Summer 8-2014

Characterizing Populations of Non-Coding RNAs in *Karenia brevis* at Different Times of the Diel Cycle

Scott Boyd Anglin
University of Southern Mississippi

The University of Southern Mississippi

CHARACTERIZING POPULATIONS OF NON-CODING RNAS IN KARENIA BREVIS AT DIFFERENT TIMES OF THE DIEL CYCLE

by

Scott Boyd Anglin

A Thesis
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Master of Science

Approved:

Dr. Timothy McLean
_____
Director

Dr. Glenmore Shearer
_____

Dr. Chaoyang Zhang
_____

Dr. Maureen Ryan
_____
Dean of the Graduate School

August 2014

ABSTRACT

CHARACTERIZING POPULATIONS OF NON-CODING RNAS IN KARENIA BREVIS AT DIFFERENT TIMES OF THE DIEL CYCLE

by Scott Boyd Anglin

August 2014

Karenia brevis is a mixotrophic, marine dinoflagellate found in the Gulf of Mexico that generates periodic, if not annual, harmful algal blooms (also known as "red tides") in certain coastal areas. In an effort to better understand the biology of this organism, a functional genomics project has been initiated. As part of that project, it has been determined that a significant number of natural antisense transcripts (NATs) as well as double-stranded RNA (dsRNA) molecules exist within the transcriptome of *K. brevis*. I hypothesize that the non-coding NATs, similar to microRNAs (miRNAs) in other organisms play a role in regulating gene expression. To test this prediction, I extracted total RNA from cells grown under different culture conditions, isolated and cloned the dsRNAs and miRNAs separately, and sequenced all transcripts from each sample. Bioinformatic analyses were used to assess the relative expression of miRNAs, NATs, and mRNAs. My determination of any differential expression between day and night conditions should either support or falsify the hypothesis of NATs and/or miRNAs regulating the expression of genes via a post-transcriptional mechanism. The miRNA analysis revealed many mature miRNA candidates, but visualization software suggests that the miRNA pathway may not be present in the *K. brevis* genome. Also, length distribution of the miRNA samples suggests that the small RNAs are too long to be bound be the Argonaute protein, which is a key factor in miRNA synthesis. Cleavage

patterns, transcript shape and read alignment patterns resemble a *cis*-Nat pathway, although it is undetermined whether this leads to siRNAs or an alternate small RNA. The RNA-seq analysis discovered that a large number of transcripts exhibited differential expression between the two time points of the diel cycle.

ACKNOWLEDGMENTS

I would like to thank Dr. Timothy McLean for guiding me in the right direction with this project, Isaac Akogwu for the hours of help with Linux and learning how to build transcriptomic assemblies, Dr. Glover George for helping me get access to the computer science servers to run these large data sets, Teresia Buza for the miRNA bioinformatic work, Dr. Alex Flynt for helping me understand small RNA pathways and genome browsers, and David Jayroe for listening to me complain daily.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF ILLUSTRATIONS

Figure

CHAPTER I

INTRODUCTION

Dinoflagellates

Dinoflagellates are an important group of unicellular, flagellated phytoplankton. They can be found in marine and freshwater habitats across the globe and their diversity can be seen through adaptation to a variety of environments. They can be divided between armored and unarmored, where the armored contain cellulose or other polysaccharides within vesicles of the cell wall, and unarmored contain a single layer of flattened vesicles making them more fragile (Hackett, Anderson, Erdner, & Bhattacharya, 2004). Many species are considered mixotrophic: having more than one means of obtaining nutrition; while others are strictly phototrophic or heterotrophic. As many dinoflagellates are capable of photosynthesis, they are major contributors to atmospheric O_2, producing a significant amount of the planet's overall O_2.

A few dinoflagellate species are known to produce algal blooms some of which can be toxic to oceanic populations of fish, shellfish, and mammals, including humans. These toxic blooms are referred to as harmful algal blooms (HABs). These toxins can cause massive fish kills and accumulate in the upper food chain over time (Backer and McGillicuddy, 2006), some of which are a source of food for humans. Concentrated amounts of toxins within the tissues of certain shellfish, if ingested, may cause a variety of poisoning syndromes such as neurotoxic shellfish poisoning (NSP), paralytic shellfish poisoning (PSP), ciguatera fish poisoning (CFP), diarrheic shellfish poisoning (DSP), and azaspiracid poisoning (AZP) (Monroe & Van Dolah, 2008). Furthermore, breaking waves can cause unarmored dinoflagellates, such as the one that produces the toxin

responsible for NSP, to be lysed and their toxins to be aerosolized. Inhalation of the aerosolized NSP-producing toxins from sea spray can result in respiratory irritation and other health effects to humans and mammals (Kirkpatrick et al., 2004). These HABs are collectively referred to as "red tides", but algal blooms can be found in many different colors, each produced by a different species of phytoplankton that produces its own toxin or suite of toxins.

Karenia Brevis

Karenia brevis is a mixotrophic, unarmored, marine dinoflagellate found in the Gulf of Mexico that generates periodic, if not annual, harmful algal blooms in certain coastal areas which can result in massive fish and marine mammal deaths; and neurotoxic shellfish poisoning and respiratory illness in humans (Cheng, Villareal et al., 2005). *K. brevis* produces several polyether brevetoxins (PbTxs) that are neurotoxic. They are tasteless, odorless, and heat and acid stable (Kirkpatrick et al., 2004), making them very difficult to detect and/or remove from contaminated food. Brevetoxins affect the voltage sensitive sodium channel of membranes by forcing them to stay continually open (Atchison, Luke, Narahashi, & Vogel, 1986). Studies show that not only fish and marine mammals, but birds, non-marine mammals and some amphibians can also suffer respiratory failure from the toxic effects of *K. brevis* (Steidinger, Landsberg, Flewelling, & Kirkpatrick, 2008). Due to the adverse effects of these toxins, *K. brevis* has been rigorously studied for the past 65 years. *K. brevis* HABs have been described as early as 1948 (Van Dolah et al., 2009), but fish kills have been recorded as early as 1648 (Steidinger et al., 2008), though the cause was unknown. By the 1960s the toxicology of *K. brevis* had been described, in the 1970s distinct toxic fractions had been determined by

chemical fractionation, and by the 1980s the crystal structure for brevetoxins had been described as well as the specific binding site of brevetoxins to the voltage-sodium channel (Baden, 1989). With the advancement of molecular tools, many new discoveries have been described. Brevenal, a nontoxic natural product, can protect fish from the neurotoxic effects associated with brevetoxins by competing with brevetoxins for binding to voltage-sodium channels, which may help in the development of therapeutics to relieve contamination during red tide events (Bourdelais et al., 2004).

Much of the molecular machinery and function involved in the life cycle of *K. brevis* is unknown. The complexity of the dinoflagellate genome has made advancement in the understanding of the molecular workings of these organisms difficult. What is known is that they lack nucleosomes, contain a large amount of hydroxymethyluracil that replaces some percentage of thymine (Li & Hastings, 1998), and have chromosomes that remain condensed throughout the cell cycle (Brunelle and Van Dolah, 2011). They also have unusually large genomes (100,000 Mb) making whole genome sequencing difficult (Lin, Zhang, Zhuang, Tran, & Gill, 2010). With the discovery of a 22-nt conserved spliced leader at the 5´ end of all dinoflagellates mRNAs, a new method for separating coding from non-coding transcripts is available (Lin et al., 2010). To date, no promoter sequences have been discovered in dinoflagellate genes, and because of this, the mechanisms associated with gene regulation are unknown (Brunelle and Van Dolah, 2011). The above evidence suggests that replication and gene expression machineries may contain unique properties, requiring new molecular techniques to discover and understand the processes involved in the life cycle of dinoflagellates (Li & Hastings, 1998).

Significance

Scientific interest in dinoflagellates has drastically increased due to the frequency of toxic blooms and because of the importance of the organisms in relation to coral reef health (Hackett et al,. 2004). They are a complex group of organisms that have adapted to the majority of salt and fresh water habitats in the world and play a key role in these environments. Also, they are a key component in oceanic food webs throughout the globe, and they are responsible, in part, for producing large quantities of O_2. A multitude of research has been conducted in relation to *K. brevis* over the past 60 years and many insights into its cellular structure and functions have been discovered, but very little has been accomplished in the form of molecular research. This is partially due to the intractability of working with the genome, primarily because of its size as well as many of its unique characteristics, and the lack of tools available. But with recent developments in sequencing, it is hoped that sequencing large sections of its transcriptome or genome may be possible to give a better understanding into the life cycle of this organism.

Recently, large numbers of long, antisense RNAs, defined as natural antisense transcripts (NATs), were found and characterized in *K. brevis* transcriptomes (McLean & Pirooznia, 2011) It is the working hypothesis of the McLean laboratory that the NATs regulate gene expression at the post-transcriptional level. One way to test this hypothesis would be to investigate the effects of the diel cycle on NATs and other antisense RNA regulators such as micro RNAs (miRNAs) from populations of *K. brevis* through development of nucleic acid purification protocols, high-throughput sequencing, and bioinformatics analysis of raw data. The purpose of this investigation was to find

additional antisense RNAs and determine the differences in their expression levels (as well as the differences in expression of the RNAs that they regulate) at time points where I predicted to see multiple genetic differences, namely in the middle of the light period and the middle of the night period.

CHAPTER II

REVIEW OF RELATED LITERATURE

Identification

Dinoflagellates are a diverse group of unicellular eukaryotes that include both photosynthetic and non-photosynthetic members, and include autotrophs, mixotrophs and heterotrophs (Taylor, Hoppenrath, & Saldarriaga, 2008). Most dinoflagellates are free-living, but some are endosymbionts such as the zooxanthellae of reef building corals (Taylor et al., 2008), some can be parasitic such as *Blastodinium* which live in the intestines of copepods (Skovgaard, 2005), some are bioluminescent such as *Gonyalaux polydra* found off the coast of British Columbia (Abrahams & Townsend, 1993), and some species are toxic and can form monospecific blooms such as *Karenia brevis* (Monroe & Van Dolah, 2008). Most dinoflagellates have a pair of unequal flagella (called a dinokont arrangement) for propulsion, a posterior flagellum for controlling the direction of movement and a transverse flagellum that causes the cell to rotate and move in a forward direction (Figure 1) (Lewis et al., 2006).

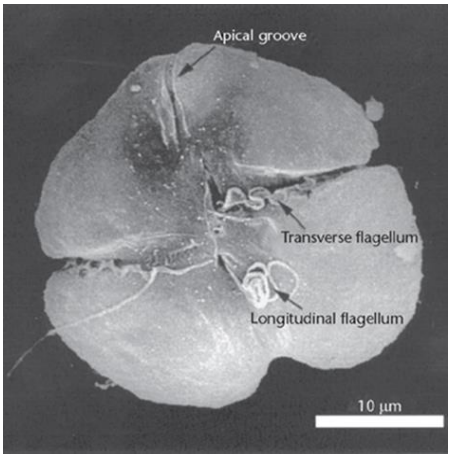


Figure 1. Scanning Electron Micrograph of *K. brevis*. *K. brevis* cell with apical groove and transverse and longitudinal flagella (Heimann, 1999).

The dinoflagellate nucleus is different from other eukaryotes due to its permanently condensed chromosomes, lack of histones and extranuclear spindle that passes through cytoplasmic channels (Hoppenrath & Leander, 2010).

The *Karenia* genus was implemented in 2000 by G. Hansen and Moestrup. The reason for the new genus was a result primarily from rDNA sequencing of the large subunit. Morphological, chloroplast pigment and toxin production differences found in these organisms versus *Gymnodinium* (the genus into which *Karenia* species were previously assigned) helped in the creation of the new genus (Daugbjerg, Hansen, Larson, & Moestrup, 2000). Dinoflagellates in the *Karenia* genus possess a unique apical groove unlike the characteristic *Gymnodinium* apical groove (Figure 2), a chloroplast pigmentation containing fucoxanthin instead of peridinin, and have several toxins unique to the *Karenia* genus (Daugbjerg et al., 2000). By 2007 the genus contained 15 species, five that occur in the Gulf of Mexico (Steidinger, 2009).

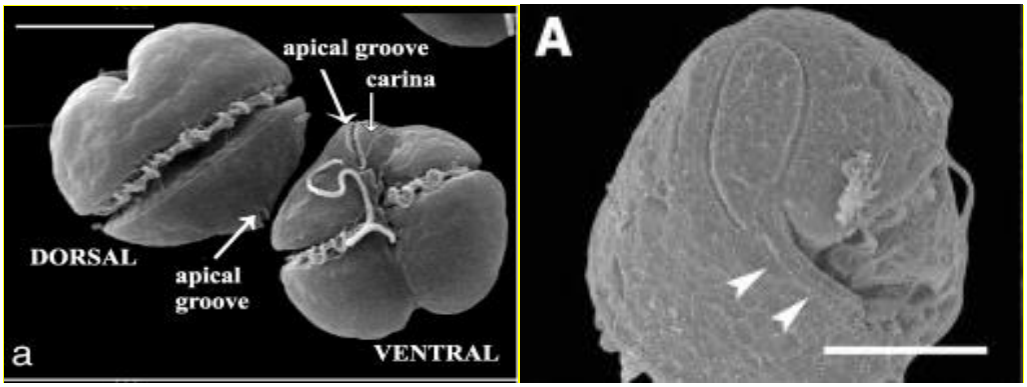


Figure 2. (Left) Scanning electron micrographs of *Karenia brevis* and *Gymnodinium litoralis*. The image on the left shows the dorsal and ventral views of straight apical groove typical of *Karenia* genus (Haywood et al., 2004). The image on the right shows the apical view of elongated, anticlockwise loop of apical groove of *Gymnodinium litoralis* (Reñé et al., 2011).Scanning electron micrographs of *Karenia* genus (Haywood et al., 2004). (Right) SEM. Apical view of elongated, anticlockwise loop of apical groove of *Gymnodinium litoralis* (Reñé et al., 2011).

Major morphological characteristics among the *Karenia* genus include a dorso-ventrally flattened cell ranging from 18-32µm long and 18-48µm wide, a linear apical groove (Figure 2), rounded epitheca and carina (Figure 2), dinokont arrangement for flagella (Figure 1), and a round nucleus located in the posterior left quadrant (Steidinger & Penta, 1999; Haywood et al., 2004).

Formerly known as *Gymnodinium breve* and *Ptychodiscus brevis* (Steidinger, 2009), *Karenia brevis* is a photosynthetic, marine dinoflagellate that is known for its monospecific, toxic blooms that occur periodically in the warm temperate to tropical waters of the Gulf of Mexico (Figure 3) (Örnólfsdóttir, Pinckney, & Tester, 2003).

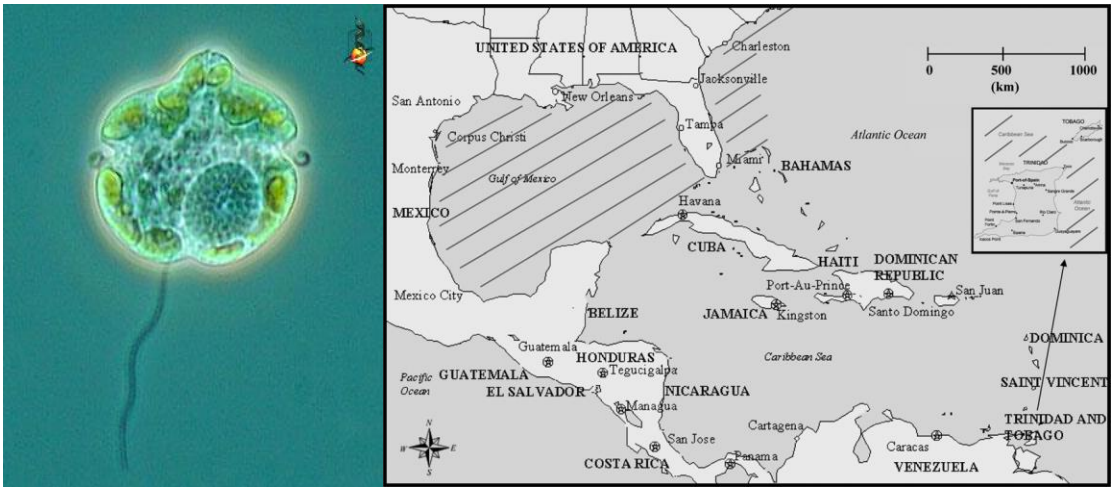


Figure 3. *Karenia brevis* and its biogeography. The light microscope image on the left shows *Karenia brevis*. Picture by Bob Andersen and D. J. Patterson. On the right is a map showing the distribution (diagonal lines) for *Karenia brevis* in the Gulf of Mexico and North Atlantic (Steidinger, 2009).

Harmful Algal Blooms

Of the more than 4000 phytoplankton species only ~300, which include diatoms, dinoflagellates, silicoflagellates, prymnesiophytes and raphidophytes, are known to produce algal blooms (Smayda, 1997). Accumulations of these organisms can stain the water different colors and can deplete oxygen levels through excessive respiration or

decomposition (Sellner, Doucette, & Kirkpatrick, 2003). Of the 300 phytoplankton that can cause algal blooms only 60-80 produce harmful algal blooms; 75% of which are dinoflagellates (Smayda, 1997). These dinoflagellates can be found across the globe and produce a multitude of unique toxins with different forms of poisoning syndromes. HABs are a natural phenomenon that can trigger economic losses to aquaculture, fisheries and tourism and are also associated with major environmental impacts and adverse human health effects (Hallegraeff, 1993). In recent years many of these organisms have been producing HABs with increasing frequency, intensity and geographical diversity (Figure 4). This increase is theorized to be caused by anthropogenic influences such as eutrophication and artificial spread (Taylor et al., 2008); however, alternate explanations have been examined, e.g., changing ocean currents, temperatures, and weather patterns (Kirkpatrick et al., 2004).

