INVESTIGATING THE BIOLOGY OF TOXIN-PRODUCING *KARENIA* SPECIES: A TRANSCRIPTOMICS APPROACH

A Dissertation

by

DARCIE E. RYAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Lisa Campbell
Committee Members,	Alan Pepper
	Robert Hetland
	Daniel Thornton
Head of Department,	Debbie Thomas

December 2016

Major Subject: Oceanography

ABSTRACT

The dinoflagellate *Karenia brevis* is a prominent bloom-forming harmful algae species in the Gulf of Mexico. *K. brevis* produces two ladder-frame polyketide brevetoxins, PbTx-1 and PbTx-2. PbTx-1, PbTx-2, and their derivatives bind to neurotoxin receptor site 5 of voltage-gated Na^+ channels and prevent channel deactivation. Through the depolarizing activity of brevetoxins, *K. brevis* blooms kill fish and may sicken humans who eat shellfish from the bloom region. Despite these risks, the biological function of brevetoxins is poorly characterized, including the genes that participate in PbTx synthesis. Large and repetitive, the *K. brevis* genome has not been sequenced. However, with *de novo* transcriptomics, genomic analyses of *K. brevis* are possible. During this dissertation study, the transcriptomes of multiple *Karenia* species, including three strains of *K. brevis* (SP1, SP3, and Wilson), cytotoxin-producing *Karenia mikimotoi*, and PbTx-2-producing *Karenia papilionacea*, were assembled and analyzed. Analyses included comparative transcriptomics among *Karenia* species and *K. brevis* strains, putative protein annotation, ortholog prediction, gene tree construction, and single nucleotide polymorphism (SNP) prediction.

Through the comparison of multiple *de novo* transcriptome assembly methods, this study developed a pipeline to produce highly complete dinoflagellate reference transcriptomes. Thousands of *Karenia* transcripts were annotated with potential functions and gene ontology terms, including highly conserved putative voltage-gated cation channel genes. Because both Na^+ and Ca^{2+} channels were identified, our work

suggests that *Karenia* species are capable of selective transmembrane ion transport. It also highlights the need for biochemical research to investigate the interaction, if any, between brevetoxins and dinoflagellate voltage-gated Na^+ channels.

The *Karenia* ortholog detection step identified 4799 genes that were expressed by brevetoxin-producing *K. brevis* and *K. papilionacea*, but not *K. mikimotoi*. Transcripts involved with "heterocycle production" were overrepresented in the 4799 "unique" orthologs, including five putative polyketide synthases. These genes represent interesting targets for further brevetoxin production research. Additionally, a novel transcript with high homology to multimodular type 1 PKSs was identified in the *K. brevis* transcriptome and RNA from field samples *of K. brevis*. The multimodular PKS, which has never been characterized before, indicates that *K. brevis* synthesizes polyketides and/or other secondary metabolites using both type I (multimodular) and type I-like (single domain) proteins.

DEDICATION

To Acro "Little Dude" Ryan, whose daily routines brought joy to his family . . . all except Mikey the Chihuahua.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Lisa Campbell, and my committee members, Dr. Alan Pepper, Dr. Robert Hetland, and Dr. Daniel Thornton, for their guidance and support throughout the course of this research.

I am grateful for my colleagues and the Oceanography Department faculty and staff for making my time at Texas A&M University a great experience.

Finally, thanks to my mother, father, and friends for their support and love.

NOMENCLATURE

aa	amino acid
ACP	acyl carrier protein
AT	acyltransferase
bp	base pairs
BLAST	Basic Local Alignment Search Tool
CEGMA	Core Eukaryotic Genes Mapping Approach
DH	dehydratase
DNA	deoxyribonucleic acid
ER	enoylreductase
EST	expressed sequence tag
KR	keto reductase
KS	keto synthase
MIP	major intrinsic protein
MMETSP	Marine Microbial Eukaryote Transcriptome Sequencing Project
NMR	nuclear magnetic resonance
ORF	open reading frame
PbTx	brevetoxin
RNA	ribonucleic acid
RNA-Seq	RNA sequencing
SLS	spliced leader sequence

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Karenia brevis

In the spring and summer of 1947, a then-unnamed algae species discolored the Florida Gulf, causing fish kills and marine animal mortalities as the algal cell concentration reached 60 million cells per liter (Davis 1948; Gunter et al. 1948; Williams et al. 1947). Living samples from the bloom were collected and examined by light microscopy. Morphologically similar to known dinoflagellate species from the *Gymnodinium* genera, the bloom-forming harmful algae species was named *Gymnodinium brevis* (Davis 1948). After one century of documented harmful algal blooms in the Gulf of Mexico, each with a legacy of marine animal mortalities (Gunter et al. 1948), the culprit organism had been identified; however, our knowledge about its biology was – and still is – incomplete. In fact, subsequent phylogenetic (rDNA), pigment, and ultrastructure analyses of *G. brevis* prompted its reclassification as "*Karenia brevis*," a member of a new dinoflagellate genus that was named after the researcher Karen Steidinger (Daugbjerg et al. 2000).

Most known harmful, bloom-forming, eukaryotic phytoplankton species belong to the Dinophyta (as dinoflagellates) or Heterokont (as diatoms) phyla (Granéli and Turner 2006). Based on morphology and behavior, the dinoflagellate group is diverse, with ~130 genera and ~1200 described single-celled species ranging from 1 µm to 2mm wide (Spector 1984). Typically, dinoflagellates have two flagella; they may be

autotrophs or heterotrophs, free-living or symbiotic (Dodge 1984). The first described dinoflagellate, *Noctiluca* (Baker 1753), emits stimulation-induced bioluminescence through the activity of dinoflagellate luciferase (Haddock et al. 2010). In fact, heterotrophic and autotrophic dinoflagellates from 18 known genera are bioluminescent (Baker et al. 2008). Some dinoflagellates are "armored," possessing a cell surface covered by cellulose-filled thecal plates, while others, such as *Karenia* spp., are called "unarmored" or "naked" and lack plates (Granéli and Turner 2006).

Karenia brevis is a free-living, unarmored dinoflagellate (Daugbjerg et al. 2000). The brevetoxins that *K. brevis* produces have caused fish kills and other marine animal mortalities (Landsberg 2002). Additionally, shellfish bioaccumulate brevetoxins when they filter feed on *K. brevis* cells. People who ingest shellfish from the Gulf during or immediately after a bloom may become sick with neurotoxic shellfish poisoning. Neurotoxic shellfish poisoning is characterized by gastrointestinal and neurological symptoms including nausea, dizziness, vomiting, and partial paralysis (Watkins et al. 2008). Furthermore, brevetoxins aerosolized by the breaking surf may cause eye irritation and respiratory distress in people near the shore (Backer et al. 2003). With nearly annual blooms off the coast of Texas or Florida (Steidinger et al. 1998), *K. brevis* is a prominent – if not the prominent – harmful algae species in the Gulf of Mexico. One bloom during summer 2000 killed over 2 million fish (Magaña et al. 2003) and caused an estimated $16 to $18 million damages in the city of Galveston alone (Evans and Jones 2001). Thus, there is an impetus to better understand the biological function and synthesis pathway of brevetoxins.

Brevetoxin

K. brevis synthesizes two forms of neurotoxic brevetoxin, PbTx-1 and PbTx-2. Extracellular breakdown products of PbTx-1 and PbTx-2 include PbTx-3 through PbTx-8, metabolites that also act as neurotoxins in vertebrates (Baden 1989; Bourdelais et al. 2005; Poli et al. 1986). In 1981, the structure of brevetoxin PbTx-2 (then known as brevetoxin B) was successfully visualized and refined through a combination of X-ray crystallography, the chiral dibenzoate method, and NMR spectroscopy (Lin et al. 1981). Although previous studies had characterized brevetoxin toxicity with mice bioassays (McFarren et al. 1965; Spikes et al. 1968), Lin et al. was the first group to report that PbTx-2, as a ladder-frame trans-fused ether compound, was chemically unique from other known dinoflagellate toxins (Lin et al. 1981). The structures of PbTx-1 and PbTx-3 through -10 were subsequently described (Baden 1989; Chou and Shimizu 1982; Golik et al. 1982; Shimizu et al. 1986). All brevetoxins are ladder-frame polyketides that can be distinguished by their polyether backbones and R groups (Figure I-1). PbTx-1, -7, and -10 have type-2 polyether backbones, while PbTx-2, -3, -5, -6, -8, and -9 have type-1 polyether backbones (Baden 1989). The variation in PbTx chemical composition impacts toxicity. In fact, PbTx-1 consistently demonstrates the highest measured toxin potency, based on half maximal effective concentration (EC_{50}) (Berman and Murray 1999; Cao et al. 2008).

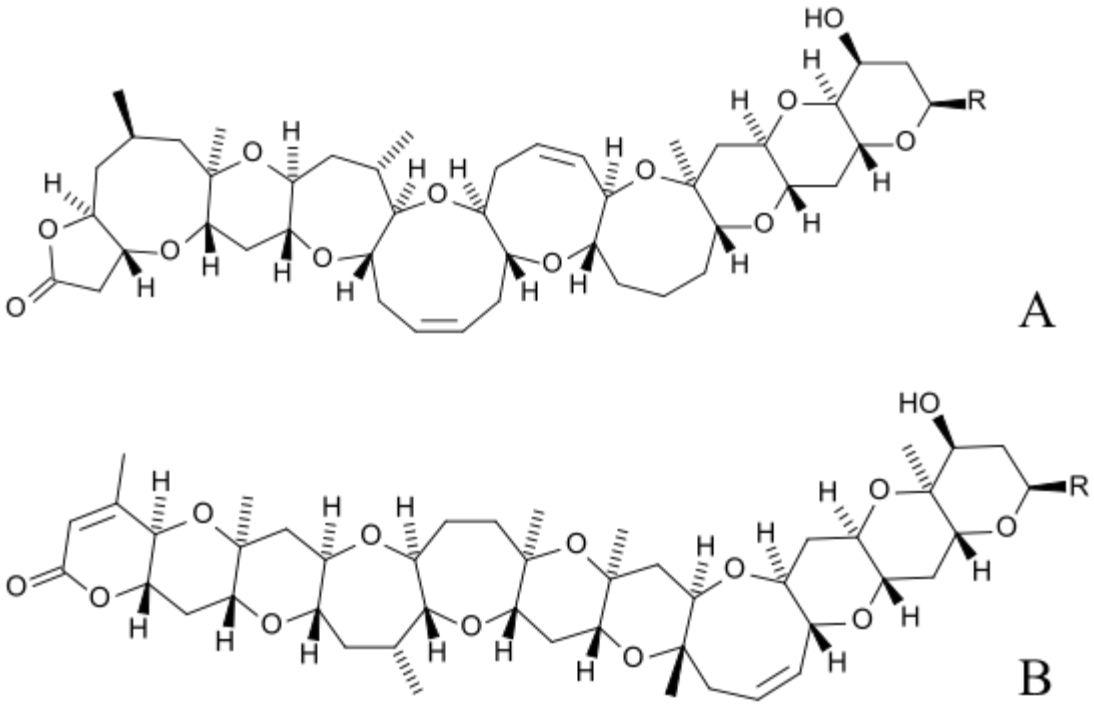

Figure I-1: Backbone structure of type-2 (A) and type-1 (B) brevetoxins. A was adapted from Nicolaou et al. 1998, and B was adapted from Matsuo et al. 2004.

Brevetoxins bind to neurotoxin receptor site 5 on the voltage-gated sodium channel (Baden 1989). The vertebrate voltage-gated sodium channel forms a transmembrane pore through which Na^+ ions selectively pass. Each channel contains four α homologous domains (α1 through α4) with six transmembrance helices (S1 through S6) each (Catterall 2000). Toxins that affect voltage-gated sodium channel activity may bind to one of six known α-domain neurotoxin receptor sites and block the channel pore, increase channel activation, and/or slow the rate of channel inactivation (Catterall et al. 2005). Brevetoxins bind to neurotoxin site 5, which is located on the α1m transmembrane helix 6 (IS6) and α4, transmembrane helix 5 (IVS5) regions (Catterall

and Gainer 1985; Poli et al. 1986; Trainer et al. 1994). Notably, a competitive inhibitor of brevetoxin – brevenal – is also synthesized by *K. brevis* cells (Bourdelais et al. 2004).

Although the channel-binding mechanism responsible for neurotoxic shellfish poisoning in humans is well-supported, the biological function of brevetoxins in *K. brevis* is subject to debate and widespread study. Recently, a 25% increase in PbTx production was measured in laboratory-cultured *K. brevis* cells after exposure to hypo-osmotic stress (Errera and Campbell 2012). It was hypothesized that brevetoxin production is related to osmoacclimation in *K. brevis*, possibly through a novel interaction with *K. brevis* voltage-gated Na^+ channels (Errera and Campbell 2012). Therefore, there is an impetus to characterize both ion transport and brevetoxin production in *K. brevis*. Similarly, the genetic pathway underlying brevetoxin synthesis is unknown.

Polyketide synthesis

Brevetoxin is one of many polyketide toxins produced by dinoflagellates (Rein and Borrone 1999). Polyketides, structurally and functionally diverse secondary metabolites, are synthesized through a carbon chain elongation process (Hopwood and Sherman 1990). Enzymes called polyketide synthases (PKSs) catalyze polyketide biosynthesis through the Claisen condensation elongation method. In brief, the combined activity of ketosynthetase (KS), acyl transferase (AT), β-keto-reductase (KR), dehydratase (DH), and/or acyl carrier protein (ACP) domains are necessary for chain

elongation with optional β-keto reduction, while AT+ACP domains function as the loading module (Shen 2003).

According to the paradigm, there are two major groups of PKSs: type I and type II. Type I PKSs are multimodular, possessing multiple PK-assembly catalytic domains in one protein, whereas type II PKSs, which are primarily expressed by prokaryotes, function through the interaction of several distinct single-domain proteins (Fischbach and Walsh 2006). In sequenced unicellular algae genomes, potential type I and type II PKS genes have been identified (Shelest et al. 2015).

Dinoflagellates produce complex polyketides, including the neurotoxins maitotoxin, brevetoxin, and ciguatoxin (Rein and Borrone 1999). Notably, when *K. brevis* cDNA libraries were first searched for sequences with homology to annotated PKSs in the NCBI protein database, the results defied the two-type paradigm. Eight "type I-like" (also called "type I modular") PKS transcripts, each containing a single predicted catalytic domain (KS, ACP, or KR) with high sequence similarity to type I proteins, were identified in *K. brevis* (Monroe and Van Dolah 2008). Type I-like PKS transcripts are also expressed by other dinoflagellate genera, including *Gambierdiscus, Alexandrium,* and *Amphidinium* (Murray et al. 2016).

De novo transcriptome assembly

Early methods to characterize mRNA expression included the construction of expressed sequence tag (EST) libraries and microarrays (Martin and Wang 2011). ESTs are short (200-800 bp), randomly sequenced fragments of cDNA (Parkinson and Blaxter

2009), while microarrays enable gene expression analysis through the hybridization of probe-labelled mRNA to DNA-spotted chips (Schena et al. 1995). Initial *K. brevis* EST and microarray work by Lidie and Van Dolah showed that *K. brevis* gene expression is complex. Like other dinoflagellate species, *K. brevis* produces vast amounts of unique transcripts, including mRNAs that are modified by *trans*-splicing processes (Lidie et al. 2005; Lidie and Van Dolah 2007; Van Dolah et al. 2007). First characterized in trypanosomes (Sutton and Boothroyd 1986), *trans*-splicing is the combination of two unlinked transcripts, a protein-coding pre-mRNA and a 5' spliced leader RNA, to produce a single mature mRNA (Agabian 1990).

Early *K. brevis* EST libraries identified ~12,000 gene clusters (Lidie et al. 2005; Lidie and Van Dolah 2007), including 80 mature *K. brevis* mRNA capped by all or part of a 22-nucleotide dinoflagellate spliced leader sequence (SLS) (Lidie and Van Dolah 2007). Although conserved within the dinoflagellate group, the dinoflagellate SLS (5′-DCCGTAGCCATTTTGGCTCAAG-3′) is not homologous to SLSs used by other eukaryotic phyla (Lidie and Van Dolah 2007; Zhang et al. 2007). To date, the function of *trans*-splicing in dinoflagellates is not known, but use of SLSs in trypanosomes has been implicated in post-transcriptional regulation to control protein translation (Brunelle and Van Dolah 2011; Morey et al. 2011). Indeed, microarray studies have observed a high percent of expressed genes are related to RNA post-transcriptional processing and protein processing in *K. brevis* (Van Dolah et al. 2007).

Recent developments in next-generation DNA/RNA sequencing technology have enabled researchers to study transcriptomes and genomes through the analyses of

millions of short reads (50 to >1000 bp long) (Metzker 2010). Platforms like Illumina HiSeq or MiSeq, Pac Bio, Roche/454, and SOLiD ABI use massively parallel sequencing to produce relatively inexpensive, high-depth data. Transcriptomes – here defined as the complete set of mRNA in a given sample at a specific time – are commonly assembled by aligning next-generation RNA-Seq reads to a reference genome (Wang et al. 2009). However, this approach limits analyses to model organisms, neglecting species without complete genomes, including most dinoflagellates. As of 2016, the only dinoflagellate with a fully assembled nuclear genome is *Symbiodinium minutum*, a symbiotic zooxanthellae species that lives within coral polyps (Shoguchi et al. 2013). To enable analyses of non-model organisms, *de novo* transcriptome assembly programs like Trinity and Velvet-Oases divide RNA-Seq reads into k-mers and assemble full transcripts using de Bruijn graph algorithms (Grabherr et al. 2011; Schulz et al. 2012; Zerbino and Birney 2008). The resulting data provides a more complete view of nonmodel transcriptomics than EST libraries alone. For example, this dissertation identified between 80,000 and 90,000 unique *K. brevis* transcripts, whereas the number of gene clusters in early *K. brevis* EST libraries were seven times less numerous.

Dissertation aims and published work

In order to characterize gene expression in toxin-producing *Karenia* species, this dissertation had four primary aims:

- Identify an efficient, effective *de novo* transcriptome assembly method for dinoflagellates with large, highly repetitive genomes.

- Predict highly conserved genes in dinoflagellates. If dinoflagellates share a pool of highly conserved, consistently expressed genes, the sequences can be used to assess transcriptome completeness.
- Identify the potential genes that underlie osmoacclimation or toxin production in harmful, bloom-forming dinoflagellate *K. brevis*, especially putative voltage-gated cation channels and polyketide synthases.
- Predict genetic variance among two *K. brevis* laboratory clones and three *Karenia* species: *K. brevis*, *K. mikimotoi*, and *K. papilionacea*.

Three papers describe the research that supports each aim. Two, "*De novo* assembly and characterization of the transcriptome of the toxic dinoflagellate *Karenia brevis*" (Ryan et al. 2014) and "Comparative transcriptomic analysis of three toxin-producing *Karenia* species" (Ryan and Campbell 2015) have been published, and a third, "Novel genes in toxin-producing *Karenia* species," has been submitted. The papers have been included as dissertation chapters.

CHAPTER II

DE NOVO ASSEMBLY AND CHARACTERIZATION OF THE TRANSCRIPTOME OF THE TOXIC DINOFLAGELLATE *KARENIA BREVIS**

Synopsis

Background

Karenia brevis is a harmful algal species that blooms in the Gulf of Mexico and produces brevetoxins that cause neurotoxic shellfish poisoning. Elevated brevetoxin levels in *K. brevis* cells have been measured during laboratory hypo-osmotic stress treatments. To investigate mechanisms underlying *K. brevis* osmoacclimation and osmoregulation and establish a valuable resource for gene discovery, we assembled reference transcriptomes for three clones: Wilson-CCFWC268, SP3, and SP1 (a low-toxin producing variant). *K. brevis* transcriptomes were annotated with gene ontology terms and searched for putative transmembrane proteins that may elucidate cellular responses to hypo-osmotic stress. An analysis of single nucleotide polymorphisms among clones was used to characterize genetic divergence.

Results

K. brevis reference transcriptomes were assembled with 58.5 (Wilson), 78.0 (SP1), and 51.4 million (SP3) paired reads. Transcriptomes contained 86,580 (Wilson),

*Reprinted from *De novo* assembly and characterization of the transcriptome of the toxic dinoflagellate *Karenia brevis*, by Darcie E. Ryan, Lisa Campbell, and Alan E. Pepper. 2014. *BMC genomics* 15(1):888. BMC Genomics is an open access journal. Chapter II is formatted in accordance with BMC guidelines.

93,668 (SP1), and 84,309 (SP3) predicted transcripts. Approximately 40% of the transcripts were homologous to proteins in the BLAST nr database with an E value ≤ $1.00x10^{-6}$. Greater than 80% of the highly conserved CEGMA core eukaryotic genes were identified in each transcriptome, which supports assembly completeness. Seven putative voltage-gated Na^{+} or Ca^{2+} channels, two aquaporin-like proteins, and twelve putative VATPase subunits were discovered in all clones using multiple bioinformatics approaches. Furthermore, 45% (Wilson) and 43% (SP1 and SP3) of the *K. brevis* putative peptides >100 amino acids long produced significant hits to a sequence in the NCBI nr protein database. Of these, 77% (Wilson and SP1) and 73% (SP3) were successfully assigned gene ontology functional terms. The predicted single nucleotide polymorphism (SNP) frequencies between clones were 0.0028 (Wilson to SP1), 0.0030 (Wilson to SP3), and 0.0028 (SP1 to SP3).

Conclusions

The *K. brevis* transcriptomes assembled here provide a foundational resource for gene discovery and future RNA-seq experiments. The identification of ion channels, VATPases, and aquaporins in all three transcriptomes indicates that *K. brevis* regulates cellular ion and water concentrations via transmembrane proteins. Additionally, >40,000 unannotated loci may include potentially novel *K. brevis* genes. Ultimately, the SNPs identified among the three ecologically diverse clones with different toxin profiles may help to elucidate variations in *K. brevis* brevetoxin production.

Background

The dinoflagellate *Karenia brevis* blooms almost annually in the Gulf of Mexico and is the region's major harmful algal species (Steidinger et al. 1998). *K. brevis* produces two ladder-frame polyether brevetoxin compounds, PbTx-1 and PbTx-2, which bind to receptor site 5 of voltage-gated Na^+ channels (Baden 1989; Dechraoui et al. 1999). Both parent compounds and their derivatives inhibit channel deactivation (Huang et al. 1984). Because they affect voltage-gated Na^+ channel activity, brevetoxins are responsible for neurotoxic shellfish poisoning (NSP), may cause marine animal mortalities during blooms, and have been implicated in fish kills (Landsberg 2002). Additionally, *K. brevis* cells that are damaged by the breaking surf have been shown to release enough aerosolized brevetoxins to cause eye irritation and respiratory distress in humans near the shore (Backer et al. 2003).

