to be similar to the divergence between fish and mammals, which occurred 550 million years ago [8].

The presence of small RNA in *T. pseudonana* unlocks the potential to discover small RNA gene regulation mechanisms and gene regulatory networks in the diatom. Understanding the gene regulation mechanisms of carbon fixation, nitrogen, iron, and silicon utilization in the diatom creates the possibility to manipulate these processes to help mitigate detrimental environmental issues, such as global warming and harmful algal blooms, as well as provide tools for the development of nanotechnology and the production of alternative fuel sources. Because diatoms resulted from double endosymbiosis events, further study can give clues to the evolutionary history of regulatory small RNAs.

2.5 Materials and Methods

Experimental Methods

Cell culture. *Thalassiosira pseudonana* strain CCMP1335 was obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences (West Boothbay Harbor, ME, USA), and maintained in artificial seawater (ASW) medium [74], supplemented with biotin and vitamin B_{12}, each at 1 ng · L^{-1}. Cell cultures were maintained at 18˚C – 20˚C in continuous light at an intensity of 150 µmol photons · m^{-2} · s^{-1}. The culture was magnetically stirred and aerated using sterile techniques. The *T. pseudonana* culture was grown to exponential growth phase in an 8-liter glass bottle to a density of 1.8 x 10^6 cells · ml^{-1}. Cells were concentrated with filtration using a 2 µm (142 mm) polycarbonate membrane (GE Osmonics #K20CP14220) filter, rinsed from the filter and harvested via centrifugation at 4,000 x g for 5 min.

Small RNA cDNA library construction. Total RNA was extracted with TriReagent (Sigma) as previously described [75]. Following total RNA extraction, PEG/NaCl precipitation was performed to separate low and high molecular weight RNAs in order to generate a small RNA library suitable for deep sequencing [41].

Subsequent processing included size-selection using gel electrophoresis and ligation of specific linkers. Briefly, microRNAs and small RNAs that are generated by an RNase III mechanism have 5’P and 3’OH terminal groups, which is the opposite of

mRNA degradation products. This feature enables sequential ligation of appropriate linkers specifically to the small RNAs. Selected linkers have been successfully used in several other laboratories [33, 34]. The 5' linker (called "Nelson", ATCGTAGGCACCTGAAA), which has a hydroxyl group at the 3' end, and can only ligate to RNAs with 5' phosphates, was ligated onto the small RNA fraction using T4 RNA ligase. The ligation product was size-selected and purified using denaturing PAGE as described [41]. The 5' end of the 3' linker (called "Modban", ACTGTAGGCACCATCAAT), contains a 5' phosphate that enables ligation to 3'OH group of the small RNA, and a 3' di-deoxyC group that acts as a block to prevent further ligation. Therefore, the small RNAs that are generated from the protocol only exist in the configuration: Nelson linker, small RNA, Modban linker. Following purification of the second ligation product, the ligated small RNAs were subjected to RT-PCR with Superscript III RT (Invitrogen), using primers corresponding to the linker sequences. A 4% Metaphor agarose (Lonza) gel purification was used to isolate amplified small RNAs of the correct size.

TOPO cloning (TOPO TA for Sequencng kit, Invitrogen) of the purified small cDNA fraction was performed, and approximately 50 transformants were grown for DNA preparation using a Qiagen QIAprep-spin Miniprep kit, and sequenced to confirm the quality of the library. Briefly, the purified PCR product was used to transform One Shot TOP10 Competent cells, and plated onto LB-ampicillin plates. The plates were incubated overnight at 37˚C. Colonies were picked and placed into

tubes containing 5ml of 2XYT media and 5 µl of 50 mg/ml ampicillin. The tubes were shaken overnight at 37˚C. Cloned DNAs were sequenced by the Sanger method using a service provided by SeqXcel (San Diego, CA).

After verification of the quality of the TOPO library, amplified cDNA material was quantified on the Agilent Bioanalyzer and approximately 160 ng of material was prepared for one half plate of 454 FLX sequencing [42] at the J. Craig Venter Institute in Rockville, MD.

Northern blot analysis. Polyacrylamide gel electrophoresis (PAGE) Northern blot methods were used to detect the presence of miRNA in the RNA samples [76, 77]. For a positive control on the Northern blots, a range of amounts (5.0, 1.7, 0.5, 0.18, and 0.06 ng) of oligonucleotides which were perfectly complementary to the probe sequence were loaded in adjacent lanes to 20µg of *T. pseudonana* total RNA. DNA oligonucleotides (Integrated DNA Technologies, Inc.) were designed with reverse complementary sequence to putative mature miRNA sequences. The oligonucleotides were end-labeled with γ^{32}P ATP 6000 Ci/mmol using established procedures [78]. The hybridization solution was normalized to 2 x 10^6 cpm/ml and added to the Northern blot in hybridization bottles. Hybridization was carried out overnight at 50°C. After washing, the blots were exposed for 24-48 hours at room temperature and imaged on an Amersham Biosciences Typhoon 9400 PhosphorImager.

Computational Methods

Computational search for RNA interference (RNAi) machinery in diatoms. The capacity of *Thalassiosira pseudonana* to employ RNAi-related mechanisms was evaluated bioinformatically. To assess sequence similarity to RNAi-related genes, the filtered gene model predicted transcripts for *T. pseudonana* was downloaded from the JGI website. A FASTA file was created from known RNAi-related protein sequences of interest from the UniProt site [79]([http://www.pir.uniprot.org/]). A BLASTp alignment [80] was performed between the gene models and the test set of protein sequences. The results were sorted by alignment length and E-value significance.

For evidence of pattern/motif similarity, the motifs lists in Additional File 1 were used to create a test set of Pfam identifiers from the Pfam database [81]([http://pfam.sanger.ac.uk/]). The Hidden Markov Model for Pfam software, hmmpfam, was used to search for the Pfam motifs in the filtered gene model predicted transcripts for *T. pseudonana*. Custom perl code was written to apply appropriate HMM cutoffs to the results. Gene models containing matches to the domains of interest were manually examined for completeness and, if necessary, extended. The extended gene models were then subjected to an expanded search.

Finally, a keyword search was performed using the following steps. The filtered gene model predicted transcripts for *T. pseudonana* were aligned with the

BLAST program against the proteins in the non-redundant database, “nr” [82]([http://www.ncbi.nlm.nih.gov/]). The top match for each gene model was chosen as the functional description. The description text for each match was then searched for the presence of the known RNAi-related protein and motif names.

Data Files. Computational analysis of the 454 small RNA sequence data was performed with the *Thalassiosira pseudonana* genome, version 3.0 [38]([http://genome.jgi-psf.org/Thaps3/Thaps3.home.html]). The unmasked version of the genome was used in the study. The GFF formatted filtered gene models were processed to define the locations of the introns, exons, and intergenic regions used in the small RNA analysis. The InterPro functional mapping was obtained from the Thaps3_chromosomes_domaininfo_FilteredModels2.tab file available on the JGI *T. pseudonana* website. The Thaps3 JGI website also contained the chloroplast and mitochondrial sequences in the file organelle.fasta. Extra sequences that could not be assembled into the genome were available as bottom_drawer.fasta. The coordinates for the repetitive regions of the *T. pseudonana* genome were obtained from JGI as a RepeatMasker file.

The one half plate of 454 FLX sequencing generated 305,484 sequences, which were deposited into a fasta formatted file.

Initial processing of 454 data. Figure 2.3 contains a flow chart of the computational analysis steps that were performed for the 454 sequence data. Custom

perl code was written to match and remove the 5' and 3' adapter sequences from the ends of the insert sequences. Sequences that did not possess both adapters (with less than 4 missing nucleotides) were removed from further consideration. Of the 305,484 original sequences, 17,924 sequences did not meet this criterion, leaving 287,560 sequences in the "candidate pool".

To reduce redundancy, the sequences were subjected to clustering using the program CD-hit [83, 84]([http://www.bioinformatics.org/cd-hit/]) at 100% identity for a length similarity of at least 80%. From this step, the candidate pool was reduced to a set of 19,890 "unique consensus sequences".

rRNA is degraded by an RNaseIII mechanism [85, 86], therefore, rRNA degradation products of the selected size are expected to be cloned along with authentic small RNAs. Since the *T. pseudonana* genome is not annotated for non-coding RNA, such as rRNA and tRNA, the consensus sequences were aligned with the non-redundant database "nr" and the RNA family database "Rfam" ([http://www.ncbi.nlm.nih.gov/][82]) ([87, 88][http://rfam.sanger.ac.uk/]). Consensus sequences which aligned with these databases with at least 70% alignment and whose best match had functional descriptions of "rRNA", "tRNA", or "ribosomal", were considered to be degraded non-coding RNA and were removed from further consideration. Some degradation products may remain in the candidate pool, since the sequences may have matched "nr" at less than 70% alignment, or the best match to

"nr" may not have contained a functional description indicating ribosomal or transfer RNA origin.

The remaining consensus sequences (called "small RNA candidates") were aligned with the *T. pseudonana* nuclear genome's unmasked chromosomes. 4,287 sequences which contained 1 mismatch or less were retained as potential small RNA candidates. The sequences were also aligned with the *T. pseudonana* organelle (chloroplast and mitochondria) data, and with the bottom drawer sequence data. Sequences were classified as chloroplast, mitochondria, or bottom drawer if they aligned with one of these categories, but not with the assembled *T. pseudonana* nuclear genome. At each step, the sequences were separated into non-redundant versus redundant sequences, and statistics were collected for the counts, lengths, and first nucleotide of each of the sequences.

A Matlab program was written to place the location and number of occurrences of the sequences into bins along each chromosome, along with their presence in either the plus or minus strands. A bin-size of 10000 was used, unless otherwise noted. Code was also written to generate a BED file, so that custom tracks containing the small RNA library data were available in the *T. pseudonana* Genome Browser.

Prediction of microRNA candidates. The prediction of microRNA candidates was performed with the small RNA candidate consensus sequences that

aligned with the *T. pseudonana* genome, organelle, or bottom drawer data, and did not contain sequence similarity to degraded non-coding RNA. Several different size classes of genomic sequence surrounding the small RNA sequences were extracted, to enable evaluation of previously established [45] plant and animal precursor lengths. Specifically, the regions of the genome extracted upstream and downstream of the small RNA sequence locations were as follows: +/- 100 nucleotides, +/- 200 nucleotides, +/- 300 nucleotides, +15/-65 nucleotides, and +65/-15 nucleotides. Plant precursors generally fall into the first three specified length categories, and animal precursors into the last two length categories. The putative precursor sequences were folded with the RNA folding software *mfold* [89]. To determine if the lowest energy form was a hairpin structure, custom software was written to parse the *mfold* text output and apply several criteria to the structures. Specifically, the stem region of the hairpin was constrained to contain the putative mature miRNA sequence with 65-80% base pairing, and the loop of the hairpin was constrained to be at least 10 nucleotides in length [47]. Structures that passed these criteria were manually curated. These sequences were also folded with the RNA folding software *RNAfold* [90], to confirm that the hairpin structure was produced by this program as well. MiRNA candidates that passed these tests were further classified as intron, exon, or intergenic, by their location in the predicted transcripts. For the miRNA candidates that were located in the introns of *T. pseudonana* genes, the precursors were further examined for the possibility of mirtrons. For the miRNA candidate to be considered a mirtron, the precursor must occupy the entire intron [50]. For each miRNA candidate, the small

RNA sequence pool was searched for sequences in the same polarity, located at a distance equal to the terminal loop length of the hairpin as predicted from the folded structure. The occurrences of such sequence pairs were further examined for the potential to form miRNA:miRNA* duplexes.

The small RNA library sequences were also aligned with the known mature miRNA sequences in the Sanger microRNA database, miRBase [91]. The query sequences were aligned using the FASTA/ssearch34 alignment program [92]. Software that was written and used previously for a published miRNA search [48] was utilized to examine the results. Matches of 90% identity or better for the first 10 nucleotides of the known miRNAs, were considered "seed" matches [93], and were retained. For the sequences that exhibited seed matches to the known miRNAs, the folded structures obtained in the first part of the miRNA prediction above, were examined for the presence of a hairpin structure.

Prediction of endogenous siRNA candidates. Three types of endogenous siRNA classes were investigated in this study: repeat-associated siRNA, natural antisense transcribed siRNA, and trans-acting siRNA.

Repeat-associated endogenous siRNA candidates were characterized by aligning the small RNA sequence data with the RepeatMasker repetitive elements in

the *T. pseudonana* genome. The matches to each repetitive region family were grouped by transposable element family [94] and tabulated.

The small RNA library consensus sequences were analyzed for the presence of endogenous siRNAs, possibly regulating genes via *cis* and *trans* mechanisms. In this study, custom perl code was written to determine the orientation of the small RNA transcripts in relation to the genomic DNA and to the coding mRNA. BLAST alignments of the transcripts with the genomic DNA resulted in a reference orientation for the small RNA sequences. The genomic coordinates of the small RNA transcripts were then combined with the gene locations and orientation for the predicted *T. pseudonana* genes in the GFF format file. The data was then processed with a perl script to determine if the orientation was in the same or opposite orientation. If the small RNA sequence fell into an intron or exon, it was considered "sense" if the orientation to the genomic and transcribed gene were the same. The sequence was considered "antisense" if the orientations were opposite. Intergenic matches were not categorized as sense or antisense. The results were further compared to determine sequences that fell into more than one category.

The prediction of trans-acting siRNAs (tasiRNAs) was also attempted using previously reported tasiRNA analysis software [54].

Sequence homology searches to the *Phaeodactylum tricornutum* genome. The most recent version of the *P. tricornutum* genome is version 2.0, was

downloaded from the JGI website ([http://genome.jgi-psf.org/Phatr2/Phatr2.home.html][95]). Following alignment, the results were filtered for matches of several levels of stringency to the *P. tricornutum* genome. The closest matches were examined for similarity in origin and genomic type.

Prediction of target genes. The prediction of target genes for the small RNA library consensus sequences was performed using a few modifications to previously published target prediction software [48]. Briefly, the process aligns the small RNA candidate sequences against potential mRNA target sequences, filters the output data according to observed target binding characteristics, assigns gene functional descriptions, and determines the binding location. The target binding characteristics were adjusted to allow the classic animal miRNA to mRNA binding characteristics, and a plant-like miRNA binding, which is more stringent. In particular, for the animal miRNA binding characteristics, the reverse complement matches were then classified according to the length of their seed binding, retaining *6-mer* matches with perfect complementary for nucleotides 2 – 7 of the 5' end of the small RNA. For the plant-like binding, the entire length of the small RNA sequence was examined for a minimum of 80% similarity.

The potential target gene data sets used for the target prediction consisted of the *T. pseudonana* filtered gene model transcripts and the *T. pseudonana* EST sequences which had annotated evidence in the GFF format file of a start codon and

stop codon. These ESTs represented 3,932 of the 11,890 predicted transcripts in the *T. pseudonana* genome. It was then possible to assign the match locations to occur in the 3'UTR, coding section (CDS), or 5'UTR. The functional descriptions were mapped to the matches based upon their transcript identifier in the InterPro (domaininfo) description file.

The Gene Ontology (GO) terms for the mRNA targets were assigned based on the GO annotation file for the *T. pseudonana* filtered gene models. The GO identifiers were then input into the Gene Ontology (GO) Terms Classification Counter [96], using the EGAD2GO classification filter for higher level grouping. The entire set of GO terms for all of the *T. pseudonana* filtered gene models was also processed with the EGAD2GO classification filter for comparison to the target gene function results.

2.6 Acknowledgements

Funding for Trina Norden-Krichmar was provided by a National Science Foundation Graduate Research Fellowship. Funding for Mark Hildebrand and supplies were provided by AFOSR grant FA9550-08-1-0178. Funding for Andy Allen was provided by the J. Craig Venter Institute and NSF grants 0727997 and 0722374 in Biological Oceanography and Environmental Genomics. Special thanks to Tina McIntosh for performing protocol optimization and the 454 sequencing at the J. Craig Venter Institute in Rockville, MD. Thanks also to Sheila Podell for advice and code contribution in the RNAi protein annotation section.

Chapter 2, in full, has been submitted for publication. Norden-Krichmar, Trina M.; Allen, Andrew E.; Gaasterland, Terry; and Hildebrand, Mark. “Characterization of the small RNA transcriptome of the diatom, Thalassiosira pseudonana”. The dissertation author was the primary researcher and author of this paper.

2.7 References

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

2. Allen AE, Vardi A, Bowler C: **An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms**. *Curr Opin Plant Biol* 2006, **9**(3):264-273.

3. Falkowski PG, Barber RT, Smetacek VV: **Biogeochemical Controls and Feedbacks on Ocean Primary Production**. *Science* 1998, **281**(5374):200-207.

4. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P: **Primary production of the biosphere: integrating terrestrial and oceanic components**. *Science* 1998, **281**(5374):237-240.

5. Sheehan J, Dunahay, T., Bennemann, J., and Roessler, P.: **A Look Back at the U.S. Department of Energy's Aquatic Species Program: Biodiesel from Algae. Close Out Report.** In.: U.S. Department of Energy's Office of Fuels Development, National Renewable Energy Laboratory; 1998.

6. Round FE, Crawford, R.M., and Mann, D.G.: **The diatoms: Biology and morphology of the genera**: Cambridge University Press; 1990.

7. Sandhage KH, Allan SM, Dickerson MB, Gaddis CS, Shian S, Weatherspoon MR, Cai Y, Ahmad G, Haluska MS, Snyder RL *et al*: **Merging biological self-assembly with synthetic chemical tailoring: The potential for 3-D genetically engineered micro/nano-devices (3-D GEMS)**. *International Journal of Applied Ceramic Technology* 2005, **2**(4):317-326.

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

9. Bowler C, Vardi, A., and Allen, A. E.: **Oceanographic and Biogeochemical Insights from Diatom Genomes**. *Annual Review of Marine Science* 2010, **2**:429-461.

10. **The Diatom EST Database [http://www.biologie.ens.fr/diatomics/EST]**.

11. Dunahay TG, Jarvis EE, Roessler PG: **Genetic transformation of the diatoms Cyclotella cryptica and Navicula saprophila**. *Journal of Phycology* 1995, **31**(6):1004-1012.

12. Poulsen N, Chesley PM, Kroger N: **Molecular genetic manipulation of the diatom Thalassiosira pseudonana (Bacillariophyceae)**. *Journal of Phycology* 2006, **42**(5):1059-1065.

13. Siaut M, Heijde M, Mangogna M, Montsant A, Coesel S, Allen A, Manfredonia A, Falciatore A, Bowler C: **Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum**. *Gene* 2007, **406**(1-2):23-35.

14. De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A: **Gene silencing in the marine diatom Phaeodactylum tricornutum**. *Nucleic Acids Res* 2009, **37**(14):e96.

15. Poulsen N, Kroger N: **A new molecular tool for transgenic diatoms: control of mRNA and protein biosynthesis by an inducible promoter-terminator cassette**. *Febs J* 2005, **272**(13):3413-3423.

16. Thamatrakoln K, Hildebrand M: **Analysis of Thalassiosira pseudonana silicon transporters indicates distinct regulatory levels and transport activity through the cell cycle**. *Eukaryot Cell* 2007, **6**(2):271-279.

17. Molnar A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC: **miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii**. *Nature* 2007, **447**(7148):1126-1129.

18. Zhao T, Li G, Mi S, Li S, Hannon GJ, Wang XJ, Qi Y: **A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii**. *Genes Dev* 2007, **21**(10):1190-1203.

19. Kloosterman WP, Plasterk RH: **The diverse functions of microRNAs in animal development and disease**. *Dev Cell* 2006, **11**(4):441-450.

20. Carthew RW: **Gene regulation by microRNAs**. *Curr Opin Genet Dev* 2006, **16**(2):203-208.

21. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, **120**(1):15-20.

22. Kim VN: **MicroRNA biogenesis: Coordinated cropping and dicing**. *Nat Rev Mol Cell Bio* 2005, **6**(5):376-385.

23. Bonnet E, Van de Peer Y, Rouze P: **The small RNA world of plants**. *New Phytol* 2006, **171**(3):451-468.

24. Millar AA, Waterhouse PM: **Plant and animal microRNAs: similarities and differences**. *Funct Integr Genomics* 2005, **5**(3):129-135.

25. Ambros V: **The functions of animal microRNAs**. *Nature* 2004, **431**(7006):350-355.

26. Jones-Rhoades MW, Bartel DP, Bartel B: **MicroRNAS and their regulatory roles in plants**. *Annu Rev Plant Biol* 2006, **57**:19-53.

27. Kim VN: **Small RNAs: Classification, biogenesis, and function**. *Mol Cells* 2005, **19**(1):1-15.

28. Ambros V, Chen XM: **The regulation of genes and genomes by small RNAs**. *Development* 2007, **134**(9):1635-1641.

29. Axtell MJ, Jan C, Rajagopalan R, Bartel DP: **A two-hit trigger for siRNA biogenesis in plants**. *Cell* 2006, **127**(3):565-577.

30. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J *et al*: **Antisense transcription in the mammalian transcriptome**. *Science* 2005, **309**(5740):1564-1566.

31. Beiter T, Reich E, Williams RW, Simon P: **Antisense transcription: a critical look in both directions**. *Cell Mol Life Sci* 2009, **66**(1):94-112.

32. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs**. *Science* 2001, **294**(5543):853-858.

33. Lau NC, Lim LP, Weinstein EG, Bartel DP: **An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans**. *Science* 2001, **294**(5543):858-862.

34. Lee RC, Ambros V: **An extensive class of small RNAs in Caenorhabditis elegans**. *Science* 2001, **294**(5543):862-864.

35. Rajagopalan R, Vaucheret H, Trejo J, Bartel DP: **A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana**. *Gene Dev* 2006, **20**(24):3407-3425.

36. Vogel J, Sharma CM: **How to find small non-coding RNAs in bacteria**. *Biol Chem* 2005, **386**(12):1219-1238.

37. Gottesman S: **Micros for microbes: non-coding regulatory RNAs in bacteria**. *Trends Genet* 2005, **21**(7):399-404.

38. **DOE Joint Genome Institute, Thalassiosira pseudonana v3.0 genome [http://genome.jgi-psf.org/Thaps3/Thaps3.home.html]**.

39. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**(4):903-919.

40. Cerutti H, Casas-Mollano JA: **On the origin and functions of RNA-mediated silencing: from protists to man**. *Current Genetics* 2006, **50**(2):81-99.

41. Lu C, Meyers BC, Green PJ: **Construction of small RNA cDNA libraries for deep sequencing**. *Methods* 2007, **43**(2):110-117.

42. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

43. Hinas A, Reimegard J, Wagner EG, Nellen W, Ambros VR, Soderbom F: **The small RNA repertoire of Dictyostelium discoideum and its regulation by components of the RNAi pathway**. *Nucleic Acids Res* 2007, **35**(20):6714-6726.

44. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T: **The small RNA profile during Drosophila melanogaster development**. *Dev Cell* 2003, **5**(2):337-350.

45. Bartel DP: **MicroRNAs: Genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.

46. Ho T, Wang H, Pallett D, Dalmay T: **Evidence for targeting common siRNA hotspots and GC preference by plant Dicer-like proteins**. *Febs Lett* 2007, **581**(17):3267-3272.

47. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen XM, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M *et al*: **A uniform system for microRNA annotation**. *RNA* 2003, **9**(3):277-279.

48. Norden-Krichmar TM, Holtz J, Pasquinelli AE, Gaasterland T: **Computational prediction and experimental validation of Ciona intestinalis microRNA genes**. *Bmc Genomics* 2007, **8**:445.

49. Zeng Y, Yi R, Cullen BR: **Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha**. *Embo J* 2005, **24**(1):138-148.

50. Ruby JG, Jan CH, Bartel DP: **Intronic microRNA precursors that bypass Drosha processing**. *Nature* 2007, **448**(7149):83-86.

51. Carthew RW, Sontheimer EJ: **Origins and Mechanisms of miRNAs and siRNAs**. *Cell* 2009, **136**(4):642-655.

52. Maumus F, Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M., and Bowler, C.: **Potential impact of stress activated retrotransposons on genome evolution in a marine diatom**. *BMC Evolutionary Biology (in review)* 2009.

53. Allen E, Xie ZX, Gustafson AM, Carrington JC: **microRNA-directed phasing during trans-acting siRNA biogenesis in plants**. *Cell* 2005, **121**(2):207-221.

54. Chen HM, Li YH, Wu SH: **Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis**. *Proc Natl Acad Sci U S A* 2007, **104**(9):3318-3323.

55. Axtell MJ, Bowman JL: **Evolution of plant microRNAs and their targets**. *Trends Plant Sci* 2008, **13**(7):343-349.

56. Chapman EJ, Carrington JC: **Specialization and evolution of endogenous small RNA pathways**. *Nat Rev Genet* 2007, **8**(11):884-896.

57. Song JJ, Liu J, Tolia NH, Schneiderman J, Smith SK, Martienssen RA, Hannon GJ, Joshua-Tor L: **The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes**. *Nat Struct Biol* 2003, **10**(12):1026-1032.

58. Voinnet O: **Origin, biogenesis, and activity of plant microRNAs**. *Cell* 2009, **136**(4):669-687.

59. Axtell MJ, Snyder JA, Bartel DP: **Common functions for diverse small RNAs of land plants**. *Plant Cell* 2007, **19**(6):1750-1769.

60. Kim YK, Kim VN: **Processing of intronic microRNAs**. *Embo J* 2007, **26**(3):775-783.

61. Reinhart BJ, Bartel DP: **Small RNAs correspond to centromere heterochromatic repeats**. *Science* 2002, **297**(5588):1831.

62. Chen J, Sun M, Kent WJ, Huang X, Xie H, Wang W, Zhou G, Shi RZ, Rowley JD: **Over 20% of human transcripts might form sense-antisense pairs**. *Nucleic Acids Res* 2004, **32**(16):4812-4820.

63. Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK: **Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis**. *Cell* 2005, **123**(7):1279-1291.

64. Lu C, Tej SS, Luo SJ, Haudenschild CD, Meyers BC, Green PJ: **Elucidation of the small RNA component of the transcriptome**. *Science* 2005, **309**(5740):1567-1569.

65. Mochizuki K, Fine NA, Fujisawa T, Gorovsky MA: **Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena**. *Cell* 2002, **110**(6):689-699.

66. Mochizuki K, Gorovsky MA: **A Dicer-like protein in Tetrahymena has distinct functions in genome rearrangement, chromosome segregation, and meiotic prophase**. *Genes Dev* 2005, **19**(1):77-89.

67. Seto AG, Kingston RE, Lau NC: **The coming of age for Piwi proteins**. *Mol Cell* 2007, **26**(5):603-609.

68. Kim VN: **Small RNAs just got bigger: Piwi-interacting RNAs (piRNAs) in mammalian testes**. *Genes Dev* 2006, **20**(15):1993-1997.

69. Lippman Z, Martienssen R: **The role of RNA interference in heterochromatic silencing**. *Nature* 2004, **431**(7006):364-370.

70. Kim VN: **Sorting out small RNAs**. *Cell* 2008, **133**(1):25-26.

71. Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C *et al*: **Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide**. *Cell* 2008, **133**(1):116-127.

72. Montgomery TA, Howell MD, Cuperus JT, Li D, Hansen JE, Alexander AL, Chapman EJ, Fahlgren N, Allen E, Carrington JC: **Specificity of**

ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. *Cell* 2008, **133**(1):128-141.

73. Abbott AL, Alvarez-Saavedra E, Miska EA, Lau NC, Bartel DP, Horvitz HR, Ambros V: **The let-7 microRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans**. *Dev Cell* 2005, **9**(3):403-414.

74. Darley WM, Volcani BE: **Role of silicon in diatom metabolism. A silicon requirement for deoxyribonucleic acid synthesis in the diatom Cylindrotheca fusiformis Reimann and Lewin**. *Exp Cell Res* 1969, **58**(2):334-342.

75. Hildebrand M, Dahlin K: **Nitrate transporter genes from the diatom Cylindrotheca fusiformis (Bacillariophyceae): mRNA levels controlled by nitrogen source and by the cell cycle**. *Journal of Phycology* 2000, **36**(4):702-713.

76. Pasquinelli AE, McCoy A, Jimenez E, Salo E, Ruvkun G, Martindale MQ, Baguna J: **Expression of the 22 nucleotide let-7 heterochronic RNA throughout the Metazoa: a role in life history evolution?** *Evol Dev* 2003, **5**(4):372-378.

77. Lee RC, Feinbaum RL, Ambros V: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, **75**(5):843-854.

78. Sambrook J, Fritsch, E.F., and Maniatis, T.: **Molecular Cloning: A laboratory manual**. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989.

79. **UniProt (Universal Protein Resource) [http://www.pir.uniprot.org/]**.

80. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

81. **Pfam: Protein Family Database [http://pfam.sanger.ac.uk/]**.

82. **The National Center for Biotechnology Information website [http://www.ncbi.nlm.nih.gov/]**.

83. Li WZ, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.

84. **CD-HIT: Cluster Database at High Identity with Tolerance [http://www.bioinformatics.org/cd-hit/]**.

85. Hsu D, Shih LM, Zee YC: **Degradation of rRNA in Salmonella strains: a novel mechanism to regulate the concentrations of rRNA and ribosomes**. *J Bacteriol* 1994, **176**(15):4761-4765.

86. Elbashir SM, Lendeckel W, Tuschl T: **RNA interference is mediated by 21- and 22-nucleotide RNAs**. *Genes Dev* 2001, **15**(2):188-200.

87. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database**. *Nucleic Acids Res* 2003, **31**(1):439-441.

88. **Rfam: RNA Family Database [http://rfam.sanger.ac.uk/**.

89. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction**. *Nucleic Acids Res* 2003, **31**(13):3406-3415.

90. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **FAST FOLDING AND COMPARISON OF RNA SECONDARY STRUCTURES**. *Monatshefte Fur Chemie* 1994, **125**(2):167-188.

91. Griffiths-Jones S: **The microRNA Registry**. *Nucleic Acids Res* 2004, **32**:D109-D111.

92. Hudson D: **FASTA software**. 1995.

93. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: **Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C.elegans**. *Cell Cycle* 2006, **127**:1193-1207.

94. Kohany O, Gentles AJ, Hankus L, Jurka J: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor**. *Bmc Bioinformatics* 2006, **7**:474.

95. **DOE Joint Genome Institute, Phaeodactylum tricornutum v2.0 genome [http://genome.jgi-psf.org/Phatr2/Phatr2.home.html]**.

96. Hu Z, Bao, J., and Reecy, JM: **A Gene Ontology (GO) Terms Classification Counter (http://www.animalgenome.org/bioinfo/tools/countgo)**. *Plant and Animal Genome XV Conference, San Diego, CA* 2007.

Chapter 3. Differential expression of small RNAs in the diatom, *Thalassiosira pseudonana*, identified by SOLiD sequencing

3.1 Abstract

Background

This study presents the first comparative analysis of the differential expression of endogenous small RNAs under nutrient limitations in a diatom, *Thalassiosira pseudonana*. Diatoms are unicellular, eukaryotic, phytoplankton that are major contributors to global biogeochemical processes. Additionally, this will be the first reported use of the Applied Biosystems SOLiD sequencing platform to investigate endogenous short interfering RNAs and novel microRNAs in a nonstandard model organism. Small RNAs play important roles in regulating gene expression.

Results

Small RNA cDNA libraries were constructed for five different culture conditions of *T. pseudonana*, and then processed with high throughput ABI SOLiD sequencing. In particular, small cDNA libraries were constructed from *T. pseudonana* cells cultured under 1) two different conditions of normal exponential growth, 2) silicon starvation, 3) nitrogen starvation, and 4) iron starvation. From the analysis of 153,812,217 sequences, compelling evidence of differential expression was found between the conditions. Further, a core subset of small RNAs was expressed across all conditions. The silicon-starvation library exhibited the most pronounced differences in expression, showing different length profiles and percentages of small RNAs in the repetitive element classes.

Conclusions

The existence of differential expression of the small RNA profiles in *T. pseudonana* has potential implications in the role of diatoms in ocean nutrient cycling and global carbon fixation, and provides a starting point for investigations into the gene regulatory mechanisms. The computational methodology developed in this study can be used to analyze SOLiD small RNA data from any organism with a sequenced genome, and to examine the genome-wide expression profile.

3.2 Background

Diatoms are photosynthetic, unicellular organisms that play major environmental roles in biogeochemistry, nutrient cycling, and carbon fixation, and are primary organisms in the marine food web. It is estimated that approximately 20% of global carbon fixation is the result of diatom photosynthesis [1, 2]. Diatoms are abundant lipid producers and have been targeted for use in the production of algal biofuels [3]. Additionally, due to their unique metabolism and utilization of silicon for their cell wall structure, diatoms are also being examined for applications in nanotechnology [4, 5]. As a model diatom, *Thalassiosira pseudonana* has a completely sequenced and compact genome of 34.5 MB [6], and an available EST sequence database ([7][http://www.biologie.ens.fr/diatomics/EST]).