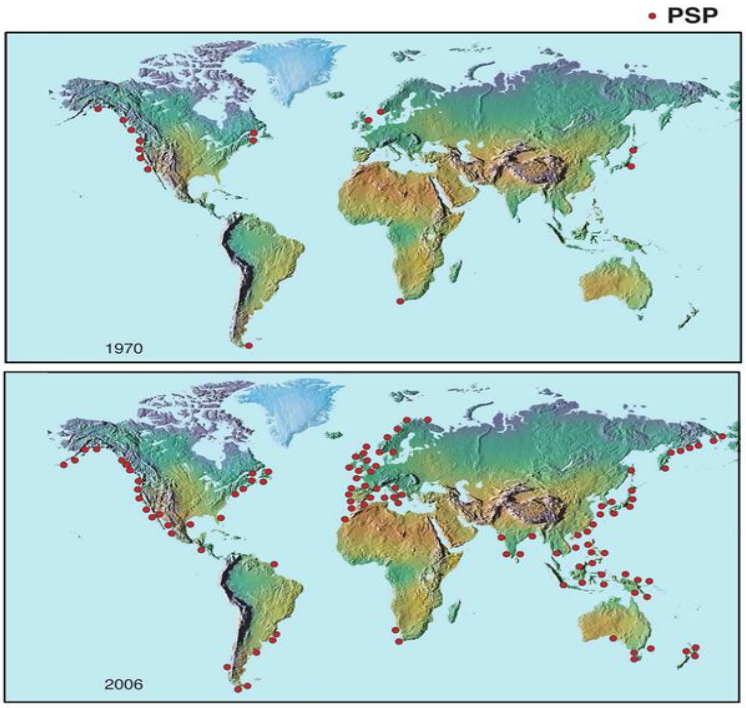


Figure 4. Known global distribution of paralytic shellfish poisoning (PSP) in 1970 (top) and 2006 (bottom) (Harmful Algae, 2012).

Several HAB forming dinoflagellates reside in the Gulf of Mexico; however, *K. brevis* is the leading cause with blooms impacting all five Gulf coast states. Although the entire Gulf coast is at risk of an event, HABs occur more often along the coast of Florida (annually) and Texas (near annually) (Stumpf et al., 2003). These blooms can take place when there is an increase in the concentration (from 10-100cells/L up to 2x107 cells/L) of a *K. brevis* population, typically occurring during the late summer to early fall in offshore waters and afterwards transported inshore by winds and tidal currents (Vargo et al., 2008). Toxic blooms can cause massive fish kills and bioaccumulation of neurotoxins in shellfish. When humans ingest contaminated shellfish, they risk getting neurotoxic shellfish poisoning, which can cause nausea, diarrhea, severe muscle aches, and numbness around the mouth (Monroe & Van Dolah, 2008). Furthermore, aerosolized toxins can cause irritation and burning in the throat and the upper respiratory tract, involuntary coughing and sneezing, and rhinorrhea (Hackett et al., 2004; Cheng, Zhou et al., 2005).

K. brevis produces low molecular weight, lipid soluble polyether neurotoxins, known as brevetoxins (PbTxs) (Steidinger, 2009). More than nine different brevetoxins have been described, however, PbTx-2 is the major toxin produced and is followed by PbTx-1 (Figure 5), both of which are considered to be the parents from which all other brevetoxins are derived (Bourdelais & Baden 2004; Kirkpatrick et al., 2004). Brevetoxin research on rat brain synaptosomes has shown that the toxins alter cellular processes by attacking voltage-sensitive sodium channels through binding to a specific site (titled as site 5) keeping the channels continually open (Poli, Mende, & Baden, 1986). Binding causes the sodium channels to remain open, allowing uncontrolled Na^+ flow into the cell

and prevents sodium channel inactivation, which can lead to repetitive firing in nerves (Kirkpatrick et al., 2004; Catterall and Gainer, 1985).

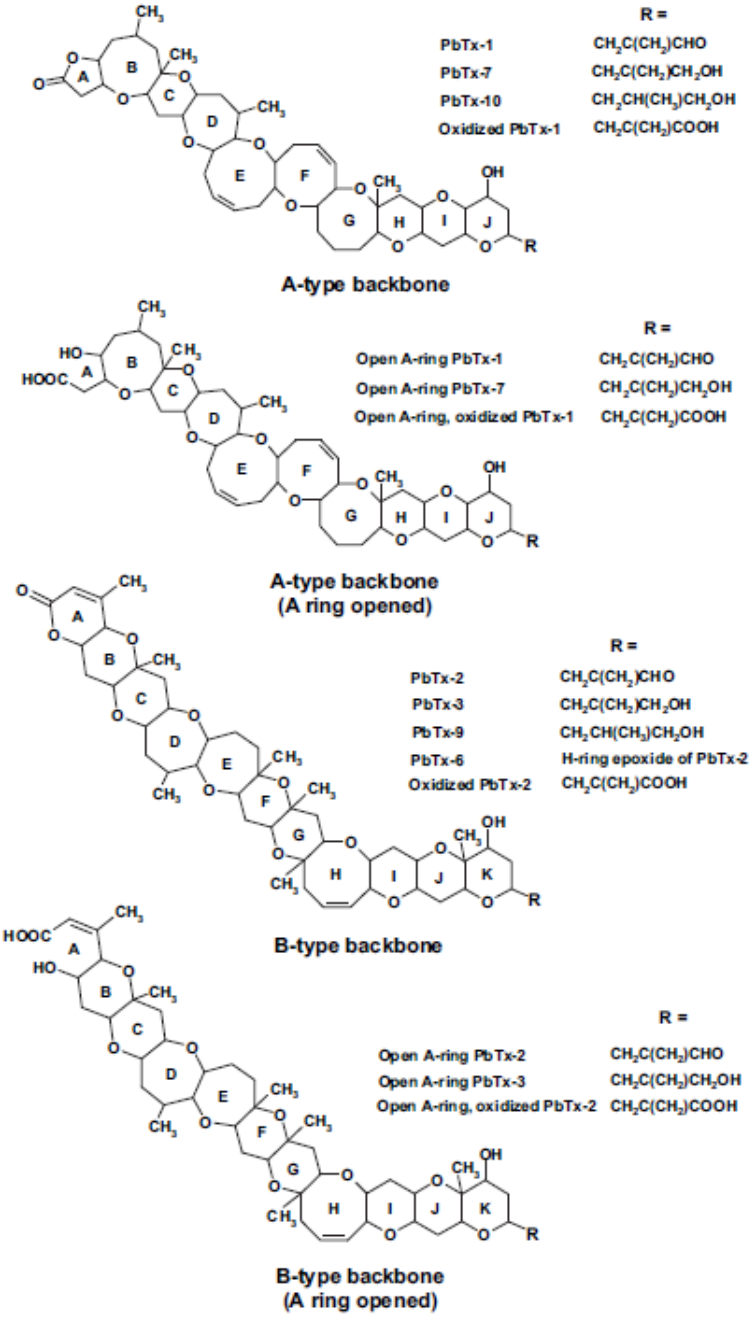


Figure 5. Parental and derived brevetoxins from *K. brevis* cultures and blooms (Roth, Twiner, Zhihong, Bottein, & Doucette, 2007).

Diel Cycle

In response to rhythmic night and day cycles caused by our planet's rotation, most organisms have evolved endogenous clocks (Pittendrigh, 1993). Circadian clocks or rhythms are oscillating regulators that operate on a 24 hour cycle which control diverse biological processes such as behavior, physiology, and biochemical reactions (Brunelle, Hazard, Sotka, & Van Dolah, 2007). These biological clocks can also continue without environmental cues, be reset by environmental cues and have a constant cycle regardless of temperature (Yacobovitch, Benayahu, & Weis, 2004). In algae and many free-living marine dinoflagellates, circadian oscillations have been identified and found to control numerous processes such as cell division, bioluminescence, motility, organelle migration and photosynthesis (Sorek et al., 2013).

In *K. brevis*, cell division is entrained to the diel phased cell cycle or photoperiod, which means that cells only divide during a narrow window of time between the light or dark phase, progressing under the control of a circadian clock (Van Dolah & Leighfield, 1999). A previous study showed that *K. brevis*' cell cycle is under circadian control with the cell cycle entrained by controlling the entry into S phase (Brunelle et al., 2007). S phase begins approximately six hours after the arrival of daylight and mitosis beginning 18 to 22 hours after dawn, with the cell cycle finishing by the beginning of the next day (Van Dolah et al., 2009). Cell cycle advancement is controlled by transcriptional and post-transcriptional regulation; furthermore, results from one *K. brevis* study suggest that the expression levels of S phase specific proteins are independent of transcription upon entry into the S phase (Brunelle & Van Dolah, 2011).

Post-Transcriptional Regulation

The regulation of cell survival, adaptation to stress, homeostasis, cell fate, and differentiation in living organisms require the dynamics of gene expression to react to environmental cues (Keene, 2007). The regulation of gene expression is a highly interconnected multi-step program (Figure 6) that is fundamental to coordinating synthesis, assembly and localization of macromolecular structures of cells (Halbeisen, Galgano, Scherrer, & Gerber, 2008). In prokaryotes, the gene expression machinery for both transcription and translation are physically coupled into polyribosomes. However, in eukaryotes, transcription and translation are temporally and spatially separated; transcription occurs in the nucleus and translation in the cytoplasm (Glisovic, Bachorik, Yong, & Dreyfuss, 2008). This separation by compartmentalization has allowed eukaryotes to highly diversify methods of gene regulation.

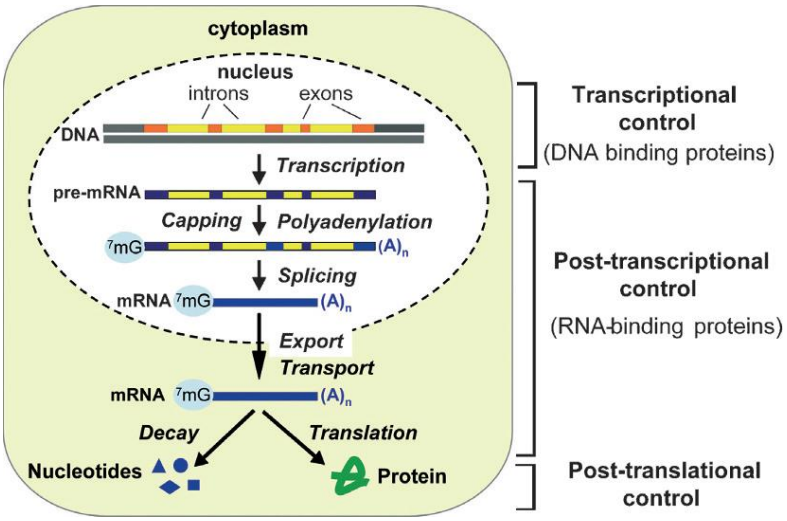


Figure 6. Regulation of gene expression at different, multiple levels (Halbeisen et al., 2008).

After transcription, many events are necessary for the synthesis of proteins, which fall in the category of post-transcriptional regulation. The discovery of post-

transcriptional regulation peaked interest in the scientific community because it was determined that the mechanisms associated with it were ubiquitous among all domains of life (Cogoni & Macino, 2000). Furthermore, post-transcriptional processes and their regulation contain hundreds of proteins and non-coding RNAs (ncRNAs) (Eulalio, Behm-Ansmant, & Izaurralde, 2007). Additionally, variations in these processes produce significant mRNA and protein diversity (Akker, Smith, & Chew, 2001). Some of these processes share messenger RNAs as a substrate, which are not all translated immediately; some are sustained in a translationally repressed state for later use, while quality control and regulatory mechanisms can degrade or repress others (Eulalio et al., 2007). Other processes include antisense regulation, which demonstrate the ability to affect RNA stability, nuclear processing, export and translation (Munroe & Zhu, 2006). Post-transcriptional regulation encompasses a large group of RNAs (ncRNAs) that do not code for proteins but instead play a large role in gene expression through highly complex mechanisms that may slow down or prevent the synthesis of proteins.

Non-Coding RNA

RNA interference (RNAi) is a post-transcriptional gene expression mechanism that disrupts or inhibits mRNAs, effectively preventing protein synthesis. The origin of RNAi is believed to have evolved from the need to prevent viral invasion in eukaryotic organisms (Montgomery, Xu, & Fire, 1998). RNA interference was first identified by researchers working with *C. elegans* who determined several key features: silencing is efficiently triggered by dsRNA but only weakly by the antisense or sense strand alone; silencing is specific for mature mRNA that is homologous to dsRNA, which is indicative of a post-transcriptional mechanism; very few dsRNA molecules are necessary to achieve

silencing, and they can spread between tissues and even to the next generation; and mRNA is degraded (Fire et al.,1998). These features derive from small silencing dsRNAs that prevent mRNAs from being expressed through complementary annealing.

Double-stranded RNAs are comprised of complementary sense and antisense strands of RNA that are bound to each other, and the formation of dsRNAs can contribute to the regulation of gene expression depending. Two of the most widely studied dsRNAs, small interfering RNAs (siRNAs) and micro RNAs (miRNAs), share many similarities in production and function, but miRNAs are thought to be more of a gene regulator whereas siRNAs are defenders against viral attacks (Carthew & Sontheimer, 2009). Beginning as larger strands of dsRNA, they are recognized by the enzyme Dicer, an RNAse III type enzyme, which cleaves them into 21-23 nucleotide (nt) long products. Most of the machinery is shared between siRNA and miRNA, but miRNA utilizes an enzyme called Drosha. This enzyme is similar to Dicer in that it is an RNase III type enzyme but is specific to pre-miRNAs which cleaves them prior to cleavage by the Dicer enzyme (Han et al., 2004). Another important piece of machinery is a complex called the RNA-induced silencing complex (RISC), which is comprised of Dicer, an Argonaute protein and a double-stranded RNA binding protein (dsRBP) (Gregory, Chendrimada, Cooch, & Shiekhattar, 2005). This complex determines the mRNA's fate, either inhibition or degradation.

Non-coding RNAs (ncRNAs) have several, sometimes confusing names. To give a better understanding of the terminology used throughout this manuscript it is necessary to outline these differences. Non-coding RNA is a general term for many types of RNAs such as rRNAs, tRNAs, siRNAs, miRNAs, asRNAs, etc. Antisense RNAs or natural

antisense transcripts refer to single RNA transcripts complementary to the sense mRNA transcripts. Antisense RNAs can be transcribed as *trans*-antisense RNAs (e.g., miRNAs and siRNAs), meaning that they are produced at separate, non-overlapping loci from the mRNAs with which they share complementary sequences, or they can be *cis*-antisense RNAs which are produced from overlapping loci on complementary strands of the DNA. The former NATs contain relatively short regions of base pairing dsRNAs while the latter instead form large regions of perfectly matched dsRNAs (Munroe & Zhu, 2006).

Natural Antisense Transcripts

Natural antisense transcripts (NATs) first discovered in 1981 (Brantl, 2002), are RNA molecules containing sequences that are complementary to other endogenous RNAs that have a known function (sense transcripts) (Vanhée-Brossollet and Vaquero, 1998). NATs can be *cis* or *trans*-acting. *Cis*-NATs are transcribed by a promoter located on the DNA strand opposite the same DNA molecule and therefore have perfect complimentarity to their target RNAs and are involved in post-transcriptional inhibition of specific RNA functions (Brantl, 2002; Faghihi & Wahlestedt, 2009; Thomason & Storz, 2010). *Trans*-NATs are transcribed from non-overlapping, separate loci and have only partial complementarity to the sense transcript, allowing them to target many different sense transcripts and form complex regulation networks (Lapidot & Pilpel, 2006). Micro-RNAs, small interfering RNAs and small nucleolar RNAs all belong to the *trans*-NAT group. NATs are not uniformly dispersed throughout the entirety of the gene but can be found covering the 5´ end, 3´ end, middle or even the entire gene allowing for both ends of the gene to have a propensity for natural antisense transcription (Faghihi & Wahlestedt, 2009; Lasa et al., 2011). By blocking access of the translational machinery

to the 5´ end the NAT reduces the level of protein synthesis, but in eukaryotes, compartmentalization has allowed for numerous and complex effects from NATs that are not possible in prokaryotes (Kumar & Carmichael, 1998). NATs were first described in prokaryotes as part of the general mechanism for gene expression, involved in biological functions such as transposition, phages and plasmid replication, and the down-regulation of gene expression in the sense transcripts (Vanhée-Brossollet and Vaquero, 1998). NATs in eukaryotes were initially found by accident, and much of their regulatory roles and mechanisms have not been described (Brantl, 2002). In the past ten years research has established that the mammalian genome contains large amounts of transcribed genes that are not protein coding genes and that many are transcribed in the antisense direction (Finocchiaro et al., 2007). Theories to why regulation from NATs could be beneficial over other types of regulation include natural antisense transcripts providing an advantage when protein levels need to be repressed securely and expressed under specific circumstances or when they are subject to broad regulation they may provide another level of control (Thomason & Storz, 2010). To better understand how NATs regulate gene expression it is necessary to discuss the mechanisms that control their outcome.

Natural Antisense Transcript Mechanisms

Four main mechanisms associated with NATs have been outlined by previous research. First is transcription-related modulation, which suggests that transcription in the antisense direction, and not the asRNA itself, controls transcription of sense RNAs (sRNAs). This mechanism is divided into two parts: transcriptional collision and genomic arrangements. Secondly, RNA-DNA interactions are associated with epigenetic regulation of transcription by alteration of DNA and chromatin, e.g., alteration of

promoter access, genomic imprinting, and X chromosome inactivation. The third mechanism for NAT gene regulation is nuclear RNA duplex formation resulting in alternative splicing and/or termination of the associated mRNA. Additionally, it has been proposed that NATs regulate mRNA transport or retention. The last mechanism associated with NATs is cytoplasmic RNA duplex formation that can change mRNA stability and translation efficiency, mask miRNA binding sites, and form endogenous siRNAs (Faghihi & Wahlestedt, 2009). It has also been suggested that overlapping transcription may affect the expression of a target gene at different levels independent of the mechanism that produced it by affecting stability of target RNA, inducing changes in the structure of mRNAs, preventing RNA polymerase from binding or extending and affecting protein synthesis by blocking or promoting ribosomal binding (Lasa et al., 2011). These mechanisms all have a distinct fate or consequence for the molecules and cells they share association with. Functions of mechanisms for NATs include transposition inhibition by reducing transposase levels, regulation of the synthesis of transcriptional regulators either positively or negatively, and regulation of the expression of some metabolic enzymes (Thomason & Storz, 2010). Many new NATs and their mechanisms are being described due to the improvement in sequencing technologies. These advances in technology will hopefully allow for insight into NAT mechanisms and functions that make *K. brevis* such a unique organism.

