ABSTRACT

Title of Dissertation: TOXIN-ANTITOXIN SYSTEMS AND OTHER STRESS RESPONSE ELEMENTS IN PICOCYANOBACTERIA AND THEIR ECOLOGICAL IMPLICATIONS.

Daniel Fucich, Doctor of Philosophy, 2020

Directed By: Dr. Feng Chen, Professor,
Institute of Marine and Environmental Technology, University of Maryland Center for Environmental Science

Picocyanobacteria (mainly *Synechococcus* and *Prochlorococcus*) contribute significantly to oceanic primary production. Unlike *Prochlorococcus,* which is mainly constrained to the warm and oligotrophic ocean, *Synechococcus* has a ubiquitous distribution. *Synechococcus* is present in freshwater, estuarine, coastal, and open ocean habitats. They have also been found in polar regions and hot springs. Endemic to the hot and the cold, the saline and the fresh, and every condition in between, *Synechococcus* appears to have the capability to adapt and tolerate nearly any environment and climate. This ability to adapt to any aquatic environment is possible through their genome plasticity, a character that is not present in the *Prochlorococcus*. Due to the differential distribution of the genera, *Synechococcus* is considered a generalist and

Prochlorococcus is considered a specialist in ecological theory. More than 400 picocyanobacterial genomes have now been sequenced, and this large genomic resource enables comprehensive genome mining and comparison. One possibility is to study the prevalence of Toxin-Antitoxin (TA) systems in picocyanobacterial genomes. TA systems are present in nearly all bacteria and archaea and are involved in cell growth regulation in response to environmental stresses. However, little is known about the presence and complexity of TA systems in picocyanobacteria.

By querying 77 complete genomes of freshwater, estuarine, coastal and ocean picocyanobacteria, Type II TA systems (the most well studied TA family) were predicted in 27 of 33 (81%) *Synechococcus* strains, but no type II TA genes were predicted in any of the 38 *Prochlorococcus* strains. The number of TA pairs varies from 0 to 80 in *Synechococcus* strains, with a trend for more type II TA systems being predicted in larger genomes. A linear correlation between the genome size and the number of putative TA systems in both coastal and freshwater *Synechococcus* was established. In general, open ocean *Synechococcus* contain no or few TA systems, while coastal and freshwater *Synechococcus* contain more TA systems. The type II TA systems inhibit microbial translation *via* ribonucleases and allow cells to enter the "dormant" stage in adverse environments. Inheritance of more TA genes in freshwater and coastal *Synechococcus* could confer a recoverable persister state which would be an important mechanism to survive in variable environments.

Different genotypes of *Synechococcus* are present in the Chesapeake Bay in winter and summer. Winter isolates of *Synechococcus* have shown high tolerance to cold conditions and other stressors. To explore their potential genetic capability, complete genomes of five

representative winter *Synechococcus* strains CBW1002, CBW1004, CBW1006, CBW1107, and CBW1108 were fully sequenced. These five winter strains share many homologs that are unique to them and not shared with pelagic *Synechococcus.* Winter *Synechococcus* genomes are enriched with particular desaturases, chaperones, and transposases. Similar amino acid sequences and annotated features were not found in distantly related *Synechococcus* from Subcluster 5.1. These shared genomic features between the winter strains imply that maintaining membrane fluidity, protein stability, and genomic plasticity are important to cold adaption of *Synechococcus.*

The winter strains also contain genes that are not traditionally considered with the canonical bacterial cold shock response. They contain a particularly high abundance of Type II TA pairs with complex association networks. They feature promiscuous toxins, like VapC, that pair with multiple antitoxins, which support the mix and match hypothesis. Winter strains also contain more monogamous toxins, such as BrnT, which tend to pair with their traditionally named antitoxin, BrnA. Expression of certain TA transcripts in response to environmental stress has been observed in the model strain CB0101, and the activity of one TA pair in CB0101 for growth arrest has been experimentally confirmed *via* heterologous expression in *E. coli*. My thesis work has identified interesting genetic systems related to niche partitioning of picocyanobacteria, particularly among the Chesapeake Bay *Synechococcus*. Future studies are paramount to understand the functional role of TA systems, desaturases, chaperons, and transposases of picocyanobacteria under various environmental stressors.

ABUNDANCE AND DIVERSITY OF TOXIN-ANTITOXIN SYSTEMS IN *SYNECHOCOCCUS* AND THEIR ECOLOGICAL IMPLICATIONS

by

Daniel Fucich

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Feng Chen, Chair
Professor Russell Hill
Professor Dail Laughinghouse
Professor Yantao Li
Professor Allen Place

Dedication

My family, for giving me life, unconditional support, love, and opportunity.

Acknowledgements

First, I would not have the opportunity to write these words without the opportunity from Feng Chen, PhD. He has been an incredible advisor from the transition from prospective student to candidate to graduate. I have appreciated his guidance through my academic process of formulating poignant scientific questions, experimental design, and academic writing.

I would like to thank my committee for their time, attentive nature, relevant research suggestions, and stimulating inquisition throughout the doctoral journey.