Despite the health risks associated with brevetoxins, their biological function within *K. brevis* is currently unknown. Notably, recent evidence suggests that PbTx-1 and PbTx-2 production increases in response to hypo-osmotic stress. Within 24 hours after a rapid media salinity reduction (35 to 27), the *K. brevis* Wilson clone (CCFWC268) produced ~25% more total brevetoxin per cell. It was therefore hypothesized that toxin production facilitates the response of *K. brevis* to salinity variations in the natural environment (Errera and Campbell 2012). As populations move from offshore oceanic to coastal waters, cells experience a range of environmental salinities. For example, cells of this oceanic species have even been observed in the hyposaline Mississippi River Delta (Maier Brown et al. 2006).

To better characterize brevetoxin production and osmoacclimation in *K. brevis*, the reference transcriptomes of three *K. brevis* clones, Wilson-CCFWC268, SP1, and SP3 were assembled. These clones represent diverse geographic origins, duration of years in culture, and toxin profiles. Wilson was initially collected off the coast of Florida Gulf in 1953. In contrast, the SP1 and SP3 clones originate from the Texas coast in 1999. While SP1 produces low, often undetectable amounts of PbTx-1 and PbTx-2, the total brevetoxin in SP3 and Wilson consistently exceeds 10 pg $cell^{-1}$ (Errera and Campbell 2012; Errera et al. 2010). Thus, differences among the three transcriptomes may improve our understanding of brevetoxin production.

Brevetoxins are ladder-frame polyethers that belong to the polyketide family. Polyketide synthase (PKS) genes have been isolated from Wilson cultures (Snyder et al. 2003). In 2008, Monroe *et al.* identified four novel *K. brevis* PKS mRNA sequences in *K. brevis* clones Wilson, C2, NOAA-1, and SP2 that were not present in other dinoflagellates, including closely-related *Karenia mikimotoi* and *Karlodinium veneficum*. Because these four PKS sequences are unique to *K. brevis* and are structurally novel, it was hypothesized that they participate in the brevetoxin biosynthetic pathway (Monroe and Van Dolah 2008). As part of this project, we searched the SP1, SP3, and Wilson transcriptomes for novel *K. brevis* PKS genes to determine if the known PKS genes were transcribed in all three clones.

Because brevetoxins increase voltage-gated Na^+ channel activity, determining whether *K. brevis* expresses these channel proteins may elucidate the physiological function of this toxin. Previously constructed *K. brevis* EST libraries contained ~12,000

unique genes (Lidie et al. 2005; Lidie and Van Dolah 2007), but no complete Na^+ channel protein-coding region was identified in this gene set. Further, channel-based Ca^{2+}, K^+, and Na^+ transport across cell membranes is one known method that plants and algae use to maintain homeostasis (Glass 1983). The presence of ion channel sequences in *K. brevis* transcriptomes would suggest that this dinoflagellate is capable of osmoregulation through selective transmembrane ion transport.

Although aquaporins and VATPases have not been implicated in brevetoxin binding, they might affect osmoacclimation efficiency in *K. brevis*. Aquaporins were originally discovered in human red blood cells (Agre et al. 1993) but have since been found in taxa belonging to the bacterial, archaeal, and eukaryotic domains (Heymann and Engel 1999). These bidirectional transport proteins belong to the major intrinsic protein (MIP) family and move water and/or glycerol molecules across lipid membranes more quickly than diffusion (Agre et al. 2002; Borgnia et al. 1999). Similarly, VATPases generate pH gradients that trigger secondary ion transport (Nelson et al. 2000). They have been identified in diverse eukaryotes and may participate in osmoregulation. For example, in *Arabidopsis thaliana*, Ca^{2+}, Na^+, and K^+ starvation induced transcript-level downregulation of VATPase family genes (Maathuis et al. 2003), and inhibition of plasma H^+-ATPases in green alga *Dunaliella salina* prevented cell volume recovery after hyper-osmotic stress (Maathuis et al. 2003).

The haploid *K. brevis* genome is estimated to be 1×10^{11} base pairs (bp) (Kamykowski et al. 1998; Kim and Martin 1974; Rizzo et al. 1982; Sigee 1984; Van Dolah et al. 2009) and has not been sequenced. This exceedingly large genome size

highlights the crucial need for a reference transcriptome in order to initiate serious genomic analyses of this species. The Wilson, SP1, and SP3 *K. brevis* transcriptomes are among the first dinoflagellate transcriptomes to be assembled. Because no reference genome was available, a goal of this study was to evaluate different techniques to determine the *de novo* assembly method best suited for our data. Among eukaryotes, dinoflagellates are unique in a number of ways. A 22-nucleotide spliced leader sequence (SLS) has been identified in nuclear mRNA from all dinoflagellate species, including *K. brevis* (Lidie and Van Dolah 2007; Zhang et al. 2007). Though the dinoflagellate SLS is conserved within the dinoflagellate group, it is not homologous to SLSs used by other eukaryotic phyla (Zhang et al. 2007). Additionally, dinoflagellate chromosomes are permanently condensed and nuclear genomes often contain a high quantity of repetitive, non-coding DNA (Lin 2011; Rizzo et al. 1982; Rizzo 2003). These characteristics make dinoflagellates a biologically interesting target for transcriptome characterization, analysis of the gene complement, and gene expression studies. Here we report the results of a search for *K. brevis* PKSs, voltage-gated ion channels, aquaporins, and VATPases and describe transcriptome sequence differences in these genes among ecologically diverse clones.

Results

Transcriptome assembly

After trimming for quality and length, 58.5 million, 78.0 million, and 51.4 million paired reads were used to assemble the *K. brevis* Wilson, SP1, and SP3 reference

transcriptomes, respectively (Table II-1) using both the Velvet-Oases (Schulz et al. 2012; Zerbino and Birney 2008) and ABySS (Simpson et al. 2009) assembly methods. Based on the N50 length and mean transcript length (Figure II-1), the transcriptomes produced by the Oases merged assembly (MA) technique with K-values of 29 bp (Wilson and SP1) or 33 bp (SP3) were considered optimal and used during all subsequent analyses. Further, the transcript analysis software TRAPID (Van Bel et al. 2013) identified 1549 more full-length open reading frames (ORFs) in the SP1 MA Oases transcriptome, compared to the SP1 MA ABySS transcriptome. These results support the choice of Velvet-Oases, since this assembler produced more transcripts with complete protein-coding regions.

The reference transcriptomes had ~ 90,000 loci each (Table II-1). Of these, 34% (Wilson), 35% (SP1), and 85% (SP3) contained two or more isoforms that were collapsed to a single representative transcript (Table II-1). Approximately 87% (Wilson and SP1) or 74% (SP3) of the transcripts were less than 2,500 bp long (Figure II-2). Based on BLASTn results, 4.3% (Wilson), 8.8% (SP1), and 3.1% (SP3) of the loci in each transcriptome were present in only one of the clones (Table II-1). However, when reads were aligned to the assembled transcriptomes, <1% of the transcripts only matched reads from one clone (Table II-2). This suggests that the BLASTn results overestimate the number of loci apparently unique to Wilson, SP1, or SP3.

Table II-1: A comparison of *K. brevis* transcriptome read number, locus number, apparently clone-unique locus number, N50 length, and mean locus length values.

Clone	# Reads	# Loci	% Isoforms	# Unique Loci	N50 (bp)	Mean Locus Length (bp)
Wilson	58,535,595	86580	34	3712	2038	1340
SP1	77,994,379	93668	35	8202	2124	1376
SP3	51,363,303	84309	85	2606	3424	1941

The % isoforms column lists the percentage of loci that were assigned two or more isoforms by Velvet-Oases.

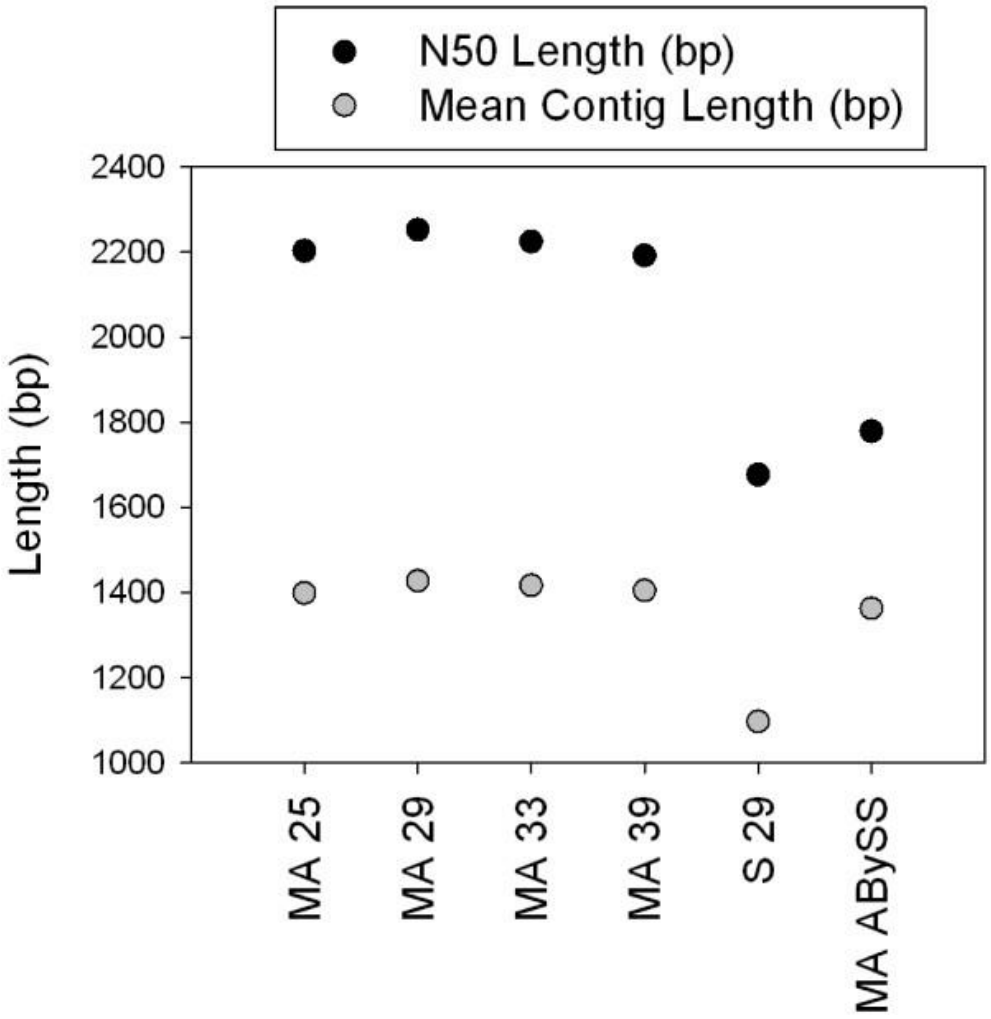

Figure II-1: N50 and mean transcript length values for merged (MA) and single k-mer (S) Velvet-Oases and ABySS SP1 transcriptome assemblies.

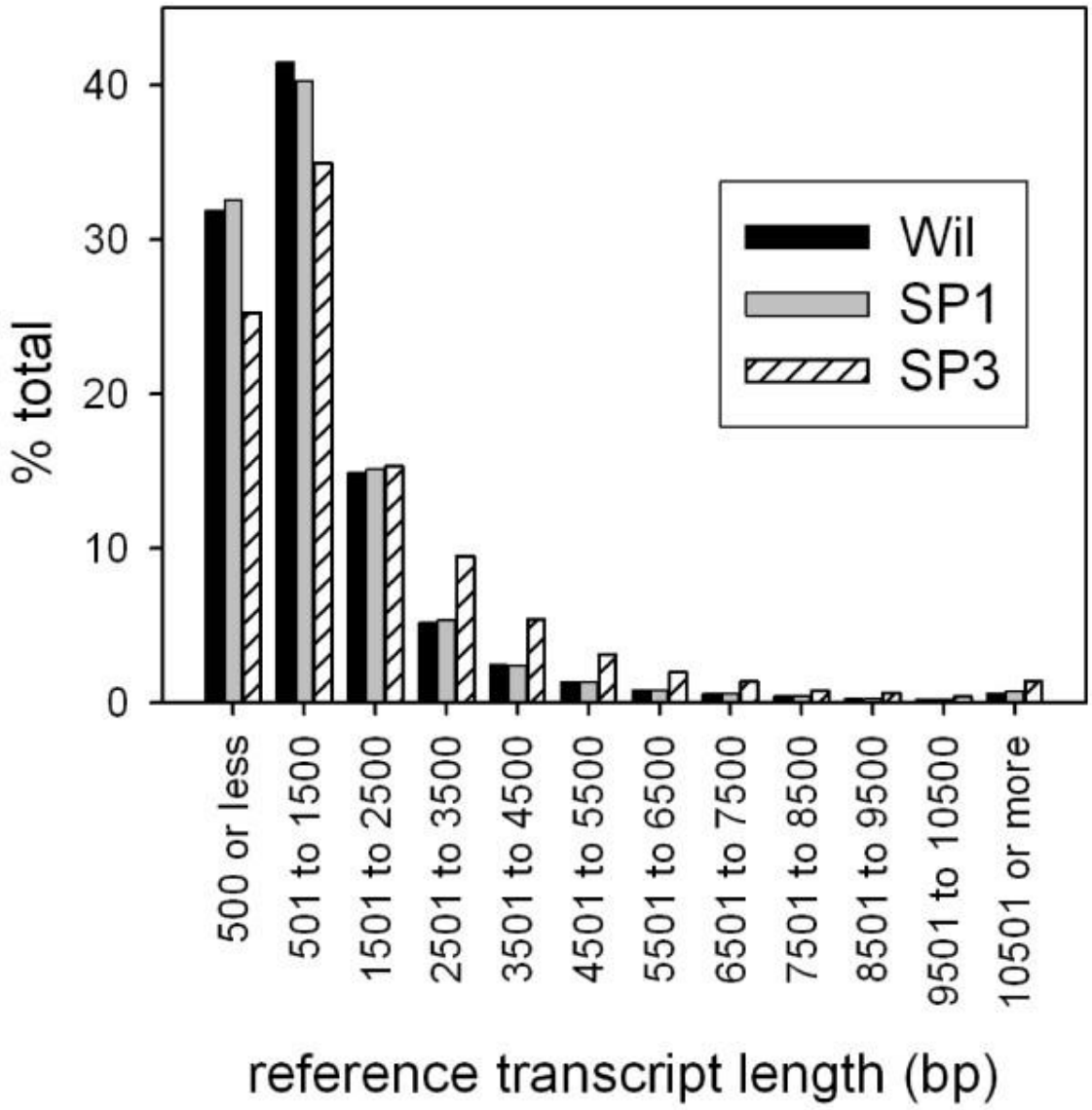

Figure II-2: *K. brevis* transcriptome reference transcript length histogram.

Table II-2: Short read alignment results.

Transcriptome	# Transcripts without Wilson alignments	# Transcripts without SP1 alignments	# Transcripts without SP3 alignments
Wilson	0	472 (0.55%)	438 (0.51%)
SP1	751 (0.80%)	0	718 (0.77%)
SP3	610 (0.72%)	564 (0.67%)	0

Whole-transcriptome annotation

The number of predicted ORFs and their length distribution was similar among the three transcriptomes (Figure II-3). Complete (start to stop codon) and partial (no start codon) ORFs longer than 300 bp were considered to be possible protein-encoding transcripts. During the BLASTp search against the nr database, 45% (Wilson), 43% (SP1), and 43% (SP3) of the *K. brevis* putative peptides >100 aa significantly hit a sequence (E value $\leq 1.00 \times 10^{-6}$). Of those transcripts with positive BLAST alignments, 77% (Wilson), 77% (SP1), and 73% (SP3) were annotated with GO terms (Figure II-4).

CEGMA and TRAPID analyses

Core eukaryotic genes (CEGs) represent an unbiased set of proteins that are expressed and conserved within diverse eukaryotes (Parra et al. 2007), and therefore CEG identification helps gauge transcriptome assembly completeness. With CEGMA, we found 81% (Wilson), 84% (SP1), and 82% (SP3) (Additional file 1) of the complete

highly conserved core genes described by Parra *et al*. (Parra et al. 2007). In comparison, 74% and 90% of the complete CEGs were identified in the dinoflagellate *Karlodinium micrum CCMP2283* transcriptome and diatom *Thalassiosira pseudonana* genome, respectively. Thus, as a metric of transcriptome assembly success, the identification of complete CEGs in our transcriptomes indicates a high level of completeness in their coverage. The analysis also identified many conserved proteins that can be used to refine dinoflagellate phylogenetic relationships.

The "missing" CEGs that were not identified in any *K. brevis* reference transcriptome may represent genes that are not used by dinoflagellates. It is also possible that "missing" *K. brevis* CEG orthologs are not highly conserved by CEGMA criteria. To investigate these options, the SP1 transcriptome was used to search the CEG protein list using BLASTx. SP1 transcripts hit 235 of the 248 core CEGs with an E value $<1.00x10^{-6}$. Therefore, homologs of 25 of the 38 "missing" genes were detected in the CEGMA dataset using BLASTx.

The ratio of full-length/quasi full-length to partial coding regions predicted by TRAPID (Van Bel et al. 2013) characterizes protein coding region completeness. The Wilson, SP1, and SP3 Oases reference transcriptomes were compared against three databases: the OrthoMCLDB 5.0 alveolate clade db, the PLAZA 2.5 green plants clade db, and the OrthoMCLDB 5.0 *T. pseudonana* CCMP1335 species db. The ratios of complete to partial coding regions identified in the reference transcriptomes were 7.3:1 (Wilson), 13.5:1 (SP1), and 12.5:1 (SP3). Over 90% of the *K. brevis* transcripts that

significantly matched sequences in the TRAPID databases had ORFs that fell within two standard deviations of the mean ORF length in their gene families.

PKS, Aquaporin, voltage-gated ion channel, and VATPase identification

The Wilson, SP1, and SP3 transcriptomes each included positive hits to the four novel *K. brevis* PKS sequences (gi # 148536478, 148536480,148536474, and 148536472), with nucleotide similarity values >99% over the aligned regions. No unique non-synonymous SNPs were identified in the SP1 PKS ORFs.

The same seven putative voltage-gated Na^{+} or Ca^{2+} channel genes were identified in all three *K. brevis* transcriptomes (SP1 transcriptome locus # 394, 12559, 19932, 26784, 30595, 36263, and 64946). Each sequence contained the voltage-sensing motif and four Pfam00520 domains with ~ six predicted transmembrane regions. During the BLASTx search against the nr database, the putative proteins very significantly (E value $\leq 1.00x10^{-50}$) matched voltage-gated Na^{+} and Ca^{+2} channels from a range of organisms, including mammals.

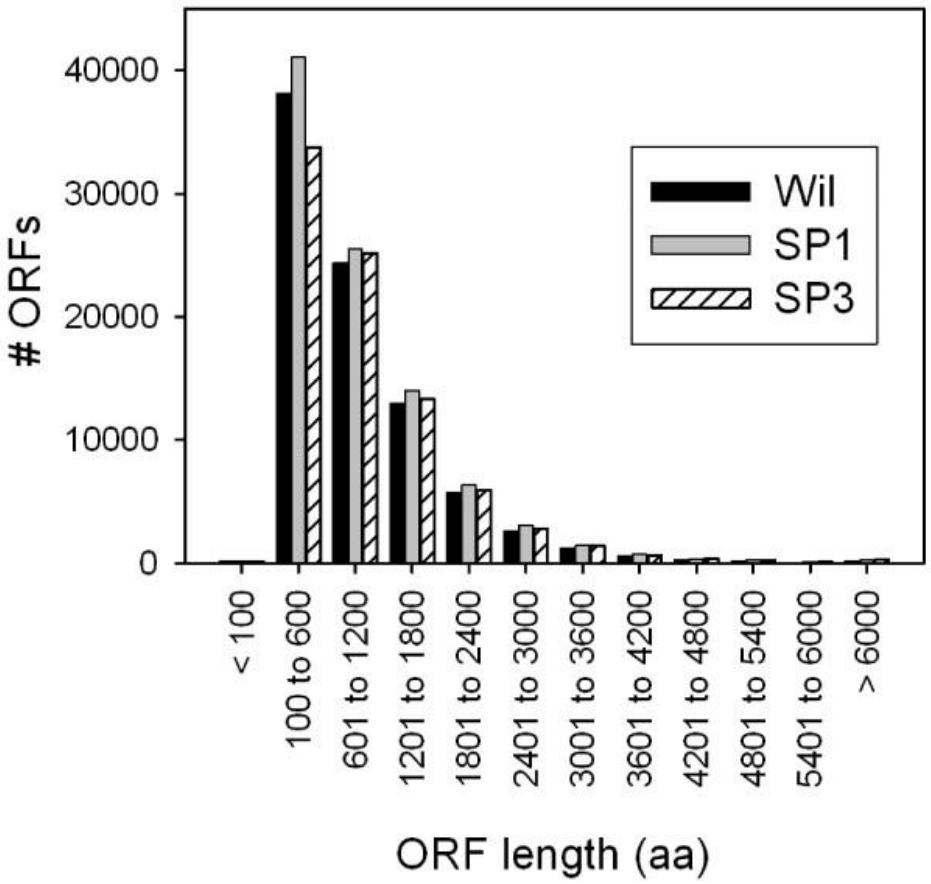

Figure II-3: Predicted ORF length distribution in the Wilson, SP1, and SP3 transcriptomes. Length values are represented in #aa, or #bp in ORF divided by three.