Micro-RNA

MicroRNAs are small, non-coding RNAs that regulate gene expression through the RNAi pathway and are also known for their roles in growth and development. The synthesis of long primary miRNAs (pri-miRNAs) begins miRNA production. These pri-miRNAs can range from 700 nucleotides to several kilobases in size (Denli, Tops, Plasterk, Ketting, & Hannon, 2004) and have a hairpin-like structure. They are processed by the Microprocessor complex inside of the nucleus. This cleavage complex contains two parts: Drosha, an RNase III type enzyme and a dsRBD protein. Together they cleave the pri-miRNA into 60-70 nucleotide precursor mi-RNA (pre-miRNAs) (Gregory et al., 2005).The pre-miRNAs can now be transported into the cytosol. Once into the cytosol the pre-miRNAs can be recognized and cleaved by the Dicer enzyme, which yield mature ~22 nucleotide miRNAs (Hutvagner et al., 2001). These mature miRNAs can now be denatured and recognized by a ribonucleoprotein effector complex known as RISC. A single strand known as the guide RNA is incorporated into RISC (Hammond, Boettcher, Caudy, Kobayashi, & Hannon, 2001). The guide strand directs RISC to its target based on complementarity between the miRNA and the mRNA. The endonuclease of RISC then cleaves the mRNAs into pieces for degradation if the miRNA and mRNA are perfect matches or inhibited translation if nearly perfect matches (Elbashir, Harborth, Weber, & Tuschl, 2002).

Micro-RNA Structure and Function

Micro-RNAs and siRNAs are highly conserved, important regulators of gene expression (Rhoades et al., 2002) but have unique structures and functions within the cell. Biogenesis, development and assemblage into their RISC complexes are also different

which suggests different functions as well (Bartel, 2004). There are many key structural and production differences found between the two dsRNAs. Micro-RNAs originate from genomic loci that are unique from other recognized genes, are produced from transcripts that can form RNA hairpin structures, generate only one double-stranded miRNA complex from each miRNA precursor molecule, and are typically conserved in related organisms. In contrast, siRNAs derive from mRNAs, transposons, viruses or heterochromatic DNA. They are produced from long duplexes of RNA (either dsRNA formed from two separate RNAs or extended intramolecular hairpins) or a single siRNA precursor molecule can generate many siRNA duplexes, and they are rarely conserved (Bartel & Bartel, 2003).

Functionally, miRNAs are mainly associated with the regulation of gene expression, which makes them significant for growth and development (Rhoades et al., 2002), but have also been found to play a role in cell proliferation, cell death, and fat metabolism in different organisms (Xu, Vernooy, Guo, & Hay, 2003). In recent years miRNAs have been shown to play many different roles throughout organisms. Alternatively, siRNAs are produced from viruses or repetitive sequences introduced by genetic engineering, but can also be produced by transposons (Tang, 2005), which are DNA elements that can jump to different locations. siRNAs function in antiviral defense, silencing mRNAs that are overproduced, and guard the genome form disruption from transposons. These differences show that despite some similarity, the two dsRNAs are very different in structure as well as functionality.

The Spliced Leader

Dinoflagellates possess enormous genomes with many cellular and molecular features atypical to eukaryotes, including chromosomes that remain permanently condensed into liquid crystal structures throughout the cell cycle and the lack of nucleosomes, TATA boxes and promoters associated with transcriptional regulation (Lin et al., 2010; Monroe & Van Dolah, 2008). Because of these features, the mechanisms associated with gene regulation are unknown, which indicate that replication and gene regulation contain unique properties that require new molecular mechanisms or explanations to elucidate the life cycle of *K. brevis* and related dinoflagellates (Li and Hastings, 1998). An additional unusual molecular attribute of dinoflagellates is the process of spliced leader *trans*-splicing, which has been found in a small but diverse number of organisms including euglenozoa, nematodes, platyhelminthes, cnidarians, rotifers, ascidians, appendicularia, and dinoflagellates (Zhang & Lin, 2009). This process allows the translation of polycistronically (mRNA that can be translated into more than one polypeptide) transcribed nuclear genes (Zhang, Campbell, Sturm, & Lin, 2009) as has been described for dinoflagellate genes and mRNAs (find some of those references for long, tandemly-arrayed dino genes).

Spliced leader (SL) *trans*-splicing was discovered in trypanosomes by Murphy, Watkins, and Agabian in 1986. Spliced leader (SL) *trans*-splicing produces mature mRNAs from pre-RNAs by utilizing a short non-coding RNA fragment (SL RNA) that is *trans*-spliced at a splice receptor site located at the 5´untranslated region (UTR) of each gene on a polycistronic message (Van Dolah et al., 2009; Zhang et al., 2009). This process results in polycistronic primary transcripts being developed into individual

monocistronic mRNAs with a common 5´ sequence (Lidie & Van Dolah, 2007) (Figure 7). In 2007, two papers were published: one by Lidie and Van Dolah that found 87 *K. brevis* mRNAs that contain the 5´ SL RNA *trans*-spliced sequence (Figure 8) and one by Zhang et al. that found a conserved 22nt SL sequence that trans-splices nuclear-encoded genes in all dinoflagellate species, from ancestral to derived lineages.

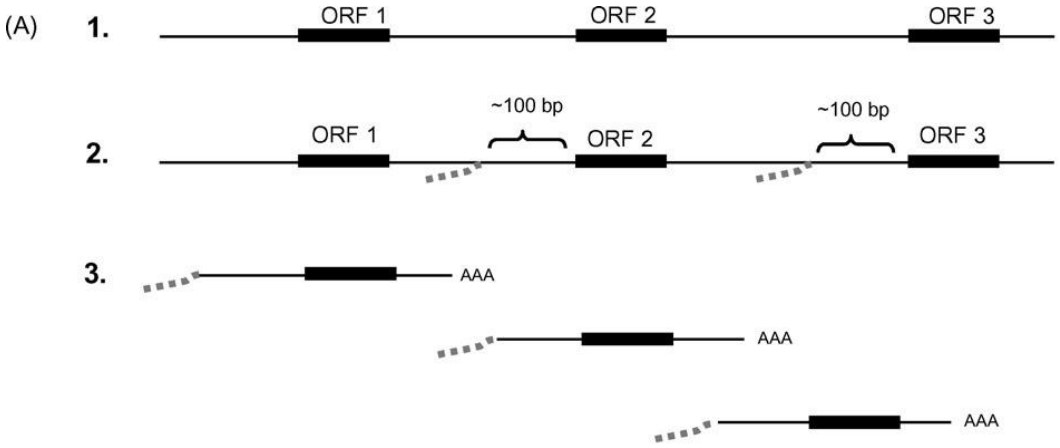


Figure 7. The spliced leader trans-splicing mechanism in trypanosomes. (1) Polycistronic message with three open reading frames (ORFs; thick bars) and intergenic regions (thin lines). (2) The spliced leader (dashed line) is added at a splice signal located 100 bp upstream from the start codon for each ORF. (3) Simultaneous addition of a poly-A tail results in mature messages containing an identical 5´cap and spliced leader, 5´ UTR, coding sequence, 3´ UTR, and poly-A tail (Van Dolah et al., 2009).

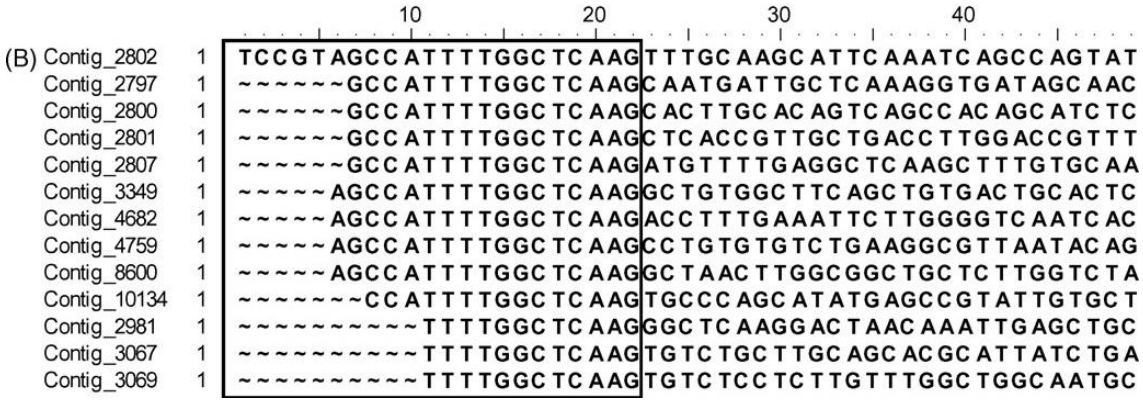


Figure 8. *K. brevis* ESTs containing 5´ spliced leader. Twelve ESTs possessing identical 5´ ends that represent the spliced leader (Van Dolah et al., 2009).

This biological process has several functions; it can clean-up the 5´ end of mRNAs, stabilize mRNAs, regulate gene translation, and generate monocistronic mRNAs. Other studies have shown that the spliced leader can enhance translational efficiency and mediate polysome association in specific organisms (Lidie & Van Dolah. 2007). This unique SL could become a new tool in separating and profiling lineage specific dinoflagellate transcriptomes, but the exact role of SL *trans*-spliced in dinoflagellates has not yet been elucidated (Lin et al., 2010; Van Dolah et al., 2009).

CHAPTER III

EXPERIMENTAL DESIGN AND METHODOLOGY

Natural Antisense Transcript Studies

Karenia brevis *Cultures*

Cultures of *Karenia brevis* were grown under a 12 hour day, 12 hour night cycle (6am-6pm day; 6pm-6am night) at 21°C in L1 medium, which is a general purpose marine medium for growing coastal algae (Guillard & Hargraves, 1993). L1 medium was prepared with 996.5mL of filtered seawater, 1mL $NaNO_3$, 1mL $NaH_2PO_4 \cdot H_2O$, 1mL trace element solution, and 0.5mL vitamin solution. Next, the medium was autoclaved with a 20 minute cycle. Cultures were given media every two weeks at a 4(culture):1(media) ratio.

RNA Extraction

The purification process of total RNA from *Karenia brevis* began with centrifuging 200ml of *K. brevis* culture at 1500 rcf for 5 minutes, discarding the supernatant. The precipitate was then used for total RNA extraction using a Qiagen total RNA extraction kit following manufacturer's instructions with the exception of increasing the elution step to two spins of 50µl each to make up for RNA loss due to DNA digestion and multiple precipitations. These samples were then stored at -20°C overnight in 2.5x volume 100% EtOH and 0.1x volume 3M NaAc. Traces of DNAs were removed from the dsRNAs by DNA digestion using an RQ1 RNase-free DNase kit from Promega following manufacturer's instructions. The samples were stored again at -20°C overnight in 2.5x EtOH 100% and .1x 3M NaAc. Finally, samples were Nano-dropped and bio-analyzed for quantity and quality.

Sequencing

A total of 6 samples were collected for sequencing total RNA. All samples taken were grown together under the same conditions. RNA was extracted from one sample at 12pm and 12am each day over a 72 hour period. In preparation for transcriptomic libraries being sequenced through Illumina Sequencing Services, the 6 total RNA samples were concentrated to 100ng/µl in nuclease free H_2O in a minimum volume of 50µl (5ug total). Libraries taken in the same 24 hour period were pooled and sequenced in individual lanes for a total of 4 lanes. This was done to increase the number of reads with the intent of acquiring a larger, more complete data set. The RNAseq libraries were prepared with Illumina's "TruSeq Stranded RNAseq Sample Prep kit" and were quantitated by qPCR and sequenced on four lanes for 101 cycles from each end (paired-end) of the fragments on a HiSeq2000 using a TruSeq SBS sequencing kit version 3. Fastq files were generated with the software Casava 1.8.2 (Illumina).

Pre-processing

Raw data reads from Illumina received as fastq files were concatenated into four files (day forward, day reverse, night forward, and night reverse) and quality checked using FastQC software. To pre-process the raw reads Fastx (fastx_trimmer) was used to remove the 13nt adaptors (Table 1) on the paired end samples. Reads were then run through FastQC again to re-evaluate quality. Next, kmer values were determined using KmerGenie.

Assembly

Once the quality of the reads and kmer value was acceptable, the reads (Table 2) were assembled into day and night assemblies using two assemblers: SOAPdenovo-trans

and IDBA-Tran. For both assemblers several steps were necessary to run properly. The purpose of using two different assemblers was to determine if the same quality data could be returned.

Post-processing

Post-processing of the assembled sequences began with statistical analysis of the assemblies to validate that they ran correctly and to determine which assembler to use in the downstream pipeline. The Perl script, assemblathon_stats.pl, was used to calculate basic metrics from the assemblies. Based on these statistics both assemblers ran exceptionally well. The IDBA-Tran assembly was chosen due to several factors, including n50 scores and size and number of contigs and scaffolds.

Table 1

Associated Data from Sequencing

Reads are 100nt in length
Average cDNA fragment size: 250nt (range from 80nt to 580nt)
Sequence of adaptors used to make the TruSeq libraries:
Adaptor sequence in read1:
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG (NNNNNN= 6 nt index)
Adaptor sequence in read2:
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCAT

Alignments

Paired-end sequences from day samples were aligned to the day assembly and the night paired-end sequences were aligned to the night assembly using the bowtie2 suite of tools. Bowtie2 allows users to align samples both in the forward and reverse directions; since NATs bind as the reverse compliment to mRNAs this function was used. For each assembly two alignments were created (forward and reverse) and comparisons made. The

output of this process was in the SAM (Sequence Alignment/Map) format, which is used to store large nucleotide sequence alignments (Li et al., 2009). Samtools is a suite of tools used manipulate SAM files. Samtools was used to index the reference transcriptome, convert the SAM file to a BAM file (Binary Alignment /Map), sort the Bam file, and run statistical analyses.

Differential Expression

The differential expression analysis was complete by using the IGV genome browser which allows the visualization of genomic or transcriptomic data. A fasta file of the assembly and two BAM files that consist of alignments of day and night RNA-seq data were pasted into the browser for visualization.

Micro-RNA Studies

Karenia brevis *Cultures*

Cultures of *Karenia brevis* were grown under a 12 hour day, 12 hour night cycle (6am-6pm day; 6pm-6am night) at 21°C in L1 medium, which is a general purpose marine medium for growing coastal algae (Guillard & Hargraves, 1993). L1 medium was prepared with 996.5mL of filtered seawater, 1mL $NaNO_3$, 1mL $NaH_2PO_4 \cdot H_2O$, 1mL trace element solution, and 0.5mL vitamin solution. Next, the medium was autoclaved with a 20 minute cycle. Cultures were given media every two weeks at a 4(culture):1(media) ratio.

miRNA extraction

Purification of total RNA from *Karenia brevis* was accomplished by centrifuging 100ml of *K. brevis* culture at 1500 rcf for 5 minutes, discarding the supernatant. Micro-RNA samples to be sequenced were prepared by following the organic extraction, total RNA isolation and isolation of small RNAs from total RNAs protocols from the mirVana™ miRNA Isolation Kit by Life Technologies. These samples were then stored at -20°C overnight in 2.5x volume 100% EtOH and 0.1x volume 3M NaAc. Traces of DNAs were removed from the dsRNAs by DNA digestion using an RQ1 RNase-free DNase kit from Promega following manufacturer's instructions. The samples were stored again at -20°C overnight in 2.5x EtOH 100% and 0.1x 3M NaAc. Finally, samples were Nano-dropped and bio-analyzed for quantity and quality.

Sequencing

Six samples were collected for sequencing miRNAs. All samples taken were grown together under the same conditions. RNA was extracted from one sample at 12pm and 12am each day over a 72 hour period. In preparation for transcriptomic libraries being sequenced through HiSeq high-throughput sequencing, the 6 miRNA samples were kept at individual concentrations (all exceeding the minimum 100ng/µl each) in nuclease free H_2O in a 10µl volume. Libraries taken in the same 24 hour period were pooled and sequenced in individual lanes for a total of 4 lanes.

Pre-processing

After sequencing, the raw reads were filtered by removing adaptor sequences, and removing contamination and low-quality reads from raw reads. Raw data reads received

as fastq files were concatenated into two files (day and night miRNAs). Quality checking was done using FastQC software.

Assembly

This assembly was run with IDBA-Tran as a single-end assembly using the day and night fasta files from the RNA-seq assemblies, the day and night miRNA fasta files and a fasta file containing all *Karenia brevis* ESTs from NCBI. This allowed for a deeper sequencing in an attempt to increase the pool of potential novel small RNAs

Post-processing

Assemblathon_stats.pl was used to gather statistical analyses from the assembly to determine how well it performed.

Alignments

Both the day and night miRNA reads were individually aligned to the IDBA-Tran assembly. In single-end alignments it was unnecessary to give the software a command to run in reverse, it automatically checked for both forward and reverse alignments. The output of this process was in the SAM format. Samtools was used to index the reference transcriptome, convert the SAM file to a BAM file, sort the Bam file, and run statistical analyses.

Differential Expression

The differential expression analysis was complete by using the IGV genome browser which allows the visualization of genomic or transcriptomic data. A fasta file of the assembly and two BAM files that consist of alignments of day and night miRNAs that were aligned to the assembly in the fasta file were used. Only the top 12 largest transcripts were visualized.

CHAPTER IV

ANALYSIS OF DATA

Natural Antisense Transcript Analysis

Quantity

Raw sequence reads received from Illumina were categorized by the sample time where D1 is the first day sample, D2 is the second day sample, etc., and N1 is the first night sample, etc. By determining where the 6 nucleotide barcode is on each raw sequence read, each could be categorized as the forward (R1) or reverse (R2) read of the paired-end sequence (see Table 2).