I would like to thank Mrs. Carole Ratcliffe, Nick Hammond, and the Philip E. and Carole R. Ratcliffe Foundation for supporting my research and career development *via* the Ratcliffe Environmental Entrepreneurs Fellowship (REEF) program. With this opportunity I have been able to learn new skills, realize new possibilities and create jobs in the Baltimore, Maryland region.

I would like to thank my lab members Pervaiz Ali, David Marsan, Yuanchao Zhan, Mengqi Sun, Ana Sosa, Yufeng Jia, and Hualong Wang for help on projects, support through graduate school, and surviving international travel.

An explicit thank you to other scientists who have specific contributions to this dissertation. These contributions include: Tsvetan Bachvaroff for attention and brainstorming around bioinformatic questions, methodology, and suggestions on experimental design. Ana Sosa for her work on heterologous gene expression of *Synechococcus* toxin-antitoxin systems in *E. coli* and subsequent growth curves in Chapter 4. David Marsan for initial discovery of toxin-

antitoxin systems in *Synechococcus* CB0101 and transcriptomic data during oxidative stress response.

Finally, none of this would be possible without the unconditional support and love from my family. Thank you, Mary Ann, William, Lisa, Bill, Colleen, and Brendan. You have provided a model toward which I can orient my life.

TABLE OF CONTENTS

Chapter I: Introduction

Marine Picocyanobacteria

Small unicellular cyanobacteria (<3 µm), defined as picocyanobacteria, are widely distributed in the marine environment and make up a significant portion of phytoplankton biomass in the ocean (Holt et al., 1994; P. W. Johnson & Sieburth, 1979; Waterbury, 1986). In most of the world's oceans, picocyanobacteria are principally comprised of *Prochlorococcus* and *Synechococcus*. Together, these two cyanobacterial genera can contribute about 25% of carbon fixation via photosynthesis or net primary productivity in the ocean (Flombaum et al., 2013). *Prochlorococcus* has a relatively smaller cell size compared to *Synechococcus* (Morel et al., 1993). While *Prochlorococcus* and *Synechococcus* can co-occur in the ocean, they have differential distribution patterns. The distribution of *Prochlorococcus* is largely constrained to pelagic environments between the 40ºN and 40ºS latitudinal transects where temperature is usually above 10ºC (Buck et al., 1996; Z. I. Johnson, Zinser, Coe, Mcnulty, et al., 2006). *Prochlorococcus* proliferates in oligotrophic oceans with relatively warm temperature, while *Synechococcus* is ubiquitous (Mackey et al., 2017) and more prevalent in nutrient rich coastal areas (Frédéric Partensky et al., 1999). In locations where they cohabitate, *Prochlorococcus* generally has a higher cell concentration than *Synechococcus* but is limited to open ocean environments where the two genera are endemic. *Synechococcus* transcends not only latitude and temperature (Zwirglmaier et al., 2008), but a wide variety of environmental conditions such as salinity, turbidity, and many other environmental factors (Callieri, 2008; Callieri & Stockner, 2002). Because of their abundance in such vast oceanic environments and contributions to global primary production, extensive studies have focused on marine picocyanobacteria in open ocean environments over the past 30 years (Coleman & Chisholm, 2007; Flombaum et al., 2013;

Hess, 2004; Frédéric Partensky et al., 1999; D J Scanlan et al., 2009; David J. Scanlan, 2012; David J. Scanlan & West, 2002) . Niche partitioning of *Prochlorococcus* in the ocean (both horizontally and vertically) is better understood compared to *Synechococcus* which has a much wider distribution from coastal to open ocean (Scanlan 2012). The genus *Synechococcus* also includes many species or strains present in freshwater and estuarine ecosystems. When the broad group of *Synechococcus* (from freshwater to marine habitats) is considered, much less is known for taxonomy, diversity, genomics, and niche partitioning of *Synechococcus* compared to *Prochlorococcus*.

Synechococcus are the dominant picocyanobacteria in estuarine and coastal waters, as they seem to have the capability to adapt and survive in diverse environments. Due to their global distribution and relatively larger cell size, *Synechococcus* can contribute more CO_2 fixation than *Prochlorococcus* (Jardillier et al., 2010). When considering hourly carbon fixation rates, estimations suggest that *Synechococcus* and *Prochlorococcus* contribute 16.7% and 8.5% to global ocean primary productivity, respectively (Flombaum et al., 2013). The abundance of marine picocyanobacteria appear to be positively correlated with water temperature. Increasing water temperature over time may make cyanobacteria more competitive with other phytoplankton. It has been predicted that by the end of 21st century, cellular concentrations of *Prochlorococcus* and *Synechococcus* are predicted to increase 29% and 14%, respectively, in the global ocean.