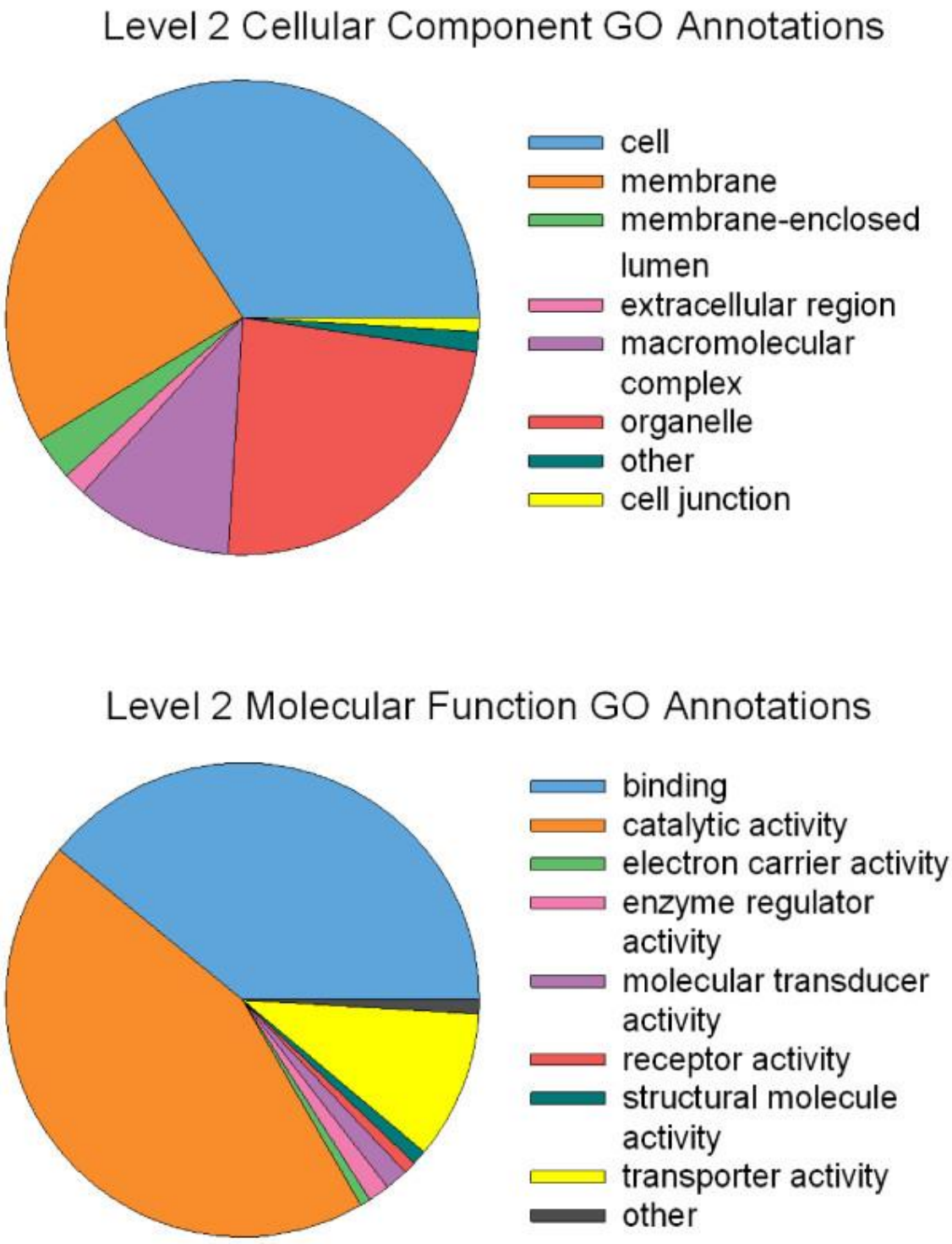

Figure II-4: Distribution of second-level cellular component and molecular function GO annotations in annotated *K. brevis* reference transcripts. The percent distribution is identical in all three clones.

The transcriptomes contained twelve different putative VATPase subunits with V-ATPase conserved domains. Two MIP genes were also identified in all the transcriptomes. The 267 and 405 aa putative proteins (SP1 transcriptome locus # 102528 and 12778, respectively) each contained six predicted transmembrane regions and conserved domains belonging to the MIP family. The aquaporin Asn-Pro-Ala (NPA) water selectivity filter motif was also identified, with conservation of the Asn in each occurrence. The *K. brevis* MIP sequences were compared to the nr protein database. Each returned over 100 hits with E values $\leq 1.00 \times 10^{-20}$. Of these, ~70% were aquaporins or predicted MIPs. The complete ORF of MIP #12778 was located in Wilson, SP1, and SP3, with almost 100% aa sequence conservation among clones, with just one aa variation in SP3, a Val to Met substitution at aa position 270. The ORF of MIP # 102529 was incomplete in the Wilson and SP3 transcriptomes.

SNP identification

When the Wilson and SP3 short reads were aligned to the SP1 transcriptome with a conservative 20-fold minimum coverage cutoff, 186,075 SNP locations were identified in 30,227 loci. Of these, 75,176 (40%) were exclusively in Wilson, 65,867 (35%) were exclusively in SP3, and 45,032 (25%) were in both Wilson and SP3. The SNP frequencies between Wilson and SP1, Wilson and SP3, and SP1 and SP3 were 0.0023, 0.0024, and 0.0022, respectively (Table II-3). The 10-fold threshold analysis identified 312,723 potential SNP locations in 58,051 loci. Among these, 117,714 (38%) were exclusively in Wilson, 111,660 (36%) were exclusively in in SP3, and 83,349 (26%)

were present in both Wilson and SP3. At this coverage threshold, the estimated Wilson and SP1, Wilson and SP3, and SP1 and SP3 SNP frequencies had increased to 0.0028, 0.0030, and 0.0028 (Table II-3).

SNPs were analyzed in the seven putative voltage-gated Na^+ or Ca^{2+} channels to examine variations among similar genes. The channel ORF length ranged from 4401 to 7908 bp, with a mean length of 6110 bp. Because the mean coverage of each sequence exceeded 10-fold, SNP rates were determined by the 10-fold threshold analysis. The mean SNP rate of all seven channels was 0.0032 (SP1 to Wilson), 0.0037 (SP3 to Wilson), or 0.0027 (SP1 to SP3) (Table II-4). Notably, clone-to-clone SNP rate varied among voltage-gated cation channel sequences. Channel 19932 was ~100% identical in Wilson, SP1, and SP3, containing only one predicted SNP. In contrast, pairwise comparisons of channel 394 predicted a SNP ~ 1 out of every 160 nucleotides (Table II-4). Non-synonymous SNPs that altered the amino acid sequence in the putative Na^+ or Ca^{2+} channels were less common than synonymous SNPs, occurring at frequencies ranging from 0.0018 to 0.00033, with mean frequencies of 0.00095 (Wilson to SP1), 0.000867 (Wilson to SP3), and 0.000667 (SP1 to SP3) (Table II-4). Variations in the non-synonymous SNP prevalence among putative cation channels suggest that the channels may be subject to different selective constraints. Furthermore, because Wilson, not SP1, is the most divergent clone overall, toxin production does not appear to affect cation channel gene selection.

Table II-3: SNP detection results.

	Clones	# SNPs	Mean SNP Rate	# Transcripts with SNPs
20X	Wilson vs SP1	120208	0.0023 (1 / 442)	25123
	Wilson vs SP3	141044	0.0024 (1 / 421)	28427
	SP1 vs SP3	110899	0.0022 (1 / 465)	22794
10X	Wilson vs SP1	201063	0.0028 (1 / 358)	37486
	Wilson vs SP3	229374	0.0030 (1 / 339)	41937
	SP1 vs SP3	195009	0.0028 (1 / 364)	36144

The mean SNP rate was calculated in transcripts with at least one SNP.

Table II-4: SNPs in putative voltage-gated Na^{+} or Ca^{2+} channel sequences.

	Channel ID	# SNPs in Transcript	Mean SNP Rate	Length transcript	Length ORF	Non-synonymous / Synonymous SNPs in ORF
Wilson vs SP1	394	45	0.0065	6893	6686	0.23
	12559	30	0.0046	6475	6213	0.39
	19932	1	0.0002	6464	6165	1.00
	26784	25	0.0041	6034	5862	0.27
	30595	23	0.0032	7274	5532	—
	36263	20	0.0025	8014	7908	0.26
	64946	6	0.0013	4532	4401	0.80
Wilson vs SP3	394	44	0.0064	6893	6686	0.29
	12559	35	0.0054	6475	6213	0.25
	19932	0	0	6464	6165	—
	26784	27	0.0045	6034	5862	0.23
	30595	24	0.0033	7274	5532	0.00

Table II-4 continued.

	Channel ID	# SNPs in Transcript	Mean SNP Rate	Length transcript	Length ORF	Non-synonymous / Synonymous SNPs in ORF
	36263	19	0.0024	8014	7908	0.20
	64946	19	0.0042	4532	4401	0.19
SP1 vs SP3	394	39	0.0057	6893	6686	0.13
	12559	17	0.0026	6475	6213	0.29
	19932	1	0.0002	6464	6165	1.00
	26784	22	0.0036	6034	5862	0.19
	30595	5	0.0007	7274	5532	—
	36263	13	0.0016	8014	7908	0.18
	64946	21	0.0046	4532	4401	0.44

Discussion

Our *K. brevis* transcriptomes are among the first dinoflagellate transcriptomes to be assembled, so it is not possible to make comparisons with closely related species. Lacking a reference genome sequence, several metrics, including transcript length, TRAPID-predicted full-length genes, and the identification of CEGs, were employed to gauge the completeness of the transcriptome assembly. Each of these criteria suggests that the transcriptomes are highly complete. First, the estimated mean protein-coding gene length of 19 model eukaryotes with sequenced genomes, including *Arabidopsis thaliana*, *Caenorhabditis elegans*, and red alga *Cyanidioschyzon merolae*, is 1,346 bp (Xu et al. 2006). This value is similar to the Wilson, SP1, and SP3 mean reference transcript lengths of 1340, 1376, and 1941 bp, respectively; therefore, our *de novo* assemblies of the *K. brevis* transcriptome appear to yield transcripts of a reasonable length.

Next, a 66% CEG identification rate was reported when Parra *et al*. analyzed the *Toxoplasma gondii* genome with CEGMA (Parra et al. 2007). Apicomplexan *T. gondii* and dinoflagellate *K. brevis* both belong to the alveolate group and are close phylogenetic relatives, based on rRNA analyses (Van de Peer and De Wachter 1997). Identification of >80% of the highly conserved CEGs from each *K. brevis* reference transcriptome provides additional support for the completeness of the assembly, since this value exceeds the expected percent based on *T. gondii* results. TRAPID results also indicated that more transcripts contained complete or mostly complete ORFs than partial

ORFs. Of the transcriptomes similar to TRAPID alveolate, gene families, over 92% were within two standard deviations of the expected length.

Only ~40% of the genes in the *K. brevis* transcriptomes are homologous to sequences in the nr protein database with a hit significance $\leq 1.00 \times 10^{-6}$. Because little is known about dinoflagellate genomes, the high percentage of unknown loci was expected. The Blast2GO annotation step did not annotate enough hits to allow conclusions about the total gene ontology distribution of the transcriptomes. The low annotation rate is a result of low (<50%) similarity scores between *K. brevis* sequences and annotated proteins. This may be the result of the phylogenetic uniqueness of dinoflagellates combined with limited phylogenetic representation in the Blast2GO databases.

The four novel *K. brevis* PKS sequences (Monroe and Van Dolah 2008) were found and expressed in all three clones. No unique non-synonymous SNPs were identified in the SP1 PKS ORFs. It is therefore possible that SP1 expresses the genes involved in brevetoxin production, though cellular PbTx-1 and PbTx-2 are often undetectable in this clone. This result is consistent with prior work investigating transcriptional and post-transcriptional regulation in *K. brevis*. Microarray studies have observed a high percent of expressed genes related to RNA post-transcriptional processing and protein processing in *K. brevis*, thus suggesting that this dinoflagellate species is highly reliant on post-transcriptional regulation (Van Dolah et al. 2009).

Some phenotypic variance between clones may result from SNPs that alter gene function. SNPs affecting 25123 (Wilson to SP1), 28427 (Wilson to SP3), or 22794 (SP1

to SP3) expressed genes were identified in laboratory-cultured *K. brevis* clones. Based on overall nucleotide divergence rates between Wilson and SP1, Wilson and SP3, and SP1 and SP3 (Table II-3), SP1 and SP3 were the most similar clones. This may be the result of time in culture. SP1 and SP3 were isolated in 1999, while Wilson-CCFWC268 has been in culture since 1953 (Errera et al. 2010).

Typically, eukaryotic algae respond to osmotic stress by differential metabolite production rates and/or the transmembrane flux of ions and water (Wegmann 1986). The identification of putative MIPs, VATPases and voltage-gated Na^+ or Ca^{2+} channels in *K. brevis* supports the hypothesis that cells osmoregulate with transmembrane channels. In *K. brevis*, aquaporins may facilitate quick responses to changes in the osmotic pressure gradient. Further, putative voltage-gated Na^+ or Ca^{2+} channels with voltage-sensing motifs may facilitate cation transport across the cell membrane in response to ion gradients. The high interspecies similarity of these protein sequences indicates functional conservation among *K. brevis* clones that show varying brevetoxin profiles. Future experimental work will need to confirm that aquaporins, VATPases, or ion channels are involved with osmoacclimation in *K. brevis*. Potential experiments may measure cell volume post hypo-osmotic stress with and without specific protein blockers.

Conclusions

Our discovery of putative ion channel, aquaporin, and VATPase sequences supports the hypothesis that *K. brevis* cells use transmembrane proteins during osmoregulation and osmoacclimation. In the future, clone-to-clone and treatment-to-

treatment comparisons of the expression of these and other novel genes using RNA-seq (Mortazavi et al. 2008) will elucidate *K. brevis* responses to osmotic stress. The transcriptomes assembled during this study also provide a foundational reference for future differential expression and protein discovery work. Thousands of *K. brevis* transcripts have been assigned GO functions (Additional file 2, Additional file 3, Additional file 4, Additional file 5, Additional file 6 and Additional file 7), and over 40,000 *K. brevis* loci containing unknown hypothetical protein-coding regions >100 aa long (~11 kDa) were assembled. Even if just a fraction of these transcribed loci encode functional proteins, our datasets identify a vast number of novel genes and gene-products. Future analyses of these genes will yield insights into the unique biology of *K. brevis*.

Methods

Cell culturing and RNA sequencing

K. brevis clones Wilson, SP1, and SP3 were maintained in L1 medium (Guillard and Hargraves 1993) at salinity 35. The medium was prepared with filtered (0.2 μm pore) and autoclave-sterilized sea water from the Flower Garden Banks region, Gulf of Mexico. For each clone, triplicate 1-L sterile glass bottles were inoculated with cells from laboratory cultures and maintained on a 12:12 hour light:dark cycle at 25°C. Cell counts were performed by light microscopy every other day to monitor growth rates.

During the late exponential growth phase, 500 ml were concentrated by centrifugation (800 × g, 10 min) and RNA was immediately extracted from the pellets in

40 μL of nuclease-free water with the Qiagen RNEasy Mini Kit (Qiagen Inc., Valencia, CA), in accordance with the manufacturer's protocol. After final elution of RNA in 40 μL of nuclease-free water, samples were stored at −80°C until library preparation. The RNA concentration and purity of each extraction was estimated by measuring the absorption spectra of 2 μL sample aliquots with a NanoDrop Spectrophotometer.

The remaining culture in each bottle was diluted from a salinity of 35 to a salinity of 27 with nutrient-enriched Milli-Q water to simulate hypo-osmotic stress. Stressed cultures were incubated for one hour before RNA was extracted, as described above. RNA was shipped on dry ice to the National Center for Genome Resources Sequencing Lab (Santa Fe NM, USA) for paired-end Illumina sequencing (Illumina, San Diego CA, USA). Libraries were prepared with the TruSeq RNA Sample Preparation Kit (Illumina) using 2 μg RNA. Paired-end 50 bp reads were sequenced with the Illumina Hi-Seq 2000 platform.

Reference transcriptome assembly

Paired-end reads are available at the NCBI SRA repository. Reads were trimmed for quality and filtered for length with CLC Genomics Workbench 5.5.1 (CLC Bio, Aarhus, Denmark); the minimum read length permitted was 45 bp. Reads were trimmed based on Phred quality scores at the probability threshold of $p = 0.05$.

Reference transcriptomes for each clone were assembled with reads pooled from both control and salinity stress treatments. Pooling the reads allowed the assembly of genes that may only be expressed in one of the two treatments. Assembly was completed

with Velvet-Oases, which uses a de Bruijn graph algorithm to build transcripts *de novo* (Schulz et al. 2012; Zerbino and Birney 2008). Coverage cutoff values were chosen automatically, and the edge fraction cutoff was increased from the default 10% to 50%. For each clone, single k-mer assemblies (k-mer lengths 21, 25, 29, 33, 37, and 41 bp) were merged into a non-redundant consensus transcriptome assembly. A 250 bp minimum transcript length threshold was enforced. For comparison purposes, an SP1 reference transcriptome was also assembled with the ABySS *de novo* paired-end assembler (Simpson et al. 2009). Single k-mer assemblies (odd lengths, 25 to 45 bp) were merged into a final transcriptome with "bubble popping" enabled.

Oases may output several transcript isoforms belonging to the same predicted locus. When multiple isoforms were present, we removed all but one representative sequence based on the following criteria. Isoforms were searched using BLAST (Altschul et al. 1997) against the other *K. brevis* transcriptomes, with the E value significance threshold 1.00×10^{-6}. The isoform that hit another sequence with the longest alignment length and highest significant bit score was retained. In the event that a locus containing multiple isoforms did not return a significant hit, the longest transcript was chosen to represent its locus. This technique retains the isoform that is most abundant across all clones, when multiple isoforms were present.

The mean locus (unigene) length and N50 value were calculated using a transcript length list. To identify putative complete genes, the transcriptomes were analyzed with TRAPID. Transcripts were assigned gene families by a comparison with the TRAPID alveolate clade database. ORFs within two standard deviations of the mean

gene family length were considered "full-length." For each *K. brevis* clone, the transcriptome assembly method that produced the greatest N50 length and the most full-length genes was considered optimal and used during subsequent analyses.

Identification of core eukaryotic proteins

To assess transcriptome completeness, loci were analyzed with the Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline. CEGMA was developed to identify a subset of 248 highly conserved core eukaryotic genes (CEGs) in eukaryotic genomes. The CEGs were derived from six diverse model organisms: *Homo sapiens, Drosophila melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Parra et al. 2007). For comparison purposes, the *T. pseudonana* genome was downloaded from the Joint Genome Institute and analyzed with CEGMA. *T. pseudonana* is a eukaryotic oceanic alga, for which a complete genome is available. Additionally, the *K. micrum* CCMP2283 transcriptome was downloaded from the CAMERA Data Distribution Center and analyzed with CEGMA. *Karlodinium* dinoflagellates are close phylogenetic relatives of *K. brevis*, based on rRNA analyses (Fensome et al. 1999).

Assessing gene completeness with TRAPID

Full-length, quasi full-length, and partial protein coding regions were predicted in the Wilson, SP1, and SP3 Oases reference transcriptomes and the SP1 ABySS merged assembly. All the transcriptomes were compared against sequences in the OrthoMCLDB

5.0 alveolate clade database, while the Oases reference transcriptomes were also compared against sequences in the PLAZA 2.5 green plants clade database and the OrthoMCLDB 5.0 *T. pseudonana* CCMP1335 species database. The similarity search considered hits that yielded an E value $<1.00x10^{-5}$ to be significant, and transcripts were annotated according to best hit values. Transcripts that hit one or more sequences in the TRAPID databases were qualified as "full-length," "quasi full-length," or "partial" based on the ORF length. ORFs that were more than two deviations shorter than the average ORF length of their assigned gene family (excluding the 10% longest and shortest sequences within the family) are "partial." ORFs that are longer than the mean minus two standard deviations are "full length" if they also contain a start and stop codon and "quasi full-length" if they lack a stop and/or start codon (Van Bel et al. 2013).

Predicting unique assemblies and SNP locations in the transcriptomes

To identify similarities and possible differences among clones, the transcriptomes were searched for transcripts that are present in just one or two out of the three clones. Unique assemblies may be caused by differences in transcript coverage or transcriptome assembly artifacts. First, each Velvet-Oases assembled *K. brevis* transcriptome was converted to a searchable BLAST nucleotide database. All transcriptomes were searched against each other with BLASTn. Only hits with E values $\leq 1.00x10^{-50}$ were considered as significant matches. Next, Wilson, SP1, and SP3 paired-end reads were aligned to each transcriptome with CLC Genomics Workbench 6.5.

During mapping, 85% similarity fraction and 90% length fraction cutoffs were enforced, as well as conservative mismatch (3), insertion (3), and deletion (3) costs.

Reads from Wilson and SP3 were mapped to the SP1 transcriptome and analyzed for SNPs with the CLC quality-based variant detection function. The following quality filtering criteria were enforced: neighborhood radius = 5, maximum gap and mismatch count = 2, minimum neighborhood quality = 15, and minimum central quality = 20. Variants also had to be present in both forward and reverse reads. Additionally, non-specific matches and broken pairs were ignored, and the program enforced a 20-fold minimum coverage threshold and 90% minimum variant frequency. The analysis was also run with less conservative but more inclusive 10-fold coverage threshold.

Whole-transcriptome annotation and targeted gene discovery

The longest putative ORF was identified in each transcript and translated to amino acids (aa) via the longorf.pl Bioperl script (Kortschak 2002). Using BLASTp, peptides were compared to the NCBI non-redundant protein database, which was downloaded from the NCBI FTP site on September 20, 2013. Significant (E value ≤ $1.00x10^{-6}$) hits were uploaded to Blast2GO and annotated with possible gene ontology (GO) terms using a 35% minimum similarity cutoff (Conesa et al. 2005).

In order to identify novel *K. brevis* PKS mRNA sequences, BLASTn was used to search the Wilson, SP1, and SP3 transcriptomes for the four novel *K. brevis* mRNA sequences identified by Monroe and Van Dolah (Monroe and Van Dolah 2008). Hits were aligned against each other with the Clustal Omega alignment program (Sievers et

al. 2011) to investigate sequence similarity among the clones and original PKS sequences.

To identify possible voltage-gated cation channels, VATPase subunits, and aquaporin transcripts, the transcriptomes were compared to annotated voltage-gated Na^+ channel alpha subunit and aquaporin sequences using BLASTx.

During the targeted search for these particular genes, loci with significant (E value $\leq 1.00x10^{-6}$) matches to one or more proteins from the databases were translated into ORFs. Any locus containing an ORF <300 bp long was discarded.

Probable transmembrane domains within the complete ion channel and aquaporin ORFs were identified with the Center for Biological Sequence Analysis TMHMM Server, version 2.0, which uses a hidden Markov model approach to predict transmembrane, intracellular, and extracellular regions in proteins (Krogh et al. 2001). Since cation channels and aquaporins each contain domains with six transmembrane regions, sequences containing at least six predicted transmembrane helices were analyzed with two additional tools. First, they were compared to the NCBI nr database with BLASTx to identify homologous sequences from a range of eukaryotes. Next, conserved protein domains from the NCBI Conserved Domain Database (Marchler-Bauer et al. 2011) were identified with CD-search (Marchler-Bauer and Bryant 2004). In particular, ion channel transcripts were expected to contain domains belonging to the Pfam ion transport protein family (Pfam00520), which includes Na^+, K^+, and Ca^{2+} channels. Aquaporin transcripts were expected to contain MIP family conserved

domains. VATPases were searched for V-ATPase domains, including Walker motifs, which are conserved among VATPase A subunits (Forgac 1989).