Table 2

Total Number of Raw RNA-seq Reads

Sample	Name of Fastq	# of Reads	Sample	Name of Fastq	# of Reads
D1	D1_ATCACG_L004_R1_001.fastq	30552347	D1	D1_ATCACG_L005_R1_001.fastq	30329120
	D1_ATCACG_L004_R2_001.fastq	30552347		D1_ATCACG_L005_R2_001.fastq	30329120
D2	D2_CGATGT_L004_R1_001.fastq	29010731	D2	D2_CGATGT_L005_R1_001.fastq	28836656
	D2_CGATGT_L004_R2_001.fastq	29010731		D2_CGATGT_L005_R2_001.fastq	28836656
D3	D3_TTAGGC_L004_R1_001.fastq	35945210	D3	D3_TTAGGC_L005_R1_001.fastq	35709203
	D3_TTAGGC_L004_R2_001.fastq	35945210		D3_TTAGGC_L005_R2_001.fastq	35709203
D4	D4_TGACCA_L004_R1_001.fastq	33594977	D4	D4_TGACCA_L005_R1_001.fastq	33374999
	D4_TGACCA_L004_R2_001.fastq	33594977		D4_TGACCA_L005_R2_001.fastq	33374999
N1	N1_ACAGTG_L004_R1_001.fastq	31895958	N1	N1_ACAGTG_L005_R1_001.fastq	31677337
	N1_ACAGTG_L004_R2_001.fastq	31895958		N1_ACAGTG_L005_R2_001.fastq	31677337
N2	N2_GCCAAT_L004_R1_001.fastq	33645489	N2	N2_GCCAAT_L005_R1_001.fastq	33495762
	N2_GCCAAT_L004_R2_001.fastq	33645489		N2_GCCAAT_L005_R2_001.fastq	33495762
N3	N3_CAGATC_L004_R1_001.fastq	33783728	N3	N3_CAGATC_L005_R1_001.fastq	33596715
	N3_CAGATC_L004_R2_001.fastq	33783728		N3_CAGATC_L005_R2_001.fastq	33596715
	Total Reads:	**456856880**		**Total Reads:**	**454039584**
D1	D1_ATCACG_L006_R1_001.fastq	30513177	D1	D1_ATCACG_L007_R1_001.fastq	20386841
	D1_ATCACG_L006_R2_001.fastq	30513177		D1_ATCACG_L007_R2_001.fastq	20386841
D2	D2_CGATGT_L006_R1_001.fastq	29003839	D2	D2_CGATGT_L007_R1_001.fastq	19894031
	D2_CGATGT_L006_R2_001.fastq	29003839		D2_CGATGT_L007_R2_001.fastq	19894031
D3	D3_TTAGGC_L006_R1_001.fastq	35956845	D3	D3_TTAGGC_L007_R1_001.fastq	25034036
	D3_TTAGGC_L006_R2_001.fastq	35956845		D3_TTAGGC_L007_R2_001.fastq	25034036
D4	D4_TGACCA_L006_R1_001.fastq	33565147	D4	D4_TGACCA_L007_R1_001.fastq	23105353
	D4_TGACCA_L006_R2_001.fastq	33565147		D4_TGACCA_L007_R2_001.fastq	23105353
N1	N1_ACAGTG_L006_R1_001.fastq	31872907	N1	N1_ACAGTG_L007_R1_001.fastq	21881200
	N1_ACAGTG_L006_R2_001.fastq	31872907		N1_ACAGTG_L007_R2_001.fastq	21881200
N2	N2_GCCAAT_L006_R1_001.fastq	33671974	N2	N2_GCCAAT_L007_R1_001.fastq	23461774
	N2_GCCAAT_L006_R2_001.fastq	33671974		N2_GCCAAT_L007_R2_001.fastq	23461774
N3	N3_CAGATC_L006_R1_001.fastq	33778324	N3	N3_CAGATC_L007_R1_001.fastq	23268815
	N3_CAGATC_L006_R2_001.fastq	33778324		N3_CAGATC_L007_R2_001.fastq	23268815
	Total Reads:	**456724426**		**Total Reads:**	**314064100**
				Lane 4	**456856880**
				Lane 5	**454039584**
				Lane 6	**456724426**
				Lane 7	**314064100**
				Combined Total Reads	**1681684990**

FastQC Analysis Of RNAseq Raw/Processed Illumina Reads

A number of various quality checks were run on the sequence reads to find and eliminate poor quality reads or process the reads to eliminate regions such as adaptors. The base sequence quality check shows an overview of the quality scores across the length of the read. Good quality calls fall within the green shaded area, reasonable quality calls fall within the orange shaded area, and calls of poor quality are in the red shaded area (Figure 9).

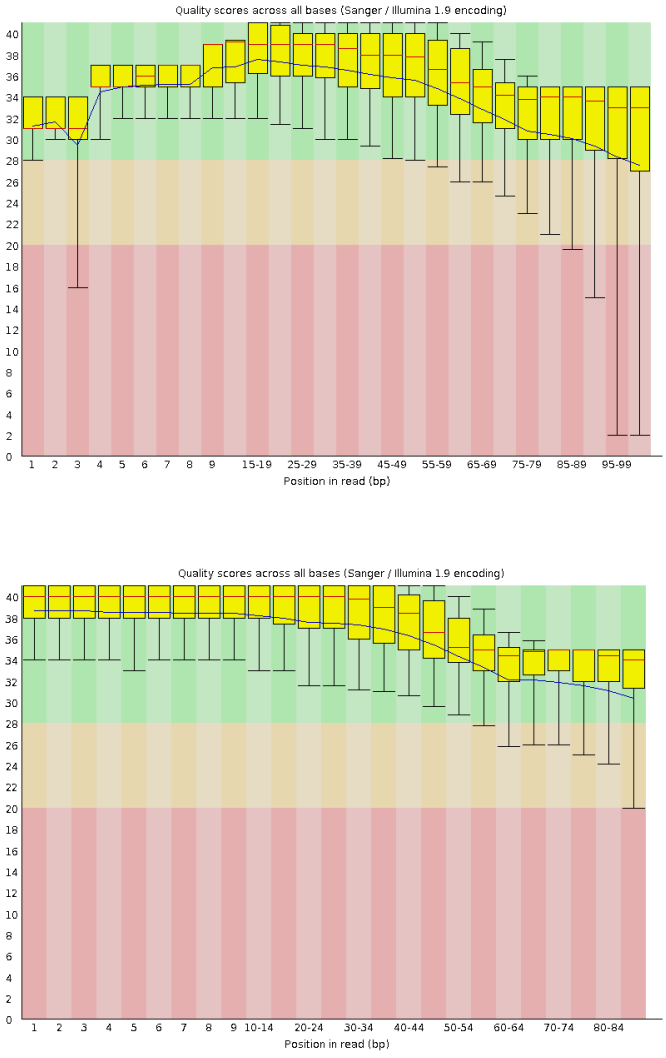
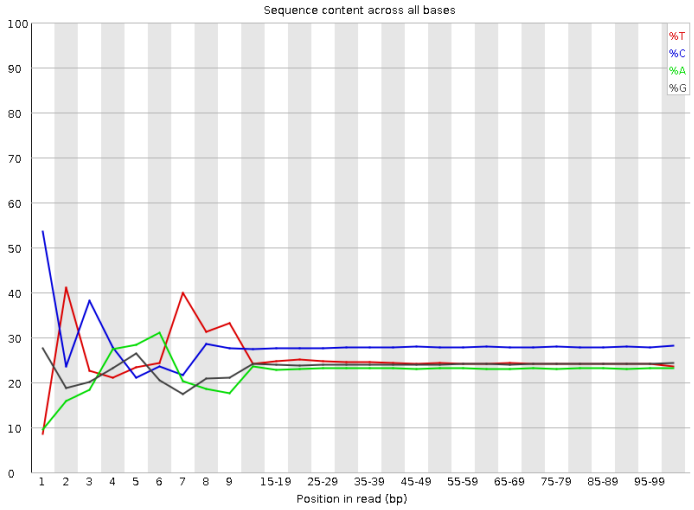


Figure 9. Base sequence quality check before and after processing of the raw RNA-seq data. The top panel shows the quality of base-calling at each nucleotide position or range

of nucleotides for the raw RNA-seq reads before applying FastQC analysis. The bottom panel shows the quality of the new nucleotide positions after applying the analysis. The y-axis shows quality scores. The central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points and the blue line represents the mean quality.

The base sequence content check determined if the ATGC content is in proportion. In a genomic/transcriptomic library there should be little to no difference between the different bases of a sequence run. If strong biases occur between bases then overrepresented sequence are contaminating the library. If there is a difference greater than 20% between base sequence content, the quality check will fail. The first 13 nucleotides contain the adapter sequence used in sequencing and should be biased as shown in the upper panel of Figure 10. After processing, i.e. removal of the adapter, the remaining sequences are unbiased along their whole lengths (bottom panel of Figure 10).



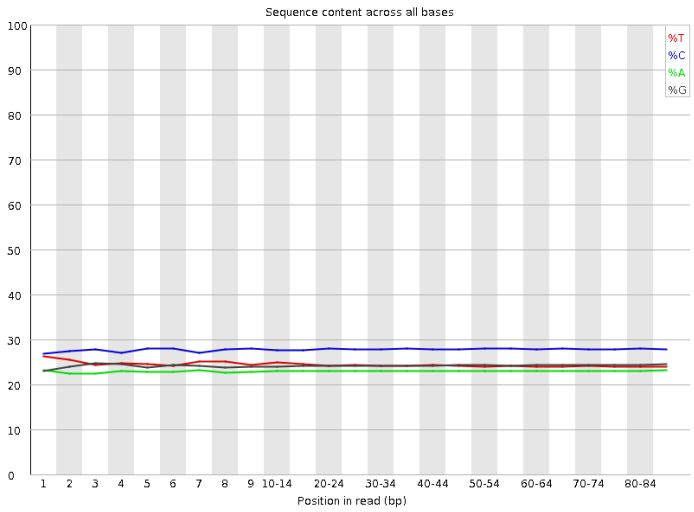
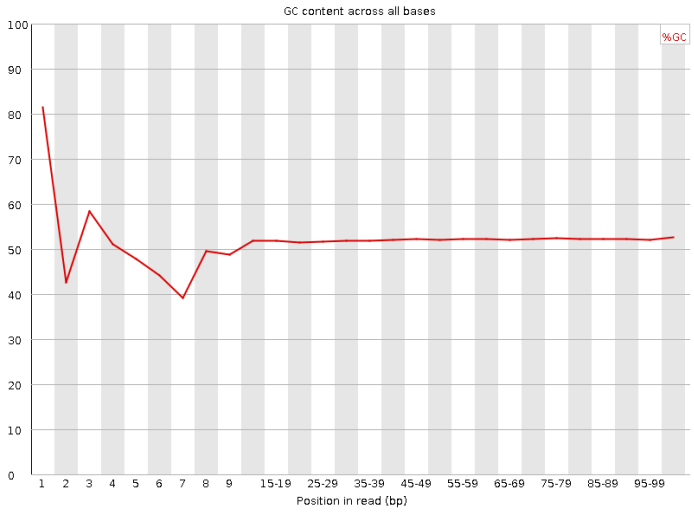


Figure 10. Base sequence content before and after processing raw RNA-seq data. The top panel shows the presence of an adapter sequence causing bias. The bottom panel shows that removing the adapter removes the bais.

Similar to the per base sequence content check, the per base GC content check should contain little to no difference in base content of a sequence run, so the line in this plot should run horizontally across the graph. If the GC content is more than 10% from the mean content the quality check will fail. Similar to the above data, the first 13 nucleotides show a GC bias prior to processing (top panel of Figure 11), whereas after processing (bottom panel Figure 11) the GC bias is no longer present.



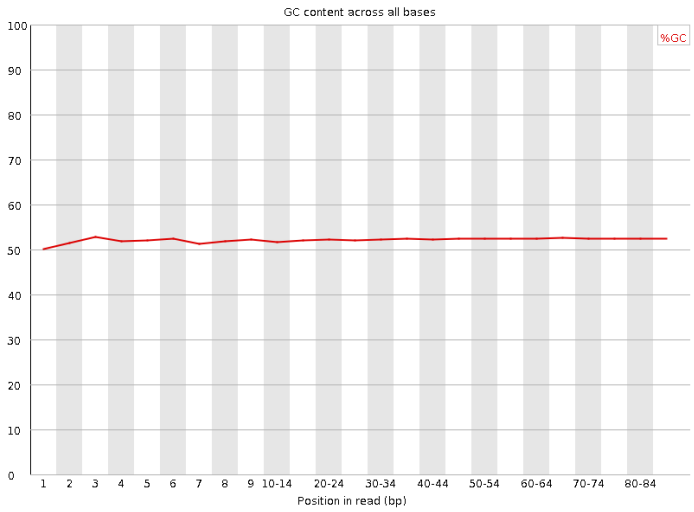
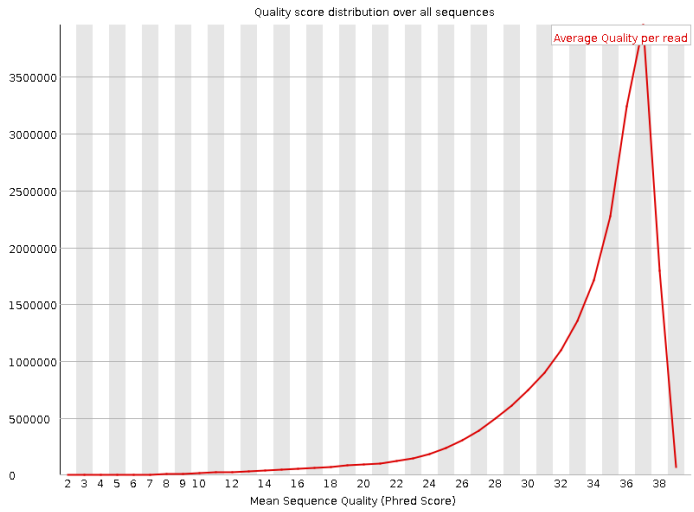


Figure 11. Base GC content before and after processing of raw RNAseq data. The top panel shows the biased GC content. The bottom panel shows normal GC content after removal of the adapter sequence.

The per sequence quality score check shows if sequences have universally low quality values. A warning is raised if the most frequently observed mean quality is below 27 (0.2% error rate). The check will fail if the most frequently observed mean quality is below 20 (1% error rate). Figure 13 shows that the average quality score per read is well above the warning cut-off for both the pre-processed and post-processed sequence reads.



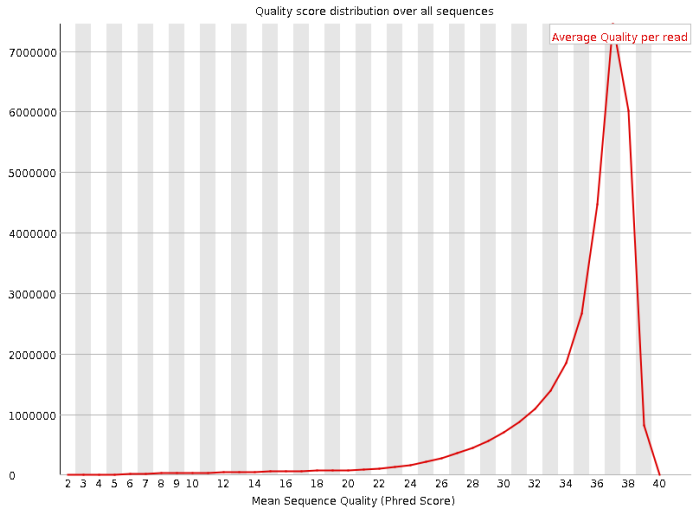


Figure 12. Sequence quality score before and after processing of raw RNAseq data for. The top panel shows that before processing the raw reads were already of high quality. The bottom shows that after processing the raw reads remained high quality.

Assembly Statistics

After cleaning and quality checking all of the raw data, it was assembled using two different assemblers (SOAPdenovo-trans and IDBA-Tran). SOAPdenovo-trans and IDBA-Tran were chosen for their speed and accuracy of assembly. Statistical analysis of the two assemblies was performed by using the Perl script assemblathon_stats.pl. The output of each assembly was categorized in a variety of ways for comparison purposes (Table 3).

Table 3

Statistical variances between two transcriptomic assemblers

Soapdenovo-trans	Day		Night		IDBA-tran	Day		Night	
Scaffolds					**Scaffolds**				
Number of scaffolds	313718		311612		Number of scaffolds	201661		204386	
Total size of scaffolds	197134166		196309134		Total size of scaffolds	302367796		304371577	
Longest scaffold	24793		18738		Longest scaffold	26076		18315	
Shortest scaffold	100		100		Shortest scaffold	300		300	
Number of scaffolds > 1K nt	70492	22.50%	70607	22.7%	Number of scaffolds > 1K nt	126021	62.5%	127332	62.3%
Number of scaffolds > 10K nt	174	0%	100		Number of scaffolds > 10K nt	354	0.2%	215	0.1%
Mean scaffold size	628		630		Mean scaffold size	1499		1489	
Median scaffold size	174		176		Median scaffold size	1290		1288	
N50 scaffold length	1686		1677		N50 scaffold length	1952		1941	
L50 scaffold count	37039		37370		L50 scaffold count	51367		52394	
% of assembly in scaffolded contigs	27.0%		27.4%		% of assembly in scaffolded contigs	0.0%		0.0%	
% of assembly in unscaffolded contigs	73.0%		72.6%		% of assembly in unscaffolded contigs	100.0%		100.0%	
Average number of contigs per scaffold	1.1		1.1		Average number of contigs per scaffold	1		1	
Ave. L of break (>25 Ns) b/w contigs in scaffold	55		55		Ave. L of break (>25 Ns) b/w contigs in scaffold	0		0	
Contigs					**Contigs**				
Number of contigs	351399		350065		Number of contigs	201661		204386	
Number of contigs in scaffolds	66021		67372		Number of contigs in scaffolds	0		0	
Number of contigs not in scaffolds	285378		282693		Number of contigs not in scaffolds	201661		204386	
Total size of contigs	195055880		194191759		Total size of contigs	302367796		304371577	
Longest contig	24793		15032		Longest contig	26076		18315	
Shortest contig	100		100		Shortest contig	300		300	
Number of contigs > 1K nt	66708	19.00%	66541	19%	Number of contigs > 1K nt	126021	62.5%	127332	62.3%
Number of contigs > 10K nt	43		23		Number of contigs > 10K nt	354	0.2%	215	0.1%
Mean contig size	555		555		Mean contig size	1499		1489	
Median contig size	188		189		Median contig size	1290		1288	
N50 contig length	1428		1417		N50 contig length	1952		1941	
L50 contig count	42712		43095		L50 contig count	51367		52394	

Figures 13-16 highlight the key characteristics that I used in choosing one assembly over the other for the downstream pipeline.