To survive life in the ocean, diatoms must adapt to rapidly changing environmental conditions and nutrient availability. For photosynthesis and growth, diatoms require carbon dioxide, light, nitrogen, silicon, vitamins, and trace elements, such as iron [8]. In the current study, we focused on the effects of nitrogen, iron, and silicon nutrient limitation on diatoms.

Nitrogen (N) is a major limiting nutrient in the ocean. Diatoms are capable of utilizing inorganic and organic forms of nitrogen for incorporation into the cell as proteins and nucleic acids [8], and diatoms have been found to outcompete most other marine phytoplankton in nitrate-replete areas of the ocean [9]. In addition to plant-like

nitrogen metabolic genes, and components derived from bacteria by lateral gene transfer, surprisingly, diatoms also contain enzymes necessary for the metazoan urea cycle [6, 9]. These features help to explain the diatom's exceptional ability to assimilate nitrogen. Nitrogen, in the form of nitric oxide (NO), is also involved in signaling stress conditions in diatom cell populations [10]. In a recent study, stress induction by nitrate limitation was found to increase the amount of LTR retrotransposons in the diatom, *Phaeodactylum tricornutum* [11].

Iron (Fe) is an important micronutrient in the marine environment, confirmed by iron-enrichment experiments to be the limiting factor for the huge high-nutrient, low-chlorophyll (HNLC) regions of the ocean [12]. Iron is involved in the components of photosynthesis and the electron transport chain. The response of the diatom, *Phaeodactylum tricornutum*, to iron starvation was examined using gene expression microarrays, qRT-PCR, and gas chromatography-mass spectroscopy. Iron limitation was found to cause significant decreases in carbon fixation, cell volume, chlorophyll per cell, and photosynthetic efficiency [13], consistent with its important role.

Silicon (Si) is an essential nutrient for most diatoms, because it is required for the production of the cell wall [14]. Silicon limitation arrests diatoms at specific stages in the cell cycle, depending on the species [15-17]. For *T. pseudonana*, this occurs in the G1 phase [17]. Diatoms starved for silicon carry out most other

metabolic processes normally [16]. Silicon limitation in diatoms has also been used to induce lipid accumulation for biofuels productions [3]. Silicon limitation offers two advantages to studying diatom cellular metabolism: 1) since silicon is not tightly linked to the metabolism of other cellular nutrients [18], it reduces the complexity of the cellular response to lipid induction, and 2) re-addition of silicon enables synchronized progression through the cell cycle [17], which enables distinction between direct responses and cell-cycle effects to environmental and other changes.

In several cases, the interrelation of these essential nutrients in diatoms has also been demonstrated. In a recent study, a whole genome tiling array for *T. pseudonana* was used to compare the differential expression of genes under variations in pH, temperature, and the nutrients Si, N, and Fe [19]. The study identified a set of 84 genes that were upregulated in Si and Fe limitation. However, not considered in this study was the effect of the cell cycle. In a study involving the pennate diatom *Pseudo-nitzchia*, iron deficiency produced changes in cell morphology, growth rate, and the proportions of C, N, and Si in the cell [20].

Investigations have also been undertaken to determine the effects of environmental stresses in brown algae, other than those based on nutrient limitation. Copper stress, similar to what is caused by coastal pollution, produced a rapid elevation of a set of 16 genes in *T. pseudonana* [21]. The pH and amount of light was found to influence the growth rates of the diatom, *Phaeodactylum tricornutum* [22].

Other abiotic stresses, such as hyposalinity, hypersalinity, and oxidative stress, have also been demonstrated to influence gene expression in the brown algae *Ectocarpus siliculosus* [23]. Although changes to the cells and the gene expression have been demonstrated in diatoms for nutrient and environmental stressors, the mechanisms and metabolic pathways for many of these processes are still unknown. It is known that some of the responses are extremely rapid [21].

Regulation by small RNAs may provide some answers about the metabolic and cellular changes that occur in diatoms during nutrient limitation. Small RNAs have been found to influence cellular processes by several mechanisms, including the ability to control and fine-tune transcription and translation of genes, alter chromatin structure, or silence repetitive elements in the genome [24-26]. There is direct evidence for a functional small RNA system in diatoms. Genes encoding proteins involved in small RNA synthesis and processing have been identified [27, 28]. Translational regulation, which is a process that can be controlled by small RNAs, has been demonstrated in diatom nitrate reductase [29] and silicon transporter [30] genes. In a recent study using 454 sequencing with a small RNA cDNA library for *T. pseudonana* [28], evidence of multiple classes of small RNA was present.

Small RNAs in eukaryotic cells range in size from 19-31 nucleotides [31]. They are divided into two major classes based upon their biogenesis: microRNA (miRNA) and short interfering endogenous RNA (siRNA) [31]. MicroRNAs, which

are usually approximately 21-22 nucleotides in length, are formed from a double-stranded RNA (dsRNA) that has a hairpin precursor. Endogenous siRNAs are formed from dsRNA without a hairpin intermediate, but are further characterized by their functions. Repeat-associated siRNAs (rasiRNA), which are usually 24-27 nucleotides in length, silence the repetitive areas of the genome, such as transposons and retrotransposons [31, 32]. Sense-antisense siRNAs are transcribed in the opposite direction to coding genes, can range in length from ~21 nucleotides to several thousand nucleotides, and act by base-pairing to the genes to cause inhibition [33].

MicroRNAs and siRNAs have been found in organisms ranging from multicellular plants and animals, to the unicellular green algae *Chlamydomonas reinhardtii* [31, 34, 35]. In *T. pseudonana* [28], it was found that there was evidence of miRNAs, repeat-associated, and sense-antisense siRNAs. Additionally, these small RNAs and their predicted targets exhibited plant-like small RNA characteristics.

Small RNAs are expressed in a temporal and state-dependent manner. In animals, the small RNA class of miRNAs are expressed specifically and differentially during developmental stages, such as neuronal, muscle and germline development [36], and in diseases such as cancer [37]. In plants, miRNAs have also been shown to exhibit differential expression during development in *Arabidopsis thaliana* [38], bread wheat [39], rice [40], and grapes [41]. Differential expression in plants has also been observed during stress conditions, such as sulfate starvation [42], phosphate starvation

[43], and even mechanical stresses [44]. Additionally, miRNAs and endogenous siRNAs were discovered when *Arabidopsis thaliana* seedlings were exposed to dehydration, salinity, temperature stress, or the plant stress hormone abscisic acid [45]. This result was further strengthened by the report of natural cis-antisense endogenous siRNAs linked to the regulation of salt tolerance in *Arabidopsis thaliana* [46]. Therefore, it is highly likely that differential expression of small RNAs exists in diatoms undergoing nutrient stress conditions.

Applied Biosystems Inc (ABI) SOLiD next-generation sequencing is a recent technology that may aid in the discovery of small RNAs, but has only been used in a handful of studies. The SOLiD (Supported Oligonucleotide Ligation and Detection) platform utilizes a sequencing-by-ligation method, which involves iterations of hybridization and ligation, on a glass slide support, using probes labeled with four different fluorescent dyes [47, 48]. Each dye encodes a two-nucleotide pair, generating sequence data represented in "colorspace" format, rather than in nucleotide "base space" data format. In most cases, SOLiD sequencing data is used to fill in areas of a genome to a greater depth [49, 50]. The promise of applications for transcriptomic analyses, however, has brought this next generation sequencing technology to the forefront [47, 51]. Recently, ABI has introduced the SOLiD Small RNA Expression Kit to provide a means of identifying the small RNA component of an organism. To date, there has been only one published study using SOLiD for identifying miRNAs in human embryonic stem cells [52]. There have been no

reported studies on the use of SOLiD sequencing to identify miRNAs or other types of endogenous siRNAs in a nonstandard model organism, such as the diatom. For SOLiD data analysis, the standard ABI SOLiD data processing pipeline includes a step whereby the data is filtered by a comparison to the Sanger miRBase database of known miRNAs [53, 54]. As the majority of miRNAs in the diatom appear to be novel [28], filtering by the known Sanger miRBase may have undesirable effects. Therefore, this study reports the development of a new methodology to process SOLiD data to extract the entire small RNA population, which can then be examined to identify and predict novel miRNAs and endogenous siRNAs.

Furthermore, this is the first reported comparative study of differential expression of small RNAs using SOLiD sequencing data. Previous miRNA expression profiling studies have utilized microarrays, a technology that suffers from several drawbacks, such as the requirement to hybridize to known sequences, and the inability to handle the possibility of variance in sequence or methylation at the 3' end [55]. Profiling of miRNAs has been performed with the 454 and Solexa next-generation sequencing platforms [39, 56, 57], but currently has not been reported for SOLiD. A comparison of the SOLiD small RNA libraries to a 454 data set produced from the same cell culture sample is also discussed, which is the first reported comparison of a small RNA library using these two sequencing methods.

3.3 Results

Processing of the small RNA libraries of SOLiD sequence data. Five different growth or harvesting conditions were examined. Two separate cell cultures were grown under exponential growth conditions, but were harvested using either iterative centrifugation (Tp-EC), or using filtration (Tp-EF). The other three cell cultures were grown under nutrient limitations, either iron-starvation (Tp-Fe), nitrogen-starvation (Tp-N), or silicon-starvation conditions (Tp-Si). The five small RNA libraries derived from these treatments were prepared for sequencing with the ABI SOLiD Small RNA Expression Kit. Each sample was divided into two equal portions, each of which was ligated to either Adapter Mix A or Adapter Mix B from the SOLiD kit. Adapter Mix A produces sequences starting from the 5' end of the small RNA, while Adapter Mix B produces sequences starting from the 3' end. Creating both an Adapter Mix A and an Adapter Mix B library enables the identification of small RNA sequences larger than 35 nucleotides, and provides greater confidence in the data when the small RNA sequence is recovered in both sets. The Agilent Bioanalyzer representation of the small RNA library samples prior to sequencing is shown in Figure 3.1. Because the adapters are 89 nucleotides in length, the small RNA library is present as a band at approximately 108-150 nucleotides. Each sample condition was barcoded as follows: Tp_EF was barcoded G00032, Tp_EC was barcoded G31013, Tp_Fe was barcoded G01130, Tp_N was barcoded G01221, and Tp_Si was barcoded G21302.

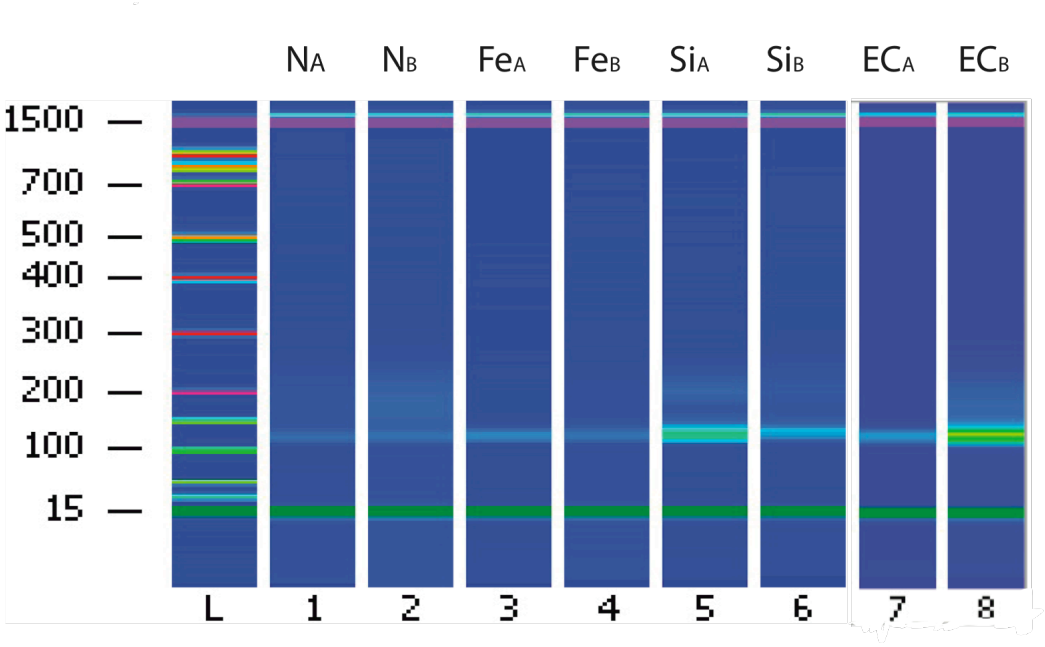

Figure 3.1. Agilent Bioanalyzer representation of the *T. pseudonana* small RNA library samples for each condition

Lane "L" contains the ladder. Lanes 1 through 8 contain the *T. pseudonana* ABI SOLiD small RNA library samples in the following order: Nitrogen-starved ligated library with A adapter mix (N_A), Nitrogen-starved library ligated with B adapter mix (N_B), Iron-starved library ligated with A adapter mix (Fe_A), Iron-starved library ligated with B adapter mix (Fe_B), Silicon-starved library ligated with A adapter mix (Si_A), Silicon-starved library ligated with B adapter mix (Si_B), exponential growth harvested by centrifugation ligated with A adapter mix (EC_A), and exponential growth harvested by centrifugation with B adapter mix (EC_B).

The libraries were sequenced on an ABI SOLiD platform. Duplicates of the Adapter Mix A samples were run on quadrants 1 and 2 of the ABI SOLiD slide, and duplicates of Adapter Mix B samples were run on quadrants 3 and 4. Sequencing of all four quadrants on the slide generated a total of 153,812,217 sequences in colorspace format. Initial processing of the data using CLCbio's CLC NGS Cell reference assembly software yielded an average of only 6.9% reads assembled to the *T. pseudonana* genome. Additionally, because this program was not able to align any sequence less than 27 nucleotides in length, and many small RNAs are in this size range, it had to be abandoned in this study.

Figure 3.2 contains a flow chart of the computational analysis methodology that was designed and implemented in the current study to process the SOLiD sequence data. Briefly, an approach was necessary to extract the small RNA sequences from the constant 35 nucleotide colorspace format SOLiD data, convert the colorspace data to its basespace equivalent, and map the sequences to the reference genome. The colorspace data, which is a numerical representation of the color produced during sequencing for each successive two-nucleotide pair, was first converted to its basespace equivalent using CLCbio's tofasta software. The basespace format sequences were then aligned to the *T. pseudonana* reference genome with BLAST [58], acting to simultaneously determine the alignment locations and trim the spurious adapter nucleotides from the ends of the small RNA sequences.

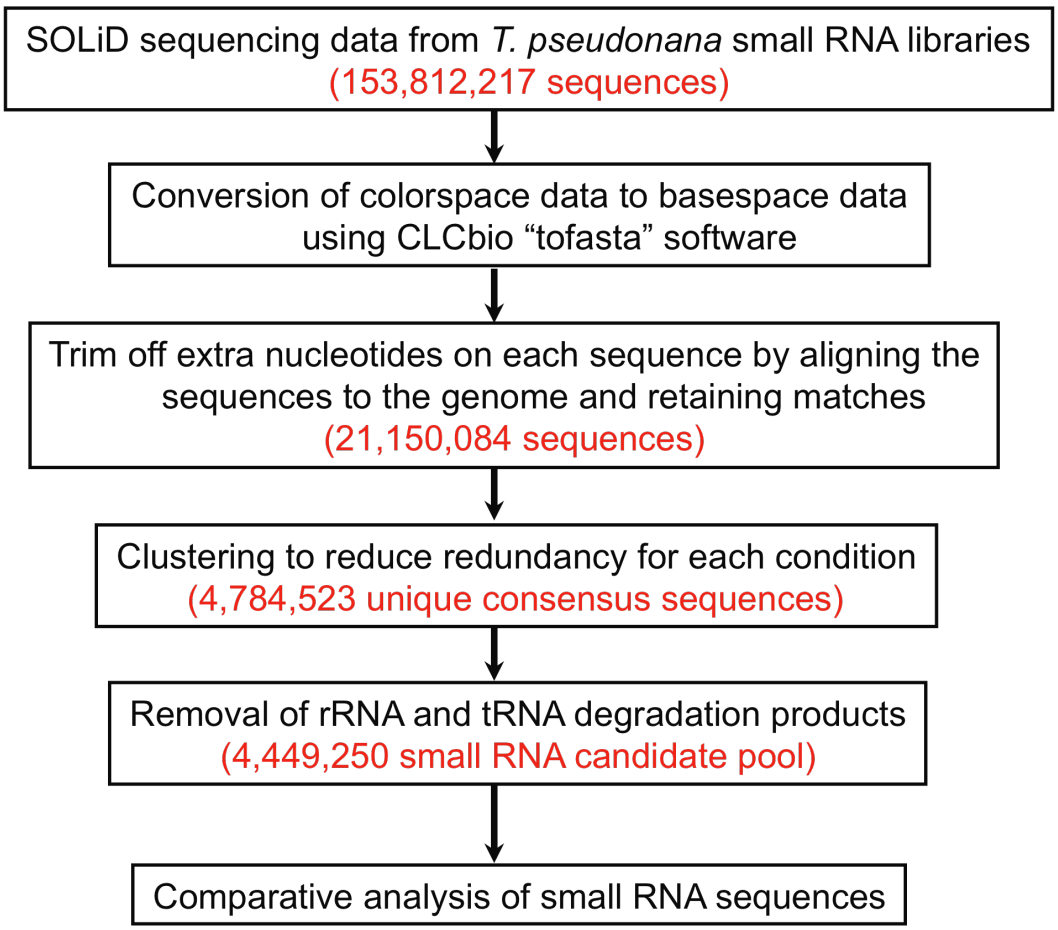

Figure 3.2. Flow chart of the computational analysis steps performed on the *T. pseudonana* SOLiD data sequences

Sequences that did not match the genome at one mismatch or less were removed from further consideration. Of the 153,812,217 original sequences, only 21,150,084 unique sequences remained in the candidate pool. The individual libraries contained different percentages of aligned sequences, ranging from 6.7% to 25.5%, with an average of 13.8% (Table 3.1). As mentioned above, the CLC NGS Cell reference assembly software yielded an average recovery rate of 6.9%. In a previous study [28] of the *T. pseudonana* small RNA library sequenced with the 454 platform, approximately 60% of the reads aligned with the *T. pseudonana* genome.