Potential ion channel transcripts were also searched for the voltage-gated Na^+ channel S4 voltage-sensing motif. This motif, a repeated triad containing one positively charged and two hydrophobic aa, is highly conserved and helps regulate conformational changes that occur during channel activation and inactivation (Catterall 2000).

Availability of supporting data

The short read data supporting the results of this article are available in the NCBI SRA repository as of March 2014 under BioProject PRJNA214017, accession IDs SRX363776, SRX363775, and SRX361898. Short read data is also available at the publically accessible camera repository (project IDs MMETSP0573, MMETSP0574, MMETSP0648, MMETSP0649, MMETSP0527 and MMETSP0528). The reference transcriptomes can be accessed through LabArchive (DOI 10.6070/H44F1NPC, 10.6070/H40P0X0H, 10.6070/H4VX0DH6).

Additional supporting data files are downloadable from BMC Genomics as "electronic supplementary material" associated with the original article. These additional files include the following text documents:

Additional file 1: CEGMA CEG prediction results. This document contains the CEGMA output describing the CEG analysis in the Wilson, SP1, and SP3 transcriptomes.

Additional file 2: SP1 transcriptome cellular component GO annotation. This document contains a full list of the Blast2GO GO cellular component terms assigned to SP1 transcripts. Data are arranged into six tab-separated columns: LevelGO, Term (Acc), Term (Name), #Seq, Score, Parents (Acc), Parents (Name). The Term (Name) and Term (Acc) columns contain the cellular component name and gene ontology ID number, respectively. #Seq lists the number of transcripts that were assigned the cellular component term. LevelGO, Parents (Acc), and Parents (Name) are all related to the hierarchal arrangement of gene ontology terms, where parents on lower levels branch into more specific, higher-level child terms. LevelGO therefore describes the specificity of the cellular component, where higher values are more specific, and the name and ID numbers of all its parents are listed in the Parents (Name) and Parents (Acc) columns.

Additional file 3: SP1 transcriptome molecular function GO annotation. This document contains a full list of the Blast2GO GO molecular function terms assigned to SP1 transcripts. Data are arranged into six tab-separated columns: LevelGO, Term (Acc), Term (Name), #Seq, Score, Parents (Acc), Parents (Name). The Term (Name) and Term (Acc) columns contain the molecular function name and gene ontology ID number, respectively. #Seq lists the number of transcripts that were assigned the cellular component term. LevelGO, Parents (Acc), and Parents (Name) are all related to the hierarchal arrangement of gene ontology terms, where parents on lower levels branch into more specific, higher-level child terms. LevelGO therefore describes the specificity

of the molecular function, where higher values are more specific, and the name and ID numbers of all its parents are listed in the Parents (Name) and Parents (Acc) columns.

Additional file 4: Wilson transcriptome cellular component GO annotation. This document contains a full list of the Blast2GO GO cellular component terms assigned to Wilson transcripts. Data is arranged as described in Additional file 1.

Additional file 5: Wilson transcriptome molecular function GO annotation. This document contains a full list of the Blast2GO GO cellular component terms assigned to Wilson transcripts. Data is arranged as described in Additional file 2.

Additional file 6: SP3 transcriptome cellular component GO annotation. This document contains a full list of the Blast2GO GO cellular component terms assigned to SP3 transcripts. Data is arranged as described in Additional file 1.

Additional file 7: SP3 transcriptome molecular function GO annotation. This document contains a full list of the Blast2GO GO molecular function terms assigned to SP3 transcripts. Data is arranged as described in Additional file 1.

CHAPTER III

COMPARATIVE TRANSCRIPTOMIC ANALYSIS OF THREE TOXIN-PRODUCING *KARENIA* SPECIES*

Synopsis

Transcriptomic data for three *Karenia* species *(K. brevis, K. papilionacea, K. mikimotoi)* were compared to identify potential *Karenia* orthologs and investigate putative peptides involved in brevetoxin biosynthesis. Recent results have shown that *K. papilionacea*, like *K. brevis*, produces brevetoxin (PbTx-2). In contrast, *K. mikimotoi* does not make brevetoxin but instead produces gymnocin, another type of ladder-frame polyether. Reference transcriptomes for each species were assembled using high-throughput sequencing technology and the *de novo* assemblers Velvet-Oases and Trinity. Orthologous putative proteins were identified among *Karenia* transcriptomes using the reciprocal BLAST method and annotated with the NCBI non-redundant protein database and InterProScan.

We identified twenty-one type I-like putative polyketide synthases and one putative epoxide hydrolase-like peptide that were expressed in *K. brevis* and *K. papilionacea*, but not *K. mikimotoi*. These enzymes represent potential steps in the brevetoxin synthesis pathway. Additionally, a database of 3,495 "apparently unique" *K. brevis* and *K. papilionacea* orthologous genes was created by querying the

*Reprinted with permission from Comparative transcriptomic analysis of three toxin-producing *Karenia* species, by Darcie E. Ryan and Lisa Campbell. 2015. *Marine and Fresh-Water Harmful Algae*:229.

transcriptomes of twenty phytoplankton species. The unique orthologs provide valuable insight into the biology of brevetoxin-producing dinoflagellates.

Introduction

Karenia brevis, a bloom-forming dinoflagellate, is among the most prominent harmful algae species in the Gulf of Mexico. *K. brevis* cells produce ladder-frame polyether polyketide compounds called brevetoxin (PbTx) (Lin et al. 1981; Shimizu et al. 1986). PbTx-1, PbTx-2, and their derivatives bind to neurotoxin receptor site 5 in mammalian voltage-gated sodium channels, thereby inhibiting channel deactivation (Baden 1989; Dechraoui et al. 1999; Huang et al. 1984). *K. brevis* blooms have caused neurotoxic shellfish poisoning incidents, fish kills, and marine animal deaths along the Gulf coast (Landsberg 2002). Despite the human health, environmental, and economic risks associated with brevetoxin, their biological function in *K. brevis* is poorly characterized, and the genes associated with brevetoxin production are currently unknown. It is hypothesized that standard polyketide synthase (PKS) acyl transferase (AT), ketosynthetase (KS), β-keto-reductase (KR), dehydratase (DH), enoylreductase (ER) and acyl carrier protein (ACP) catalytic domains participate in brevetoxin synthesis (Monroe and Van Dolah 2008; Shimizu et al. 1986). An limonene epoxide hydrolase-like enzyme may participate in polyether ring formation, much like the monesin model (Gallimore 2009; Gallimore and Spencer 2006). However, the brevetoxin biosynthetic mechanism is still under debate, particularly since recent radiolabelling work suggests

that an oxidative reaction produces PbTx-1 and PbTx-2 from alcohol intermediates (Calabro et al. 2014).

PbTx-2 has been measured in *Karenia papilionacea* (Fowler et al. 2015). To investigate potential genes underlying brevetoxin production, we assembled, compared, and functionally annotated the reference transcriptomes of *K. brevis*, *K. papilionacea*, and *Karenia mikimotoi* clones using high-throughput sequencing technology. A close phylogenetic relative to *K. brevis* and *K. papilionacea* (Haywood et al. 2004), *K. mikimotoi* does not produce brevetoxin, and was therefore an ideal control species during this experiment. By identifying apparently unique orthologs expressed by *K. brevis* and *K. papilionacea*, we aimed to elucidate the unique biology of brevetoxin-producing dinoflagellates. In particular, we searched for unique PKS and epoxide hydrolase sequences, because of their potential role in PbTx-2 biosynthesis.

Material and methods

K. brevis Wilson, *K. papilionacea* CAWD91, and *K. mikimotoi* C22 cultures were maintained in L1 medium (Guillard and Hargraves 1993) at salinity 35 *(K. brevis* and *K. papilionacea*) or salinity 30 (*K. mikimotoi*). Triplicate 150-mL cultures of each species were cultured on a 12:12 hour light:dark cycle at 25 °C (*K. brevis*) or 20 °C (*K. papilionacea* and *K. mikimotoi*).

Table III-1: MMETSP CAMERA data used during this project. The MMETSP sample IDs of each transcriptome are listed in parentheses.

Species	**Species**
Alexandrium fundyense CCMP1719 (0196, 0347)	*Isochrysis galbana* CCMP1323 (0944, 0943, 0595)
Amphidinium carterae CCMP1314 (0258, 0398, 0259)	*Karlodinium micrum* CCMP2283 (1016, 1015, 1017)
Aureococcus anophagefferens CCMP1850 (0916, 0914, 0917, 0915)	*Lingulodinium polyedra* CCMP1738 (1034, 1032, 1035, 1033)
Ceratium fusus PA161109 (1075, 1074)	*Oxyrrhis marina* LB1974 (1426, 1424, 1425)
Chaetoceros neogracile CCMP1317 (0754, 0752, 0751, 0753)	*Perkinsus marinus* ATCC50439 (0922)
Crypthecodinium cohnii Seligo (0325, 0326, 0324, 0323)	*Prorocentrum minimum* CCMP1329 (0053, 0055, 0057, 0056)
Ditylum brightwellii GSO104 (1010, 1013, 1012)	*Pseudo-nitzschia australis* 10249_10_AB (0139, 0140, 0141, 0142)
Dunaliella tertiolecta CCMP1320 (1126, 1128, 1127)	*Scrippsiella hangoei* SHTV5 (0361, 0359, 0360)
Emiliania huxleyi CCMP370 (1155, 1154, 1156, 1157)	*Symbiodinium kawagutii* CCMP2468 (0132)
Fragilariopsis kerguelensis L2_C3 (0906, 0909, 0907, 0908)	*Thalassiosira oceanica* CCMP1005 (0971, 0972, 0970, 0973)

During the late exponential growth phase, cells in each bottle were pelleted via centrifugation. RNA was extracted from the pellets with the Qiagen RNEasy Mini Kit (Qiagen Inc., Valencia, CA) in accordance with kit protocol and stored at -80 °C until sequencing. The sample with the highest RNA concentration, as determined by NanoDrop Spectrophotometer, was sent on dry ice overnight to the Michigan State University Research Technology Support Facility (RTSF). RTSF prepared sequencing libraries with the Illumina Stranded mRNA Library Prep Kit LT, and 150 bp short reads were sequenced with the llumina HiSeq 2500 Rapid Run flow cell (v1). Base calling was

performed using Illumina Real Time Analysis software v 1.17.21.3. Short reads were trimmed for quality and length with CLC Genomics Workbench v 6.5 (CLC Bio, Aarhus, Denmark). A Phred quality threshold of 0.05 and minimum length threshold of 100 bp were enforced. Trimmed short reads were processed by the Trinity (default parameters) (Simpson et al. 2009) and Velvet-Oases (k-mer length 45) (Schulz et al. 2012; Zerbino and Birney 2008) *de novo* transcriptome assemblers. Trinity and Velvet-Oases assemblies, including all predicted isoforms, were combined into complete reference transcriptomes for *K. brevis*, *K. mikimotoi*, and *K. papilionacea*. The longest potential open reading frame (ORF) in each transcript was extracted and converted to amino acids with longorf.pl (Kortschak 2002), thus producing peptide databases. Redundant peptides (≥ 99% similar) were trimmed with CD-HIT v 4.5.4 (Fu et al. 2002).

To identify orthologs among *Karenia* species, the peptide databases were compared with reciprocal BLASTp (Altschul et al. 1997), using a maximum E value of $1.00x10^{-20}$, according to protocol provided by the Harvard FAS Center for Systems Biology. Orthologs unique to *K. brevis* and *K. papilionacea* were annotated with a BLASTp search against the NCBI non-redundant database (maximum E value of $1.00x10^{-20}$) and the complete application suite included in InterProScan 5 (Mitchell et al. 2014). Orthologs in *K. brevis* and *K. papilionacea* were further compared to 20 phytoplankton reference transcriptomes from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Table III-1) using reciprocal BLAST, as above.

Results and discussion

After removing redundancies with CD-HIT, the *K. brevis*, *K. mikimotoi*, and *K. papilionacea* peptide databases contained 147,200, 179,645, and 180,963 sequences, respectively. Approximately 80% of the putative peptides were >100 aa long (Figure III-1A), from continuous ORFs >300 bases. Short ORFs are more likely to occur randomly, to be incorrectly annotated during a protein BLAST search, or yield no statistically significant annotation results (Linial 2003). But predicted ORFs were not removed from the datasets based on length, to best support the goal of novel protein identification. After CD-HIT concatenation, more than 70% of the nonredundant peptide sequences were from the Oases assembler, thus suggesting that Velvet-Oases created more complete ORFs than Trinity.

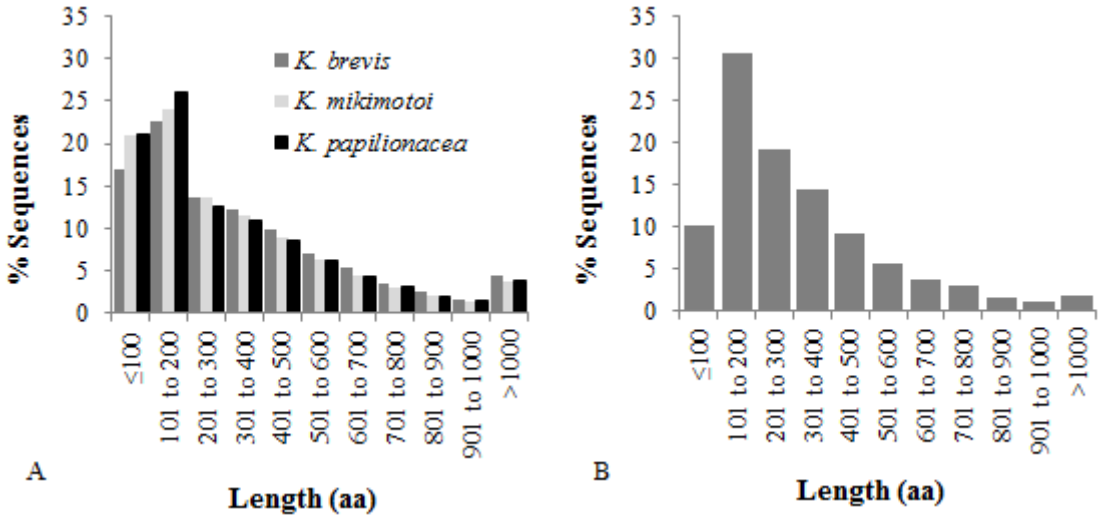

Figure III-1: Length distribution of (A) the complete *K. brevis*, *K. mikimotoi*, and *K. papilionacea* putative peptide databases and (B) the 3,495 apparently unique *K. brevis* and *K. papilionacea* peptides. Lengths in (B) are graphed based on the *K. brevis* ortholog data.

The reciprocal BLASTp search identified 70,032 sequences in the *K. brevis* transcriptome with potential orthologs expressed by *K. mikimotoi* and/or *K. papilionacea* (Figure III-2). Of these, 6,561 orthologs were expressed by *K. brevis* and *K. papilionacea*, but not *K. mikimotoi*. After querying the twenty MMETSP transcriptomes, the pool of "apparently unique" *K. brevis* and *K. papilionacea* orthologs decreased to 3,495. Over 90% of the "apparently unique" peptide sequences were >100 aa long (Figure III-1B). Only 8.24% of the "apparently unique" peptides significantly matched one or more nr sequences with an E value $\leq 1.0x10^{-20}$. This low annotation rate is expected from a database of *K. brevis* proteins with no orthologs in close phylogenetic relative *K. mikimotoi* or any of the MMETSP representative species. In contrast, InterProScan successfully assigned a protein family, repeat, domain, and/or site to 64.12% of the apparently unique putative proteins, suggesting that 1,953 unique sequences with no nr annotation may nevertheless contain short conserved protein motifs.

K. brevis expressed 21 putative PKSs in common with *K. papilionacea*, but not *K. mikimotoi*, orthologs. Of these, four had a predicted AT domain, one had a predicted KS domain, three had a predicted KR domain, one had a predicted DH domain, 10 had a predicted ER domain, and five had a predicted ACP domain.

Only the KS-containing ortholog was an "apparently unique" sequence, since the other PKSs had at least one orthologous match in the group of MMETSP phytoplankton transcriptomes. The KS domain catalyzes carbon bond formation (Claisen condensation) in polyketide skeletons and is highly conserved among eukaryotes (Keatinge-Clay

2012). We identified the KS cysteine (TACSSS) and histidine (EAHGTG and KSNIGHT) motifs (Keatinge-Clay 2012) in the apparently unique *K. brevis* and *K. papilionacea* PKS (Figure III-3).

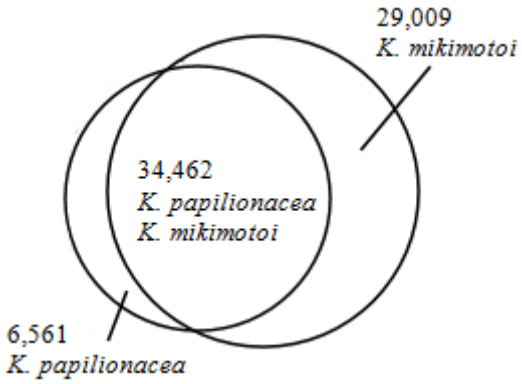

Figure III-2: Number of peptides from the *K. brevis* reference transcriptome with probable orthologs in one, both, or neither of the *K. mikimotoi* and *K. papilionacea* transcriptomes. Orthology was predicted with a reciprocal BLASTp search.

One "apparently unique" epoxide hydrolase-like sequence was expressed by *K. brevis* and *K. papilionacea*. With significant BLAST similarity to bacterial limonene epoxide hydrolases in the nr database, it was 394 aa long and 76.9% identical between the two *Karenia* species. InterProScan identified a conserved N-terminus epoxide hydrolase protein motif (pfam ID 06441) from amino acid 19 to 131.

Based on comparative transcriptomics, the *K. brevis* total transcriptome contains more sequences that are orthologous to genes in the *K. mikimotoi* transcriptome than the *K. papilionacea* transcriptome (Figure III-2). This result is not unexpected, since *K. mikimotoi* is the closer phylogenetic relative to *K. brevis*, based on rDNA sequences

(Haywood et al. 2004). Nevertheless, this study identified >3000 predicted orthologs coding putative peptides >100 aa long that were "apparently unique" to *K. brevis* and *K. papilionacea*, based on reciprocal BLAST searches against 21 phytoplankton species, including *K. mikimotoi*. Of particular interest are the 21 PKS sequences that were expressed by *K. brevis* and *K. papilionacea*, but not *K. mikimotoi*, including the KS domain-containing peptide without an identified ortholog among the MMETSP transcriptomes. Most of the PKS transcripts had a single catalytic domain, similar to eight type I-like PKS sequences that have been identified previously in *K. brevis* (Monroe and Van Dolah 2008). The apparently unique putative limonene epoxide hydrolase-like peptide is another intriguing target for future research as a step in brevetoxin synthesis. The novel genes identified in this comparative transcriptomic study of brevetoxin-producing dinoflagellates has produced a wealth of genes for further study.

```
K. brevis           FHCDTACSSTNVT
K. papilionacea     QHIDTACSSSNVA
cysteine            ----TACSSS---

K. brevis           TTCELHGTGTALG
K. papilionacea     TTTELHGTGTALG
histidine1          ---EAHGTG----

K. brevis           SAGKSNSNHNELC
K. papilionacea     SSGKSNSNHQELA
histidine2          ---KSNIGHT---
```

Figure III-3: Cysteine and histidine catalytic regions in the *K. brevis* and *K. papilionacea* "apparently unique" KS domain-containing PKS. *Karenia* sequences are aligned to each other and the highly conserved concensus motif.

CHAPTER IV

IDENTIFYING ORTHOLOGOUS GENES IN TOXIN-PRODUCING *KARENIA* SPECIES

Synopsis

The reference transcriptomes of the dinoflagellates *Karenia brevis*, *Karenia papilionacea*, and *Karenia mikimotoi* were assembled and analyzed to identify orthologous genes, including potential steps in the brevetoxin synthesis pathway. Predicted orthologs were annotated for putative function and assigned gene ontology (GO) terms. The transcriptomes were also compared to a field bloom metatranscriptome from a *K. brevis* bloom in the Gulf of Mexico. *K. papilionacea* and *K. brevis* each synthesize brevetoxin, while *K. mikimotoi* produces gymnocin, another ladder-frame polyether polyketide compound. Genes that are unique to *K. papilionacea* and *K. brevis*, but not the closely related *K. mikimotoi*, represent intriguing targets for toxin biosynthesis inquiries, particularly if they are expressed during *Karenia* blooms. Based on the annotation results, each transcriptome contained about 100 putative type I-like polyketide synthase (PKS) transcripts with single catalytic domains. In fact, the 4799 orthologs that were "unique" to *K. brevis* and *K. papilionacea* included five PKSs with predicted acyltransferase or ketoacyl synthase functional regions. In addition, a novel transcript with homology to multimodular type I PKSs and hybrid nonribosomal peptide synthetase PKSs was identified in the *K. brevis* transcriptome and field bloom metatranscriptome. The diversity of PKS transcripts expressed by *K. brevis*, *K.*

papilionacea, and *K. mikimotoi* suggest that toxin-producing dinoflagellates use a combination of type I and type I-like activity to synthesize polyketides and/or other secondary metabolites.

Introduction

Karenia brevis is a red tide dinoflagellate species that blooms almost annually in the Gulf of Mexico (Steidinger et al. 1998). *K. brevis* cells produce the brevetoxins PbTx-1 and PbTx-2. Brevetoxins are ladder-frame polyketides with trans-fused polycyclic ether rings (Rein and Borrone 1999) that bind to neurotoxin receptor site 5 in voltage-gated sodium channels and impede channel inactivation (Baden 1989; Dechraoui et al. 1999; Huang et al. 1984). Brevetoxin may cause neurotoxic shellfish poisoning and fish kills during *K. brevis* blooms (Landsberg 2002). However, the function of brevetoxin in *K. brevis* is unknown, and the brevetoxin synthesis pathway has not been characterized.

Previously, the transcriptomes of three *K. brevis* strains (Wilson, SP3, and SP1) with different PbTx-1 and PbTx-2 profiles were assembled and compared (Ryan et al. 2014). Although ~100 different putative *K. brevis* polyketide synthase (PKS) genes were identified in the transcriptomes, no differences in gene expression was measured among Wilson, SP3, and SP1. This result suggests that brevetoxin production is subject to post-transcriptional regulation (Ryan et al. 2014) and complements microarray studies that identified abundant genes related to RNA post-transcriptional regulation (Van Dolah et al. 2009).