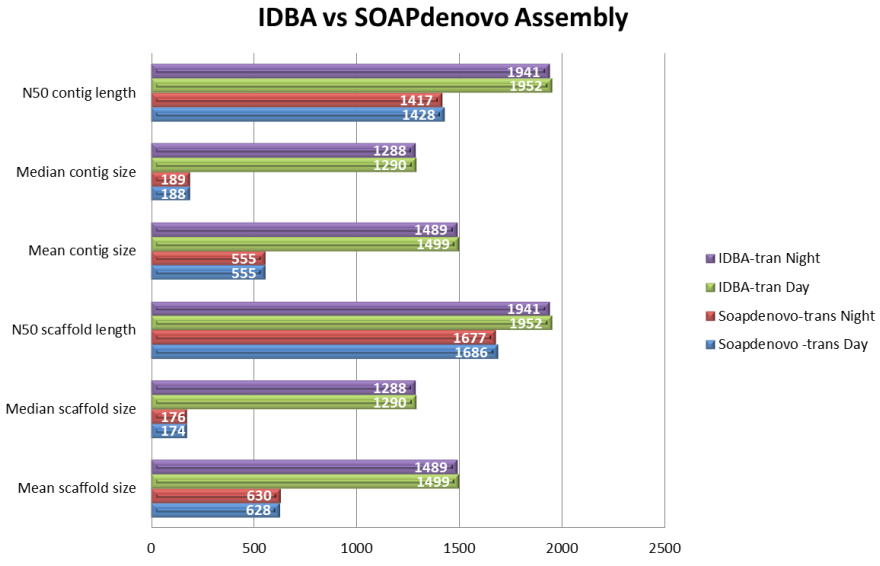


Figure 13. Assembly analysis of the mean, median and N50 scores for contigs and scaffolds. The x-axis is length in nucleotides.

The N50 score represents contig length, where 50% of a de novo assembly lies in blocks this size or larger. The larger the N50 score means larger overall contig and scaffold length, which suggests a better built assembly.

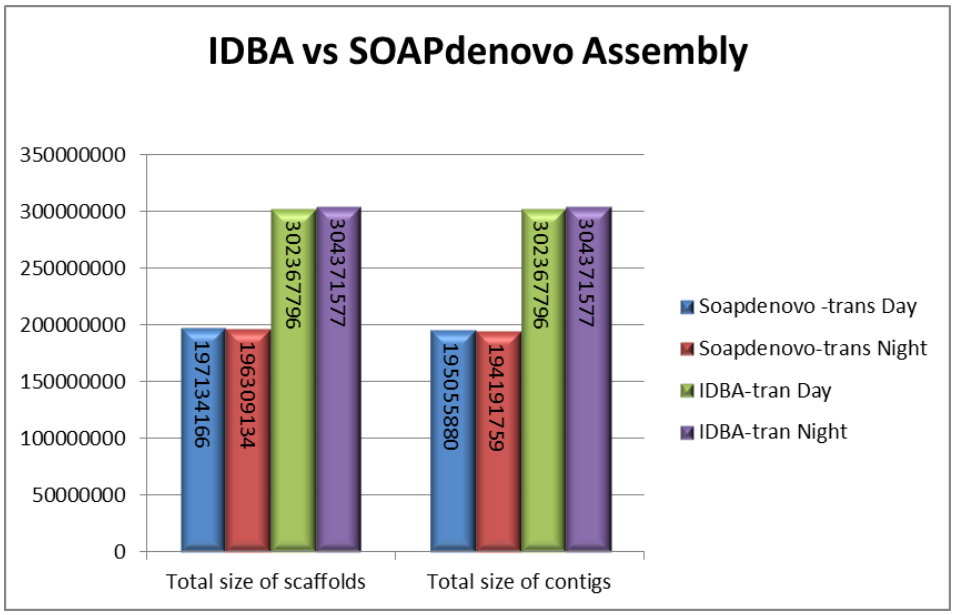


Figure 14. Assembly analysis of the total size of scaffolds and contigs. IDBA was able to create an assembly with much larger scaffolds and contigs.

In all categories, IDBA produced larger values than those for SOAPdenovo (Figure 13). Similarly, IDBA produced larger scaffolds and larger contigs (Figure 14). The scaffold and contig number is larger in the SOAPdenovo assembly, but the IDBA assembly has many more scaffolds and contigs that are 1K in size or larger which is preferred over numbers of scaffolds and contigs that may be smaller than 1K (Figure 15). The IDBA assembly also contains many more scaffolds and contigs that are 10K in size or greater (Figure 16). In total, many factors suggested that the IDBA assembly was preferable over the SOAPdenovo assembly for further analysis.

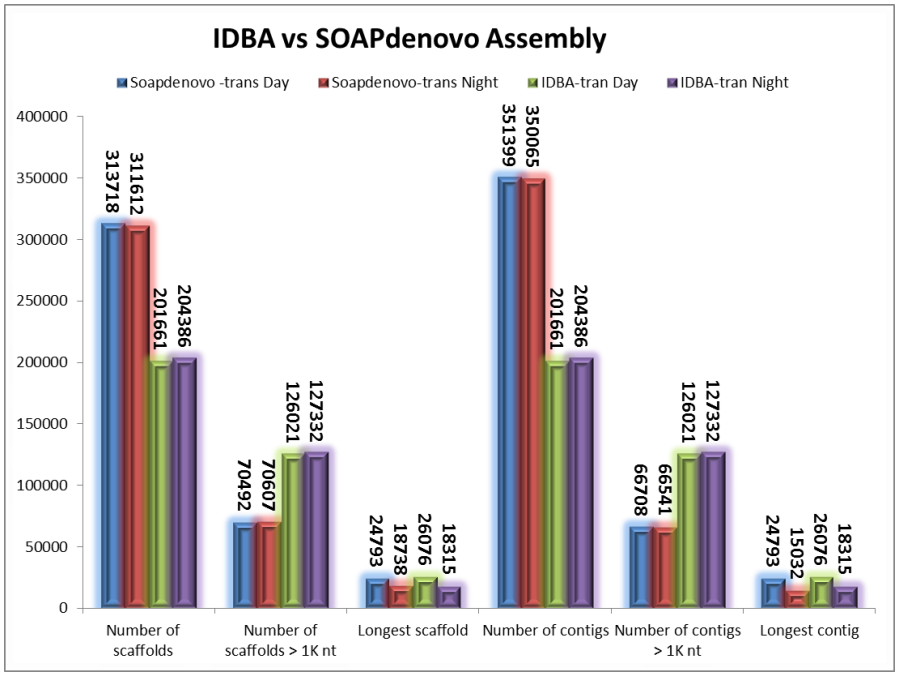


Figure 15. Assembly analysis of the number of scaffolds and contigs and longest scaffold and contigs. The scaffold and contig numbers are larger for SOAPdenovo, while IDBA has many more that are 1K or larger.

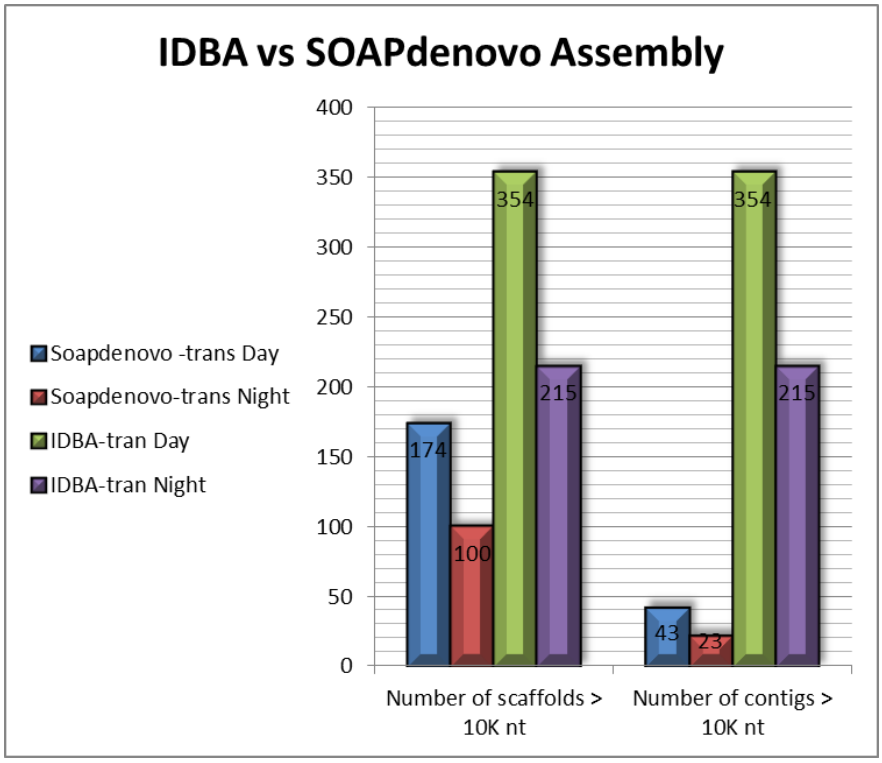


Figure 16. Assembly analysis of scaffolds and contigs larger than 10,000 nucleotides.

Analysis of Alignments with the SAM and BAM formats

Using the IDBA day and night assemblies, each data set was aligned back to the respective assembly using Bowtie. Basic alignment statistics (Table 4) are printed out in the SAM format at the end of each Bowtie alignment run and are useful in determining what types of alignments are present in the build. Concordant alignments are those that align within the expected mate orientation and the expected range of distances between mates. In contrast discordant alignments do not meet paired-end expectations but both mates will have unique alignments, which can be desirable when seeking structural variants (Lapidot & Pilpel, 2006). Reads may also be separated into mates in an attempt to align them individually if they do not align as a pair.

Table 4

Alignment Statistics

Day_IDBA.sam	Night_IDBA.sam
351172036 reads; of these:	366029983 reads; of these:
351172036 (100.00%) were paired; of these:	366029983 (100.00%) were paired; of these:
57433688 (16.35%) aligned concordantly 0 times	53980337 (14.75%) aligned concordantly 0 times
133146586 (37.91%) aligned concordantly exactly 1 time	138792915 (37.92%) aligned concordantly exactly 1 time
160591762 (45.73%) aligned concordantly >1 times	173256731 (47.33%) aligned concordantly >1 times
57433688 pairs aligned concordantly 0 times; of these:	53980337 pairs aligned concordantly 0 times; of these:
5100353 (8.88%) aligned discordantly 1 time	4281765 (7.93%) aligned discordantly 1 time
52333335 pairs aligned 0 times concordantly or discordantly; of these:	49698572 pairs aligned 0 times concordantly or discordantly; of these:
104666670 mates make up the pairs; of these:	99397144 mates make up the pairs; of these:
78889093 (75.37%) aligned 0 times	74548124 (75.00%) aligned 0 times
8048952 (7.69%) aligned exactly 1 time	8391360 (8.44%) aligned exactly 1 time
17728625 (16.94%) aligned >1 times	16457660 (16.56%) aligned >1 times
88.77% overall alignment rate	89.82% overall alignment rate

This set of statistics comes from the original paired-end reads being aligned back to the assembly (being used as the reference genome). During the course of processing many reads are discarded (i.e. low quality, trimmed ends, and adaptors), causing the final alignments to not have a 100% overall alignment rate. Also if paired-end reads don't

meet specific requirements (set by the Bowtie aligner) the reads don't align (Langmead & Salzberg, 2012). Reads that align concordantly 0 times signifies that the percentage given in that row (16.35%) is the percentage of reads that did not align concordantly. The next two fields show that 84% did align concordantly. Next, out of the 16.35% that did not align concordantly, 8.88% of them aligned discordantly (Table 4). The remaining percentage of reads that did not aligned concordantly or discordantly are separated from their mates and aligned individually along the assembly and a percentage of how many aligned or did not align is given (Table 4). Figures 17 and 18 show that the day and night reads aligned with an overall alignment of 90%.

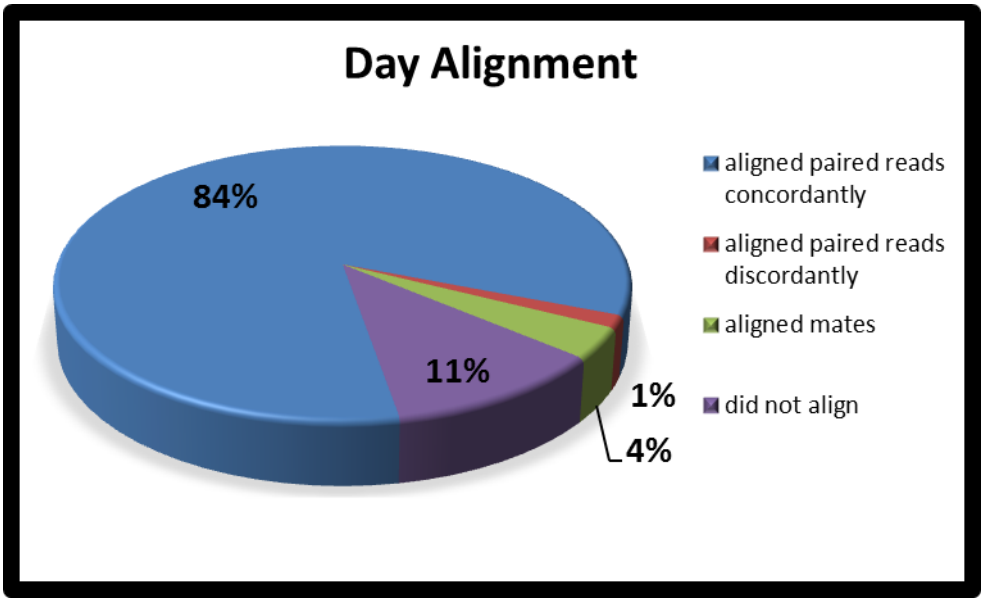


Figure 17. Bowtie paired-end alignment of the day IDBA assembly.

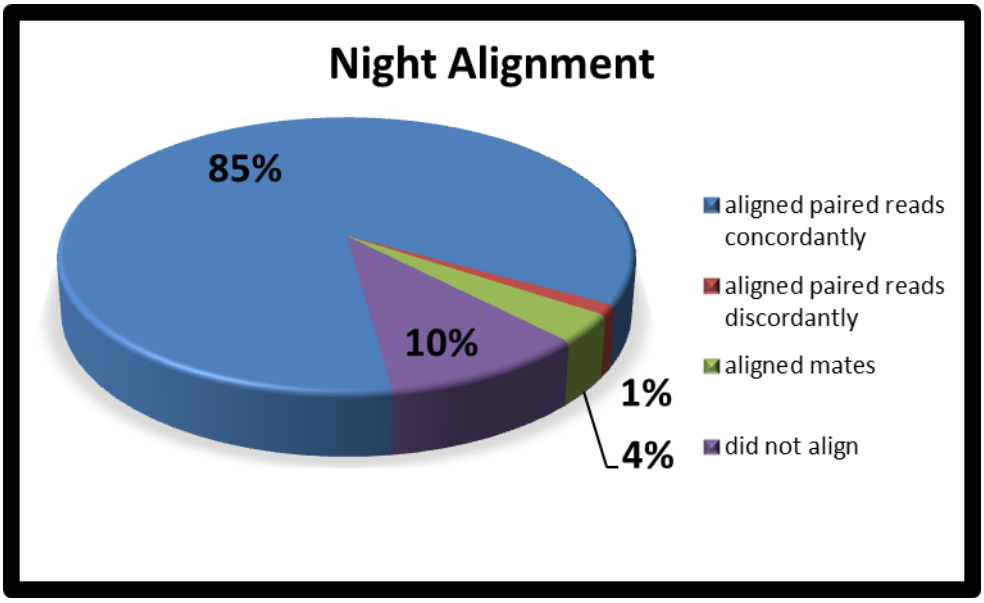


Figure 18. Bowtie paired-end alignment of the night IDBA assembly.

Once samples are converted to the BAM format, they are ready for use in several pipelines such as gene annotation, genome browsers, SNP calling, and differential expression. Before moving on to these final stages, it is necessary to quality check the alignments (Table 5). Flagstat is part of the Samtools suite of tools that quality checks the Bam files. This checks number of reads total, number of reads mapped, individual mates per pair, and singletons mapped. Most importantly, it checks the number of proper pairs mapped to the assembly (Table 5). MAPQ (map quality) is an important parameter to set before converting a Sam file to a BAM file. By default this is set to 0. Having a MAPQ = 0 means that the read maps may map to multiple locations. By setting this to 10 the likelihood of getting a unique transcript is increased. The values discussed show that the BAM files are of sufficient quality to continue with additional analyses.

BAM files from the day and night alignments were used in the IGV genome browser. Analysis revealed differential expression among several transcripts. A few of the most widely diverging expression profiles are shown in Table 6. Some of the

transcripts are more highly expressed at night and some during the day. It is not unexpected that expression for transcripts is found at both times since the vast majority of dinoflagellate mRNAs are continuously being expressed (Morey et al.,2011 and Van Dolah et al.,2007).

Table 5

Day and Night BAM file quality check

Day_IDBA.bam	Night_IDBA.bam
314166879 + 0 in total (QC-passed reads + QC-failed reads)	327133270 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates	0 + 0 duplicates
314166879 + 0 mapped (100.00%:-nan%)	327133270 + 0 mapped (100.00%:-nan%)
314166879 + 0 paired in sequencing	327133270 + 0 paired in sequencing
157420227 + 0 read1	163940153 + 0 read1
156746652 + 0 read2	163193117 + 0 read2
297110114 + 0 properly paired (94.57%:-nan%)	311219418 + 0 properly paired (95.14%:-nan%)
309496544 + 0 with itself and mate mapped	322049384 + 0 with itself and mate mapped
4670335 + 0 singletons (1.49%:-nan%)	5083886 + 0 singletons (1.55%:-nan%)
4167216 + 0 with mate mapped to a different chr	4558904 + 0 with mate mapped to a different chr
4167216 + 0 with mate mapped to a different chr (mapQ>=5)	4558904 + 0 with mate mapped to a different chr (mapQ>=5)

Table 6

Differential Expression Analysis

	Name	Length	# of reads	% difference
Day	transcript-63_100796	1623	624	55%
Night	transcript-63_100796	1623	1394	
Day	transcript-63_4798	7777	2296	45%
Night	transcript-63_4798	7777	1259	
Day	transcript-63_5015	8513	1415	45%
Night	transcript-63_5015	8513	784	
Day	transcript-63_100591	477	2415	45%
Night	transcript-63_100591	477	4423	
Day	transcript-63_399	8668	4021	45%
Night	transcript-63_399	8668	7281	

Transcript 15992 shows specific, regulated degradation of the night transcript. This amount of degradation would not allow for translation of the transcript.