Clustering was performed to reduce redundancy, resulting in a set of 4,784,523 unique consensus sequences (Table 3.2), which represents an average of 3.1% of the total sequences in the original data set. This value fell into a similar range as the previous 454 sequencing study [28] at 6.5% of the total sequences.

Following removal of RNA degradation products, a pool of 4,449,250 unique sequences was retained for further processing as potential small RNA candidates (Table 3.2). The sequences for each barcoded condition were also separated into non-redundant versus redundant sequences (Table 3.3).

Table 3.1. Counts and percentages of sequences before clustering for each barcode and slide quadrant

Quadrant	Barcode	Library	Number of sequences in original set	Number of sequences aligning to genome at 1 mismatch or less	Percentage of sequences that align to genome
PoolA_R1	G00032	Tp-EF	1856132	472464	25.5%
	G31013	Tp-EC	6163428	1346704	21.8%
	G01130	Tp-Fe	13842679	1466321	10.6%
	G01221	Tp-N	6689612	1581617	23.6%
	G21302	Tp-Si	8188760	1576599	19.3%
PoolA_R2	G00032	Tp-EF	2131294	514446	24.1%
	G31013	Tp-EC	6896391	1484281	21.5%
	G01130	Tp-Fe	14660812	1598363	10.9%
	G01221	Tp-N	6911700	1689234	24.4%
	G21302	Tp-Si	8947807	1533830	17.1%
PoolB_R3	G00032	Tp-EF	2597380	378499	14.6%
	G31013	Tp-EC	2219798	402458	18.1%
	G01130	Tp-Fe	22307717	1536416	6.9%
	G01221	Tp-N	5773460	892874	15.5%
	G21302	Tp-Si	5895493	778015	13.2%
PoolB_R4	G00032	Tp-EF	2427350	370673	15.3%
	G31013	Tp-EC	2100212	385380	18.3%
	G01130	Tp-Fe	22820569	1521753	6.7%
	G01221	Tp-N	5636514	851313	15.1%
	G21302	Tp-Si	5745109	768844	13.4%
Totals			**153812217**	**21150084**	**13.8%**

Table 3.2. Counts and percentages of sequences after clustering and RNA degradation product removal for each barcode and pool

Pool	Barcode	Library	Number of seqs in combined pool	Number of unique seqs after clustering	% unique seqs after clustering	Number of unique seqs after removal of RNA degradation products	% unique seqs in candidate pool
Pool_A12	G00032	Tp-EF	3987426	139873	3.5%	110635	2.8%
	G31013	Tp-EC	13059819	535298	4.1%	502462	3.8%
	G01130	Tp-Fe	28503491	854604	3.0%	818123	2.9%
	G01221	Tp-N	13601312	497180	3.7%	468909	3.4%
	G21302	Tp-Si	17136567	643601	3.8%	598519	3.5%
Total in A12			**76288615**	**2670556**	**3.5%**	**2498648**	**3.3%**
Pool_B34	G00032	Tp-EF	5024730	127181	2.5%	98726	2.0%
	G31013	Tp-EC	4320010	236552	5.5%	216533	5.0%
	G01130	Tp-Fe	45128286	958018	2.1%	913230	2.0%
	G01221	Tp-N	11409974	385002	3.4%	355550	3.1%
	G21302	Tp-Si	11640602	407214	3.5%	366563	3.1%
Total in B34			**77523602**	**2113967**	**2.7%**	**1950602**	**2.5%**
Overall Totals			**153812217**	**4784523**	**3.1%**	**4449250**	**2.9%**

Table 3.3. Counts of redundant and nonredundant sequences after clustering and RNA degradation product removal for each barcode and pool

Pool	Barcode	Library	Reads	Redundant sequences	Nonredundant sequences	Total sequences
Pool_A12	G00032	Tp-EF	total	99638	99624	199262
			unique	11011	99624	110635
	G31013	Tp-EC	total	758031	382128	1140159
			unique	120334	382128	502462
	G01130	Tp-Fe	total	722162	709070	1431232
			unique	109053	709070	818123
	G01221	Tp-N	total	934272	350840	1285112
			unique	118069	350840	468909
	G21302	Tp-Si	total	853791	495041	1348832
			unique	103478	495041	598519
Pool_B34	G00032	Tp-EF	total	81413	89991	171404
			unique	8735	89991	98726
	G31013	Tp-EC	total	177891	193179	371070
			unique	23354	193179	216533
	G01130	Tp-Fe	total	779753	791750	1571503
			unique	121480	791750	913230
	G01221	Tp-N	total	559859	296467	856326
			unique	59083	296467	355550
	G21302	Tp-Si	total	435362	314595	749957
			unique	51968	314595	366563
Total unique sequences				**726565**	**3722685**	**4449250**

Comparison of SOLiD data to 454 small RNA sequence data. To validate the SOLiD data, a comparison was done with the small RNAs from exponentially growing *T. pseudonana* identified using 454 sequence data in a previous study [28], using biological and technical replicates. Approximately 86% of the putative miRNA candidates from the 454 data were found in the SOLiD data set, and for 64% of these candidates, a corresponding miRNA* (the other base-paired half of the precursor miRNA duplex) was identified. In the 454 data set, only one of the 175 putative miRNA:miRNA* duplexes was found. In addition, in the 454 data set, only 4,287 unique sequences were identified, thus the SOLiD approach has identified 1000-fold more small RNA candidates than by 454. The higher coverage provided by SOLiD sequencing appears to facilitate the discovery of small RNA precursor components present at low levels. As shown in Table 3.4, the majority of the 454 small RNA library sequences for the other classes of small RNA types were also found in the SOLiD data set. It is interesting to note that because not all of the 454 small RNA candidates are represented in the SOLiD data set, which had a much larger sample size, it is apparent that there are procedural biases in one or the other high throughput approaches. The 454 sequence data set was also included as a standard reference in all of the comparative analyses in the current study, and will be discussed in each separate analysis.

Table 3.4. Counts and percentages of sequences from the 454 data set that were found in the SOLiD data set

Data type	Unique ids in 454 file	Unique ids found in SOLID	Percentage of 454 data found in SOLiD data set
miRNA	175	152	86.9%
miRNA*	175	127	72.6%
miR/miR*	175	112	64.0%
repeats	812	808	99.5%
SAS	3636	3173	87.3%
Seqs without degraded rRNA	4287	3584	83.6%
Seqs with degraded rRNA	2665	2651	99.5%

Comparative analyses of the small RNA libraries. The small RNA libraries were compared according to the length distribution, the nucleotide frequency at the 5' end of the sequence, and the location and abundance profile along the *T. pseudonana* chromosomes.

Length distribution. The length distribution of the SOLiD small RNA library sequences from the different conditions exhibits both similarities and differences (Figure 3.3). The most noticeably similar trend in the Adapter A pool (A12) of data is the presence of dominant peaks at 21 nucleotides and 27 nucleotides in the redundant sequences of the exponential growth library, the iron-starved library, and the nitrogen-starved library. Among the libraries with dominant 21 and 27 nucleotide peaks, the exponential growth and nitrogen-starved libraries had the largest peak at 21 nucleotides, while the iron-starved library had the largest peak at 27 nucleotides. The nonredundant sequences in all libraries had less pronounced peaks. Only the silicon-starved library had a more muted distribution across all lengths in both redundant and nonredundant sequence sets, with only a slightly higher abundance at 27 nucleotides. In all SOLiD libraries, the abundance of sequences exhibits a sharp decline at lengths greater than 27 nucleotides. For the Adapter B pool (B34) of data, there were also slight indications that the 21 and 27 nucleotide peaks were largest, but they were less distinct than in the A12 pool. This may have been caused by a bias in the Adapter B Mix in the protocol.

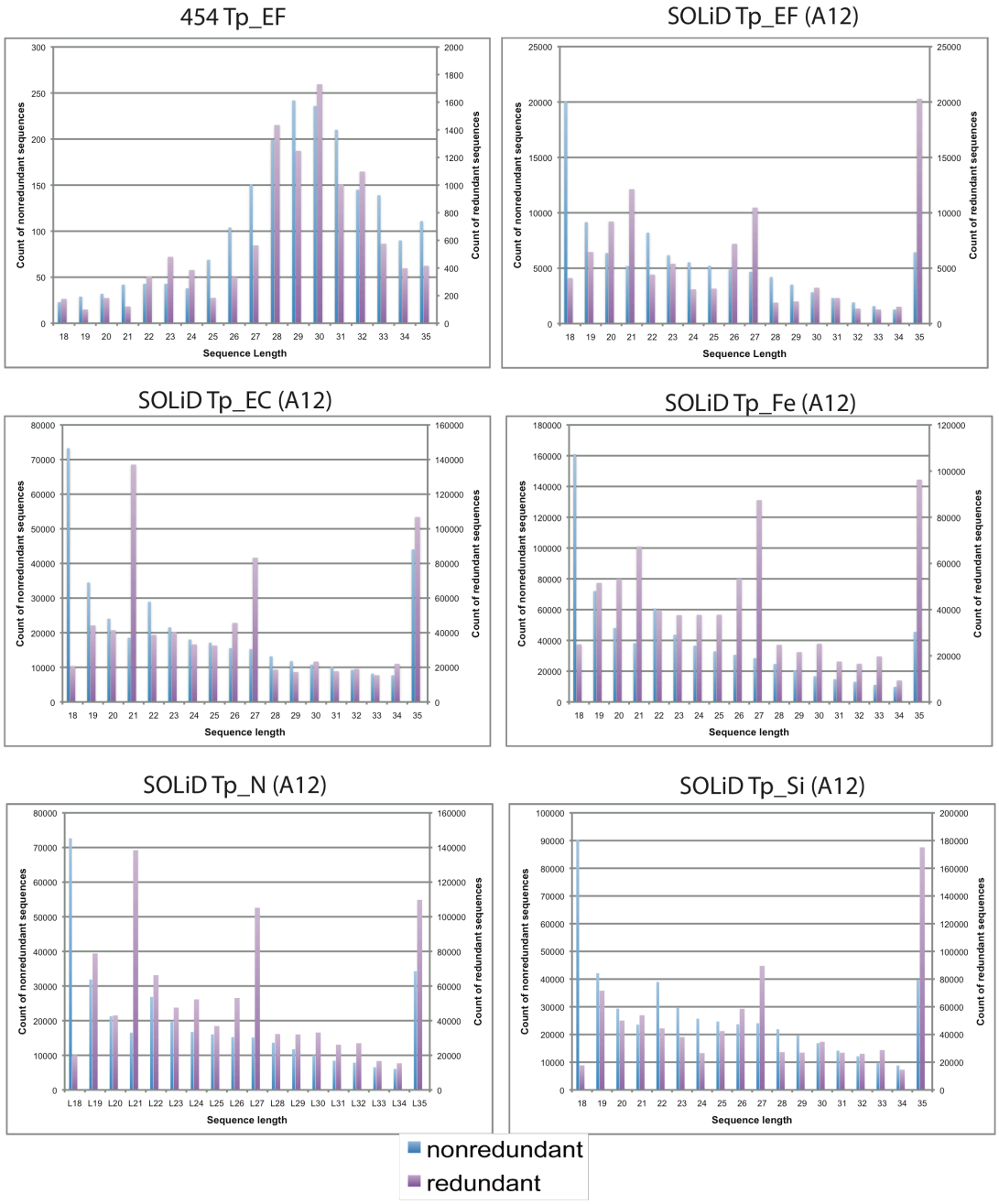

Figure 3.3 Length distribution of *T. pseudonana* small RNA candidate sequences

Length distribution was calculated after alignment with the *T. pseudonana* genome and removal of RNA degradation products.

A) Plots of 454 data and each of the barcoded condition in the SOLiD pool A, quadrants 1 and 2 (A12).

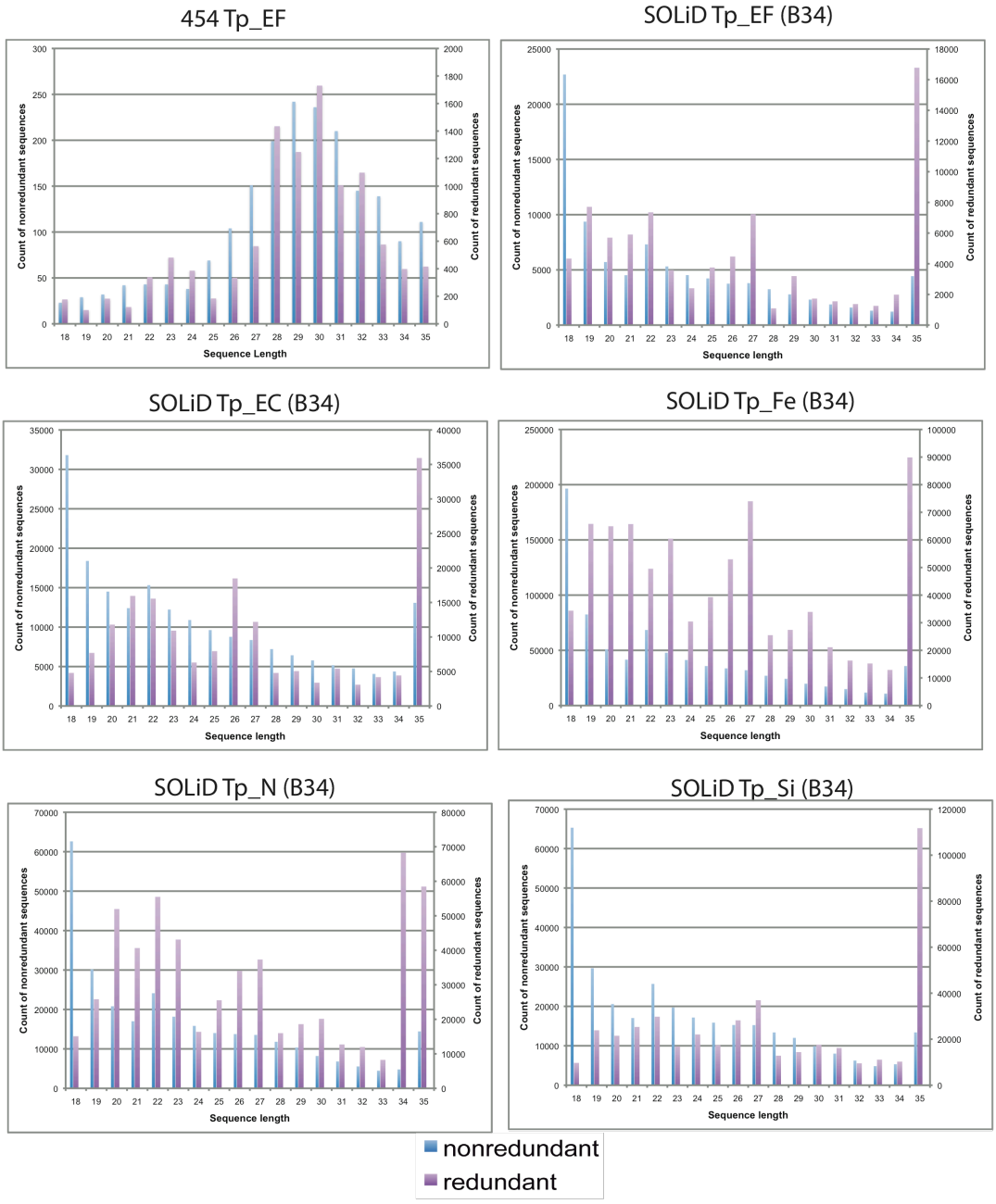

Figure 3.3, Continued. Length distribution of *T. pseudonana* small RNA candidate sequences

Length distribution was calculated after alignment with the *T. pseudonana* genome and removal of RNA degradation products.