To investigate the genes involved with brevetoxin synthesis, we assembled, compared, and functionally annotated the reference transcriptomes of *K. brevis*, *Karenia papilionacea*, and *Karenia mikimotoi* clones using high-throughput sequencing technology. PbTx-2 has been measured in *K. papilionacea*, so *K. brevis* and *K. papilionacea* might share brevetoxin synthesis orthologs (Fowler et al. 2015). A close phylogenetic relative to *K. brevis* (Haywood et al. 2004), *K. mikimotoi* does not produce brevetoxin, and is therefore an ideal control species for this experiment. Orthologous genes were identified among the *Karenia* species using the reciprocal BLAST best-hit method. Orthologs expressed by *K. brevis* and *K. papilionacea*, but not *K. mikimotoi*, were considered "unique" and analyzed for potential PbTx synthesis function. Of particular interest were "unique" PKS sequences with acyl transferase (AT), ketosynthetase (KS), β-keto-reductase (KR), dehydratase (DH), and/or acyl carrier protein (ACP) catalytic domains, because of their potential role in polyketide formation. "Unique" sequences were compared to a field bloom metatranscriptome to confirm that *K. brevis* in the natural environment also express the transcripts.

The "unique" transcripts were further annotated for gene ontology and potential function using a variety of protein databases, and novel transcripts with open reading frames (ORFs) >300 bp were identified as intriguing targets for future brevetoxin production research.

Methods

Culturing and RNA sequencing

Triplicate 150-mL *K. brevis* Wilson, *K. papilionacea* C91, and *K. mikimotoi* C22 cultures were grown in L1 medium at salinity 35 *(K. brevis* and *K. papilionacea*) or 30 (*K. mikimotoi*) and temperature 25 °C (*K. brevis*) or 20 °C (*K. papilionacea* and *K. mikimotoi*). Cultures were incubated under a 12:12 hour light:dark cycle with ~60 μmol photons m^{-2} during each light period. Cell counts were taken every other day to monitor growth rate. Once cultures had reached the late exponential growth period, 50 mL from each replicate was pelleted through centrifugation (800 × g, 10 min) and stabilized with RNALater (Qiagen). Using the Qiagen RNEasy Mini Kit (Qiagen Inc., Valencia, CA), RNA was extracted from each pellet in 40μL of nuclease-free water, in accordance with manufacturer protocol. The RNA samples were stored at −80°C, while small (2 μL) aliquots from each sample were analyzed with a NanoDrop Spectrophotometer and Agilent 2100 Bioanalyzer to estimate RNA concentration and purity. For each *Karenia* species, the sample with the greatest purity was shipped overnight on dry ice to the Michigan State University Research Technology Support Facility (RTSF) for high-throughput RNA sequencing.

The RTSF prepared sequencing libraries with the Illumina Stranded mRNA Library Prep Kit LT. For each clone, 150-bp, paired-end short reads were sequenced with the llumina HiSeq 2500 Rapid Run flow cell (v1). Base calling was performed using Illumina Real Time Analysis software v 1.17.21.3.

Transcriptome assembly

The short reads were trimmed for quality and length with CLC Genomics Workbench v 6.5 (CLC Bio, Aarhus, Denmark), using a 100-bp minimum length threshold and a 0.005 Phred quality threshold. Trimmed reads were exported from CLC in fasta format and processed with Velvet-Oases, an assembly program that uses a de Bruijn graph algorithm to assemble transcriptomes without a reference genome (*de novo*) (Schulz et al. 2012; Zerbino and Birney 2008). During assembly, automatic expected coverage and minimum coverage cutoff values, a 50% edge fraction cutoff, and a 300-bp minimum transcript length were enforced. For each species, single k-mer assemblies (k-mer lengths 61, 65, 69, and 73 bp) were merged into a non-redundant consensus transcriptome assembly with the K-value 45.

To remove redundancies, each reference transcriptome was analyzed with CD-HIT EST (Fu et al. 2012). Redundant (>98% similarity) clusters were collapsed into the longest representative transcript sequence. After CD-Hit trimming, the longest putative open reading frame (ORF) was identified in each transcript with longorf.pl, a script written by Dan Kortschak (Kortschak 2002). ORFs >300 bp were extracted and converted to amino acids. The CLC Pfam domain search function (Pfam database v 27) (Finn et al. 2013), identified conserved domains in the putative protein sequences.

Transcriptome analysis

The entire *K. brevis*, *K. mikimotoi*, and *K. papilionacea* transcriptomes were analyzed with TRAPID, an online package with multiple *de novo* transcriptome analysis

tools, including ORF completeness and ontology prediction (Van Bel et al. 2013). Based on homology, TRAPID can assign transcripts to representative gene families and predict whether the transcript protein-coding region is "complete," "quasi-complete," or "partial," where partial transcripts are >2 standard deviations shorter than the mean gene family ORF length. This function makes TRAPID a powerful estimator of transcriptome assembly completeness. Each *Karenia* transcriptome was compared to genes in the PLAZA 2.5 database. PLAZA 2.5 contains 25 plant genomes, including 5 that belong to chlorophyte algae, with ~780,000 total protein-coding genes in 32,294 gene families (Van Bel et al. 2013). Gene family, function, and ORF completeness were inferred for all *Karenia* transcripts that significantly (E value $<1.0x10^{-5}$) matched a PLAZA gene sequence. TRAPID also measured fundamental transcriptome characteristics, including mean transcript length, mean ORF length, and potential frameshift occurrences.

Karenia *ortholog identification*

Orthologs were identified among the *Karenia* transcriptomes using the reciprocal BLAST (Altschul et al. 1997) best-hit method. The search enforced a $1.0x10^{-6}$ E value significance threshold. *K. brevis* transcripts with orthologs in *K. papilionacea* but not *K. mikimotoi* were considered "unique" and extracted. The *K. brevis* transcriptome, including "unique" sequences, was further compared to 34 phytoplankton reference transcriptomes from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Table IV-1) using reciprocal BLAST, as described above (Keeling et al. 2014). The MMETSP transcriptomes were assembled by the National Center for

Genome Resources with ABySS (Simpson et al. 2009). "Unique" proteins from *K. brevis* were aligned to their *K. papilionacea* orthologs with Clustal Omega (Sievers et al. 2011) to visualize homologous regions.

"Unique" ortholog analysis

To determine GO enrichment in the "unique" group, InterProScan was run on the complete *K. brevis* transcriptome, including the "unique" *K. brevis* transcripts. InterPro assigns potential protein domains and gene ontology terms to nucleotide sequences by drawing from the Pfam (Finn et al. 2013), PRINTS (Attwood et al. 2012), PROSITE (Sigrist et al. 2012), ProDom (Corpet et al. 1998), CATH-Gene3D (Lees et al. 2014), HAMAP (Pedruzzi et al. 2013), PANTHER (Mi et al. 2013), PIRSF (Nikolskaya et al. 2006), SMART (Letunic et al. 2012), SUPERFAMILY (de Lima Morais et al. 2010), and TIGRFAMs databases (Mitchell et al. 2014). Over- or under-represented GO terms in the "unique" transcript group were predicted using Fisher's Exact Test in Blast2GO, with the default false discovery rate (FDR) term filter mode (Conesa et al. 2005). Significant results were collapsed to the most specific GO term.

The "unique" *K. brevis* transcripts were also annotated with a BLASTx search against the NCBI nr protein database (accessed December 15, 2015). Hits with an E value $\leq 1.0 \times 10^{-6}$ were assigned gene ontology terms using the Blast2GO 3.0 mapping and annotation functions with a 30% minimum similarity cutoff (Conesa et al. 2005). Annotation validation removed redundant GO terms for each transcript, where multiple terms belonging to the same GO branch were collapsed into the most specific, highest-

level child term. KEGG pathways for annotated "unique" transcripts were generated using Blast2GO 2.8 (Conesa et al. 2005).

Putative Karenia *PKS identification*

Transcripts that were assigned any PKS-related term during the CLC Pfam search, including AT, KS, KR, DH, and/or ACP domains, were extracted and compared to the nr protein database with BLASTx. Transcripts that were significantly (E value $1.0x10^{-20}$) similar to nr PKSs were considered putative *Karenia* PKS genes.

Field bloom metatranscriptome comparison

The *K. brevis*, *K. papilionacea*, and *K. mikimotoi* transcriptomes were compared to a field bloom metatranscriptome with BLAST, using the E value cutoff $1.0x10^{-100}$. The metatranscriptome samples were collected during a *K. brevis* bloom (2015) in the Gulf of Mexico. The metatranscriptome was assembled by Dr. Darren W. Henrichs.

Results

Transcriptome assembly and analysis

After length and quality trimming, 86,626,190 (*K. brevis*), 100,920,977 (*K. papilionacea*), and 93,255,006 (*K. mikimotoi*) paired-end reads were processed by Velvet-Oases. The three final reference transcriptomes had similar locus and transcript numbers (Table IV-2). Likewise, for *K. brevis, K. papilionacea, and K. mikimotoi,* the

mean transcript length was 1353, 1406, and 1299 bp, respectively, and the mean predicted ORF length for each transcriptome was ~1000 bp (Table IV-2).

During the TRAPID analysis, 21.4% (*K. brevis*), 21.6% (*K. papilionacea*), or 19.2% (*K. mikimotoi*) of the transcriptomes were assigned a predicted gene family and analyzed for ORF completeness. There were 6.4 to 7.8 times more complete or quasi-complete ORFs than partial ORFs in the *Karenia* transcriptomes (Table IV-2).

The PLAZA species most similar to all three *Karenia* species were the five unicellular green algae species, *Micromonas* sp., *Chlamydomonas reinhardtii*, *Volvox carteri*, *Ostreococcus lucimarinus*, and *Ostreococcus tauri*. HOM000003, a pentatricopeptide repeat sequence, was the most common gene family assigned to the reference transcriptomes (Table IV-3). Previous microarray analyses have shown that pentatricopeptide repeats, involved in RNA processing, are abundant in *K. brevis* (Morey et al. 2011). When possible, transcript function was predicted based on gene family matches and associated gene ontology terms. Between 14.5 and 16 percent of the reference transcriptomes were given GO terms (Table IV-3).

Ortholog prediction and analysis

The reciprocal BLAST search identified 28,073 total *K. brevis* transcripts, 30% of the transcriptome, with at least one ortholog in the *K. mikimotoi* and *K. papilionacea* transcriptomes. Of these, 4799 orthologs (5% of the whole transcriptome) were “unique” to *K. brevis* and *K. papilionacea* (Table IV-4). Most (4,448) “unique” transcripts had no predicted orthologs in the 34 MMETSP phytoplankton transcriptomes, either. The

"unique" *K. brevis* and *K. papilionacea* orthologs encompassed 858 predicted gene families, 1,509 discrete gene ontology terms, and 1,141 different conserved protein domains in the TRAPID PLAZA database. The mean length of "unique" *K. brevis* transcripts was 1,894, with a mean ORF length of 1489 bp and 10:1 complete to partial ORF ratio. About 59% (2,811 transcripts) of the "unique" proteins matched one or more sequences in the nr protein database with an E value $<1.0x10^{-6}$, while 52% (2,489 transcripts) contained one or more significant (E value $<1.0x10^{-5}$) protein domain hits to the Pfam database.

InterPro analysis

Most (65%) of the *K. brevis* transcriptome was assigned at least one protein family, domain, repeat, and/or site by InterProScan. Of the InterPro annotated transcripts, 24,884 (27% of the whole transcriptome) had GO information. Similarly, 3492 "unique" transcripts were annotated by InterPro, and 1821 had one or more assigned GO terms. The Fisher's exact test identified fifteen GO categories that were significantly overrepresented in the "unique" compared to total transcripts (Table IV-5).

PKS discovery

The Pfam and nr BLAST searches identified 118 (*K. brevis*), 106 (*K. papilionacea*), and 110 (*K. mikimotoi*) transcripts with significant (E value $<1.0x10^{-20}$) homology to annotated nr PKS proteins and at least one Pfam-predicted KS, AT, ACP, or KR domain. The mean PKS ORF length was 2,554 bp (*K. brevis*), 2,805 bp (*K.*

papilionacea), and 2,502 bp (*K. mikimotoi*), with a range from ~300 to ~13,000 bp. Most putative *Karenia* PKS transcripts possessed a single catalytic domain. However, in *K. mikimotoi* and *K. papilionacea,* several two-domain and three-domain transcripts were identified (Table IV-6, IV-7). Specifically, *K. mikimotoi* expressed two different transcripts with KR, AT, and polyketide dehydrogenase (PS-DH) domains, one transcript with AT and KS domains, three transcripts with KR and KS domains, and two transcripts with KR, KS, and PS-DH domains (Table IV-6). *K. papilionacea* expressed one transcript with KR, KS, and PS-DH domains, two transcripts with AT and KS domains, and three transcripts with KR and KS domains (Table IV-7).

One multimodular PKS sequence was identified in the *K. brevis* transcriptome. Its ORF was 8094 bp-long and complete, based on the presence of start and stop codons. The ORF region included KR, KS (both C and N terminal regions), and AT domains (Table IV-8). The translated protein sequence was further analyzed with the NCBI CD search program with a maximum expect value threshold of $1.0x10^{-6}$. CD search identifies, characterizes, and visualizes conserved domains through a RPS-BLAST search against the NCBI conserved domain database (CDD) (Marchler-Bauer and Bryant 2004; Marchler-Bauer et al. 2011). CD search can also match the query sequence with protein groups that have similar domain architecture with the Conserved Domain Architecture Retrieval Tool (Marchler-Bauer et al. 2012). The multimodular PKS was most similar to type I polyketide synthases, multifunctional polyketide-peptide synthases, and non-ribosomal peptide synthetases, based on its domain architecture (Figure IV-1).

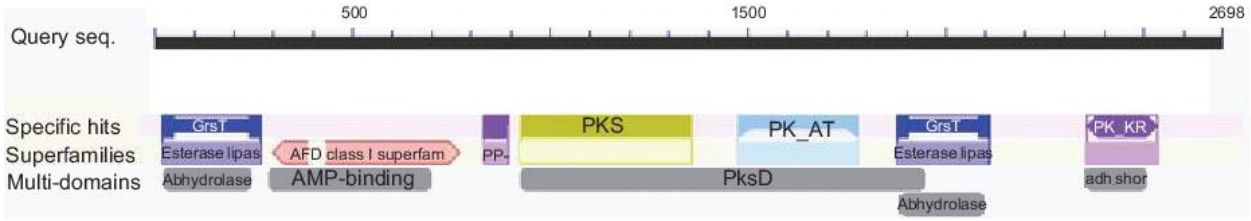

Figure IV-1: *K. brevis* multidomain PKS sequence with significant (expect value $<1.0x10^{-6}$) CD search-predicted conserved domain regions. PKS = cd00833 (polyketide synthases), PKS_AT = smart00827 (acyl transferase domain in polyketide synthase enzymes), GrsT = COG3208 (surfactin synthase thioesterase subunit), PKS_KR = smart00822 (enzymatic polyketide synthase domain that catalyses the first step in the reductive modification of the beta-carbonyl centres in the growing polyketide chain), PksD = COG3321 (acyl transferase domain in polyketide synthase enzymes), AMP-binding = pfam00501 (AMP-binding enzyme), PP- = pfam00550 (phosphopantetheine attachment site), adh_short = pfam00106 (short chain dehydrogenase), Abhydrolase_ = pfam12697 (alpha/beta hydrolase family). Figure was generated by CD search (Marchler-Bauer and Bryant 2004) .

The five putative PKS sequences in the "unique" transcript group each contained one predicted catalytic domain. Three had KS domains, and two had AT domains. In conserved protein regions, the *K. brevis* and *K. papilionacea* PKS orthologs were between 60 and 80 percent similar, based on amino acid sequence (Figure IV-2). The "unique" PKS sequences significantly (E value $<1.00x10^{-20}$) matched other PKSs in the nr database, including type I-like PKSs previously identified in other dinoflagellates, although they were not identical to any nr protein. In fact, the closest nr match, a KS-containing transcript, yielded a 43.84% identity score (Table IV-9). The complete "unique" PKSs in *K. brevis* and *K. papilionacea* also contained a novel terminal motif that has only been identified in other dinoflagellate PKS proteins (Eichholz et al. 2012).

Field bloom metatranscriptome comparison

From the BLAST search between the *K. brevis* transcriptome and field bloom metatranscriptome, there were 72,900 highly significant (E value $\leq 1.00x10^{-100}$) hits with $\geq$ 99% sequence similarity, including all the type I-like and type I PKS transcripts that were identified during this study. In addition, the complete *K. brevis* multimodular PKS ORF was also found in the field bloom metatranscriptome.

K. mikimotoi and *K. papilionacea* transcripts were also found in the field bloom metatranscriptome. The *K. mikimotoi* comparison yielded 1,388 ($\geq$ 99% similarity) highly significant hits. Of those, 501 had 100% nucleotide identity. The *K. papilionacea* comparison yielded 15 ($\geq$ 99% similarity) highly significant hits. One of the hits was 100% identical over the aligned region.

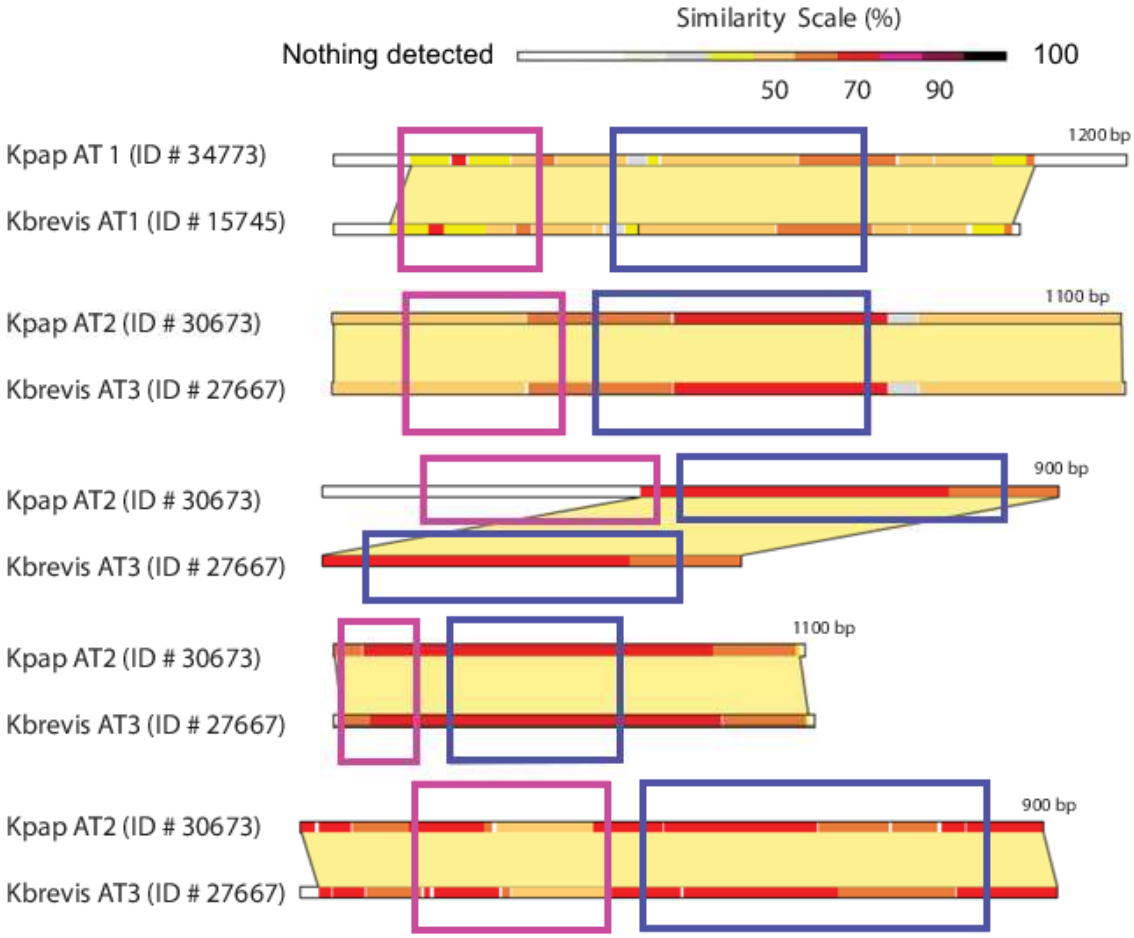

Figure IV-2: Sequence conservation between *K. brevis* and *K. papilionacea* "unique" PKS proteins. Blue boxes demarcate regions that are significantly (Evalue $<1.0x10^{-20}$) homologous to the NCBI PKS conserved domain cd00083. Pink boxes surround the TIGR04556 terminal domain that has been characterized in dinoflagellate PKS-like proteins. Sequence similarity was graphed using the LALN View program. PKSs with a predicted AT region are labelled AT1-3, and PKSs with a predicted KS region are labelled KS1-2.

Discussion

PKSs produce polyketides through the Claisen condensation elongation method, using the combined activity of KS, AT, KR, and ACP domains, with an AT+ACP loading module (Shen 2003). In sequenced unicellular algae genomes, potential type I and type II PKS genes have been identified (Shelest et al. 2015). Type I PKSs are multimodular, possessing multiple PK-assembly catalytic domains in one protein, whereas type II PKSs function through the interaction of single-domain proteins (Fischbach and Walsh 2006). Through the analysis of *K. brevis* expressed sequence tag libraries, "type I-like" PKS transcripts, each containing a single predicted catalytic domain (KS, ACP, or KR) with high homology to type I proteins, have been identified *(Monroe and Van Dolah 2008).* Subsequently assembled *K. brevis* Wilson, SP1, and SP3 transcriptomes contained dozens of full-length type I-like PKS transcripts (Ryan et al. 2014). During this study, >100 putative PKS transcripts were identified in each of the *Karenia* transcriptomes. Notably, based on gene annotation results, not all of the *Karenia* PKS transcripts encoded single-domain, type I-like PKSs. Putative multimodular PKSs were expressed by all three *Karenia* species (Tables IV-6, IV-7, IV-8). In fact, one *K. brevis* transcript with predicted beta-ketoacyl (N- and C-terminal domains), acyl transferase, and ketoreductase domains (Table IV-8) had a protein architecture similar to previously characterized type I polyketide synthases, multifunctional polyketide-peptide synthases, and non-ribosomal peptide synthetases. This multimodular PKS was also expressed with 100% ORF conservation by *K. brevis*

cells in the field bloom metatranscriptome sample, which suggests that the protein is not exclusive to the Wilson cultures.