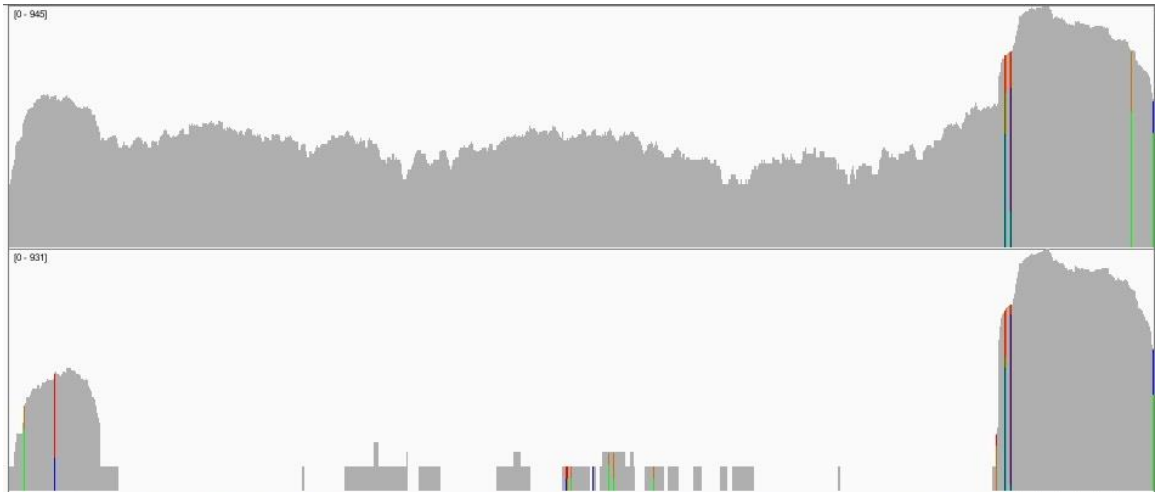


Figure 19. IGV genome browser visualization of transcript-63_15992 for both the day (top) and night (bottom) transcript. Degradation of the night transcript.

Micro-RNA Analysis

The data returned from the HiSeq run showed a large number of clean reads and clean bases (Table 7). This data along with the FastQC analysis was used in determining overall quality of the reads.

Table 7

miRNA Read Statistics Results

Sample Name	Clean reads	Clean bases	Read length (bp)	GC (%)
Dm1	17,443,655	733,362,713	49	41.8%
Dm2	8,353,359	343,042,414	49	41.8%
Dm3	8,176,634	334,601,148	49	41.8%
Nm1	13,013,878	542,834,465	49	41.7%
Nm2	6,087,649	248,376,637	49	42.0%
Nm3	13,861,404	570,604,301	49	41.9%

The length distribution for all 6 miRNA samples contained similar results. These figures showed that a large majority of the reads were 40nt long or larger (Figure 20). miRNAs are typically 20 to 24nt long.

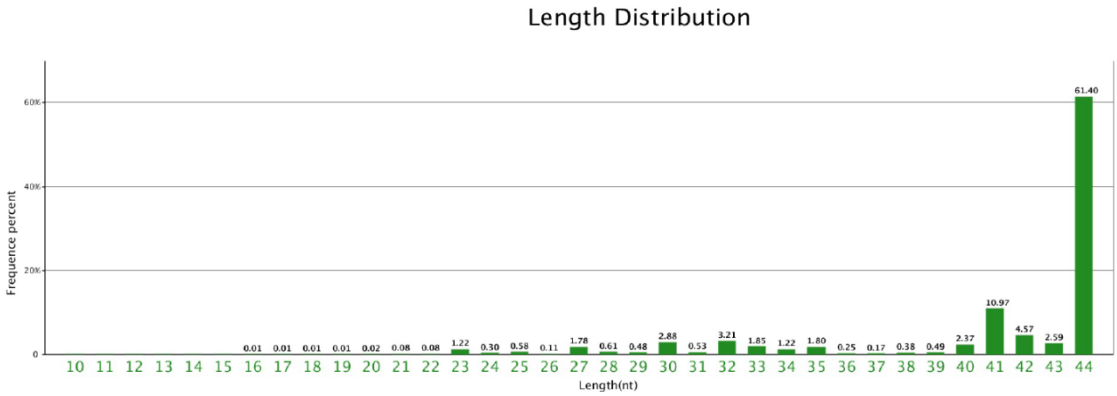


Figure 20. Length Distribution of miRNAs. This distribution shows that roughly 80% of the miRNA reads were 40nt or larger.

FastQC base quality scores show an overview of the quality scores across the length of the read. The y-axis shows quality scores, good quality calls are green, reasonable quality are orange, and calls of poor quality are red. Figure 21 shows good quality scores for the miRNA reads

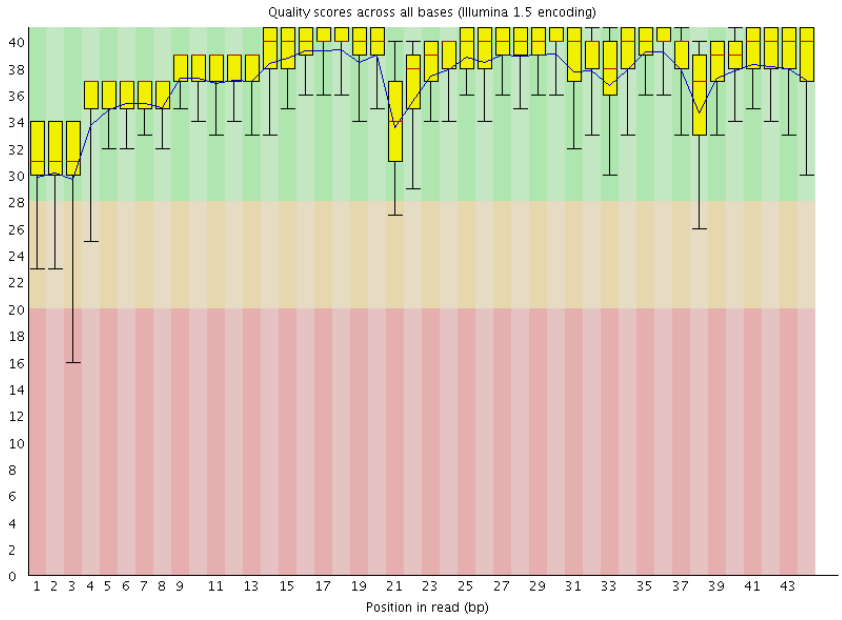


Figure 21. FastQC base sequence quality of miRNAs. This shows the quality of base-calling at each nucleotide position or range of nucleotides for miRNA reads that are processed and ready for assembly. The y-axis shows quality scores. The central red line is the median value, the yellow box represents the inter-quartile range (25-75%), the upper

and lower whiskers represent the 10% and 90% points and the blue line represents the mean quality.

The per sequence quality score check shows if sequences have universally low quality values. A warning is raised if the most frequently observed mean quality is below 27 (0.2% error rate). The check will fail if the most frequently observed mean quality is below 20 (1% error rate). The Phred score or mean quality score showed that the overall average quality per read was good (Figure 22).

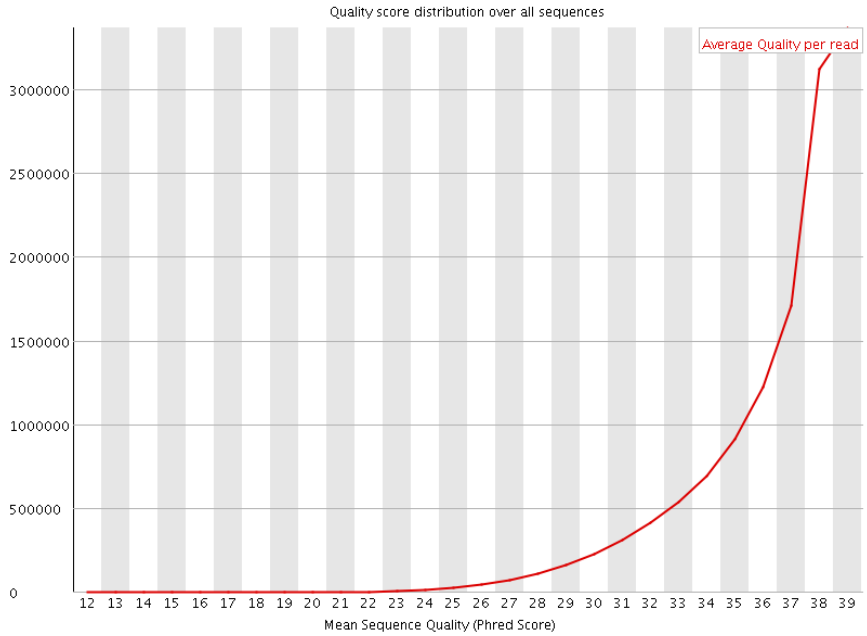


Figure 22. FastQC sequence quality score of miRNAs.

The Perl script "assemblathon_stats.pl" was used to access relevant information about the IDBA-Tran single-end assembly build. IDBA-Tran aligns contigs to form transcripts rather than contigs in scaffolds. These transcripts are later used in different de novo pipelines. This tool is strictly a de novo assembler based on sequencing RNA reads only. IDBA-Tran uses local assembly to reconstruct kmers in low-expressed transcripts and then utilizes an advanced cutoff on contigs to separate graphs into components that corresponds to a gene and contains few transcripts (Peng, Leung, Yiu, & Chin, 2010).

The IDBA assembler was used to build an assembly containing the day and night reads from the RNA-seq data set and the day and night reads from the miRNA data set and from *K.brevis* ESTs found on the NCBI website. This produced a larger assembly in the hopes of finding more locations that the miRNA reads would align back to. The details characterizing the resulting assembly are shown in Table 8.

Table 8

Statistical Analysis of the miRNA Assembly

IDBA-tran miRNA Assembly Scaffolds/Contigs		
Number of scaffolds/contigs	202163	
Total size of scaffolds/contigs	302971115	
Longest scaffold/contig	26076	
Shortest scaffold/contig	300	
Number of scaffolds/contigs > 1K nt	126298	62.5%
Number of scaffolds/contigs > 10K nt	356	0.2%
Mean scaffold/contig size	1499	
Median scaffold/contig size	1290	
N50 scaffold/contig length	1950	
L50 scaffold/contig count	51488	

The blast results for the miRNAs showed many hits to different miRNA families. Most were single hits per family but, mir-125,159, 204 and 219 showed several hits to each family (Table 9). Amongst these families, 62 mature miRNA candidates were found and are awaiting further testing for criteria matching.

Table 9

Day, Night and Common Hits among miRNA Families

Day Hits Only 13 miRNA Families	**Day&Night Common Hits** 6 miRNA Families	**Night Hits Only** 27 miRNA Families	
miR-125	miR-219	miR-15	miR-5119
miR-159	miR-4177	miR-2	miR-5292
miR-1277	miR-427	miR-204	miR-549
miR-1692	miR-466	miR-219	miR-5658
miR-219	miR-5658	miR-3168	miR-574
miR-302	miR-5831	miR-341	miR-5831
miR-4177		miR-4177	miR-6106
miR-4185		miR-4249	miR-6114
miR-427		miR-427	miR-6421
miR-466		miR-458	miR-6478
miR-5658		miR-466	miR-6905
miR-5831		miR-4680	miR-716
miR-6529		miR-50	miR-7438
		miR-5106	

Figures 23-25 are showing the highly conserved regions of mature miRNAs that matched the *K. brevis* miRNA reads. Mir-125 is showing a perfect match to highly conserved miRNAs, while mir-159 is only showing nearly perfect matches (Figure 23). Mir-204 was only found in the night samples, and the *K. brevis* analog shows near perfect alignment with other known mir-204 sequences (Figure 24). The mir-219 conserved region shows matches for both day and night samples. This would be a potential target for differential expression analysis since it is present in both data sets (Figure 25).

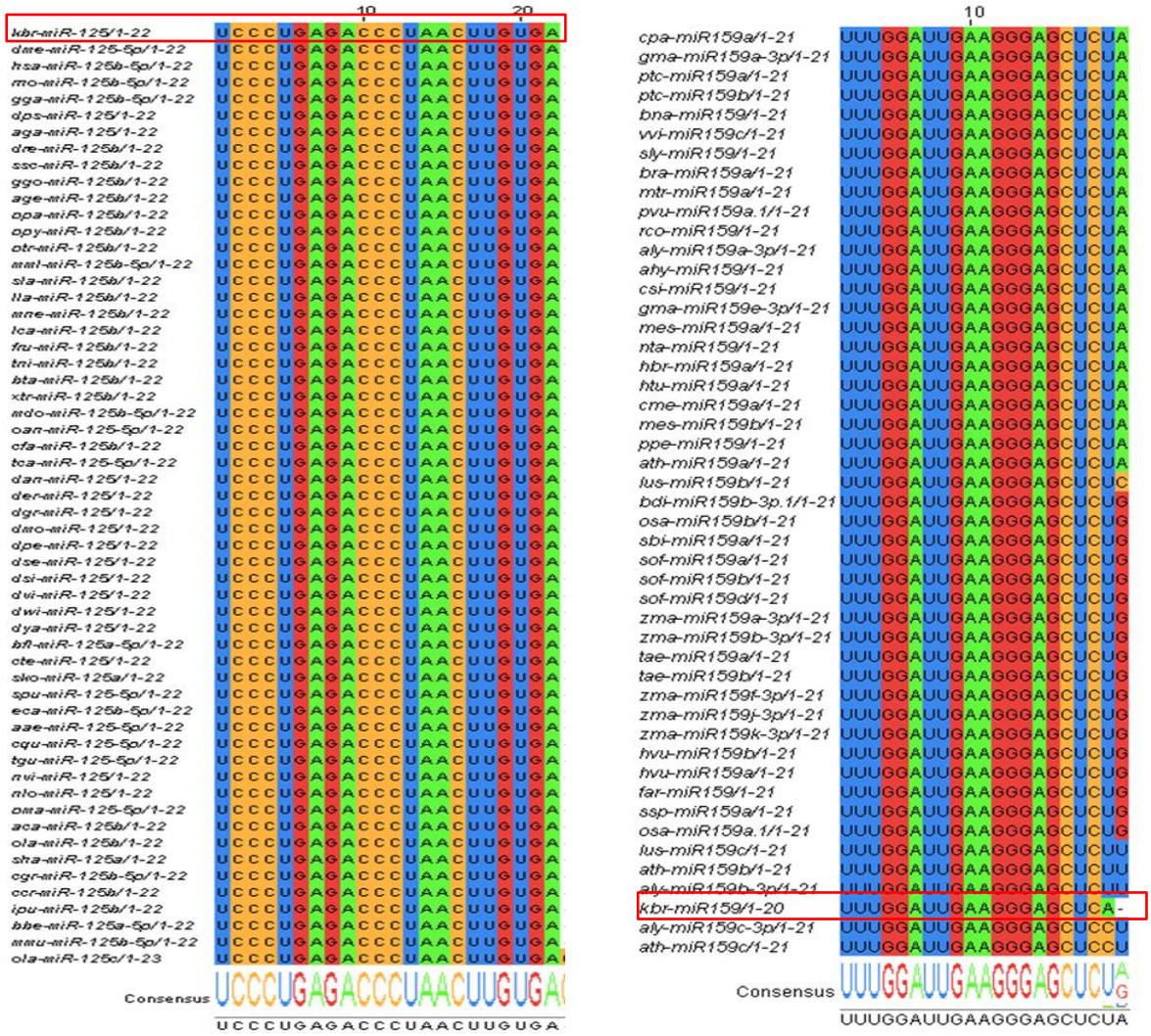


Figure 23. Alignments of mir-125 and mir-159 miRNAs including putative miRNA sequences from *Karenia brevis*. Perfect matches to highly conserved animal mature miRNA clustered in the mir-125 Family (Left). Near-perfect matches to conserved plant mature miRNA in the mir-159 Family (Right).

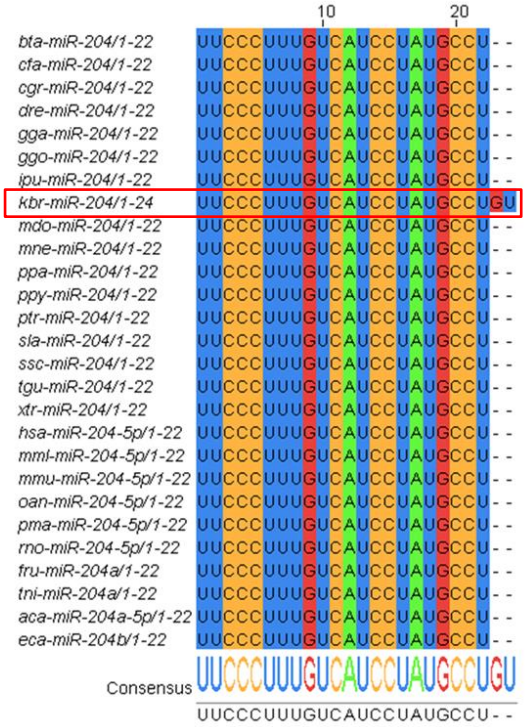


Figure 24. Matches to highly conserved miRNA clusters from night samples. Near-perfect matches to highly conserved animal mature miRNA clustered in the mir-204 Family.

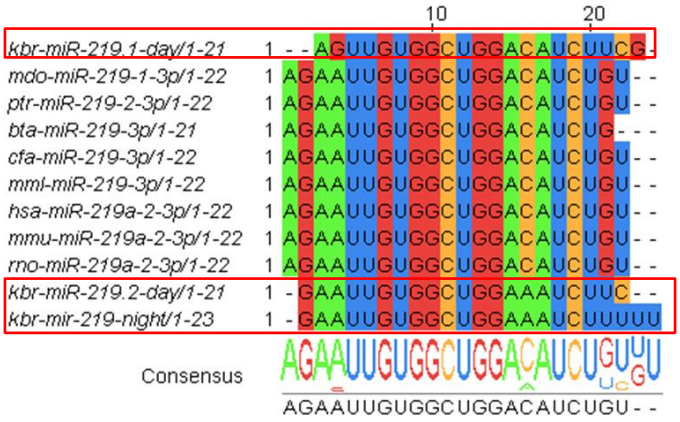


Figure 25. Multiple *Karenia brevis* miRNA sequences show near perfect alignment with the mir-219 family. The mir-219 family is a highly conserved animal mature miRNA. Taken from both day and night miRNA reads.

The IGV genome browser was utilized in an effort to determine if there is differential expression present among transcripts of small RNAs. Figure 26 shows a few things for a representative transcript. First, it shows evidence of a degradation pathway.

This is visualized by cleavage sites that make the transcript look sculpted into columns and by the mismatches that align the edges. Secondly, it shows differential expression of the day (top) and night (bottom) transcript, which can be seen in the shaded coverage area of the figure. Lastly, it also shows read alignments (red and blue) that have been transcribed by what may be convergent transcription, which may suggest the *cis*-NAT pathway.