B) Plots of 454 data and each of the barcoded condition in the SOLiD pool B, quadrants 3 and 4 (B34).

The length distribution profile for the 454 data was different from all of the SOLiD libraries, showing a large general peak between 27-31 nucleotides. This effect may be due to differences in the adapter properties or the size-selection methods in the experimental protocols between these two techniques.

Nucleotide frequency. The nucleotide present at the 5' end of a small RNA, which is most frequently a U in other organisms [24, 59], is important due to its link to sorting by the Argonaute protein [60]. The nucleotide frequency at the 5' end of the sequence did not demonstrate a consistent preference across all the SOLiD libraries (Figure 3.4). Additionally, the redundant and nonredundant sequences do not follow the same trend, even within the same SOLiD library. However, the SOLiD data illustrate the same general trend as the 454 data; that is, G was most frequently found at the 5' end, followed closely by U. In all cases, an A was found least often at the 5' end of the sequence.

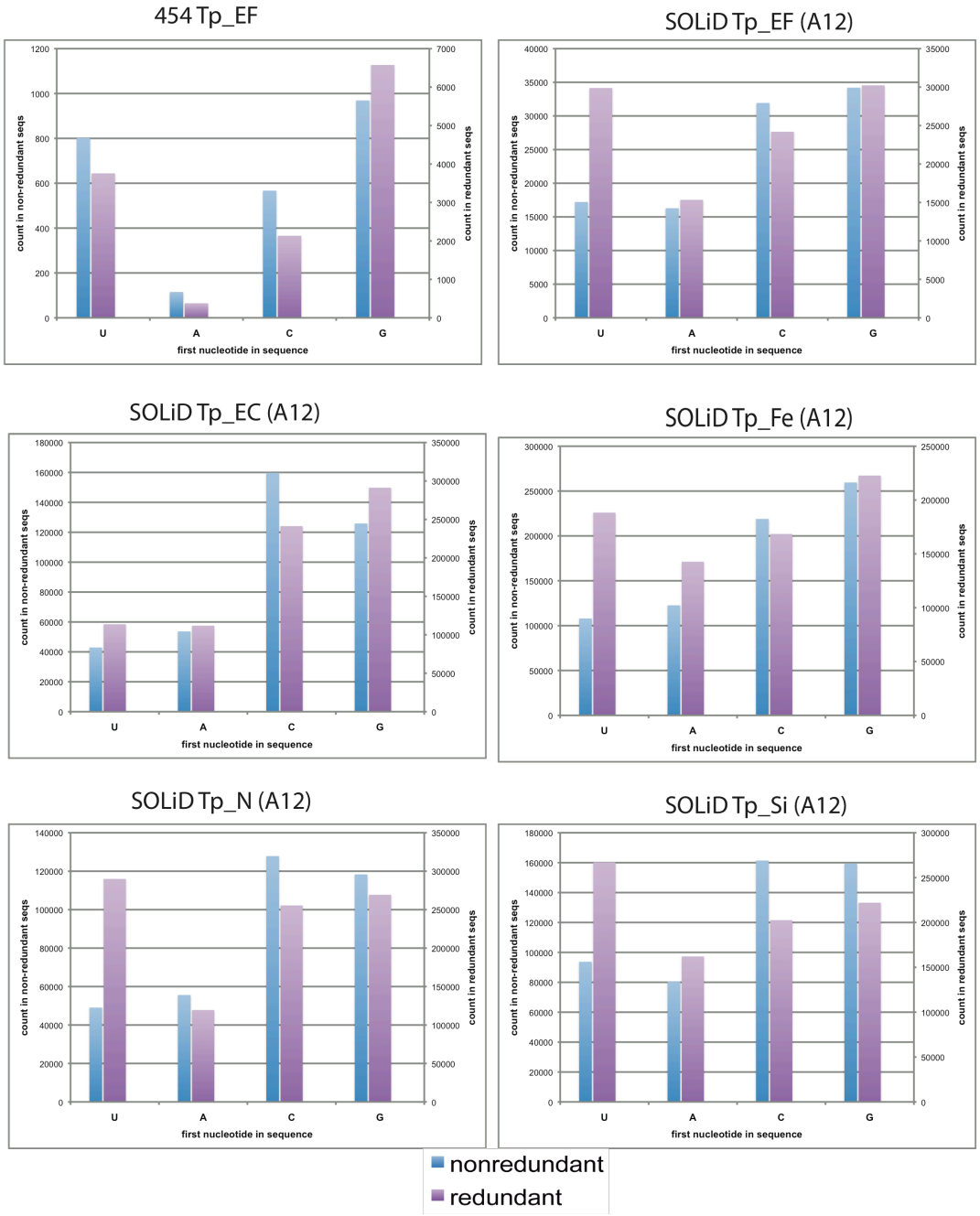

Figure 3.4. Nucleotide frequency at the 5' end of the small RNA candidate sequences

Nucleotide frequency was tabulated after alignment with the *T. pseudonana* genome and removal of RNA degradation products. Plots of 454 data and each condition in SOLiD pool A, quadrants 1 and 2 (A12).

Comparative mapping along the chromosomes. To explore the expression profile between the small RNA libraries, each set of sequences was binned, normalized, and represented in a heatmap plot along the length of the chromosomes (Figure 3.5). The intensity of the spot on a heatmap denotes the abundance of sequences generated at the particular site, with darker colors depicting higher abundance.

A cursory glance at the heatmaps allows an evaluation of the broad trends in the distribution. In particular, the small RNAs were not generated evenly along the chromosomes, but rather, were produced from certain hotspots. This is consistent with other small RNA studies [61] and with the *T. pseudonana* 454 data set. Additionally, the A12 and B34 pools of data appear extremely similar, implying that the techniques and adapter sets in the protocol for sequencing from the 3' and 5' ends of the small RNA were capable of capturing equivalent small RNAs for sequencing. More importantly, however, because the 454 data showed a similar profile to the SOLiD data, this gives strong evidence that there exists a core compliment of small RNAs expressed throughout all libraries at all times, during normal growth and during stress conditions.

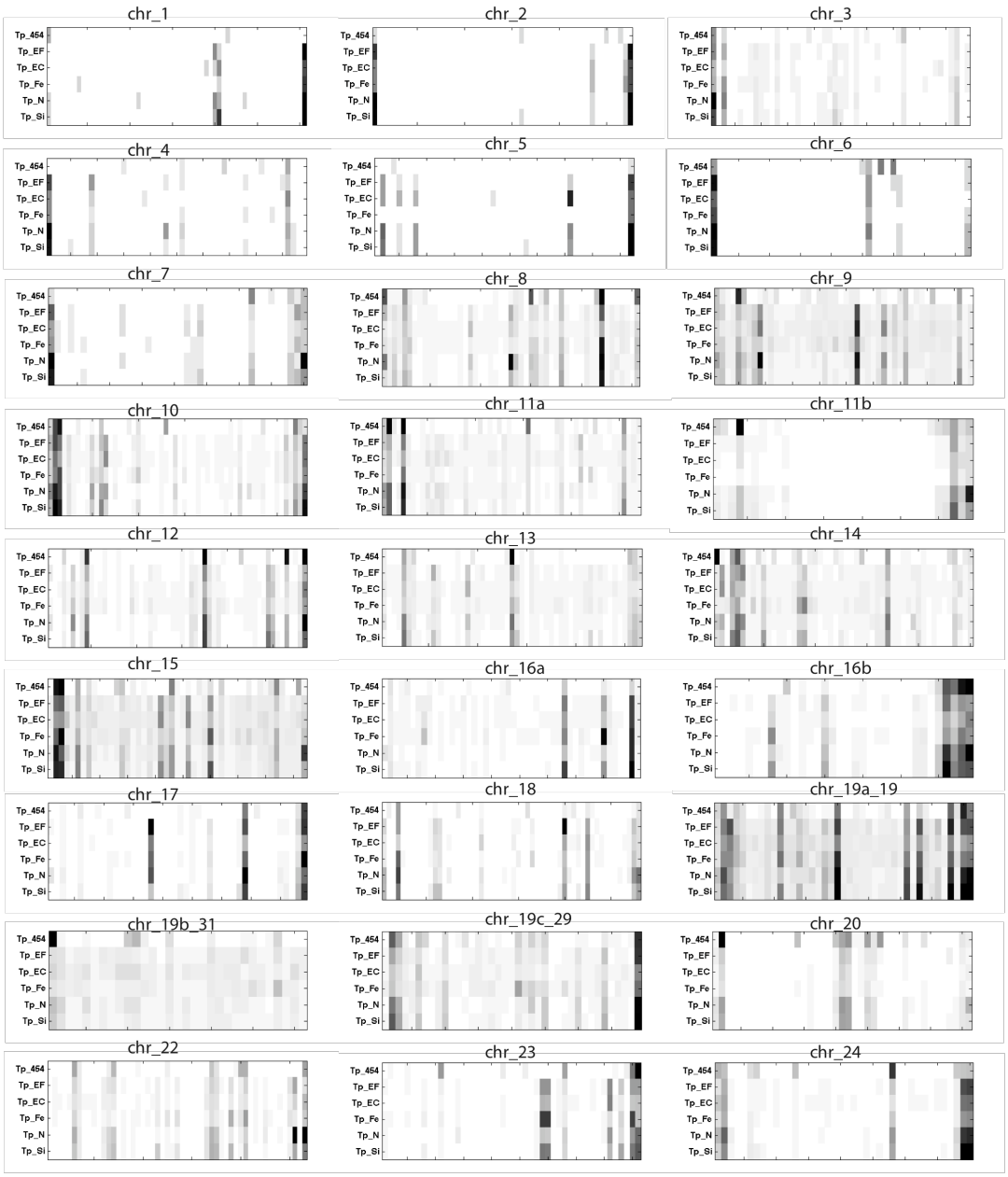

Figure 3.5. Heatmap representation of the small RNA candidate abundance mapped along the *T. pseudonana* chromosomes

All matches to the genome were binned, normalized, and then plotted along the length of the chromosome as a heatmap. Darker colors denote higher frequency counts at a particular location. Each row of the heatmap represents a different sample library in the following order: Tp_454data, Tp_EF, Tp_EC, Tp_Fe, Tp_N, and Tp_Si.

A) Heatmaps of all the SOLiD pool A12 barcoded sequences.

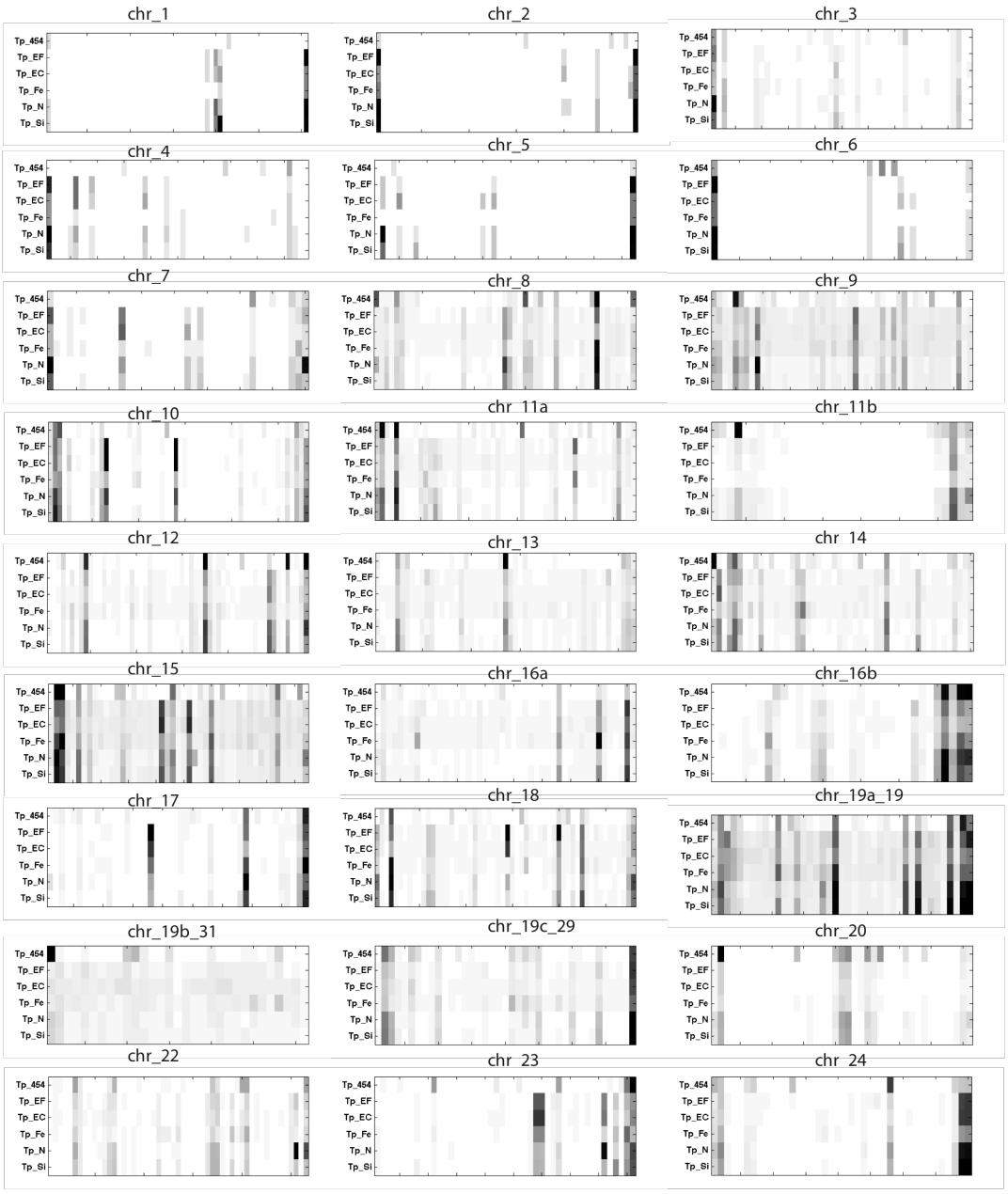

Figure 3.5, Continued. Heatmap representation of the small RNA candidate abundance mapped along the *T. pseudonana* chromosomes

All matches to the genome were binned, normalized, and then plotted along the length of the chromosome as a heatmap. Darker colors denote higher frequency counts at a particular location. Each row of the heatmap represents a different sample library in the following order: Tp_454data, Tp_EF, Tp_EC, Tp_Fe, Tp_N, and Tp_Si.

B) Heatmaps of all the SOLiD pool B34 barcoded sequences.

Upon closer examination of the heatmaps, the similarities and differences between the libraries are revealed. To illustrate this further, Figure 3.6 contains two chromosomes, chr_16a and chr_16b, that have almost identical expression across all libraries. The histograms provide a more detailed scale of the abundance. Figure 3.7, on the other hand, contains two chromosomes, chr_4 and chr_7, which show a patchwork of expression in areas other than the common hotspots. This disparity suggests that the cells under the various nutrient limitation conditions expressed a different repertoire of small RNA sequences.