Non-ribosomal peptide synthetases (NRPS) produce structurally diverse, specialized peptides through the multienzyme thiotemplate mechanism (Marahiel et al. 1997). The modular organization of NRPSs is similar to type I PKSs, although NRPSs perform C-N elongation with specific adenylation and condensation domains (Cane and Walsh 1999). Previously, a fosmid library constructed with *K. brevis* Wilson DNA contained a hybrid NRPS-PKS gene cluster with three NRPS condensation and adenylation modules proceeded by a type I PKS with KS, AT, and KR domains (López-Legentil et al. 2010). Because of the architecture similarity between the hybrid NRPS-PKS (Genbank ID FJ172507) and the multimodular *K. brevis* PKS identified during this study, the hybrid NRPS-PKS protein-coding region was downloaded from Genbank and aligned to the translated multimodular PKS with Clustal Omega (Sievers et al. 2011). NRPS-PKS #FJ172507 is 1,539 aa long, half the length of the multimodular PKS, and the aligned region has a 30.63% amino acid identity score. Thus, the hybrid NRPS-PKS is a different gene than the type I PKS described during this study.

Hyrbid NRPS-PKS #FJ172507 was converted to a searchable BLAST protein database and compared to the *K. brevis*, *K. papilionacea*, and *K. mikimotoi* transcriptomes using BLASTx. The field bloom metatranscriptome was also searched for #FJ172507. Although the KS, AT, and KR regions in #FJ172507 were significantly (E value $>1.0x10^{-20}$) similar to 70 putative PKSs in each transcriptome/field bloom metatranscriptome, the *K. brevis* transcriptome and field bloom metatranscriptome did

not contain the NRPS-PKS. Furthermore, no orthologs were present in the *K. papilionacea* and *K. mikimotoi* transcriptomes. In fact, the most significant match ($2.00x10^{-138}$) was the *K. brevis* multimodular type I PKS. It is possible that the cultures were not expressing the hybrid NRPS-PKS gene when RNA was collected, or that its expression level was too low for *de novo* assembly. The novel type I PKS that was assembled hints at complex, varied polyketide and/or nonribosomal peptide assembly in toxin-producing dinoflagellates.

Velvet-Oases gives each discrete transcript a locus ID. Loci may contain one transcript or multiple transcripts representing predicted isoforms. A previously assembled *K. brevis* Wilson transcriptome contained 86,580 loci, with a mean length of 1340 bp (Ryan et al. 2014). These values are similar to the loci number and mean transcript length of each *Karenia* reference transcriptome produced during this study (Table IV-2). Because the TRAPID analysis detected few partial ORFs, we are confident in the *Karenia* assembly completeness.

Based on reciprocal BLAST results, the *K. brevis* total transcriptome contains more sequences that are orthologous to genes in the *K. mikimotoi* transcriptome than the *K. papilionacea* transcriptome (Table IV-2). This result is not unexpected, since *K. mikimotoi* is the closer phylogenetic relative to *K. brevis*, based on rDNA. Nevertheless, this study identified 4,799 predicted orthologs that were expressed by *K. brevis* and *K. papilionacea* but not *K. mikimotoi*. TRAPID results show that the "unique" orthologs were, on average, longer and more complete than the overall *K. brevis* or *K. papilionacea* reference transcriptomes. It is therefore unlikely that the subset is only

"unique" due to atypically short transcripts or incomplete ORFs. Of particular interest are the 5 "unique" PKS transcripts.

All five "unique" PKS transcripts were type I-like, with either a single KS or AT domain. Based on our reciprocal BLAST search, the "unique" PKSs had no orthologs in the other phytoplankton, based on the MMETSP transcriptome database (Table IV-1). Because BLAST parameters such as sequence complexity-based filtering and alignment methods affect the number of orthologs predicted (Moreno-Hagelsieb and Latimer 2008), the conservative reciprocal BLAST best-hit approach we used might underestimate the number of orthologous transcripts, particularly as the phylogenetic distance between species increases. Similarly, it is possible that components of the PbTx-2 synthesis pathway were not included in the "unique" subset. For this experiment, high confidence in orthologs was valued over maximizing ortholog discovery.

To maximize the number of unique *Karenia* transcripts annotated with a putative function, protein domain, and/or GO term, we used several annotation methods with different protein databases. InterProScan yielded the most annotations, with 3492 of the 4799 transcripts assigned at least one result. In contrast, the nr database annotated 2,811 "unique" transcripts, Pfam annotated 2,489 "unique" transcripts, and the TRAPID PLAZA search annotated 1,552 "unique" transcripts. InterProScan compares transcripts to data from eleven databases, including Pfam, so its high annotation rate was not unexpected (Mitchell et al. 2014). InterProScan results were therefore used during the GO enrichment analysis step.

Fifteen GO terms were overrepresented in the "unique" transcript group. However, several related terms overlapped. For example the GO:0018130 (heterocycle biosynthetic process) and GO:1901362 (organic cyclic compound biosynthetic process) transcripts were identical, with one exception. Similarly, there was significant overlap among the microtubule motor activity, microtubule-based movement, microtubule binding, and kinesin complex transcripts. Although PKSs are cyclic compounds, none of the transcripts assigned a GO: 1901362 term were specifically related to polyketide synthesis.

"Unique" transcripts identified during this comparative transcriptomic study, including the unannotated, potentially novel sequences, provide the foundation for further research as indicators for brevetoxin production. Furthermore, the diversity of PKS and/or NRPS-PKS transcripts expressed by *K. brevis*, *K. papilionacea*, and *K. mikimotoi* suggest that toxin-producing dinoflagellates use a combination of type I and type I-like activity to synthesize polyketides and other secondary metabolites. These results are supported by an analysis of a field bloom metatranscriptome assembled from RNA collected during a Gulf bloom. Not only were all the type I-like and type I *K. brevis* PKS genes expressed in the environmental sample, our BLAST search results showed that *K. mikimotoi* and *K. papilionacea* cells contributed RNA to the field bloom metatranscriptome. The red tide event, although dominated by *K. brevis*, supported other toxin-producing *Karenia* species.

Table IV-1: Phylum, genus, and species of each MMETSP phytoplankton transcriptome analyzed.

Phylum	**Genus**	**Species**
Bacillariophyta	*Chaetoceros*	*debilis*
Bacillariophyta	*Ditylum*	*brightwellii*
Bacillariophyta	*Extubocellulus*	*spinifer*
Bacillariophyta	*Fragilariopsis*	*kerguelensis*
Bacillariophyta	*Nitzschia*	*punctata*
Bacillariophyta	*Proboscia*	*alata*
Bacillariophyta	*Pseudo-nitzschia*	*australis*
Bacillariophyta	*Skeletonema*	*dohrnii*
Bacillariophyta	*Thalassiosira*	*oceanic*
Cercozoa	*Lotharella*	*globosa*
Chlorophyta	*Dunaliella*	*tertiolecta*
Chlorophyta	*Picocystis*	*salinarum*
Chlorophyta	*Tetraselmis*	*striata*
Cryptophyta	*Rhodomonas*	sp.
Dinophyta	*Alexandrium*	*monilatum*
Dinophyta	*Azadinium*	*spinosum*
Dinophyta	*Tripos (formerly Ceratium)*	*fusus*
Dinophyta	*Durinskia*	*baltica*
Dinophyta	*Karlodinium*	*micrum*
Dinophyta	*Prorocentrum*	*minimum*
Haptophyta	*Chrysochromulina*	*polylepsis*
Haptophyta	*Emiliania*	*huxleyi*
Haptophyta	*Gephyrocapsa*	*oceanic*
Haptophyta	*Isochrysis*	*galbana*
Haptophyta	*Pleurochrysis*	*carterae*
Haptophyta	*Prymnesium*	*parvum*
Labyrinthista	*Aurantiochytrium*	*limacinum*
Ochrophyta	*Aureococcus*	*anophagefferens*
Ochrophyta	*Pelagococcus*	*subviridis*
Rhodophyta	*Rhodella*	*maculate*
Dinophyta	*Symbiodinium*	*kawagutii*
Dinophyta	*Oxyrrhis*	*marina*
Dinophyta	*Dinophysis*	*acuminate*
Ochrophyta	*Pteridomonas*	*danica*

Table IV-2: *K. brevis*, *K. papilionacea*, and *K. mikimotoi* transcriptome locus number, transcript number, mean transcript length, mean ORF length, and predicted full to partial ORF ratio.

Species	# Loci	# Transcripts	Mean transcript length	Mean ORF length	Full:partial ORF
K. brevis	85,697	93,226	1353.6	1075.7	7.1:1
K. papilionacea	88290	96,758	1406.5	1094.3	7.8:1
K.mikimotoi	86,408	100980	1299.4	974.2	6.4:1

Table IV-3: TRAPID results (total number of gene families assigned to one or more transcripts in the transcriptome, % transcriptome assigned at least one gene family, % transcriptome assigned at least one frameshift, total number of GO terms, and % transcriptome assigned at least one GO term) for the *K. brevis*, *K. papilionacea*, and *K. mikimotoi* reference transcriptomes. Gene families and gene ontology (GO) terms were determined with data from PLAZA 2.5.

Species	# Gene Families	% Assigned Gene Family	% with frameshift	# GO Terms	% Assigned GO Term
K. brevis	4127	21.4	3.4	3023	16
K. papilionacea	4144	21.6	3.7	3087	15.9
K.mikimotoi	3962	19.2	4.1	2941	14.5

Table IV-4: Number of transcripts from the *K. brevis* reference transcriptome with predicted orthologs in one, both, or neither of the *K. mikimotoi* and *K. papilionacea* transcriptomes.

	# orthologous transcripts in *K. brevis* transcriptome	% *K. brevis* transcriptome
K. papilionacea and *K. mikimotoi*	4637	5.4
K. papilionacea	4799	5.6
K. mikimotoi	18,637	21.7
Total	28,073	33

Table IV-5: Overrepresented GO terms in the "unique" group, as determined by a Fisher's exact test on the InterProScan-annotated *K. brevis* transcriptome. The % Unique Group and % Total Transcriptome columns were calculated by the equation (# transcripts with GO ID)/(# total GO-annotated transcripts).

GO ID	Term	FDR	P-Value	% Unique Group	% Total Transcriptome
GO:0005524	ATP binding	1.04×10^{-8}	7.23×10^{-11}	14.4	8.8
GO:0004672	protein kinase activity	3.27×10^{-5}	3.69×10^{-7}	6.8	3.8
GO:0006468	protein phosphorylation	3.74×10^{-5}	4.47×10^{-7}	6.6	3.7
GO:0003777	microtubule motor activity	2.20×10^{-4}	3.65×10^{-6}	2.5	1.1
GO:0007018	microtubule-based movement	4.10×10^{-4}	7.61×10^{-6}	2.5	1.1
GO:0005871	kinesin complex	8.53×10^{-4}	1.70×10^{-5}	1.7	0.6
GO:0008017	microtubule binding	3.72×10^{-3}	7.90×10^{-5}	1.6	0.7
GO:0090407	organophosphate biosynthetic process	6.56×10^{-3}	1.52×10^{-4}	1.4	0.6
GO:0018130	heterocycle biosynthetic process	1.37×10^{-2}	3.45×10^{-4}	3.2	1.8
GO:0044271	cellular nitrogen compound biosynthetic process	1.48×10^{-2}	$3.83\text{Ex}10^{-4}$	3.3	1.9
GO:1901362	organic cyclic compound biosynthetic process	1.87×10^{-2}	4.89×10^{-4}	3.3	1.9
GO:0006886	intracellular protein transport	2.99×10^{-2}	8.03×10^{-4}	0.9	0.3
GO:0003743	translation initiation factor activity	3.32×10^{-2}	9.58×10^{-4}	0.4	0.1
GO:0048285	organelle fission	3.32×10^{-2}	9.58×10^{-4}	0.4	0.1
GO:0019438	aromatic compound biosynthetic process	3.58×10^{-2}	1.05×10^{-3}	3.0	1.7

Table IV-6: Pfam results for the *K. mikimotoi* multidomain PKS sequences.

Transcript ID	Pfam domain	Domain description	E value
Locus_7602_Transcript_1/1	PF08659.5	KR domain	2.70×10^{-51}
	PF00698.16	Acyl transferase domain	3.20×10^{-50}
	PF14765.1	Polyketide synthase dehydratase	7.70×10^{-42}
Locus_238_Transcript_1/1	PF08659.5	KR domain	9.40×10^{-31}
	PF00698.16	Acyl transferase domain	6.80×10^{-22}
	PF14765.1	Polyketide synthase dehydratase	5.10×10^{-10}
	PF00550.20	Phosphopantetheine attachment site	1.90×10^{-04}
Locus_11177_Transcript_1/1	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	1.50×10^{-48}
	PF00698.16	Acyl transferase domain	5.10×10^{-37}
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	2.50×10^{-35}
Locus_24202_Transcript_1/1	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	2.90×10^{-55}
	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	1.30×10^{-50}
	PF08659.5	KR domain	9.80×10^{-47}
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	1.40×10^{-31}
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	2.60×10^{-29}
	PF00550.20	Phosphopantetheine attachment site	7.00×10^{-10}
	PF00550.20	Phosphopantetheine attachment site	1.40×10^{-8}
Locus_26402_Transcript_1/1	PF08659.5	KR domain	4.30×10^{-38}
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	1.30×10^{-25}
	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	1.70×10^{-6}
Locus_27638_Transcript_1/1	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	1.20×10^{-64}
	PF08659.5	KR domain	1.70×10^{-61}
	PF14765.1	Polyketide synthase dehydratase	5.20×10^{-37}
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	1.50×10^{-30}
Locus_35576_Transcript_1/1	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	8.90×10^{-61}
	PF08659.5	KR domain	1.40×10^{-42}
	PF14765.1	Polyketide synthase dehydratase	5.20×10^{-36}
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	7.00×10^{-32}

Table IV-6 continued.

Transcript ID	Pfam domain	Domain description	E value
	PF00107.21	Zinc-binding dehydrogenase	$3.10x10^{-19}$
	PF00550.20	Phosphopantetheine attachment site	$2.00x10^{-9}$
	PF08659.5	KR domain	$1.50x10^{-5}$
Locus_42174_Transcript_1/1	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$1.40x10^{-55}$
	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$3.70x10^{-54}$
	PF08659.5	KR domain	$2.80x10^{-40}$
	PF08659.5	KR domain	$2.60x10^{-35}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$4.00x10^{-35}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$5.60x10^{-27}$
	PF00550.20	Phosphopantetheine attachment site	$5.50x10^{-08}$

Table IV-7: Pfam results for the *K. papilionacea* multidomain PKS sequences.

Transcript ID	Pfam domain	Domain description	E value
Locus_6012_Transcript_1/2	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$1.6x10^{53}$
	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$4.6x10^{53}$
	PF08659.5	KR domain	$2.8x10^{41}$
	PF08659.5	KR domain	$6x10^{30}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$1.2x10^{26}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$9x10^{26}$
	PF14765.1	Polyketide synthase dehydratase	$7.6x10^{23}$
	PF03959.8	Serine hydrolase (FSH1)	$1.3x10^{17}$
	PF00975.15	Thioesterase domain	$1.1x10^{16}$
	PF13738.1	Pyridine nucleotide-disulphide oxidoreductase	$2.7x10^{12}$

Table IV-7 continued.

Transcript ID	Pfam domain	Domain description	E value
	PF00550.20	Phosphopantetheine attachment site	$1.8x10^{9}$
	PF00550.20	Phosphopantetheine attachment site	$2.6x10^{6}$
	PF00550.20	Phosphopantetheine attachment site	$4.8x10^{6}$
Locus_258_Transcript_1/1	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$1x10^{-37}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$1.2x10^{-30}$
	PF00698.16	Acyl transferase domain	$5.5x10^{-24}$
Locus_2405_Transcript_1/1	PF00698.16	Acyl transferase domain	$1.7x10^{-41}$
	PF01575.14	MaoC like domain	$3.9x10^{-23}$
	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$9.4x10^{-19}$
	PF08354.5	Domain of unknown function (DUF1729)	$1.2x10^{-16}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$1.1x10^{-13}$
	PF13561.1	x10noyl-(Acyl carrier protein) reductase	$1.5x10^{-10}$
	PF01648.15	4'-phosphopantetheinyl transferase superfamily	$2.1x10^{-8}$
	PF13452.1	N-terminal half of MaoC dehydratase	$1.9x10^{-5}$
Locus_9117_Transcript_1/1	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$1.2x10^{-51}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$2.9x10^{-31}$
	PF08659.5	KR domain	$2.2x10^{-10}$
	PF00550.20	Phosphopantetheine attachment site	$1.9x10^{-8}$
Locus_22382_Transcript_1/4	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$2.1x10^{-62}$
	PF08659.5	KR domain	$1.8x10^{-49}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$1.3x10^{-32}$
	PF00550.20	Phosphopantetheine attachment site	$4.4x10^{-10}$
Locus_22382_Transcript_4/4	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	$2x10^{-62}$
	PF08659.5	KR domain	$1.7x10^{-49}$
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	$1.2x10^{-32}$
	PF00550.20	Phosphopantetheine attachment site	$4.1x10^{-10}$

Table IV-8: Pfam results for the *K. brevis* multidomain PKS sequence.

Transcript ID	**Pfam domain**	**Domain description**	**E value**
Locus_4043_Transcript_1/2	PF00501.23	AMP-binding enzyme	6.10×10^{-56}
	PF00109.21	Beta-ketoacyl synthase, N-terminal domain	4.40×10^{-49}
	PF00698.16	Acyl transferase domain	1.50×10^{-48}
	PF00975.15	Thioesterase domain	2.70×10^{-35}
	PF08659.5	KR domain	1.40×10^{-29}
	PF02801.17	Beta-ketoacyl synthase, C-terminal domain	9.20×10^{-29}
	PF00975.15	Thioesterase domain	1.80×10^{-24}
	PF00550.20	Phosphopantetheine attachment site	6.90×10^{-11}
	PF13193.1	AMP-binding enzyme C-terminal domain	8.70×10^{-05}

Table IV-9: Predicted PKS catalytic domain and nr BLAST results for each "unique" PKS ortholog in the *K. brevis* and *K.papilionacea* transcriptomes. Data is based on the results for the *K. brevis* ortholog. The "top nr hit species" was determined by E value and lists the species with the most significant nr protein hit to each "unique" PKS.

Locus ID	**PKS Domain**	**Top nr hit species**	**Lowest E value**	**Highest % identity**
Locus 14861	KS	*Bacillus hemicellulosilyticus*	$1.33x10^{-43}$	33.06
Locus 5089	KS	*Azadinium spinosum*	0.00	43.84
Locus 15745	AT	*Azadinium spinosum*	$3.67x10^{-84}$	38.16
Locus 27667	AT	*Alexandrium ostenfeldii*	$2.28x10^{-153}$	49.11
Locus 38639	AT	*Alexandrium ostenfeldii*	$2.42x10^{-146}$	47.47

CHAPTER V

CONCLUSION

After developing a reliable pipeline for dinoflagellate *de novo* transcriptome assembly, this dissertation produced highly complete *Karenia* reference transcriptomes. Subsequent analyses annotated thousands of putative proteins, including voltage-gated cation channels, MIPs, VATPases, and PKSs. Of particular interest are genes expressed by the brevetoxin-producing *K. brevis* and *K. papilionacea*, but not by *K. mikimotoi*, which does not produce brevetoxin. The results that support each primary dissertation goal are described:

Identify efficient, effective *de novo* transcriptome assembly method for dinoflagellates with large, highly repetitive genomes

Because *K. brevis* is a non-model organism, this dissertation used several traditional and nontraditional metrics to gauge the completeness of each transcriptome. Commonly, values like mean transcript length, maximum transcript length, N50 length[1], and transcript number are used as proxies for transcriptome quality. High length metrics and a relatively low transcript number are considered optimal, since they ideally contain longer, more complete mRNA sequences. However, based on test transcriptomes

[1]The N50 length is calculated by arranging transcripts from longest to shortest and then identifying the "N50 transcript." The sum of transcript lengths above or below the N50 transcript should contain 50% of all bases in the transcriptome. The length of the N50 transcript is the N50 length.

assembled *de novo* from model organism data, length-based metrics, including N50, do not consistently identify perfect assemblies (O'Neil and Emrich 2013).

Additionally, without reference genomes, the expected N50, transcript number, and mean transcript length outputs for *Karenia* species can only be estimated based on genomic data from other eukaryotes, a potentially unreliable system. Even within the dinoflagellate group, genome size is highly variable (3.0x10^6 to 245.0x10^6 kbp), and the estimated protein-coding gene number ranges from 40,000 to 90,000, depending on species (Hou and Lin 2009). To date, the only dinoflagellate species with a sequenced genome are symbiotic members of the *Symbiodinium* genus (Lin et al. 2015; Shoguchi et al. 2013). *Symbiodinium* spp. genomes are, on average, over 30 times smaller than the *K. brevis* genome (Hackett et al. 2004; Shoguchi et al. 2013). Based on the positive relationship between protein-coding genes and genome size (Hou and Lin 2009), we expect *K. brevis* to express more discrete transcripts than *Symbiodinium*. Free-living *Karenia* and symbiotic *Symbiodinium* zooxanthellae also have disparate life cycles that may necessitate different types of genes.

The *Karenia* transcriptomes were also assessed with annotation-, CEG- and ORF-based methods that are well-suited for nonmodel organisms. As described in earlier chapters, TRAPID predicts ORF completeness with an annotation process. Transcripts are first processed with a RAPSearch2 similarity comparison to proteins in the PLAZA 2.5 (Van Bel et al. 2011) or OrthoMCLDB 5.0 (Chen et al. 2006) databases (Van Bel et al. 2013). During the processing step, users choose whether the transcriptome will be compared to a single species or clade in one of the databases (Table V-1). Alternatively,

the transcriptome can be compared to "gene family representatives," in which one gene per family is chosen, regardless of species or clade, based on the family connectivity protocol outlined in Van Bel et al 2011 (Van Bel et al. 2013; Van Bel et al. 2011). Transcripts receive gene family assignments based on the RAPSearch results.

Table V-1: Clades, species, and proteins in the databases used by TRAPID. Table is adapted from the article "TRAPID: an efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes" (Van Bel et al. 2013).

Reference database	**Clade**	**#Species**	**#Proteins**
OrthoMCL-DB 5.0	All	150	1,398,546
PLAZA 2.5	Alveolata	15	98,796
	Amoebozoa	4	41,930
	Archaea	16	30,233
	Bacteria	36	112,059
	Euglenozoa	9	107,034
	Eukaryota	98	1,256,264
	Fungi	24	680,778
	Metazoa	29	529,788
	Viridiplantae (green plants)	25	780,667
	Angiosperms	18	671,950
	Eudicots	13	480,106
	Monocots	5	191,844

Additionally, the longest potential ORF in each transcript is identified. Transcripts with both a gene family and ORF are assessed for completeness, as described in Chapter II, based on its length and the mean gene family ORF length (Van Bel et al. 2013).