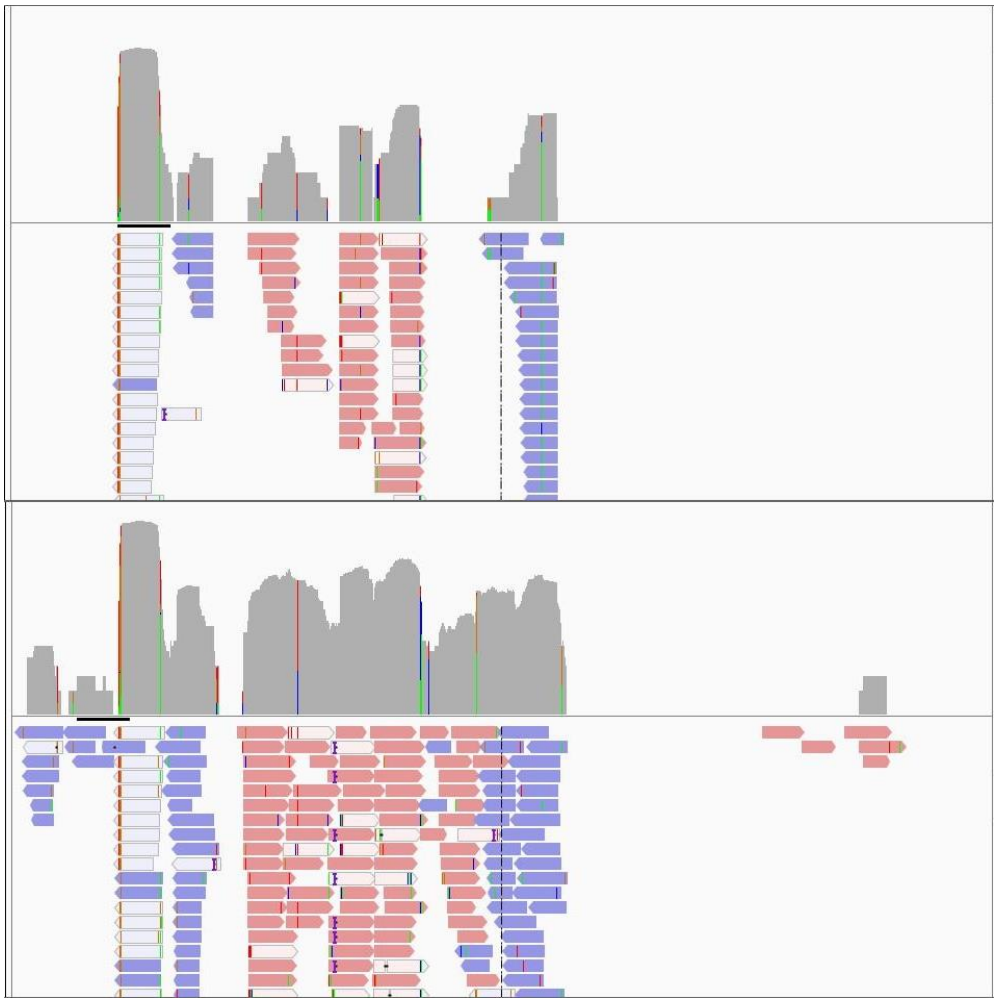


Figure 26. IGV genome browser visualization of transcript-63_15992 for both the day (top) and night (bottom) transcript. The shaded areas represent coverage, while the black bars underneath it represents more than 1000 reads to that location. The red and blue bars represent read orientation.

Transcript 53908 shows possible hairpin structures that if verified, represent precursor small RNAs such siRNAs or miRNAs (Figure 27). The columns immediately next to each other would represent the arms of the hairpin that have folded over and paired together and the empty space between would form the head of the hairpin.

Figure 27. IGV genome browser visualization of transcript-63_53908 for the day and night transcript. Possible hairpin structures.

Transcript 41968 shows possible alternate splicing (Figure 28). Looking only at the transcript from the total RNA it is clear that the transcript is expressed differently during the day and night. Also, the transcript from the small RNA data is expressed differently. The difference in the night transcript versus the day from the small RNA data could be causing alternate splicing, changing the gene that is expressed at night.

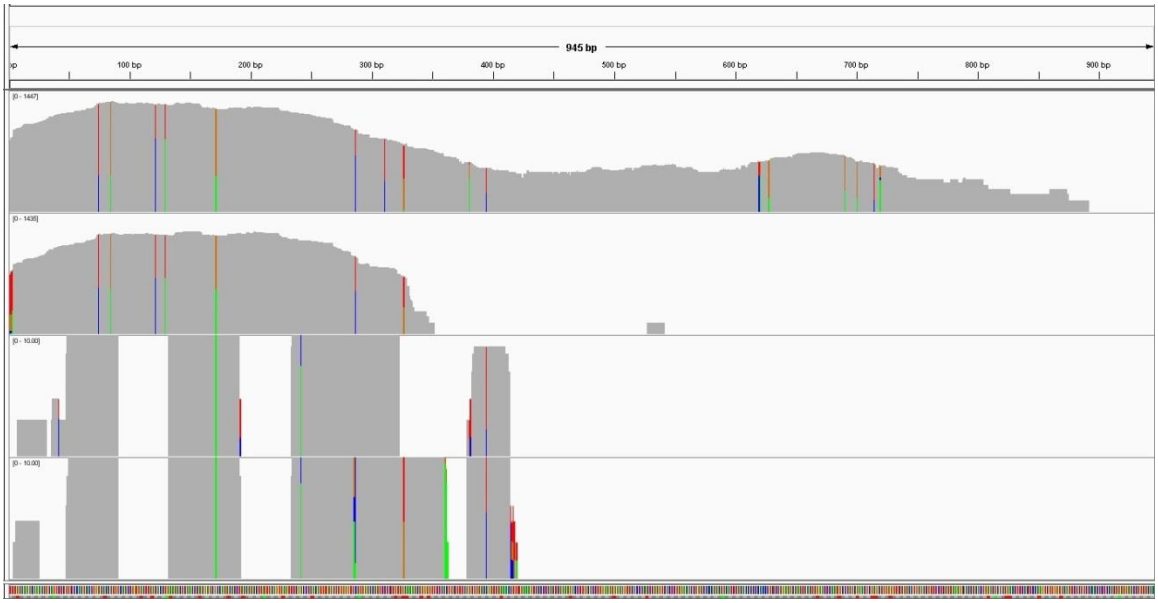


Figure 28. IGV genome browser visualization of transcript-63_41968 for the day and night transcript. The top transcript is from the day total RNA data, second from the night total RNA data, third from the day small RNA data and the bottom from the night small RNA data. Possible alternate splicing site.

CHAPTER V

SUMMARY AND CONCLUSIONS

The dinoflagellate nucleus is so strange that it was once considered mesokaryotic: a stage between prokaryotes and eukaryotes (Gornik et al., 2012). *Karenia brevis* has been studied for more than 65 years (Steidinger et al., 2008). Since then, many interesting characteristics have been discovered. Some of the more unique characteristics include their extremely large genomes, which make it difficult to sequence their genome or transcriptome, chromosomes that stay condensed throughout the cell cycle, and the lack of nucleosomes that control chromatin condensation and regulate transcription and replication activities (Costas & Goyanes, 2005). These unusual or unique features suggest an alternate or hybrid version of transcription and replication typical of other eukaryotes. Within dinoflagellate chromosomes a whorled structure called a cholesteric liquid crystal organization has been found using electron microscopy. This structure may also enforce limitations on replication and transcription (Gornik et al., 2012).

To date, no consensus promoter sequences have been found within the *Karenia* genome; specifically no TATA box (Brunelle & Van Dolah, 2011) or any known promoter elements have been found (Li & Hastings 1998). Coupled with the above information, these characteristics bring up the question of how dinoflagellates regulate transcription. Transcriptional studies suggest that 50% of the genes in dinoflagellates do not match genes documented in other organisms, and only10 -27% of dinoflagellate genes are regulated through transcription (Lin, 2011). A micro array study by Lidie, Ryan, Barbier, and Van Dolah (2005) showed that out of 8500 genes associated with the diel cycle and the circadian clock, 90% were constantly expressed. This suggests an

alternate mechanism for gene regulation. Even with all that is known about dinoflagellates the mechanism of gene regulation is still unknown. Currently, the consensus theory among those in the field is that gene expression within dinoflagellates is controlled through post transcriptional machinery (Van Dolah et al., 2009).

The purpose of this project was to find NATS and miRNAs, due to their association with post-transcriptional regulation, and determine if there was any differential expression among these antisense RNAs in an attempt to implicate them in the mechanisms controlling gene expression. Due to the unique nature and extreme size of the dinoflagellate genome, it was necessary to employ new tools and associated techniques to properly identify these molecules. RNA sequencing was beneficial to this study because of its ability to sequence transcripts without an existing genome and find ncRNAs, but RNA-seq and associated software is still in its infancy. De novo assemblies can be built with or without a reference genome, but a build without a reference genome comes with several obstacles. *K. brevis* doesn't have a reference genome simply because of its complex nature and lack of molecular testing. There are also no gene annotations to compare new assemblies too. Because of this, it was difficult to identify NATs in the total RNA dataset with any certainty and to perform a differential expression analysis.

Even with difficulties analyzing this data, after purification and sequencing, nearly 2 billion reads were recovered from the total RNA data set. These reads were run through quality check software and showed that the reads were of high quality before and after the cleaning processing. While both transcriptomic assemblers produced assemblies that contained large contigs and scaffolds, several parameters indicated that IDBA created a preferable assembly for our downstream analysis. The paired-end reads aligned

back to the transcriptome with an overall rate of 90%. The transcriptomic assembly built for this data set contained over 200,000 mostly unique transcripts. With this size transcriptome, it was increasingly difficult to find all transcripts with potential interest with our current software and algorithms. Software that will analyze the data set with these conditions will need to be found or developed. The building of the assembly with aligned reads allowed for the creation of files that could be used in a genomic/transcriptomic browser. This tool allowed for differential expression analysis by visualizing individual transcripts that were present in both day and night reads and determining their relative abundance under each condition.

A good place to begin future analyses of these transcripts would be to blast the entire transcriptome. If the blast software is downloaded to a Linux server the process would save many hours of manually testing the 200,000+ transcripts. The blast would result in obtaining a large data set of gene hits that show basic similarities. These hits could then be annotated to find more specific biological functions and structures with many options of online tools (Wit et al., 2012). Next, these annotated genes could be aligned to the assembly. This file would be the beginning of reference genome. This file would be used to compare the day and night reads from the total RNA data.

Subsequently, with Cufflinks, a differential expression analysis could be run, which would produce figures that would allow the visualization of individual genes with differential expression between the day and night reads (Trapnell et al., 2012). This would lead to a better understanding of the *K. brevis* genome and possible leads into the gene regulation mechanisms that control it.

The miRNA sequencing also returned quality reads, but the length distribution showed that 80% of the reads were 40nt or larger. MiRNAs are typically 20 to 24nt long (Chekulaeva & Filipowicz, 2009). To maximize the return on miRNA read alignments to the assembly, the RNA-seq day and night reads, miRNA day and night reads and *K. brevis* ESTs (from NCBI) were compiled in one large single-end assembly. The statistical analysis showed that the assembly contained large N50 values which were a good indicator of a proper build. After assembly miRNAs were used in two different analysis pipelines. First, the cleaned reads were analyzed via miRanalyzer for novel and conserved miRNAs, and then the assembly was analyzed through the IGV genome browser for differential expression analysis.

The miRanalyzer returned 62 hits for mature miRNA candidates among several miRNA families. The miRNA candidates with multiple hits to mir families were analyzed with Clustal Omega to find highly conserved regions. Further analysis would include blasting these highly conserved miRNA candidates against the nucleotide collection database with megablast (Altschul, Gish, Miller, Myers, & Lipman, 1990). If the hits met the blast criteria for a significant match, the transcript would be analyzed for miRNA secondary structures or hairpins via the program, RNAfold. To further validate that the candidates are real miRNAs, first the transcripts would need to show that they could fold into hairpin structures. Second, the transcripts would need to meet the criteria of minimum free energy and the base pairing probability matrix which ensures that the transcripts are precursor miRNAs (Hofacker & Stadler, 2006). These are visualized in heat maps showing the brighter color as a higher probability of base pair matches and proper folding. Sequences in transcript 53908 in Figure 27 and any others with similar

transcript coverage could be used in RNAfold to determine the likelihood of such structures to be formed. These hairpin-like structures would have to meet the same criteria as previously mentioned. Next, the sequences making up the paired-end part of the hairpin would need to be compared to the reads in the transcript to determine if the locations match. Another criterion for validating new miRNAs is pre-miRNA examination on Northern blots. When examined with reduced Dicer activity, these pre-miRNAs increase in abundance (Bartel, 2004). Such validation is beyond the scope of this project but would be required to confirm these candidates as real miRNAs.

The IGV genome browser was used to visualize the entire miRNA transcript assembly. First, the results from this analysis showed that several transcripts indicated differential expression between day and night samples. A better design for this data set would resemble the aforementioned total RNA data set. More extractions including several time points would be sequenced with HiSeq technologies utilizing several lanes. This would increase the overall size of the assembly. Next, past assemblies would be merged with the new one to increase the coverage of the transcriptome. Then, the assembly would be blasted for basic annotation and structural and functional annotation. At this point the new, merged total RNA assembly would be used as a reference genome to find differential expression and possibly increase knowledge of genes previously found.

Secondly, the data derived from the IGV genome browser showed cleavage sites within transcripts that are associated with degradation pathways. RNAi mechanisms produce dsRNAs generated by transcription of inverted repeats, resulting in RNA hairpins or by convergent transcription leading to overlapping transcripts. These dsRNAs

are processed by RNase III type endonucleases (Gullerova & Proudfoot, 2012). Running a tblast on the miRNA assembly against the RNase III peptide domain would be a good way to test if this is the degradation pathway present in both data sets. RNase III can be divided into three structural classes. The first class and simplest protein of RNase III has one RNase III domain and a dsRNA-binding domain (dsRBD) and is typically found in bacteria processing long dsRNAs (Carmell & Hannon, 2004). The second class is DROSHA, and it contains two RNase III domains a a dsRBD domain, a protein rich region (PRR), and a serine-arginine rich domain (RS) and is found in a variety of organisms excluding bacteria (Fortin, Nicholson, & Nicholson, 2002). Third class is Dicer (also found in a variety of organisms), contains two RNase III domains, a dsRBD domain, a PAZ domain that is also found in Argonaute proteins, a RNA helicase domain and a domain of unknown function (DUF283) (Carmell & Hannon, 2004). Finding any of the RNase III domains or associated domains would prove invaluable in determining the presence and type of degradation pathways in the *K. brevis* transcriptome. Once found, further annotation could determine structure and function. Next, these domains could be aligned back to the assembly. With this, it might be possible to determine expression levels of the RNase III domain. If it is found it may be possible to conclude that some small RNA is playing a role in post-transcriptional regulation. If this domain is more highly expressed during the night or day, it would be possible to conclude that the diel cycle is controlling this expression.

Lastly, the pattern of directionality of the sequences building the transcript showed that they were being transcribed from both directions facing towards each other, which suggests convergent transcription for the cis-Nat pathway. *Cis*-NATs are

transcribed from the same genomic loci as their sense transcripts but on the opposite DNA strand, but trans-NATs, such as miRNAs, are expressed from genomic regions different from those encoding their sense transcripts (Wang et al., 2005). These transcripts look like *cis*-NATs; the question is what their role is? To determine that, looking at the type of Argonaute proteins present in *K. brevis* may provide some answers. The Argonaute proteins are divided into two main classes of conserved proteins, the AGO and PIWI. These proteins bind to small RNAs smaller than 32nt long (Okamura & Lai, 2008). The length distributions of the miRNAs in this data set, as mentioned earlier, are larger than 32nt long. Because of this, it is impossible for the AGO family of proteins to bind to these small RNAs. With secondary structure testing of the miRNAs with the RNAfold web server, it would be possible to make a better conclusion of the presence or absence of miRNAs in the *K. brevis* genome. If Argonaute proteins are to bind to transcripts of this size, it may be necessary to introduce a new class of AGO proteins.

In conclusion, the small RNA data set produced interesting results that could lead to future projects. The blast data results supports the presence of miRNAs, but they should be held as preliminary since the identified miRNAs are merely candidates, and analysis from the genome browser suggests that the small RNA pathway found in eukaryotes may not be present. The analysis did show some evidence of different pathways: the *cis*-NAT and degradation pathways. The presence of differential expression within transcripts from both the day and night were visualized within the small RNA dataset using a genome/transcriptome browser. Also, the same browser was able to look at the differential expression of the total RNA data set, showing differences between day and night transcripts and possibly found some alternative splicing sites. In

summation, while the lack of molecular studies, reference genomes, gene annotations, and conserved small RNAs for *K. brevis* made it a difficult task to find NATs within the RNA-seq data set, the observation of cis-NATs and the possibility of them being involved in a RNA degradation and/or alternative splicing pathway supports our hypothesis of post-transcriptional regulation and aids in honing in on a possible mechanism to explain such regulation.

The genome browser was an excellent tool for discovery of potential transcripts for future analyses. Obviously, the browser alone is not enough to claim the presence of or lack of transcript characteristics but is a good tool for pointing a researcher in the right direction and making new hypotheses and predictions. Using the transcript characteristics found in the genome browser, further research could begin where this project ended, by trying different blasting techniques to find some of the key players in the cis-NAT pathway, degradation pathway, RNase III domains, and Argonaute domains. For any future work in NAT discovery it would be of great value to create a larger data set utilizing RNA-seq technology. Then use annotation tools to annotate the entire assembly. Next, align these annotations back to the assembly to use as a reference genome for any other downstream analyses. At this point, many avenues of analysis would be open, including differential expression with Cufflinks or DeSeq. Conserved and novel miRNA detection with miRDeep (small RNA analysis tool) would also be possible. This tool could confirm the existence of or lack of miRNAs, but also potentially find other small RNAs within the transcriptome. With the discovery of small RNAs many other hypotheses could be modeled leading to many other pathways currently not being sought after.