Prediction of endogenous repeat-associated siRNA candidates. Repeat-associated endogenous siRNA are small RNAs that have the ability to silence repetitive elements in the genome [31]. A total of 122,470 unique sequences from the SOLiD data set mapped to repetitive regions of the *T. pseudonana* genome, which is 2.75% of the total unique sequences in the small RNA candidate pool. Over both the A12 and B34 pools of data, the trends were similar to each other for each condition type (Figure 3.8). For all of the SOLiD libraries, the LTR retrotransposon class, which contains Copia and Gypsy, was expressed in the highest abundance. The DNA transposons, Harbinger and MuDR, were expressed in lower abundance. This is to be expected, since the LTR retrotransposons comprise a larger proportion of the *T. pseudonana* genome [6, 11, 62]. This result is different from the 454 data, where the

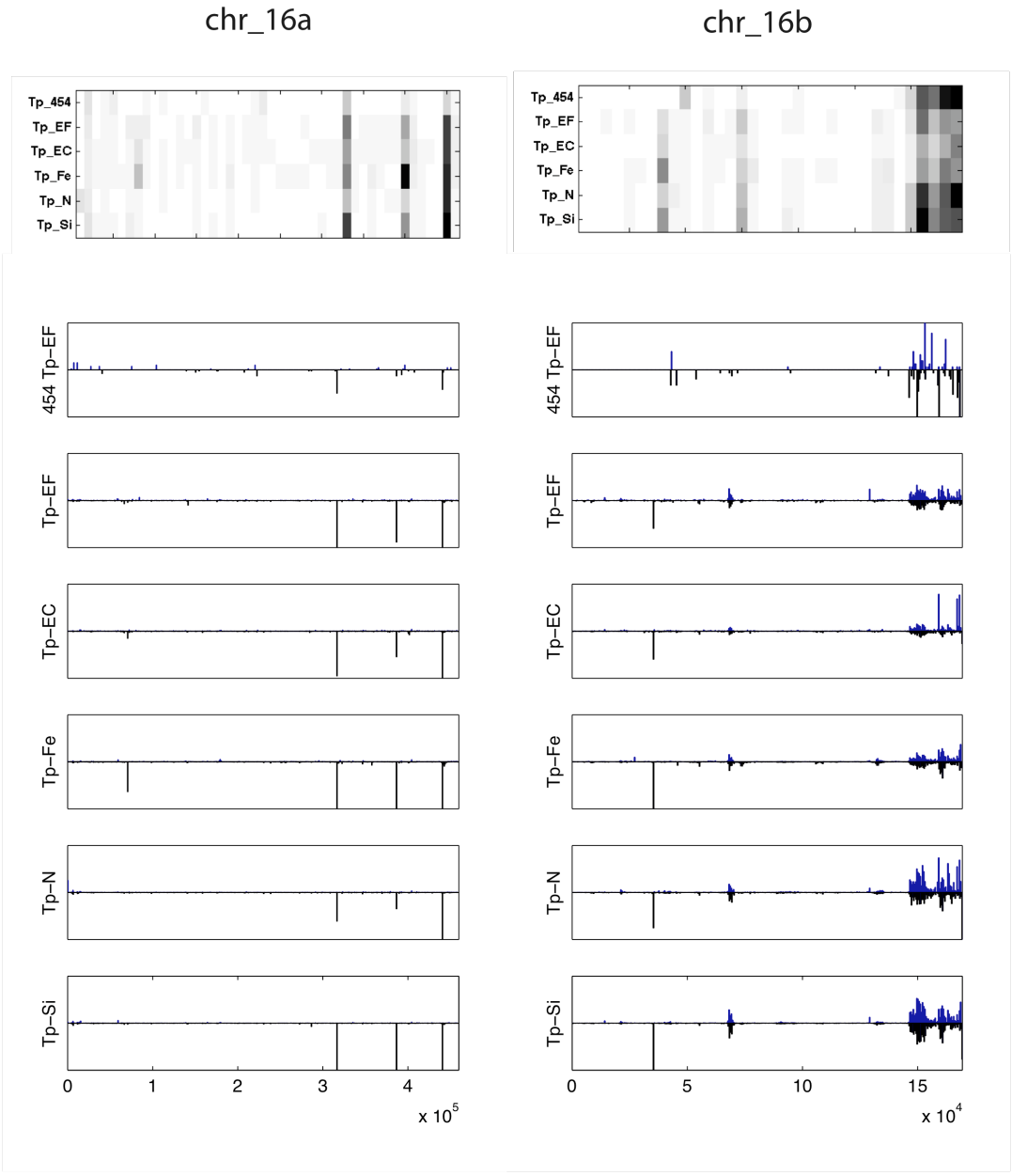

Figure 3.6. Heatmap and histogram representation of the small RNA candidate abundance mapped along the *T. pseudonana* chromosomes 16a and 16b to show an example of similarity between the libraries

All alignments to the genome were binned, normalized, and then plotted along the length of the chromosome as a heatmap and as a histogram. In the heatmap, darker colors denote higher frequency counts at a location. Each row of the heatmap represents a different sample library in the following order: Tp_454data, Tp_EF, Tp_EC, Tp_Fe, Tp_N, and Tp_Si. Bars above the line represent the plus strand and bars below the line represent the complimentary strand.

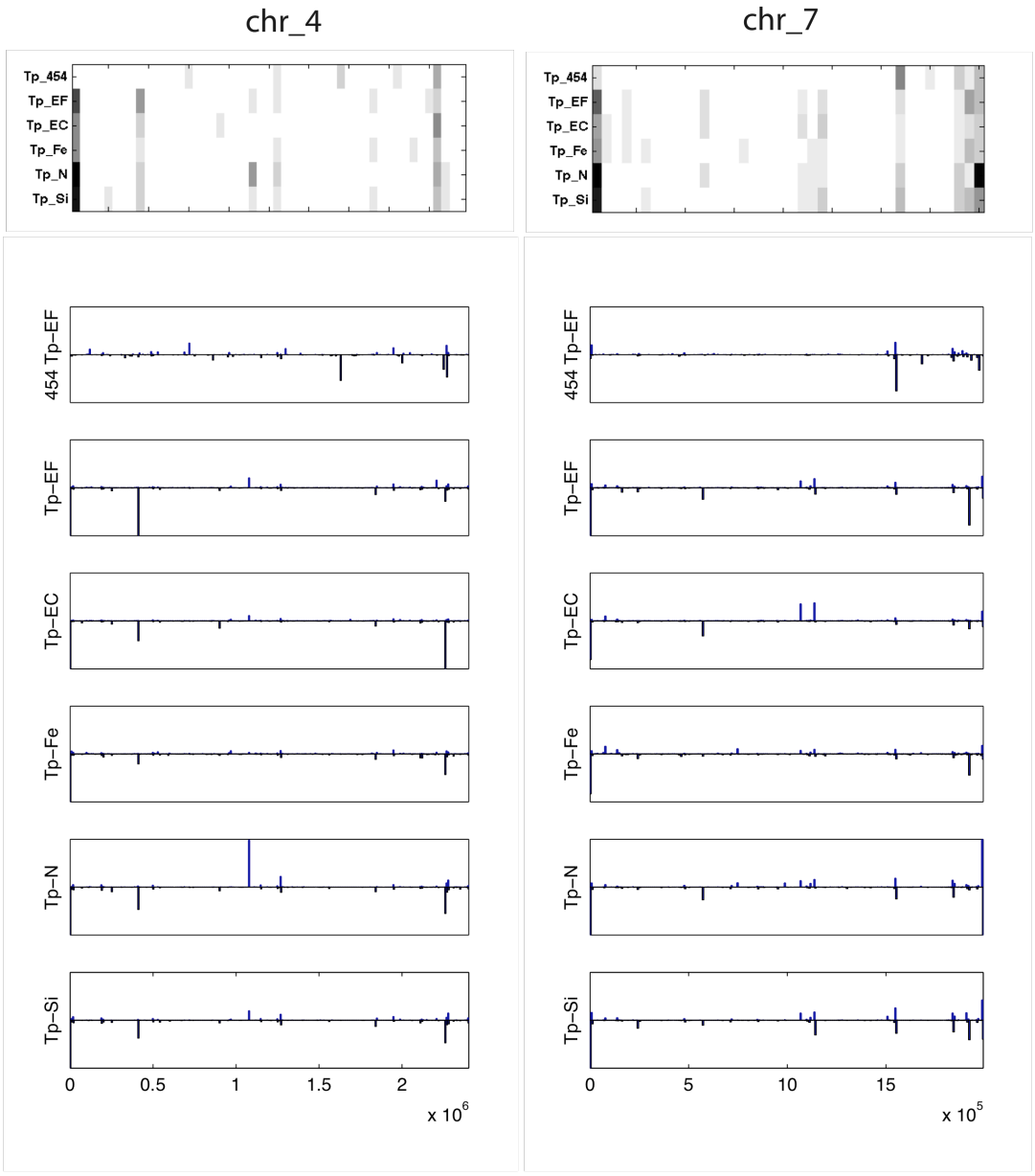

Figure 3.7. Heatmap and histogram representation of the small RNA candidate abundance mapped along the *T. pseudonana* chromosomes 4 and 7 to show an example of differences between the libraries

All matches to the genome were binned, normalized, and then plotted along the length of the chromosome as a heatmap and as a histogram. In the heatmap, darker colors denote higher frequency counts at a location. Each row of the heatmap represents a different sample library in the following order: Tp_454data, Tp_EF, Tp_EC, Tp_Fe, Tp_N, and Tp_Si. Bars above the line represent the plus strand and bars below the line represent the complimentary strand.

DNA transposons were expressed at a higher percentage than the LTR retrotransposons. This may be due to a bias in the processing or in the lower quantity of sequences in the 454 data set. That is, there were 812 unique repeat-associated candidates in the 454 data, as compared to 122,470 in the SOLiD data set.

Except for the silicon-starved library, all of the SOLiD libraries had approximately the same general percentages of small RNAs mapping to the LTR retrotransposons and the DNA transposons. That is, the common trend in the other libraries was 43-45% mapping to the LTR retrotransposons, and approximately 30% mapping to the DNA transposons. The silicon-starved library, on the other hand, demonstrated 50-51% mapping to the LTR retrotransposons, and 27-28% to the DNA transposons. Both the A12 and B34 quadrant pools of the data exhibited this trend in the silicon-starvation library.

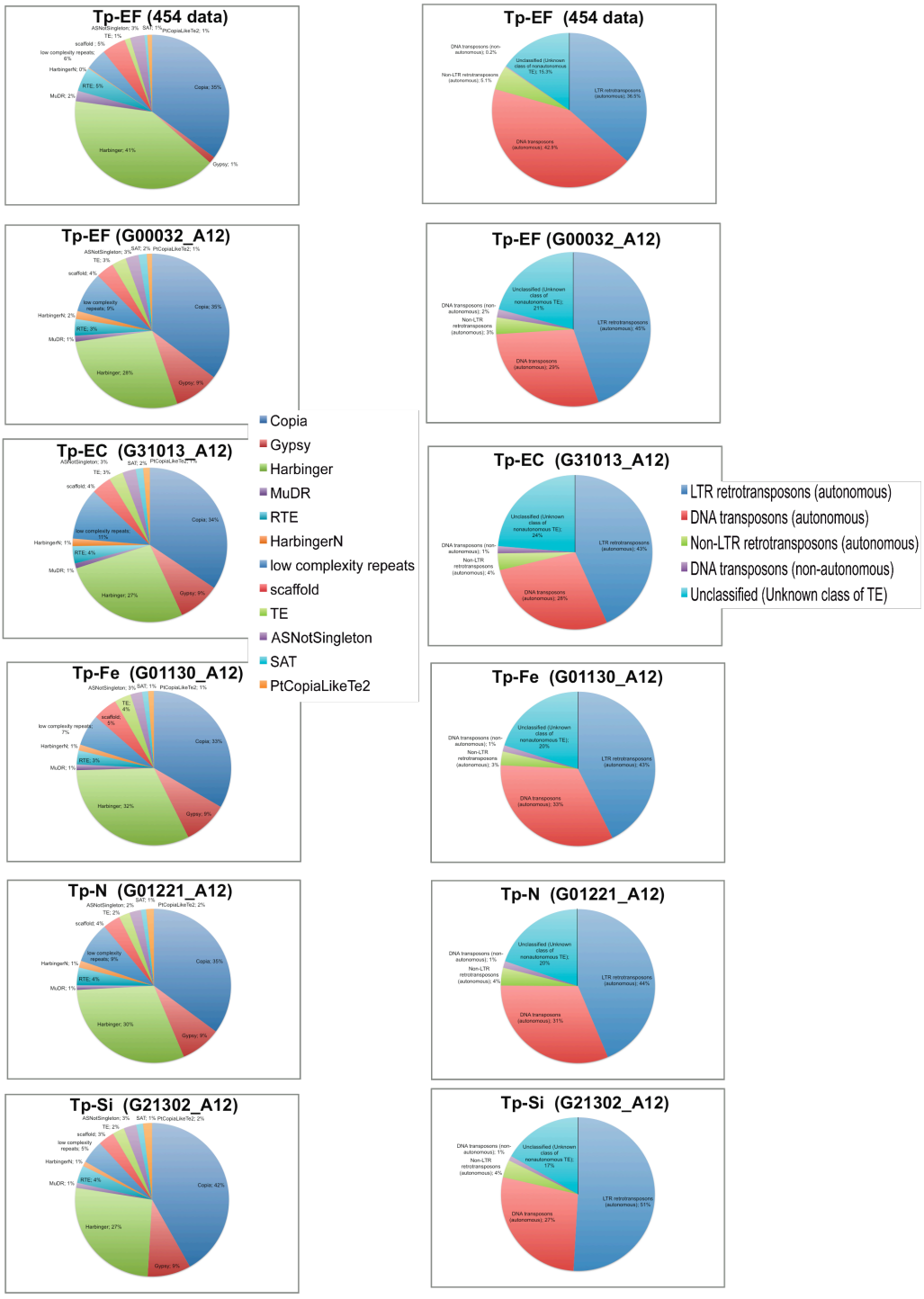

Figure 3.8. Percentage of small RNA sequences in each repetitive element class Left, percentage of small RNA sequences relative to specific subclasses of transposons. Right, percentage of small RNA sequences relative to general class of transposons

A) Pie charts of 454 data and all the SOLiD pool A12 barcoded sequences.

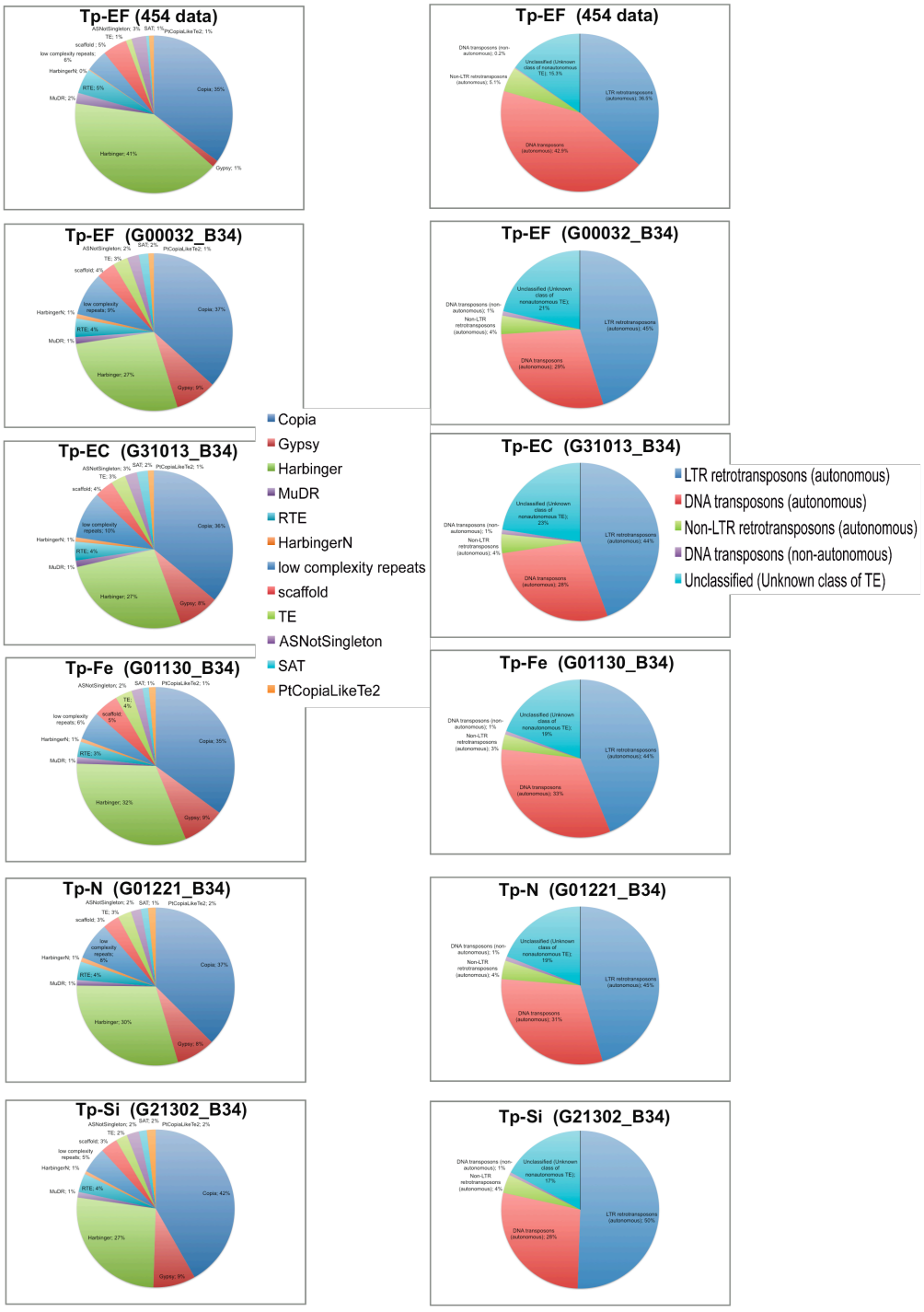

Figure 3.8, Continued. Percentage of small RNA sequences in each repetitive element class Left, percentage of small RNA sequences relative to specific subclasses of transposons. Right, percentage of small RNA sequences relative to general class of transposons

B) Pie charts of 454 data and all the SOLiD pool B34 barcoded sequences.