There are several potential complications associated with assessing a dinoflagellate transcriptome with TRAPID. First, dinoflagellates are not represented by the OrthoMCLDB 5.0 or PLAZA 2.5 databases, although the former does include the marine diatom *Thalassiosira* (Alveolata clade), and the latter includes green marine algae (Viridiplantae clade). With less than 200 species in the TRAPID databases (Table V-1), none dinoflagellates, the RAPSearch may fail to annotate less-conserved genes. Indeed, just ~20% of the *K. brevis* transcriptome was assigned a gene family by TRAPID. In contrast, a BLASTx search against the nr database, which contains proteins from >2.3x10^5 formally named species and >4.05x10^5 informally named species (Federhen 2012), annotated ~40% of the *K. brevis* transcriptome. Because TRAPID uses a relatively small species databases, "completeness" estimates may only represent a minority of nonmodel transcriptomes: specifically, more highly conserved genes.

Furthermore, the TRAPID ORF completeness analysis cannot distinguish between partial ORFs and naturally shorter ORFs in nonmodel species like *K. brevis*. When *K. brevis* transcripts were processed against the OrthoMCL database instead of the PLAZA database, the estimated complete to partial ratio was over two times higher (Table V-2). This suggests that "completeness" for nonmodel organisms, particularly species without close phylogenetic relatives in the TRAPID databases, is highly qualitative.

That said, TRAPID is a powerful tool for assessing different *de novo* assembly methods, as long as RAPSearch processing variables, such as database choice, are kept consistent. This dissertation consistently used the PLAZA database to assess *Karenia*

transcriptomes because PLAZA-processed transcripts also receive predicted gene ontology terms. To ensure that PLAZA 2.5 would not annotate fewer transcripts than the larger OrthoMCL 5 database, 1,749 randomly chosen *K. brevis* Wilson transcripts were analyzed with both databases. Based on results from the sample test, the number of transcripts that received a TRAPID-assigned gene family is similar, regardless of database choice (Table V-2).

Table V-2: TRAPID results after 1,749 sample *K. brevis* Wilson transcripts were processed with the PLAZA 2.5 or OrthoMCL-DB 5.0 databases. For each database, transcripts were either annotated against a specific clade or gene family representatives. The Viridiplantae and Alveolata clades were chosen because they contain marine phytoplankton species.

ID	**% Transcripts with Gene Family**	**% Transcripts with Protein Domain**	**Complete:Partial Ratio**
KB PLAZA 2.5 Representative gene	(42.9%)	(39.9%)	6.3:1
KB Wilson PLAZA 2.5 Viridiplantae clade	(48.4%)	(43.7%)	8.8:1
KB Wilson ORTHO Representative gene	(45.6%)	(36.6%)	17.5:1
KHB Wilson ORTHO Alveolata clade	(38.4%)	(28.5%)	22.48:1

The TRAPID "completeness" estimates were complemented by CEGMA. As described in chapter II, CEGMA searches transcriptomes for a set of 248 highly conserved eukaryotic genes originally identified in *Homo sapiens, Drosophila melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Parra et al. 2007). CEGMA identified 99% (246) of

the CEGs in the "most complete" (according to CEGMA) *K. brevis* transcriptome (Table V-3).

Table V-3: Transcript number, % CEGs, N50 length, TRAPID complete:partial ratio, % transcriptome assigned a gene family by TRAPID, mean transcript length, and mean ORF length for *Karenia* reference transcriptomes. KbWil Trinity = Trinity assembly, KbWil VO = Velvet Oases single k-mer assembly (k-mer 41), KbWil VO MA = Velvet Oases merged assembly (k-mers 21, 25, 29, 33, 37, 41), KbWil VO MA + Trinity = Velvet Oases single k-mer assembly combined with Trinity assembly.

ID	**# Transc.**	**%CEGs**	**TRAPID complete:partial**	**Mean transc. lgth (bp)**	**Mean ORF lgth (bp)**
KbWil Trinity	166,134	79.91	8.0:1	1096.6	1218.3
KbWil VO	85,697	71.70	7.1:1	1353	1075.7
KbWil VO MA	86,580	81	7.3:1	1340	961.6
KbWil VO MA + Trinity	309,289	99.12	7.0:1	1257.5	1096.5

While attempting to identify an optimal *de novo* assembly method for *K. brevis* RNA-seq data, this dissertation highlighted the difficulty inherent to using any one method to gauge assembly success. For comparison purposes, *K. brevis* transcriptomes were produced with Trinity, Velvet-Oases (single k-mer method), Velvet-Oases (merged k-mer method), and Trinity + Velvet-Oases (single k-mer method). Length-based metrics, % CEGs, and TRAPID completeness were calculated for each transcriptome (Table V-3). Velvet-Oases (single k-mer method) yielded the highest mean transcript length and lowest transcript number. Trinity yielded the highest mean ORF length and

TRAPID completeness ratio. Trinity + Velvet-Oases yielded the highest % CEG value by a considerable margin. No one assembly is "optimal" according to all—or even most—of the criteria. When all the metrics are considered together, the Velvet-Oases (merged k-mer method) has overall high length and completeness scores with a low total transcript number (Table V-3). However, for research that is dependent on assembling the most transcripts possible, it seems prudent to create transcriptomes with both Trinity and Velvet-Oases, combine the results, and remove redundancies. Although the Trinity + Velvet-Oases transcriptome was up to three times larger than the Velvet-Oases merged transcriptome, it also contained 18% more CEGs (Table V-3).

Based on the methods explored during this dissertation, a basic *de novo* transcriptome assembly (Figure V-1), assessment (Figure V-2), and annotation (Figure V-3) procedure has been proposed. The pipeline is a guide and should be modified to best fit each experiment's goals and data. Of the four *de novo* transcriptome assemblers tested during this dissertation (ABySS, CLC, Velvet-Oases, and Trinity), Velvet-Oases and Trinity produced optimal transcriptomes (see Chapter II and III) and are therefore recommended for further dinoflagellate transcriptomic research. The Velvet-Oases merged function is also recommended over the single k-mer approach.

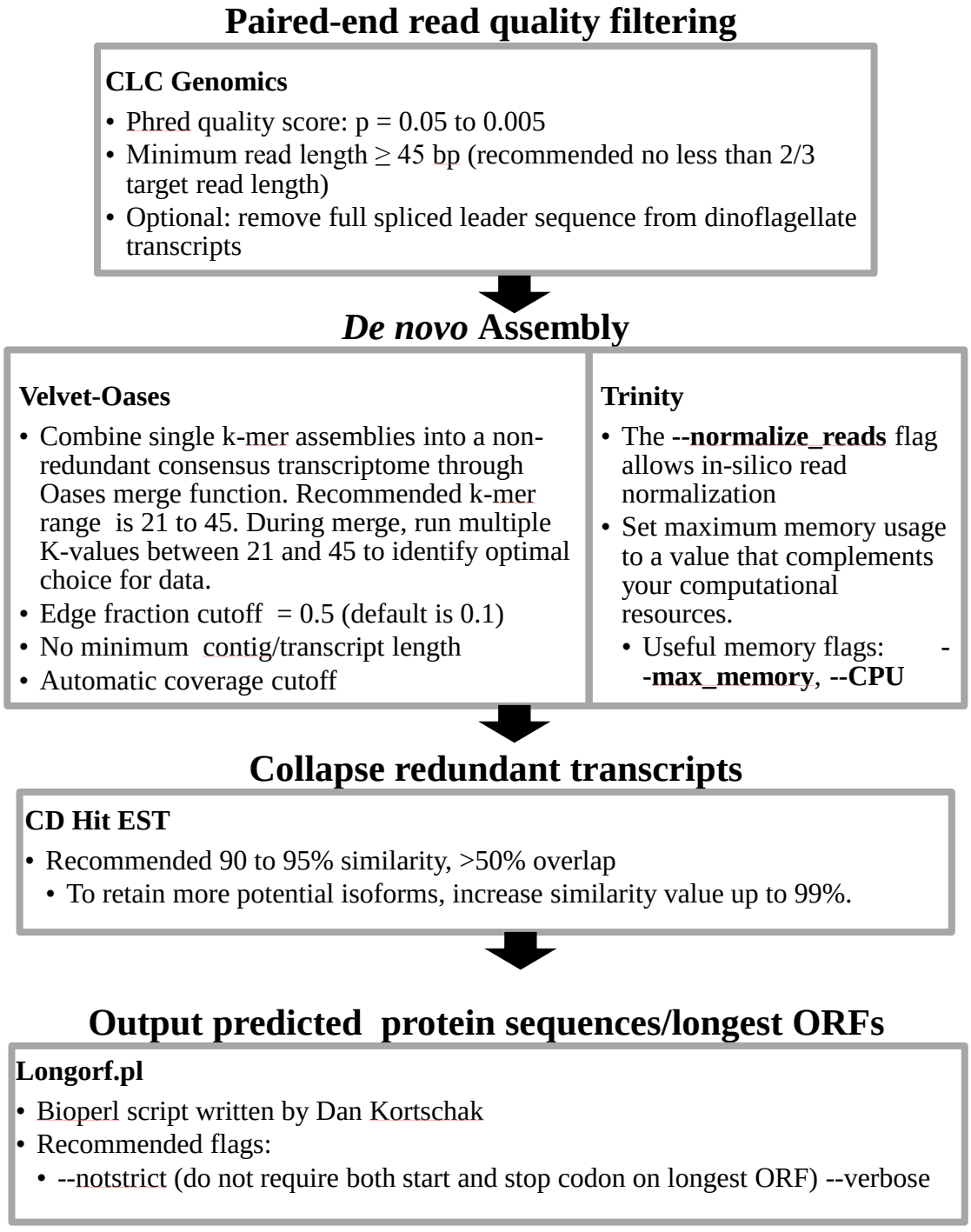

Figure V-1: *De novo* transcriptome assembly pipeline.

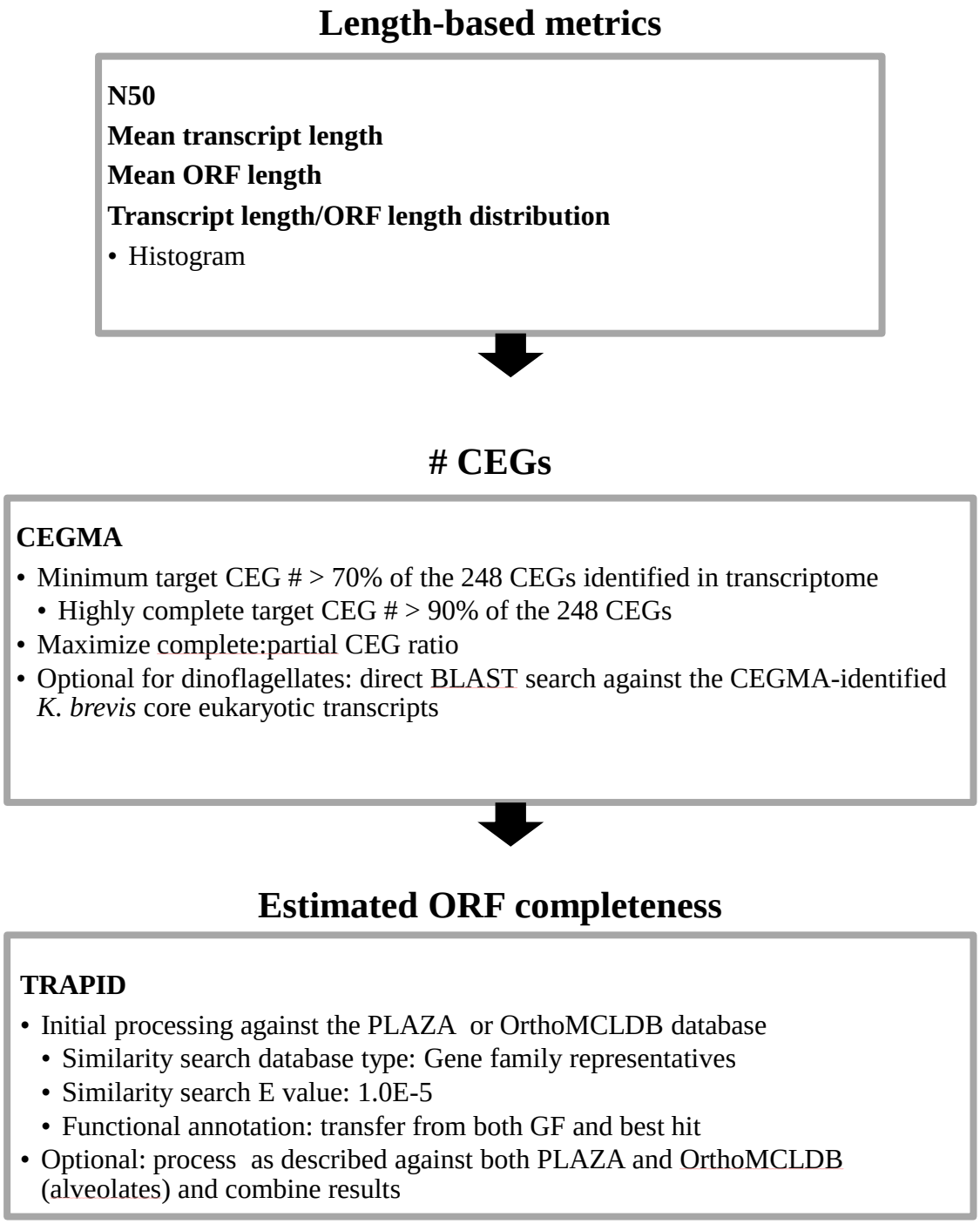

Figure V-2: *De novo* transcriptome assembly assessment pipeline.

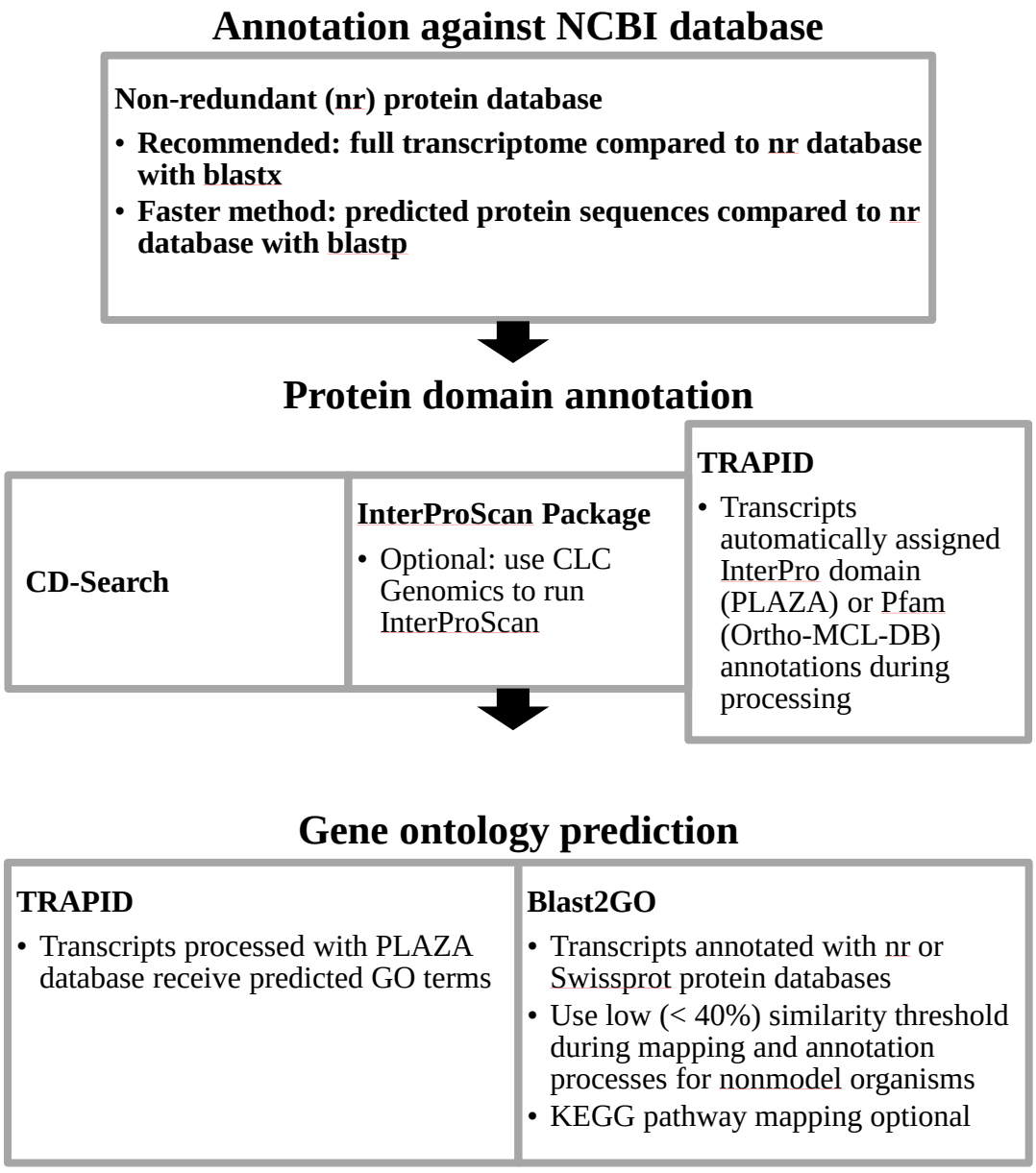

Figure V-3: *De novo* transcriptome annotation pipeline.

Predict highly conserved genes in dinoflagellates and marine eukaryotes

CEG recovery is highly sensitive to transcriptome assembly method. When only one assembler (either Velvet-Oases or Trinity) processed the RNA-seq data, between 70% and 80% of the 248 highly conserved CEGs were identified through a CEGMA search. However, by combining output from both Velvet-Oases and Trinity, 99% of the CEGs were identified among the *K. brevis* transcripts. Therefore, CEGMA output from the Velvet-Oases + Trinity combined *K. brevis* Wilson transcriptome was used to further investigate CEGs in dinoflagellates.

These *K. brevis* CEG protein sequences were converted to a searchable BLAST database and compared to *Alexandrium fundyense* CCMP1719, *Karlodinium micrum* CCMP2283, *Lingulodinium polyedra* CCMP1738, *Prorocentrum minimum* CCMP1329, and *Symbiodinium kawagutii* CCMP2468 transcriptomes from the MMETSP repository (Keeling et al. 2014). The search enforced a minimum E value ($<1.0x10^{-20}$). For comparison purposes, all ORFs >300 bp in the Velvet-Oases+Trinity transcriptome were converted to protein sequences and compared to the five dinoflagellate transcriptomes using the BLAST guidelines above. There were 147,200 total *K. brevis* ORFs >300 bp.

For every queried dinoflagellate species except *Symbiodinium*, over 95% of the *K. brevis* CEGs were identified during the BLAST search (Table V-4). The average protein similarity score for each CEG hit was ~70% (Table V-4). Each of these values is markedly higher than their counterparts yielded by the full-transcriptome BLAST search (Table V-4).

This result suggests two important conclusions. First, the CEGMA-derived core eukaryotic transcripts in *K. brevis* represent highly conserved genes among free-living dinoflagellate species. Not only were almost all of the CEG orthologs identified in *Alexandrium*, *Karlodinium*, *Lingulodinium*, and *Prorocentrum*, the sequence similarity between species was higher, on average, than the similarity between non-CEG orthologs (Table V-4).

Second, *Symbiodinium*, the only dinoflagellate genus with a sequenced genome, was also the only dinoflagellate species assessed during this dissertation that expressed few (35%) of the *K. brevis* CEGs. This highlights the need for more genomic data from non-zooxanthellae dinoflagellate species in order to better understand toxin-producing, bloom-forming dinoflagellates.

Table V-4: Results of *K. brevis* CEG BLAST against five dinoflagellate transcriptomes. The MMETSP IDs column lists the sample(s) that contributed to each transcriptome. The "% KB CEGs" and "% total KB transcripts" columns show the percent of *K. brevis* CEG or non-CEG transcripts with a predicted ortholog in the species transcriptome. The "mean CEG % sim" and "mean total sim" columns show the average % protein similarity between CEG and non-CEG orthologs.

Species	MMETSP IDs	% KB CEGs	Mean CEG % sim	% total KB transcripts	Mean total sim
Alexandrium fundyense CCMP1719	0196, 0347	98	73	4.2	56
Karlodinium micrum CCMP2283	1016, 1015, 1017	95	73	34	47
Lingulodinium polyedra CCMP1738	1034, 1032, 1035, 1033	98	73	34	47
Prorocentrum minimum CCMP1329	0053, 0055, 0057, 0056	96	70	29	45
Symbiodinium kawagutii CCMP2468	1032	35	76	3.1	58

Identify the potential genes that underlie osmoacclimation or toxin production in harmful, bloom-forming dinoflagellate *K. brevis*

In order to better understand osmoacclimation in *K. brevis*, putative transmembrane ion channels, major intrinsic proteins, and VATPases were identified in *K. brevis* Wilson, SP1, and SP3 transcriptomes. Notably, seven *K. brevis* transcripts were significantly homologous to mammalian and fish voltage-gated cation channels (Chapter II) (Ryan et al. 2014). The interaction between brevetoxins and human voltage-gated sodium channels results in neurotoxic shellfish poisoning. Simply put, brevetoxins bind to neurotoxin receptor site 5 and increase sodium transport, thus causing harmful neurological symptoms (Baden 1989; Dechraoui et al. 1999). Although fragments of ion channel-like sequences were identified in early *K. brevis* EST libraries (Thompson 2011), the transcriptomes assembled as part of this dissertation research contained the first full-length putative cation channel genes to be assembled from brevetoxin-producing *K. brevis* clones (Ryan et al. 2014).

The identification of voltage-gated sodium channels in *Karenia* species is a motivation to further investigate the interaction between brevetoxins and dinoflagellates, particularly in brevetoxin-producing *K. brevis* and *K. papilionacea*. To date, it is not known whether PbTx-1, PbTx-2, their derivatives, and/or brevenal bind to *Karenia* cation channels. If brevetoxins do alter *K. brevis* ion transport activity, the interaction may point to a biological impetus for PbTx synthesis. Alternatively, if *K. brevis* ion channels do not bind with brevetoxins, variations in ion channel orthologs between *K.*

brevis and humans may help us understand the site-5 configuration that is necessary for interactions with PbTx compounds.

In addition to transmembrane channels, this dissertation searched for putative PKSs in *Karenia* transcriptomes. *K. brevis*, *K. mikimotoi*, and *K. papilionacea* all expressed >100 distinct PKS transcripts with KS, AT, ACP, and/or KR domains (Chapter III and IV). The relatively high number of PKS genes was not uncommon. A recent study that analyzed the transcriptomes of 210 phytoplankton genera showed highly expanded PKS gene families in the dinoflagellate group (Kohli et al. 2016). On average, the 46 dinoflagellate strains expressed 56 KS domain-containing transcripts. The specific number varied from 140 (*Azadinium spinosum*) to 90 (*K. brevis*) to only 7 (*Togula jolla*) distinct KS-containing PKS transcripts (Kohli et al. 2016).