Identifying the relevant molecules and understanding the genetic regulation for such biochemical and physiological activities as growth control, nutrient acquisition, and gene expression is essential to understanding the gene-environment interactions that are so important to understanding harmful bloom dynamics for dinoflagellates. It has been hypothesized that dinoflagellates regulate their genes via a post-transcriptional mechanism(s), but no mechanisms have been sufficiently laid out and tested. My data support my original hypothesis that non-coding, anti-sense RNAs are present and likely play a post-transcriptional role in the regulation of mRNAs. More work is necessary to validate the exact nature of some of these anti-sense RNAs and the exact role that they play in regulation. This information not only increases our understanding of the basic cellular biology of this unique taxon of organisms, but it may also provide new targets to which molecules could be designed that specifically target and disrupt a dinoflagellate's abilities to grow and form blooms.

REFERENCES

Abrahams, M. V., & Townsend, L. D. (1993). Bioluminescence in dinoflagellates: A test of the burgular alarm hypothesis. *Ecology*, *74*(1), 258-260.

Akker, S. A., Smith, P. J., & Chew, S. L. (2001). Nuclear post-transcriptional control of gene expression. *Journal of molecular endocrinology*, *27*(2), 123-131.

Altschul S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of molecular biology*, *215*, 403-410.

Atchison, W. D., Luke, V. S., Narahashi, T., & Vogel, S. M. (1986). Nerve membrane sodium channels as the target site of brevetoxins at neuromuscular junctions. *British journal of pharmacology*, *89*(4), 731-738.

Backer, L., & McGillicuddy, D. (2006). Harmful algal blooms. *Oceanography*, 19(2), 94.

Baden, D. G. (1989). Brevetoxins: unique polyether dinoflagellate toxins. *The FASEB journal*, *3*(7), 1807-1817.

Bartel, B., & Bartel, D. P. (2003). MicroRNAs: at the root of plant development? *Plant Physiology*, *132*(2), 709-717.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, *116*(2), 281-297.

Bourdelais, A. J., & Baden, D. G. (2004). Toxic brevetoxin complexes are in aqueous solutions. *Toxicologist*, *78*(1-S), 807.

Bourdelais, A. J., Campbell, S., Jacocks, H., Naar, J., Wright, J. L., Carsi, J., & Baden, D. G. (2004). Brevenal is a natural inhibitor of brevetoxin action in sodium channel receptor binding assays. *Cellular and molecular neurobiology*, *24*(4), 553-563.

Brantl, S. (2002). Antisense-RNA regulation and RNA interference. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, *1575*(1), 15-25.

Brunelle, S. A., Hazard, E. S., Sotka, E. E., & Dolah, F. M. V. (2007). Characterization of a dinoflagellate cryptochrome blue-light receptor with a possible role in circadian control of the cell cycle1. *Journal of phycology*, *43*(3), 509-518.

Brunelle, S. A., & Van Dolah, F. M. (2011). Post-transcriptional regulation of s-phase genes in the dinoflagellate, Karenia brevis. *Journal of Eukaryotic Microbiology*, *58*(4), 373-382.

Carmell, M. A., & Hannon, G. J. (2004). RNase III enzymes and the initiation of gene silencing. *Nature structural & molecular biology*, *11*(3), 214-218.

Carthew, R. W., & Sontheimer, E. J. (2009). Origins and mechanisms of miRNAs and siRNAs. *Cell*, *136*(4), 642-655.

Catterall, W. A., & Gainer, M. (1985). Interaction of brevetoxin A with a new receptor site on the sodium channel. *Toxicon*, *23*(3), 497-504.

Chekulaeva, M., & Filipowicz, W. (2009). Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Current opinion in cell biology*, *21*(3), 452-460.

Cheng, Y. S., Villareal, T. A., Zhou, Y., Gao, J., Pierce, R. H., Wetzel, D., & Baden, D. G. (2005). Characterization of red tide aerosol on the Texas coast. *Harmful Algae*, *4*(1), 87-94.

Cheng, Y. S., Zhou, Y., Irvin, C. M., Pierce, R. H., Naar, J., Backer, L. C., ... & Baden, D. G. (2005). Characterization of marine aerosol for assessment of human exposure to brevetoxins. *Environmental health perspectives*, *113*(5), 638-643.

Cogoni, C., & Macino, G. (2000). Post-transcriptional gene silencing across kingdoms. *Current opinion in genetics & development*, *10*(6), 638-643.

Costas, E., & Goyanes, V. (2005). Architecture and evolution of dinoflagellate chromosomes: an enigmatic origin. *Cytogenetic and genome research*, *109*(1-3), 268-275.

Daugbjerg, N., Hansen, G., Larsen, J., & Moestrup, Ø. (2000). Phylogeny of some of the major genera of dinoflagellates based on ultrastructure and partial LSU rDNA sequence data, including the erection of three new genera of unarmoured dinoflagellates. *Phycologia*, *39*(4), 302-317.

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., & Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature*, *432*(7014), 231-235.

Elbashir, S. M., Harborth, J., Weber, K., & Tuschl, T. (2002). Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods*, *26*(2), 199-213.

Eulalio, A., Behm-Ansmant, I., & Izaurralde, E. (2007). P bodies: at the crossroads of post-transcriptional pathways. *Nature reviews Molecular cell biology*, *8*(1), 9-22.

Faghihi, M. A., & Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. *Nature reviews Molecular cell biology*, *10*(9), 637-643.

Finocchiaro, G., Carro, M. S., Francois, S., Parise, P., DiNinni, V., & Muller, H. (2007). Localizing hotspots of antisense transcription. *Nucleic acids research*, *35*(5), 1488-1500.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, *391*(6669), 806-811.

Fortin, K. R., Nicholson, R. H., & Nicholson, A. W. (2002). Mouse ribonuclease III. cDNA structure, expression analysis, and chromosomal location. *BMC genomics*, *3*(1), 26.

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, *582*(14), 1977-1986.

Gornik, S. G., Ford, K. L., Mulhern, T. D., Bacic, A., McFadden, G. I., & Waller, R. F. (2012). Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Current Biology*, *22*(24), 2303-2312.

Gregory, R. I., Chendrimada, T. P., Cooch, N., & Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell*, *123*(4), 631-640.

Guillard, R., & Ryther, J. (1962). Studies of marine planktonic diatoms. I. Cyclotella nana Hustedt, and Detonula confervacea (cleve) gran. *Canadian Journal of Microbiology*, *8*, 229.

Guillard, R. R. (1975). Culture of phytoplankton for feeding marine invertebrates. In *Culture of marine invertebrate animals* (pp. 29-60). New York, NY: Springer US.

Guillard, R. R. L., & Hargraves, P. E. (1993). Stichochrysis immobilis is a diatom, not a chrysophyte. *Phycologia*, *32*(3), 234-236.

Gullerova, M., & Proudfoot, N. J. (2012). Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. *Nature structural & molecular biology*, *19*(11), 1193-1201.

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

Halbeisen, R. E., Galgano, A., Scherrer, T., & Gerber, A. P. (2008). Post-transcriptional gene regulation: from genome-wide studies to principles. *Cellular and molecular life sciences*, *65*(5), 798-813.

Hallegraeff, G. M. (1993). A review of harmful algal blooms and their apparent global increase*. *Phycologia*, *32*(2), 79-99.

Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R., & Hannon, G. J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science*, *293*(5532), 1146-1150.

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., & Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development*, *18*(24), 3016-3027.

Harmful Algae. (2012). [Global distribution of PSP toxins recorded in 1972 and 2006]. Distribution of HABs throughout the world. Retrieved from http://www.whoi.edu/redtide/page.do?pid=14899

Haywood, A. J., Steidinger, K. A., Truby, E. W., Bergquist, P. R., Bergquist, P. L., Adamson, J., & MacKenzie, L. (2004). Comparative morphology and molecular

phylogenetic analysis of three new species of the genus Karenia (dinophyceae) from New Zealand1. *Journal of Phycology*, *40*(1), 165-179.

Heimann, K. (1999). Gymnodinium and related dinoflagellates. *Encyclopedia of Life Sciences*, (pp. 1-17). Hoboken, NJ: John Wiley & Sons.

Hofacker, I. L., & Stadler, P. F. (2006). Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics, 22*(10), 1172-1176.

Hoppenrath, M., & Leander, B. S. (2010). Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences. *PLoS One*, 5(10), e13220.

Hutvagner, G., & Simard, M. J. (2008). Argonaute proteins: key players in RNA silencing. *Nature reviews Molecular cell biology, 9*(1), 22-32.

Jones-Rhoades, M. W., Bartel, D. P., & Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biol*ogy, *57*, 19-53.

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, *8*(7), 533-543.

Kirkpatrick, B., Fleming, L. E., Squicciarini, D., Backer, L. C., Clark, R., Abraham, W., & Baden, D. G. (2004). Literature review of Florida red tide: implications for human health effects. *Harmful Algae, 3*(2), 99-115.

Kumar, M., & Carmichael, G. G. (1998). Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiology and Molecular Biology Reviews*, *62*(4), 1415-1434.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.

Lapidot, M., & Pilpel, Y. (2006). Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. *EMBO reports*, *7*(12), 1216-1222.

Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., de los Mozos, I. R., Vergara-Irigaray, M., & Gingeras, T. R. (2011). Genome-wide antisense transcription drives mRNA processing in bacteria. *Proceedings of the National Academy of Sciences*, *108*(50), 20172-20177.

Lewis, N. I., Xu, W., Jericho, S. K., Kreuzer, H. J., Jericho, M. H., & Cembella, A. D. (2006). Swimming speed of three species of Alexandrium (Dinophyceae) as determined by digital in-line holography. *Journal Information*, *45*(1).

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, L., & Hastings, J. W. (1998). The structure and organization of the luciferase gene in the photosynthetic dinoflagellate Gonyaulax polyedra. *Plant molecular biology*, *36*(2), 275-284.

Lidie, K. B., & Van Dolah, F. M. (2007). Spliced Leader RNA-Mediated trans-Splicing in a Dinoflagellate, Karenia brevis. *Journal of Eukaryotic Microbiology*, *54*(5), 427-435.

Lidie, K. B., Ryan, J. C., Barbier, M., & Van Dolah, F. M. (2005). Gene expression in Florida red tide dinoflagellate Karenia brevis: analysis of an expressed sequence tag library and development of DNA microarray. *Marine Biotechnology*, *7*(5), 481-493.

Lin, S. (2011). Genomic understanding of dinoflagellates. *Research in microbiology*, *162*(6), 551-569.

Lin, S., Zhang, H., Zhuang, Y., Tran, B., & Gill, J. (2010). Spliced leader–based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proceedings of the National Academy of Sciences*, *107*(46), 20033-20038.

McLean, T.I., & M. Pirooznia. 2011. Functional Genomics and Molecular Analysis of a Subtropical Harmful Algal Bloom Species, Karenia brevis. In. Nriagu J.O. (Ed.) *Encyclopedia of Environmental Health*. (Vol 2). (pp. 816–828). Oxford, UK: Elsevier.

Monroe, E. A., & Van Dolah, F. M. (2008). The Toxic Dinoflagellate Karenia brevis Encodes Novel Type I-like Polyketide Synthases Containing Discrete Catalytic Domains. *Protist*, *159*(3), 471-482.

Montgomery, M. K., Xu, S., & Fire, A. (1998). RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans. *Proceedings of the National Academy of Sciences*, *95*(26), 15502-15507.

Morey, J.S., Monroe, E.A., Kinney, A.L., Beal, M., Johnson, J.G, Hitchcock, G.L., & Van Dolah, F.M. (2011). Transcriptomic response of the red tide dinoflagellate, Karenia brevis, to nitrogen and phosphorus depletion and addition. *BMC Genomics*, *12*, 346

Morris, T. J., & Dodds, J. A. (1979). Isolation and analysis of double-stranded RNA from virus-infected plant and fungal tissue. *Phytopathology*, *69*(8), 855.

Munroe, S. H., & Zhu, J. (2006). Overlapping transcripts, double-stranded RNA and antisense regulation: a genomic perspective. *Cellular and Molecular Life Sciences CMLS*, *63*(18), 2102-2118.

Murphy, W. J., Watkins, K. P., & Agabian, N. (1986). Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: Evidence for Trans splicing. *Cell*, *47*(4), 517-525.

Okamura, K., & Lai, E. C. (2008). Endogenous small interfering RNAs in animals. *Nature reviews Molecular cell biology*, *9*(9), 673-678.

Örnólfsdóttir, E. B., Pinckney, J. L., & Tester, P. A. (2003). Quantification of the relative abundance of the toxic dinoflagellate, Karenia brevis (Dinophyta), using unique photopigments 1. *Journal of Phycology*, *39*, 449-457.

Peng, Y., Leung, H. C., Yiu, S. M., & Chin, F. Y. (2010). IDBA–a practical iterative de Bruijn graph de novo assembler. *Lecture Notes in Computer Science*, *6044*(2010), 426-440.

Pittendrigh, C. S. (1993). Temporal organization: reflections of a Darwinian clock-watcher. *Annual Review of Physiology*, *55*(1), 17-54.

Poli, M. A., Mende, T. J., & Baden, D. G. (1986). Brevetoxins, unique activators of voltage-sensitive sodium channels, bind to specific sites in rat brain synaptosomes. *Molecular pharmacology*, *30*(2), 129-135.

Reñé, A., Satta, C. T., Garcés, E., Massana, R., Zapata, M., Anglès, S., & Camp, J. (2011). Gymnodinium litoralis sp. nov.(Dinophyceae), a newly identified bloom-forming dinoflagellate from the NW Mediterranean Sea. *Harmful Algae*, *12*, 11-25.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., & Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, *110*(4), 513-520.

Roth, P. B., Twiner, M. J., Zhihong, W., Bottein Dechraoui, M. Y., & Doucette, G. J. (2007). Fate and distribution of brevetoxin (PbBx) following lysis of Karenia brevis by algicidal bacteria, including analysis of open A-ring derivatives. *Toxicon*, *50*(8), 1175- 1191.

Sellner, K. G., Doucette, G. J., & Kirkpatrick, G. J. (2003). Harmful algal blooms: causes, impacts and detection. *Journal of Industrial Microbiology and Biotechnology*, *30*(7), 383-406.

Skovgaard, A. (2005). Infection with the dinoflagellate parasite Blastodinium spp. in two Mediterranean copepods. *Aquatic microbial ecology*, *38*(1), 93-101.

Smayda, T. J. (1997). Harmful algal blooms: their ecophysiology and general relevance to phytoplankton blooms in the sea. *Limnology and oceanography*, *42*(5), 1137-1153.

Sorek, M., Yacobi, Y. Z., Roopin, M., Berman-Frank, I., & Levy, O. (2013). Photosynthetic circadian rhythmicity patterns of Symbiodium, the coral endosymbiotic algae. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1759), 20122942.

Steidinger, K. A. (2009). Historical perspective on Karenia brevis red tide research in the Gulf of Mexico. *Harmful Algae*, *8*(4), 549-561.

Steidinger, K. A., & Penta, H. M. (1999). *Harmful microalgae and associated public health risks in the Gulf of Mexico*. Stennis Space Center, MS: Gulf of Mexico Program Office.

Steidinger, K. A., Landsberg, J. H., Flewelling, J. L., & Kirkpatrick, B. (2008). *Toxic dinoflagellates*. New York, NY: Elsevier Science Publishers.

Stumpf, R. P., Culver, M. E., Tester, P. A., Tomlinson, M., Kirkpatrick, G. J., Pederson, B. A., ... & Soracco, M. (2003). Monitoring Karenia brevis blooms in the Gulf of Mexico using satellite ocean color imagery and other data. *Harmful Algae*, *2*(2), 147-160.

Tang, G. (2005). siRNA and miRNA: an insight into RISCs. *Trends in biochemical sciences*, *30*(2), 106-114.

Taylor, F. J. R., Hoppenrath, M., & Saldarriaga, J. F. (2008). Dinoflagellate diversity and distribution. *Biodiversity and conservation*, *17*(2), 407-418.

Thomason, M. K., & Storz, G. (2010). Bacterial antisense RNAs: How many are there, and what are they doing?. *Annual Review of Genetics*, *44*, 167-88.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., ... & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, *7*(3), 562-578.

Van Dolah, F. M., & Leighfield, T. A. (1999). Diel phasing of the cell-cycle in the Florida red tide dinoflagellate, Gymnodinium breve. *Journal of phycology*, *35*(6), 1404-1411.

Van Dolah, F. M., Lidie, K. B., Morey, J. S., Brunelle, S. A., Ryan, J. C., Monroe, E. A., & Haynes, B. L. (2007). microarray analysis of diurnal-and circadian-regulated genes in the florida red-tide dinoflagellate Karenia brevis (Dinophyceae) 1. *Journal of Phycology*, *43*(4), 741-752.

Van Dolah, F. M., Lidie, K. B., Monroe, E. A., Bhattacharya, D., Campbell, L., Doucette, G. J., & Kamykowski, D. (2009). The Florida red tide dinoflagellate Karenia brevis: New insights into cellular and molecular processes underlying bloom dynamics. *Harmful Algae*, *8*(4), 562-572.

Vanhée-Brossollet, C., & Vaquero, C. (1998). Do natural antisense transcripts make sense in eukaryotes? *Gene, 211*(1), 1-9.

Vargo, G. A., Heil, C. A., Fanning, K. A., Dixon, L. K., Neely, M. B., Lester, K., & Bell, S. (2008). Nutrient availability in support of Karenia brevis blooms on the central West Florida Shelf: What keeps Karenia blooming. *Continental Shelf Research, 28,* 73-98.

Wang, X. J., Gaasterland, T., & Chua, N. H. (2005). Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana. *Genome biology*, *6*(4), R30.

Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., ... & Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular ecology resources, 12*(6), 1058-1067.

Xu, P., Vernooy, S. Y., Guo, M., & Hay, B. A. (2003). The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Current biology: CB*, *13*(9), 790.

Yacobovitch, T., Benayahu, Y., & Weis, V. M. (2004). Motility of zooxanthellae isolated from the Red Sea soft coral Heteroxenia fuscescens(Cnidaria). *Journal of experimental marine biology and ecology*, *298*(1), 35-48.

Zhang, H., Campbell, D. A., Sturm, N. R., & Lin, S. (2009). Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements. *Molecular biology and evolution*, *26*(8), 1757-1771.

Zhang, H., Hou, Y., Miranda, L., Campbell, D. A., Sturm, N. R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proceedings of the National Academy of Sciences*, *104*(11), 4618-4623.

Zhang, H., & Lin, S. (2009). Retrieval of missing spliced leader in dinoflagellates. *PLoS One*, *4*(1), e4129.