3.4 Discussion

In this study, ABI SOLiD sequencing was utilized to discover differential expression of small RNAs in the diatom, *Thalassiosira pseudonana,* under various nutrient stress conditions. Additionally, computation techniques were developed to perform a comparative analysis of the data, providing the ability to focus on the genome-wide expression profiles of the small RNAs.

The similarities between the expression profiles of the *T. pseudonana* small RNA libraries confirm that the common small RNAs discovered in this study are valid, since they were generated from multiple cell samples, as biological replicates, and with two different protocols and sequencing approaches, as technical replicates. Virtually the entire set of small RNAs characterized in *T. pseudonana* by 454 sequencing was also present in the SOLiD libraries (Table 3.4). The agreement between the libraries is most clearly visible as hotspots of transcription for the small RNAs in the histogram and heatmap distribution plots along the chromosomes (Figures 3.5, 3.6, and 3.7). An overall similarity between the libraries for the predicted repeat-associated siRNAs can also be observed (Figure 3.8). These features demonstrate the existence of a core group of small RNAs that is constitutively expressed in the diatom, since these small RNAs are present during normal exponential growth as well as in nutrient stress conditions. This observation agrees with previous differential expression studies in other organisms, which have

demonstrated that, although some small RNAs show differential expression according to developmental stage, specific tissue, or stress condition, there are many small RNAs that show a broad expression across all conditions [38, 40, 45, 56].

The overall trends in the length distribution across the libraries were accentuated by the abundance of SOLiD sequence data. In particular, a distinct peak at 21 nucleotides and 27 nucleotides in the redundant sequence set could be strongly distinguished in the exponential growth, nitrogen-starved, and iron-starved libraries. Small RNAs of ~21 nucleotides are the characteristic length of miRNAs, but can also comprise rasiRNAs and SAS siRNAs. Small RNAs of 27 nucleotides are more typically rasiRNAs and SAS siRNAs [31] The nonredundant sequences in all libraries showed less pronounced peaks, implying there may be a subset of highly expressed small RNAs dominating certain size classes under the nutrient stress conditions. The length profile for the silicon-starved library did not exhibit the same clear trend in its redundant sequence set, and therefore, appears to contain a different profile of expression than the other libraries.

Similar to the results found with 454 data, the SOLiD sequences showed that the 5' nucleotide was most commonly G, with U as the second most frequent. In all cases, A was the least likely nucleotide to exist in the 5' position. In most small RNA studies to date, U is the most prevalent character in the 5' position [24, 59]. The 5' nucleotide has been linked to sorting by the Argonaute protein [60, 63, 64], so the G

may be related to differences in the diatom Argonaute specificity. Variations between the libraries may also be due to biases in the SOLiD protocol. In a study reported by Applied Biosystem [65] using small RNA isolated from human cell samples with the ABI SOLiD Small RNA Expression kit, it was found that a significant number of miRNAs had 5' ends that were different from those listed in the Sanger miRBase. This result was attributed in the study to a possible "permissive processing in miRNA regulation". More research may be necessary to confirm this proposed mechanism.

Differences in the chromosomal spatial distribution between the libraries can readily be observed as a patchwork of intensities in the heatmaps (Figures 3.5, 3.6, and 3.7). Without further analysis, however, it is difficult to make any global statements about the correlations between the stress conditions. That is, in the spatial distribution along all of the chromosomes, the expression profile is not consistent between the libraries. This result is not surprising. In the whole genome tiling array study [19], only 709 genes of the 11,390 predicted genes in the genome were differentially expressed. Of those 709 genes, only 89 genes, which is less than 1% of the total genes in the organism, showed similar expression between the silicon and iron starvation libraries. Therefore, correlations in small RNA expression trends may also display these very subtle trends.

In the repeat-associated siRNA analysis, the percentage of the different types of transposons was similar, with the exception of the silicon-starved library, which

contained a 5-10% higher expression of LTR retrotransposons, Copia and Gypsy. Silicon starvation has been shown to arrest the cell cycle [17], and modify the morphology of the girdle bands of the frustule of the diatom [66]; however, most other cellular metabolic processes are able to continue [16]. Therefore, it is possible these repeat-associated siRNAs are involved in specific silicon-related metabolic processes and are activated to a different degree during the other nutrient stress conditions. This case is also very interesting because silicon metabolism in the diatom is not involved in other cellular metabolic processes (Claqin 2002), so the gene regulatory pathway may be more readily discerned.

The SOLiD and 454 data sets displayed differences in the expression of the repeat-associated siRNAs, both in the percentages mapped to certain types of repetitive elements and in the overall prevalence. The SOLiD library expression trend, with the highest expression of LTR retrotransposons, is as would be expected, since LTR retrotransposons are the most abundant in the *T. pseudonana* genome [6, 62]. Therefore, one would expect that the highest number of small RNAs would exist to regulate them. Another difference between the 454 and SOLiD data for the repeat-associated siRNAs involved the prevalence of candidates derived from repetitive regions of the genome. For the 454 data, 15.7% of the small RNA candidates were derived from repetitive regions, but for the SOLiD data, the amount was 2.75%. The SOLiD data value corresponds well with the estimate that 2% of the *T. pseudonana* genome is comprised of repetitive sequence [6, 62].

The differences between the small RNA libraries demonstrate that nutrient stress conditions induce differential expression of small RNAs in the diatom. This result consistent with other observations, since differential expression of mRNAs, as well as miRNAs and siRNAs, occurs in other organisms [24, 25, 38]. This finding is significant because it is the first reported differential expression of small RNAs in a diatom, organisms that play pivotal roles in nutrient cycling and productivity in the oceans as well as in global carbon fixation. Diatoms are subject to substantial changes in environmental conditions, and having a means to regulate gene expression rapidly, such as through the intermediary of small RNAs, is likely to play a critical role in diatoms' adaptability. In addition, this study may aid in elucidating the mechanisms that underlie differential gene expression in diatoms. In particular, it has been found in multicellular organisms that the preferential expression of small RNAs in specific tissues is linked to their function. In a study of small RNAs in rice, miRNAs that were preferentially expressed in the roots were orthologous to miRNAs involved in root development in *Arabidopsis thaliana* [40]. This suggests that correlations found in expression may be linked to their function, thereby aiding in clarifying the mechanisms of cellular processes underlying the responses to nutrient limitation stress. In addition to possible correlations between the small RNAs in *T. pseudonana*, this data enables comparison between the suite of small RNAs and mRNAs expressed under similar conditions. The existence of *T. pseudonana* EST data produced from stress libraries [7, 67] and the availability of the whole genome *T. pseudonana* tiling

array data [19], creates the opportunity to mine this data for the presence of genes showing correlated expression. mRNA demonstrating differential regulation in *T. pseudonana* gene expression studies under the same nutrient stress conditions may be strong candidates for future validation as targets of the small RNAs.

The differences observed between the SOLiD and 454 sequence data sets are expected due to potential biases in the experimental and sequencing approaches. In the sample preparation prior to 454 sequencing, the total RNA was treated with a PEG/NaCl precipitation step followed by PAGE size-selection to reduce the rRNA degradation product content [68]. In the ABI SOLiD protocol, on the other hand, the total RNA was eluted through a flashPAGE device to achieve this same result. The two protocols also used different ligation adapters to selectively bind to the small RNAs [28]. The sequencing technologies are vastly different, whereby 454 involves pyrosequencing in picotiter plates [69], and SOLiD involves sequencing-by-ligation on a support [48]. The data produced by these two different sequencing platforms also requires different manipulation and analysis procedures. In 454 sequencing, the bases are determined individually, while in SOLiD sequencing, it is necessary for the base-calling procedure to interpret one color as a 2-base pair. This SOLiD colorspace data is converted to its equivalent base space format, and then mapped to a reference genome before further analysis can be performed. Finally, the quantity of data produced by 454 sequencing is magnified by approximately 100-fold compared with SOLiD. The most recent estimates claim that a typical 454 sequencing run generates

600 Mb of data [70]([www.454.com]), while SOLiD generates over 60 Gb of sequence per run [71]([http://www.appliedbiosystems.com]). Although these differences in techniques can be used to explain disparities between the 454 and SOLiD data, these biases also serve to strengthen the validity of general trends exhibited across both sets.

Computationally, the abundance of data produced from the SOLiD sequencing has both advantages and disadvantages. The massive data set is beneficial because the increase in data accentuates the trends in the data. For example, the length distribution in the libraries has sharp peaks at 21 and 27 nucleotides, which were not visible in the 454 data set because there was approximately 1000 times less data. It was also beneficial to sequence from both the 3' and 5' ends of the small RNAs, allowing greater confidence in the results. The striking agreement between the A and B pools of data show that the protocol was able to capture the same repertoire of small RNAs. The disadvantages of the large sequence data set include lengthier computational processing times and the requirement of more storage space. Additionally, in our study, we found that approximately 80% of the data did not align with the genome and had to be discarded. These disadvantages will likely become insignificant as the technologies to collect and process the data mature.

The methodology presented in this study provides the steps necessary to discover small RNA genes in next generation sequence data, and to perform a

comparative analysis of differential expression. Unlike the ABI SOLiD Small RNA data analysis pipeline, the method described here contains no filtering of the data by Sanger miRBase [53, 54], thereby freeing the analysis to pursue all types of small RNAs. By using the BLAST program [58] to align the reads to the genome, all length classes are represented and all locations of the matches are collected, while trimming the adapters from the ends of the reads. Additionally, this method assembled, on average, two or three times more reads to the genome than CLCbio's NGS Cell program, thereby producing a large data set for further analysis.

This study demonstrates that there is differential expression of small RNAs in the diatom *T. pseudonana* under different nutrient stress conditions. The methodology used in this study proves that SOLiD sequencing can be used to investigate all known types of small RNAs in a nonstandard model organism. These computational techniques can be used for any organism that has a sequenced genome. The results presented here pave the way for further elucidation of the gene regulatory pathways by small RNAs. By choosing a model organism such as the diatom for these studies, the impact of any discovery is amplified in the implications to environmental issues. In particular, it may aid in the understanding of the biological control processes of lipid synthesis, carbon fixation, and silicon utilization in the diatom.

3.5 Materials and Methods

Experimental Methods

Cell culture. *Thalassiosira pseudonana* strain CCMP1335 was obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton, Bigelow Laboratory for Ocean Sciences (West Boothbay Harbor, ME, USA), and maintained in artificial seawater (ASW) medium [16], supplemented with biotin and vitamin B_{12}, each at 1 ng · L^{-1}. Cell cultures were maintained at 18˚C – 20˚C in continuous light at an intensity of 150 µmol photons · m^{-2} · s^{-1}. Cultures were magnetically stirred and aerated using sterile techniques.

Tp-EC: Exponential growth Centrifugation cell culture. The *T. pseudonana* exponential growth culture (Tp-EC) was grown in an 8-liter glass bottle of ASW to a density of 1.14 x 10^6 cells · ml^{-1}. This sample was different from all the other conditions, because repeated centrifugation, instead of filtration, was used to harvest the cells. In this harvesting method, the cells were centrifuged at 4,000 x g for 14 minutes, and the supernatant was decanted. The cell pellet was transferred in a minimum volume of 3.5% NaCl, and centrifuged for an additional 5 minutes at 4,000 x g. The liquid was aspirated from the pellet, and then the pellet was frozen and stored at -80 C.

In all of the remaining cell culture conditions, filtration followed by centrifugation was used for harvesting the cells. Cells were concentrated with filtration using a 2 μm (142 mm) polycarbonate membrane (GE Osmonics #K20CP14220) filter, rinsed from the filter and harvested via centrifugation at 4,000 x g for 5 min.

Tp-EF: Exponential growth Filtration cell culture. The *T. pseudonana* exponential growth culture (Tp-EF) was grown in an 8-liter glass bottle of ASW to a density of 1.8 x 10^6 cells · ml^{-1}. This cell culture sample was also used in a previous study to construct a small RNA cDNA library for 454 pyrosequencing [28]

Tp-N: Nitrogen-starved cell culture. The *T. pseudonana* nitrogen-starved culture (Tp-N) was grown in a 25 ml volume of ASW that was adjusted to a 50uM concentration of nitrate (KNO_3). The cells were transferred three times to deplete the cells' internal stores of nitrogen. The cells from the transfer were then used to inoculate an 8-liter glass bottle of ASW which was adjusted to a 25 uM concentration of nitrate (KNO_3). The cells were grown to a density of 2.1 x 10^6 cells · ml^{-1} and harvested.

Tp-Fe: Fe-starved cell culture. The *T. pseudonana* iron-starved culture (Tp-Fe) was grown in a 25 ml volume of ASW that was adjusted to 50 nM Fe. The cells

were transferred six times before inoculating 8-liters of ASW media at 50 nM Fe. The cells were grown to a density of 1.47 x 10^5 cells · ml^{-1} and harvested.

Tp-Si: Silicon-starved cell culture. The *T. pseudonana* silicon-starved culture (Tp-Si) was prepared as previously described [17]. Briefly, the cells were grown in an 8-liter glass bottle of ASW to a density of 2.66 x 10^6 cells · ml^{-1}. The cells were harvested by centrifugation, and then resuspended in silicate-free media (Si-) in an 8-liter polycarbonate bottle. The cells were stirred and aerated for 24 hours, and were then harvested at a density of 0.87 x 10^6 cells · ml^{-1}.

Small RNA cDNA library construction. Total RNA from *T. pseudonana* cell cultures was extracted with TriReagent (Sigma) as previously described [72]. To enrich for the ~18-40 nt small RNA fraction, the total RNA samples were treated with the flashPAGE fractionator (Ambion #AM13100) and flashPAGE Clean-Up Kit (Ambion #AM12200). The small RNA enriched samples were then processed according to the SOLiD Small RNA Expression Kit protocol (Applied Biosystems, Life Technologies, #4399443).

The first step of the SOLiD Small RNA Expression Kit protocol involved a ligation step with Adapter Mix A or Adapter Mix B using RNA ligase, which requires an RNA with a 5' phosphate and a 3' hydroxyl group, as is characteristic of small RNAs. Adapter Mix A is used for SOLiD sequencing starting from the 5' ends of the

small RNA. Adapter Mix B is used for sequencing starting from the 3' end. We prepared both types of ligated product, A and B, separately for the RNA from each cell culture condition. Preparation of both orientations of sequences was performed to allow for the possibility of detecting small RNAs larger than 35 nucleotides by sequencing from both ends of the small RNA.

Each condition was barcoded by amplifying the library with a particular PCR primer set provided in the kit, where each primer set differed by a known 6-nucleotide sequence. Therefore, after sequencing, the samples could be identified and sorted into barcoded condition sequences. The barcode numbering scheme was as follows: Tp_EF was barcoded G00032, Tp_EC was barcoded G31013, Tp_Fe was barcoded G01130, Tp_N was barcoded G01221, and Tp_Si was barcoded G21302.

The quality and quantity of the samples was verified on the Agilent Bioanalyzer. Figure 3.1 contains the Agilent Bioanalyzer gel representation of the small RNA libraries. Approximately 200 ng of each condition and adapter mix was prepared for Applied Biosystems SOLiD next generation high throughput sequencing at the J. Craig Venter Institute in Rockville, MD. Duplicates of each sample were run on two quadrants of the slide. That is, the first quadrant (A1) and second quadrant (A2) both contained a sample of the Adapter A mix preparation. Similarly, the third quadrant (B3) and fourth quadrant (B4) both contained a sample of the Adapter B mix preparation.

Computational Methods

Data Files. Computational analysis of the SOLiD small RNA sequence data was performed with the *Thalassiosira pseudonana* genome, version 3.0 [73]([http://genome.jgi-psf.org/Thaps3/Thaps3.home.html][73]). The unmasked version of the genome was used in the study. The coordinates for the repetitive regions of the *T. pseudonana* genome were obtained from JGI as a RepeatMasker file.

The full slide of ABI SOLiD sequencing generated 153,812,217 sequences, which were sorted by slide region (A1, A2, B3, B4) and barcode (G00032, G31013, G01130, G01221, G21302) and deposited into colorspace formatted files.