Regarding PKS characterization, this dissertation made one unexpected discovery. Although most *Karenia* PKSs were type I-like, with a single catalytic domain, a multimodular PKS sequence (Table IV-8) was expressed by *K. brevis* in both laboratory and environmental samples (Chapter IV). Based on the reciprocal BLAST method, the multimodular PKS was novel to *K. brevis*, with no orthologs in the *K. mikimotoi* or *K. papilionacea* transcriptomes. However, both *K. mikimotoi* and *K. papilionacea* expressed other two-domain or three-domain putative PKS genes (Table IV-6 and IV-7). This suggests that polyketide and/or nonribosomal peptide assembly in *Karenia* species is more complex than the type I-like paradigm.

Predict genetic variance among two *K. brevis* laboratory clones and three *Karenia* species: *K. brevis*, *K. mikimotoi*, and *K. papilionacea*

Thousands of orthologs were identified among three harmful *Karenia* species: *K. brevis*, *K. mikimotoi*, and *K. papilionacea* (Chapter III and Chapter IV). Of special interest are the >4,700 transcripts unique to the brevetoxin-producing *K. brevis* and *K. mikimotoi*, including 5 type I-like PKSs. These "unique" orthologs represent prime targets for future brevetoxin biosynthesis research, such as targeted silencing with antisense oligonucloetides.

To understand genetic variance among laboratory-cultured *K. brevis* clones, we compared the transcriptomes of Wilson, SP1, and SP3 (Chapter II). Each clone expressed the same genes, but we did identify 186,075 SNP locations were in 30,227 highly expressed (20-fold minimum average read coverage) transcripts, corresponding with 0.0023 (Wilson and SP1), 0.0024 Wilson and SP3, and 0.0022 (SP1 and SP3) SNP rates (Table II-3). Based on these results, SP1 and SP3 are more similar to each other overall than SP1 to Wilson or SP3 to Wilson. It is not known whether Wilson is relatively divergent because of its long (over 60 years) time in culture. However, the strain-to-strain variance among tens of thousands of highly expressed genes suggests that experiments on laboratory-cultured *K. brevis* should use more than one strain to confirm that biological responses are representative of more than one clone with a potentially phenotype-altering mutation.

The *Karenia* transcriptomes that are described in this dissertation each contain a vast pool of expressed genes. To date, no toxin-producing dinoflagellate has a reference genome. Therefore, highly complete, *de novo* transcriptomes are critical tools to understand the biology of toxin-producing dinoflagellate species. For example, we now know that *K. brevis*, *K. papilionacea*, and *K. mikimotoi* express putative transmembrane channels and a hundreds of novel PKSs. This said, there is still much to learn from the >300,000 *Karenia* transcripts. The transcriptomes will be valuable resources for future work.

REFERENCES

Agabian N. 1990. Trans splicing of nuclear pre-mRNAs. Cell 61(7):1157-1160.

Agre P, King LS, Yasui M, Guggino WB, Ottersen OP, Fujiyoshi Y, Engel A, Nielsen S. 2002. Aquaporin water channels–from atomic structure to clinical medicine. The Journal of Physiology 542(1):3-16.

Agre P, Preston G, Smith B, Jung J, Raina S, Moon C, Guggino WB, Nielsen S. 1993. Aquaporin CHIP: The archetypal molecular water channel. American Journal of Physiology-Renal Physiology 265(4):F463-F476.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research 25(17):3389-3402.

Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A, Mitchell AL. 2012. The PRINTS database: A fine-grained protein sequence annotation and analysis resource—its status in 2012. Database 2012:bas019.

Backer LC, Fleming LE, Rowan A, Cheng Y-S, Benson J, Pierce RH, Zaias J, Bean J, Bossart GD, Johnson D. 2003. Recreational exposure to aerosolized brevetoxins during Florida red tide events. Harmful Algae 2(1):19-28.

Baden DG. 1989. Brevetoxins: unique polyether dinoflagellate toxins. The FASEB Journal 3(7):1807-1817.

Baker A, Robbins I, Moline MA, Iglesias-Rodríguez MD. 2008. Oligonucleotide primers for the detection of bioluminescent dinoflagellates reveal novel luciferase

sequences and information on the molecular evolution of this gene. Journal of Phycology 44(2):419-428.

Baker H. 1753. On some luminous water insects. Employment for the Microscope: In Two Parts. London: R. Dodsley.

Berman FW, Murray TF. 1999. Brevetoxins cause acute excitotoxicity in primary cultures of rat cerebellar granule neurons. Journal of Pharmacology and Experimental Therapeutics 290(1):439-444.

Borgnia M, Nielsen S, Engel A, Agre P. 1999. Cellular and molecular biology of the aquaporin water channels. Annual Review of Biochemistry 68(1):425-458.

Bourdelais AJ, Campbell S, Jacocks H, Naar J, Wright JL, Carsi J, Baden DG. 2004. Brevenal is a natural inhibitor of brevetoxin action in sodium channel receptor binding assays. Cellular and Molecular Neurobiology 24(4):553-563.

Bourdelais AJ, Jacocks HM, Wright JL, Bigwarfe PM, Baden DG. 2005. A new polyether ladder compound produced by the dinoflagellate *Karenia brevis*. Journal of Natural Products 68(1):2-6.

Brunelle SA, Van Dolah FM. 2011. Post-transcriptional regulation of S-phase genes in the dinoflagellate, *Karenia brevis*. Journal of Eukaryotic Microbiology 58(4):373-382.

Calabro K, Guigonis J-M, Teyssié J-L, Oberhänsli F, Goudour J-P, Warnau M, Bottein M-YD, Thomas OP. 2014. Further insights into brevetoxin metabolism by *de novo* radiolabeling. Toxins 6(6):1785-1798.

Cane DE, Walsh CT. 1999. The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. Chemistry & Biology 6(12):R319-R325.

Cao Z, George J, Gerwick WH, Baden DG, Rainier JD, Murray TF. 2008. Influence of lipid-soluble gating modifier toxins on sodium influx in neocortical neurons. Journal of Pharmacology and Experimental Therapeutics 326(2):604-613.

Catterall WA. 2000. From ionic currents to molecular mechanisms: the structure and function of voltage-gated sodium channels. Neuron 26(1):13-25.

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

Catterall WA, Goldin AL, Waxman SG. 2005. International Union of Pharmacology. XLVII. Nomenclature and structure-function relationships of voltage-gated sodium channels. Pharmacological Reviews 57(4):397-409.

Chen F, Mackey AJ, Stoeckert CJ, Roos DS. 2006. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Research 34(suppl 1):D363-D368.

Chou H-N, Shimizu Y. 1982. A new polyether toxin from *Gymnodinium breve* Davis. Tetrahedron Letters 23(52):5521-5524.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674-3676.

Corpet F, Gouzy J, Kahn D. 1998. The ProDom database of protein domain families. Nucleic Acids Research 26(1):323-326.

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.

Davis CC. 1948. *Gymnodinium brevis* sp. nov., a cause of discolored water and animal mortality in the Gulf of Mexico. Botanical Gazette 109(3):358-360.

de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, Chothia C, Gough J. 2010. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Research 39(suppl 1):D427-D434.

Dechraoui M-Y, Naar J, Pauillac S, Legrand A-M. 1999. Ciguatoxins and brevetoxins, neurotoxic polyether compounds active on sodium channels. Toxicon 37(1):125-143.

Dodge JD. 1984. Dinoflagellate Taxonomy. Orlando, Florida: Academic Press.

Eichholz K, Beszteri B, John U. 2012. Putative monofunctional type I polyketide synthase units: A dinoflagellate-specific feature? PLoS One 7(11): e48624.

Errera R, Campbell L. 2012. Correction for Errera and Campbell, Osmotic stress triggers toxin production by the dinoflagellate *Karenia brevis*. Proceedings of the National Academy of Sciences 109(43):17723-17724.

Errera RM, Bourdelais A, Drennan M, Dodd E, Henrichs D, Campbell L. 2010. Variation in brevetoxin and brevenal content among clonal cultures of *Karenia brevis* may influence bloom toxicity. Toxicon 55(2):195-203.

Evans G, Jones L. 2001. Economic impact of the 2000 red tide on Galveston County, Texas: A case study. College Station: Department of Agricultural Economics, Texas A&M University.

Federhen S. 2012. The NCBI taxonomy database. Nucleic Acids Research 40(D1):D136-D143.

Fensome RA, Saldarriaga JF, Taylor MF. 1999. Dinoflagellate phylogeny revisited: Reconciling morphological and molecular based phylogenies. Grana 38(2-3):66-80.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J. 2013. Pfam: The protein families database. Nucleic Acids Research:gkt1223.

Fischbach MA, Walsh CT. 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. Chemical Reviews 106(8):3468-3496.

Forgac M. 1989. Structure and function of vacuolar class of ATP-driven proton pumps. Physiological Reviews 69(3):765-796.

Fowler N, Tomas C, Baden D, Campbell L, Bourdelais A. 2015. Chemical analysis of *Karenia papilionacea*. Toxicon 101:85-91.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150-3152.

Gallimore AR. 2009. The biosynthesis of polyketide-derived polycyclic ethers. Natural Product Reports 26(2):266-280.

Gallimore AR, Spencer JB. 2006. Stereochemical uniformity in marine polyether ladders—implications for the biosynthesis and structure of maitotoxin. Angewandte Chemie International Edition 45(27):4406-4413.

Glass AD. 1983. Regulation of ion transport. Annual Review of Plant Physiology 34(1):311-326.

Golik J, James JC, Nakanishi K, Lin Y-Y. 1982. The structure of brevetoxin C. Tetrahedron Letters 23(25):2535-2538.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29(7):644-652.

Granéli E. 2006. Ecology of harmful algae. Ed. Jefferson T. Turner. Berlin, Heidelberg: Springer.

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

Gunter G, Williams RH, Davis CC, Smith FW. 1948. Catastrophic mass mortality of marine animals and coincident phytoplankton bloom on the west coast of Florida, November 1946 to August 1947. Ecological Monographs 18(3):310-324.

Hackett JD, Anderson DM, Erdner DL, Bhattacharya D. 2004. Dinoflagellates: a remarkable evolutionary experiment. American Journal of Botany 91(10):1523-1534.

Haddock SH, Moline MA, Case JF. 2010. Bioluminescence in the sea. Marine Science 2: 443-493.

Haywood AJ, Steidinger KA, Truby EW, Bergquist PR, Bergquist PL, Adamson J, MacKenzie L. 2004. Comparative morphology and molecular phylogenetic analysis of three new species of the genus *Karenia* (Dinophyceae) from New Zealand. Journal of Phycology 40(1):165-179.

Heymann JB, Engel A. 1999. Aquaporins: phylogeny, structure, and physiology of water channels. Physiology 14(5):187-193.

Hopwood DA, Sherman DH. 1990. Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. Annual Review of Genetics 24(1):37-62.

Hou Y, Lin S. 2009. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: Gene content estimation for dinoflagellate genomes. PLoS One 4(9):e6978.

Huang J, Wu CH, Baden DG. 1984. Depolarizing action of a red-tide dinoflagellate brevetoxin on axonal membranes. Journal of Pharmacology and Experimental Therapeutics 229(2):615-621.

Kamykowski D, Milligan EJ, Reed RE. 1998. Biochemical relationships with the orientation of the autotrophic dinoflagellate *Gymnodinium breve* under nutrient replete conditions. Marine Ecology Progress Series 167:105-117.

Keatinge-Clay AT. 2012. The structures of type I polyketide synthases. Natural Product Reports 29(10):1050-1073.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biology 12(6):e1001889.

Kim YS, Martin DF. 1974. Effects of salinity on synthesis of DNA, acidic polysaccharide, and ichthyotoxin in *Gymnodinium breve*. Phytochemistry 13(3):533-538.

Kohli GS, John U, Van Dolah FM, Murray SA. 2016. Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. The ISME Journal. 10:1877–1890

Kortschak, D. 2002. longorf.pl [perl script]. Available at https://github.com/bioperl/bioperl-live/blob/master/examples/longorf.pl

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology 305(3):567-580.

Landsberg JH. 2002. The effects of harmful algal blooms on aquatic organisms. Reviews in Fisheries Science 10(2):113-390.

Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, Yeats C, Dessailly BH, Rentzsch R, Orengo CA. 2014. Gene3D: Multi-domain annotations for protein

sequence and comparative genome analysis. Nucleic Acids Research 42(D1):D240-D245.

Letunic I, Doerks T, Bork P. 2012. SMART 7: Recent updates to the protein domain annotation resource. Nucleic Acids Research 40(D1):D302-D305.

Lidie KB, Ryan JC, Barbier M, Van Dolah FM. 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.

Lidie KB, Van Dolah FM. 2007. Spliced Leader RNA-Mediated trans-Splicing in a Dinoflagellate, *Karenia brevis*. Journal of Eukaryotic Microbiology 54(5):427-435.

Lin S. 2011. Genomic understanding of dinoflagellates. Research in Microbiology 162(6):551-569.

Lin S, Cheng S, Song B, Zhong X, Lin X, Li W, Li L, Zhang Y, Zhang H, Ji Z. 2015. The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. Science 350(6261):691-694.

Lin Y-Y, Risk M, Ray SM, Van Engen D, Clardy J, Golik J, James JC, Nakanishi K. 1981. Isolation and structure of brevetoxin B from the" red tide" dinoflagellate *Ptychodiscus brevis* (*Gymnodinium breve*). Journal of the American Chemical Society 103(22):6773-6775.

Linial M. 2003. How incorrect annotations evolve–the case of short ORFs. Trends in Biotechnology 21(7):298-300.

López-Legentil S, Song B, DeTure M, Baden DG. 2010. Characterization and localization of a hybrid non-ribosomal peptide synthetase and polyketide synthase gene from the toxic dinoflagellate *Karenia brevis*. Marine Biotechnology 12(1):32-41.

Maathuis FJ, Filatov V, Herzyk P, C Krijger G, B Axelsen K, Chen S, Green BJ, Li Y, Madagan KL, Sánchez-Fernández R. 2003. Transcriptome analysis of root transporters reveals participation of multiple gene families in the response to cation stress. The Plant Journal 35(6):675-692.

Magaña HA, Contreras C, Villareal TA. 2003. A historical assessment of *Karenia brevis* in the western Gulf of Mexico. Harmful Algae 2(3):163-171.

Maier Brown AF, Dortch Q, Dolah FMV, Leighfield TA, Morrison W, Thessen AE, Steidinger K, Richardson B, Moncreiff CA, Pennock JR. 2006. Effect of salinity on the distribution, growth, and toxicity of *Karenia* spp. Harmful Algae 5(2):199-212.

Marahiel MA, Stachelhaus T, Mootz HD. 1997. Modular peptide synthetases involved in nonribosomal peptide synthesis. Chemical Reviews 97(7):2651-2674.

Marchler-Bauer A, Bryant SH. 2004. CD-Search: Protein domain annotations on the fly. Nucleic Acids Research 32(suppl 2):W327-W331.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR. 2011. CDD: A Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research 39(suppl 1):D225-D229.

Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ. 2012. CDD: Conserved domains and protein three-dimensional structure. Nucleic Acids Research: 41(D1): D348-D352.

Martin JA, Wang Z. 2011. Next-generation transcriptome assembly. Nature Reviews Genetics 12(10):671-682.

Matsuo G, Kawamura K, Hori N, Matsukura H, Nakata T. 2004. Total synthesis of brevetoxin-B. Journal of the American Chemical Society 126(44):14374-14376.

McFarren E, Tanabe H, Silva F, Wilson W, Campbell J, Lewis K. 1965. The occurrence of a ciguatera-like poison in oysters, clams, and *Gymnodinium breve* cultures. Toxicon 3(2):111-123.

Metzker ML. 2010. Sequencing technologies—the next generation. Nature Reviews Genetics 11(1):31-46.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. 2013. Large-scale gene function analysis with the PANTHER classification system. Nature Protocols 8(8):1551-1566.

Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S. 2014. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Research 43(D1): D213-D221.

Monroe EA, Van Dolah FM. 2008. The toxic dinoflagellate *Karenia brevis* encodes novel type I-like polyketide synthases containing discrete catalytic domains. Protist 159(3):471-482.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24(3):319-324.

Morey JS, Monroe EA, Kinney AL, Beal M, Johnson JG, Hitchcock GL, Van Dolah FM. 2011. Transcriptomic response of the red tide dinoflagellate, *Karenia brevis*, to nitrogen and phosphorus depletion and addition. BMC Genomics 12(1):346.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5(7):621-628.

Murray SA, Suggett DJ, Doblin MA, Kohli GS, Seymour JR, Fabris M, Ralph PJ. 2016. Unravelling the functional genetics of dinoflagellates: A review of approaches and opportunities. Perspectives in Phycology:37-52.

Nelson N, Perzov N, Cohen A, Hagai K, Padler V, Nelson H. 2000. The cellular biology of proton-motive force generation by V-ATPases. Journal of Experimental Biology 203(1):89-95.

Nicolaou K, Yang Z, Shi G-q, Gunzner JL, Agrios KA, Gärtner P. 1998. Total synthesis of brevetoxin A. Nature 392(6673):264-269.

Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH. 2006. PIRSF family classification system for protein functional and evolutionary analysis. Evolutionary Bioinformatics Online 2:197.

O'Neil ST, Emrich SJ. 2013. Assessing *de novo* transcriptome assembly metrics for consistency and utility. BMC Genomics 14(1):1.

Parkinson J, Blaxter M. 2009. Expressed sequence tags: An overview. Methods in Molecular Biology 533:1-12.

Parra G, Bradnam K, Korf I. 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061-1067.

Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, De Castro E, Baratin D, Cuche BA, Bougueleret L, Poux S. 2013. HAMAP in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Research 41(D1):D584-D589.

Poli M, Mende TJ, Baden DG. 1986. Brevetoxins, unique activators of voltage-sensitive sodium channels, bind to specific sites in rat brain synaptosomes. Molecular Pharmacology 30(2):129-135.

Rein KS, Borrone J. 1999. Polyketides from dinoflagellates: Origins, pharmacology and biosynthesis. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 124(2):117-131.

Rizzo P, Jones M, Ray S. 1982. Isolation and properties of isolated nuclei from the Florida red tide dinoflagellate *Gymnodinium breve* (Davis) 1. Journal of Eukaryotic Microbiology 29(2):217-222.

Rizzo PJ. 2003. Those amazing dinoflagellate chromosomes. Cell Research 13(4):215-217.

Ryan DE, Campbell L. 2015. Comparative transcriptomic analysis of three toxin-producing *Karenia* species. Proceedings of the 16th International Conference on Harmful Algae:229-232.

Ryan DE, Pepper AE, Campbell L. 2014. *De novo* assembly and characterization of the transcriptome of the toxic dinoflagellate *Karenia brevis*. BMC Genomics 15(1):888.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270(5235):467-470.

Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086-1092.

Shelest E, Heimerl N, Fichtner M, Sasso S. 2015. Multimodular type I polyketide synthases in algae evolve by module duplications and displacement of AT domains in trans. BMC Genomics 16(1):1.

Shen B. 2003. Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Current Opinion in Chemical Biology 7(2):285-295.

Shimizu Y, Chou HN, Bando H, Van Duyne G, Clardy J. 1986. Structure of brevetoxin A (GB-1 toxin), the most potent toxin in the Florida red tide organism *Gymnodinium breve* (*Ptychodiscus brevis*). Journal of the American Chemical Society 108(3):514-515.

Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, Takeuchi T, Hisata K, Tanaka M, Fujiwara M. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. Current Biology 23(15):1399-1408.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7(1):539.

Sigee DC. 1984. Structural DNA and genetically active DNA in dinoflagellate chromosomes. Biosystems 16(3):203-210.

Sigrist CJ, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2012. New and continuing developments at PROSITE. Nucleic Acids Research 41 (D1):D344-D347.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol İ. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research 19(6):1117-1123.

Snyder R, Gibbs P, Palacios A, Abiy L, Dickey R, Lopez J, Rein K. 2003. Polyketide synthase genes from marine dinoflagellates. Marine Biotechnology 5(1):1-12.

Spector DL. 1984. Dinoflagellates: An introduction. Orlando, Florida: Academic Press.

Spikes JJ, Ray SM, Aldrich DV, Nash JB. 1968. Toxicity variations of *Gymnodinium breve* cultures. Toxicon 5(3):171-174.

Steidinger KA, Vargo GA, Tester PA, Tomas CR. 1998. Bloom dynamics and physiology of *Gymnodinium breve* with emphasis on the Gulf of Mexico. NATO ASI Series G: Ecological Sciences 41:133-154.

Sutton RE, Boothroyd JC. 1986. Evidence for trans splicing in trypanosomes. Cell 47(4):527-535.

Thompson NJ. 2011. Brevetoxin: How is it Made and Why? Texas A&M University.

Trainer VL, Baden DG, Catterall WA. 1994. Identification of peptide components of the brevetoxin receptor site of rat brain sodium channels. Journal of Biological Chemistry 269(31):19904-19909.

Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K. 2013. TRAPID: An efficient online tool for the functional and comparative analysis of *de novo* RNA-Seq transcriptomes. Genome Biology 14(12):R134.

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K. 2011. Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiology:111.189514.

Van de Peer Y, De Wachter R. 1997. Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. Journal of Molecular Evolution 45(6):619-630.

Van Dolah FM, Lidie KB, Monroe EA, Bhattacharya D, Campbell L, Doucette GJ, 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.

Van Dolah FM, Lidie KB, Morey JS, Brunelle SA, Ryan JC, Monroe EA, Haynes BL. 2007. Microarray analysis of diurnal- and circadian-regulated genes in the Florida red-tide dinoflagellate *Karenia brevis*. Journal of Phycology 43(4):741-752.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1):57-63.

Watkins SM, Reich A, Fleming LE, Hammond R. 2008. Neurotoxic shellfish poisoning. Marine Drugs 6(3):431-455.

Wegmann K. 1986. Osmoregulation in eukaryotic algae. FEMS Microbiology Letters 39(1):37-43.

Williams RH, Gunter GF, Smith W. 1947. Mass mortality of marine animals on the lower west coast of Florida, November 1946-January 1947. Science 105:256-257

Xu L, Chen H, Hu X, Zhang R, Zhang Z, Luo Z. 2006. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Molecular Biology and Evolution 23(6):1107-1108.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research 18(5):821-829.

Zhang H, Hou Y, Miranda L, Campbell DA, Sturm NR, Gaasterland T, Lin S. 2007. Spliced leader RNA trans-splicing in dinoflagellates. Proceedings of the National Academy of Sciences 104(11):4618-4623.