CLC processing of SOLiD data. The first attempt to process the SOLiD data was via reference assembly to the *T. pseudonana* genome using CLC NGS Cell software (CLCbio). The ungapped, colorspace, short assembly parameters of the software (clc_ref_assemble_short –c –u) were utilized.

Processing of SOLiD data. Figure 3.2 contains a flow chart of the computational analysis steps that were performed for the SOLiD sequence data.

The SOLiD colorspace data was converted to basespace data using CLCbio's tofasta program. This program does not reference a genome, so color errors were not detected. Additionally, all SOLiD sequencing reads in the data set were 35

nucleotides, even though the small RNA insert could be any length. Therefore, to remove sequences containing color errors, and to trim off extra nucleotides surrounding the small RNA insert, a methodology was developed in this study using BLAST [58] alignment to the *T. pseudonana* genome. Custom perl code was written to parse the BLAST output and only retain sequences which matched the genome perfectly for 100% identity regardless of length, or which had 1-mismatch to the genome. By retaining the coordinates of the match to extract the sequences from the data file, the sequences could be trimmed to the exact length of the small RNA insert. Sequences that did not match the genome were removed from further consideration. To reduce redundancy, the sequences were subjected to clustering using the program CD-hit ([74, 75][http://www.bioinformatics.org/cd-hit/]) at 100% identity for a length similarity of at least 80%.

RNA degradation products, such as degraded rRNA and tRNA, were removed from the pool using the locations of these entities as determined in a previous study [28]. The sequences were separated into non-redundant versus redundant sequences. The lengths and first nucleotide of each of the sequences was also tabulated.

A Matlab program was modified from a previous study in order to handle multiple data sets [28]. The program was utilized to place the location and number of occurrences of the sequences into bins along each chromosome, along with their presence in either the plus or minus strands. A second Matlab program was written to

visualize the frequency distribution of the data along the chromosomes as a heatmap, to accentuate the similarities and differences between the conditions. In both the histograms and heatmaps, the data was normalized by dividing the frequency counts in the bins by the total number of sequences for that particular condition. The histograms and heatmaps were manually examined to determine trends in the data set.

Comparison of SOLiD data to 454 small RNA sequence data. To validate the SOLiD sequence data, the sequence locations were compared to those of a *T. pseudonana* small RNA library produced by 454 pyrosequencing in a previous study [28]. The percentage of the 454 data sequences that were found in the SOLiD library were calculated over all barcoded conditions. Additionally, all figures in this current study include the 454 data results for comparative analysis as a biological and technical replicate.

Prediction of endogenous repeat-associated siRNA candidates. Repeat-associated endogenous siRNA candidates were identified by aligning the small RNA sequence data with the RepeatMasker repetitive elements [76] in the *T. pseudonana* genome.

3.6 Acknowledgements

Funding for Trina Norden-Krichmar was provided by a National Science Foundation Graduate Research Fellowship and by the graduate department at Scripps Institution of Oceanography. Thanks to Mark Hildebrand for providing assistance, supplies, and the laboratory space necessary for growing the *T. pseudonana* cultures. Thanks to Andy Allen for providing assistance and funding for the ABI SOLiD small RNA library construction and sequencing at the J. Craig Venter Institute (JCVI). The ABI SOLiD small RNA libraries were constructed by Jing Bai at JCVI. The *T. pseudonana* iron-starved cell culture was grown by Hong Zheng at JCVI. Initial code for barcode removal of the SOLiD data was provided by Jamison McCorrison at JCVI in Rockville, MD.

3.7 References

1. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P: **Primary production of the biosphere: integrating terrestrial and oceanic components**. *Science* 1998, **281**(5374):237-240.

2. Falkowski PG, Barber RT, Smetacek VV: **Biogeochemical Controls and Feedbacks on Ocean Primary Production**. *Science* 1998, **281**(5374):200-207.

3. Sheehan J, Dunahay, T., Bennemann, J., and Roessler, P.: **A Look Back at the U.S. Department of Energy's Aquatic Species Program: Biodiesel from Algae. Close Out Report.** In.: U.S. Department of Energy's Office of Fuels Development, National Renewable Energy Laboratory; 1998.

4. Round FE, Crawford, R.M., and Mann, D.G.: **The diatoms: Biology and morphology of the genera**: Cambridge University Press; 1990.

5. Sandhage KH, Allan SM, Dickerson MB, Gaddis CS, Shian S, Weatherspoon MR, Cai Y, Ahmad G, Haluska MS, Snyder RL *et al*: **Merging biological self-assembly with synthetic chemical tailoring: The potential for 3-D genetically engineered micro/nano-devices (3-D GEMS)**. *International Journal of Applied Ceramic Technology* 2005, **2**(4):317-326.

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

7. **The Diatom EST Database [http://www.biologie.ens.fr/diatomics/EST]**.

8. Bowler C, Vardi, A., and Allen, A. E.: **Oceanographic and Biogeochemical Insights from Diatom Genomes**. *Annual Review of Marine Science* 2010, **2**:429-461.

9. Allen AE, Vardi A, Bowler C: **An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms**. *Curr Opin Plant Biol* 2006, **9**(3):264-273.

10. Vardi A, Formiggini F, Casotti R, De Martino A, Ribalet F, Miralto A, Bowler C: **A stress surveillance system based on calcium and nitric oxide in marine diatoms**. *PLoS Biol* 2006, **4**(3):e60.

11. Maumus F, Allen, A. E., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M., and Bowler, C.: **Potential impact of stress activated retrotransposons on genome evolution in a marine diatom**. *BMC Evolutionary Biology (in review)* 2009.

12. Behrenfeld MJ, Bale AJ, Kolber ZS, Aiken J, Falkowski PG: **Confirmation of iron limitation of phytoplankton photosynthesis in the equatorial Pacific Ocean**. *Nature* 1996, **383**(6600):508-511.

13. Allen AE, Laroche J, Maheswari U, Lommer M, Schauer N, Lopez PJ, Finazzi G, Fernie AR, Bowler C: **Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation**. *Proc Natl Acad Sci U S A* 2008, **105**(30):10438-10443.

14. Kroger N, Poulsen N: **Diatoms-From Cell Wall Biogenesis to Nanotechnology**. *Annual Review of Genetics* 2008, **42**:83-107.

15. Coombs J, Halicki PJ, Holmhans.O, Volcani BE: **Studies On Biochemistry And Fine Structure Of Silica Shell Formation In Diatoms .2. Changes In Concentration Of Nucleoside Triphosphates In Silicon-Starvation Synchrony Of Navicula Pelliculosa (Breb) Hilse**. *Exp Cell Res* 1967, **47**(1-2):315.

16. Darley WM, Volcani BE: **Role of silicon in diatom metabolism. A silicon requirement for deoxyribonucleic acid synthesis in the diatom Cylindrotheca fusiformis Reimann and Lewin**. *Exp Cell Res* 1969, **58**(2):334-342.

17. Hildebrand M, Frigeri LG, Davis AK: **Synchronized growth of Thalassiosira pseudonana (Bacillariophyceae) provides novel insights into cell-wall synthesis processes in relation to the cell cycle**. *Journal of Phycology* 2007, **43**(4):730-740.

18. Claquin P, Martin-Jezequel V: **Uncoupling of silicon compared with carbon and nitrogen metabolisms and the role of the cell cycle in continuous cultures of Thalassiosira pseudonana (Bacillariophyceae) under light, nitrogen, and phosphorus control**. *Journal of Phycology* 2002, **38**(5):922-930.

19. Mock T, Samanta MP, Iverson V, Berthiaume C, Robison M, Holtermann K, Durkin C, Bondurant SS, Richmond K, Rodesch M *et al*: **Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses**. *Proc Natl Acad Sci U S A* 2008, **105**(5):1579-1584.

20. Marchetti A, Harrison PJ: **Coupled changes in the cell morphology and the elemental (C, N, and Si) composition of the pennate diatom Pseudo-nitzschia due to iron deficiency**. *Limnology and Oceanography* 2007, **52**(5):2270-2284.

21. Davis AK, Hildebrand M, Palenik B: **Gene expression induced by copper stress in the diatom Thalassiosira pseudonana**. *Eukaryot Cell* 2006, **5**(7):1157-1168.

22. Bartual A, Galvez LA: **Growth and biochemical composition of the diatom Phaeodactylum tricornutum at different pH and inorganic carbon levels under saturating and subsaturating light regimes**. *Botanica Marina* 2002, **45**(6):491-501.

23. Dittami SM, Scornet D, Petit JL, Segurens B, Da Silva C, Corre E, Dondrup M, Glatting KH, Konig R, Sterck L *et al*: **Global expression analysis of the brown alga Ectocarpus siliculosus (Phaeophyceae) reveals large-scale reprogramming of the transcriptome in response to abiotic stress**. *Genome Biol* 2009, **10**(6):R66.

24. Bartel DP: **MicroRNAs: Genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.

25. Bartel DP, Chen CZ: **Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs**. *Nature Reviews Genetics* 2004, **5**(5):396-400.

26. Kloosterman WP, Plasterk RH: **The diverse functions of microRNAs in animal development and disease**. *Dev Cell* 2006, **11**(4):441-450.

27. De Riso V, Raniello R, Maumus F, Rogato A, Bowler C, Falciatore A: **Gene silencing in the marine diatom Phaeodactylum tricornutum**. *Nucleic Acids Res* 2009, **37**(14):e96.

28. Norden-Krichmar TM, Allen, A. E., Gaasterland, T., and Hildebrand, M.: **Characterization of the small RNA transcriptome of the diatom, Thalassiosira pseudonana**. *(submitted)* 2009.

29. Poulsen N, Kroger N: **A new molecular tool for transgenic diatoms: control of mRNA and protein biosynthesis by an inducible promoter-terminator cassette**. *Febs J* 2005, **272**(13):3413-3423.

30. Thamatrakoln K, Hildebrand M: **Analysis of Thalassiosira pseudonana silicon transporters indicates distinct regulatory levels and transport activity through the cell cycle**. *Eukaryot Cell* 2007, **6**(2):271-279.

31. Kim VN: **Small RNAs: Classification, biogenesis, and function**. *Mol Cells* 2005, **19**(1):1-15.

32. Ambros V, Chen XM: **The regulation of genes and genomes by small RNAs**. *Development* 2007, **134**(9):1635-1641.

33. Beiter T, Reich E, Williams RW, Simon P: **Antisense transcription: a critical look in both directions**. *Cell Mol Life Sci* 2009, **66**(1):94-112.

34. Molnar A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC: **miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii**. *Nature* 2007, **447**(7148):1126-1129.

35. Zhao T, Li G, Mi S, Li S, Hannon GJ, Wang XJ, Qi Y: **A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii**. *Genes Dev* 2007, **21**(10):1190-1203.

36. Stefani G, Slack FJ: **Small non-coding RNAs in animal development**. *Nat Rev Mol Cell Biol* 2008, **9**(3):219-230.

37. Croce CM, Calin GA: **miRNAs, cancer, and stem cell division**. *Cell* 2005, **122**(1):6-7.

38. Axtell MJ, Bartel DP: **Antiquity of microRNAs and their targets in land plants**. *Plant Cell* 2005, **17**(6):1658-1673.

39. Wei B, Cai T, Zhang R, Li A, Huo N, Li S, Gu YQ, Vogel J, Jia J, Qi Y *et al*: **Novel microRNAs uncovered by deep sequencing of small RNA transcriptomes in bread wheat (Triticum aestivum L.) and Brachypodium distachyon (L.) Beauv**. *Funct Integr Genomics* 2009, **9**(4):499-511.

40. Xue LJ, Zhang JJ, Xue HW: **Characterization and expression profiles of miRNAs in rice seeds**. *Nucleic Acids Res* 2009, **37**(3):916-930.

41. Carra A, Mica E, Gambino G, Pindo M, Moser C, Pe ME, Schubert A: **Cloning and characterization of small non-coding RNAs from grape**. *Plant J* 2009, **59**(5):750-763.

42. Jones-Rhoades MW, Bartel DP: **Computational identification of plant MicroRNAs and their targets, including a stress-induced miRNA**. *Mol Cell* 2004, **14**(6):787-799.

43. Chiou TJ, Aung K, Lin SI, Wu CC, Chiang SF, Su CL: **Regulation of phosphate homeostasis by microRNA in Arabidopsis**. *Plant Cell* 2006, **18**(2):412-421.

44. Lu S, Sun YH, Shi R, Clark C, Li L, Chiang VL: **Novel and mechanical stress-responsive MicroRNAs in Populus trichocarpa that are absent from Arabidopsis**. *Plant Cell* 2005, **17**(8):2186-2203.

45. Sunkar R, Zhu JK: **Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis**. *Plant Cell* 2004, **16**(8):2001-2019.

46. Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK: **Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis**. *Cell* 2005, **123**(7):1279-1291.

47. Morozova O, Marra MA: **Applications of next-generation sequencing technologies in functional genomics**. *Genomics* 2008, **92**(5):255-264.

48. Shendure J, Porreca GJ, Reppas NB, Lin XX, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**(5741):1728-1732.

49. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M: **SHRiMP: accurate mapping of short color-space reads**. *Plos Comput Biol* 2009, **5**(5):e1000386.

50. Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH: **Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications**. *Bioinformatics* 2008, **24**(23):2776-2777.

51. Morozova O, Hirst M, Marra MA: **Applications of new sequencing technologies for transcriptome analysis**. *Annu Rev Genomics Hum Genet* 2009, **10**:135-151.

52. Goff LA, Davila J, Swerdel MR, Moore JC, Cohen RI, Wu H, Sun YE, Hart RP: **Ago2 immunoprecipitation identifies predicted microRNAs in human embryonic stem cells and neural precursors**. *PLoS One* 2009, **4**(9):e7192.

53. **The Sanger miRBase of microRNA data [http://microrna.sanger.ac.uk/]**.

54. Griffiths-Jones S: **The microRNA Registry**. *Nucleic Acids Res* 2004, **32**:D109-D111.

55. Nelson PT, Wang WX, Wilfred BR, Tang G: **Technical variables in high-throughput miRNA expression profiling: much work remains to be done**. *Biochim Biophys Acta* 2008, **1779**(11):758-765.

56. Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Carrington JC: **Genome-wide profiling and analysis of Arabidopsis siRNAs**. *PLoS Biol* 2007, **5**(3):e57.

57. Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M *et al*: **Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells**. *Genome Res* 2008, **18**(4):610-621.

58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

59. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T: **The small RNA profile during Drosophila melanogaster development**. *Dev Cell* 2003, **5**(2):337-350.

60. Kim VN: **Sorting out small RNAs**. *Cell* 2008, **133**(1):25-26.

61. Ho T, Wang H, Pallett D, Dalmay T: **Evidence for targeting common siRNA hotspots and GC preference by plant Dicer-like proteins**. *FEBS Lett* 2007, **581**(17):3267-3272.

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

63. Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C *et al*: **Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide**. *Cell* 2008, **133**(1):116-127.

64. Montgomery TA, Howell MD, Cuperus JT, Li D, Hansen JE, Alexander AL, Chapman EJ, Fahlgren N, Allen E, Carrington JC: **Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation**. *Cell* 2008, **133**(1):128-141.

65. Applied_Biosystems: **miRNA discovery and profiling with the SOLiD Small RNA Expression Kit**. *Applied Biosystems Application Note* 2008.

66. Davis AK, Hildebrand M, Palenik B: **A stress-induced protein associated with the girdle band region of the diatom Thalassiosira pseudonana (Bacillariophyta)**. *Journal of Phycology* 2005, **41**(3):577-589.

67. Maheswari U, Mock T, Armbrust EV, Bowler C: **Update of the Diatom EST Database: a new tool for digital transcriptomics**. *Nucleic Acids Res* 2009, **37**:D1001-D1005.

68. Lu C, Meyers BC, Green PJ: **Construction of small RNA cDNA libraries for deep sequencing**. *Methods* 2007, **43**(2):110-117.

69. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

70. **454 Life Sciences Roche website [http://www.454.com/]**

71. **Applied Biosystems SOLiD System website [http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiD-System-Sequencing-B/index.htm**]

72. Hildebrand M, Dahlin K: **Nitrate transporter genes from the diatom Cylindrotheca fusiformis (Bacillariophyceae): mRNA levels controlled by nitrogen source and by the cell cycle**. *Journal of Phycology* 2000, **36**(4):702-713.

73. **DOE Joint Genome Institute, Thalassiosira pseudonana v3.0 genome [http://genome.jgi-psf.org/Thaps3/Thaps3.home.html]**.

74. **CD-HIT: Cluster Database at High Identity with Tolerance [http://www.bioinformatics.org/cd-hit/]**.

75. Li WZ, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.

76. Kohany O, Gentles AJ, Hankus L, Jurka J: **Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor**. *BMC Bioinformatics* 2006, **7**:474.