Pigmentation differentiates *Synechococcus* from *Prochlorococcus* (Morel et al., 1993). *Prochlorococcus* contains divinyl chlorophyll *a* and *b* derivitives (Goericke & Repeta, 1992), and *Synechococcus* contains the primary pigment “true” chlorophyll *a*, along with a range of

phycobilisomes including phycocyanin (PC) and phycoerythrin (PE) (Wood et al., 1985). Phycobiliproteins allow *Synechococcus* to have a variety of pigmentation profiles to utilize different light spectrums (Figure 1.1) (Six et al., 2007; Waterbury, 1986). PC- and PE-rich *Synechococcus* can be distinguished and quantified under an epifluorescence microscopy (Figure 1.2) due to their differential excitation and emission profiles (Wood et al., 1985). Different pigmentations enable the use of flow cytometry to count *Prochlorococcus* and *Synechococcus* in aquatic environments, especially in the open ocean. Rapid enumeration of picocyanobacteria with flow cytometry improves cell identification because it allows for high throughput and instant pigment identification (Buck et al., 1996). Marine *Synechococcus* and *Prochlorococcus* are well studied and characterized due to their early discovery and ease of accurate enumeration (Coleman & Chisholm, 2007; Flombaum et al., 2013; Hess, 2004; Frédéric Partensky et al., 1999; David J. Scanlan, 2012; David J. Scanlan & West, 2002).

Genetic diversity of picocyanobacteria and their niche adaptation.

As molecular techniques developed, this ecological information was married with genomic data (D J Scanlan et al., 2009) to develop a multidimensional understanding of picocyanobacterial diversity in the world's oceans. A high degree of diversity exists in both the *Synechococcus* and *Prochlorococcus* groups. Dozens of clades have been identified using a variety of different methods. Novel clades were regularly found in the 2010's and these clades have differential distributions in the world's oceans that have been studied using many different genetic markers (i.e. the ITS, 16S rRNA, *narB*, *ntcA*, and *rpoC1* loci) (Ahlgren & Rocap, 2006, 2012; Huang et al., 2011). Due to their ecological significance, work has been done to understand the genetic

diversity and the distribution pattern of picocyanobacteria in the ocean. Unfortunately, the same attention was not granted to estuarine *Synechococcus*.

Ecological Significance of Estuarine Synechococcus

Estuaries are unique environments which connect freshwater river flow and saline coastal water. The mixing of freshwater and saltwater creates strong environmental gradients and estuaries are characterized by shifts in abiotic conditions like turbidity, temperature, and salinity. Estuaries and coastal waters normally contain higher nutrient concentrations compared to offshore waters due to the immediate impact of terrestrial activities as well as precipitation *via* rain and snowfall (Herbert, 1999). Estuaries vary widely in physical and chemical conditions, community composition (Brunet & Lizon, 2003), and biogeochemical cycling (J. E. Cloern et al., 2014). Carbon fixation by phytoplankton is a vital carbon source in estuarine systems which can influence productivity at higher trophic levels (J. E. Cloern et al., 2014). Phytoplankton which include picocyanobacteria are central to the biogeochemical activity in estuarine systems (James E. Cloern & Dufford, 2005). In estuarine systems, *Synechococcus* is the dominant form of picocyanobacteria, while *Prochlorococcus* are either in low abundance or undetectable (Moore et al., 2007).

Compared to pelagic picocyanobacteria, we know much less about estuarine picocyanobacteria. Most of early studies focused on abundance, growth, productivity, spatial and temporal distribution of picocyanobacteria in the estuarine environment. In Kiel Bight, picocyanobacteria (namely *Synechococcus*) are vital to the phytoplankton community, especially in the summer. Picocyanobacteria abundance peaks at 1.4-2.6 x 10^8 cells l^{-1} in the

summer and account for up to 97% of all autotrophic picoplankton carbon (Jochem, 1988). Carpenter et al., (1988) reported that in the Long Island Sound, active division and growth rate for *Synechococcus* was highest in the summer. *Synechococcus* has been studied in many different estuarine environments each with unique hydrological properties and variable *Synechococcus* abundance and distribution patterns (J E Cloern et al., 2016; Fortunato & Crump, 2011). While no estuary is identical, it is evident that *Synechococcus* is ubiquitous in estuarine environments around the globe. The Chesapeake Bay was one focus of early research on the ecological distribution of *Synechococcus* in estuarine ecosystems.

Early studies of Picocyanobacteria in the Chesapeake Bay

The Chesapeake Bay (CB) is the largest estuary in the United States, with 16 million people living and impacting the watershed which creates eutrophic conditions (Harding et al., 2016). Phytoplankton biomass in the Bay has been increasing over the second half of the 20th century due to eutrophication (Harding and Perry 1997). Picocyanobacteria were considered as part of picophytoplankton in the earlier studies of CB phytoplankton. In the Lower York river, a CB sub-estuary, picocyanobacteria accounted for 51% of the picophytoplankton biomass, and had the highest abundance at 7.2 x 10^5 cells ml^{-1} in September (Ray et al., 1989). Autotrophic picophytoplankton cell concentrations were lowest in the winter at 7.36 x 10^6 cells L^{-1} and highest in the summer at 9.28 x 10^8 cells L^{-1} (Affronti, 1990) (in this case, autotrophic picophytoplankton includes both prokaryotic and eukaryotic cells). This seasonal pattern is mirrored by picoplankton primary production being highest in July at 55.6% and lowest at 2.3% in January (Affronti & Marshall, 1994). These studies suggest that picocyanobacteria have

higher cell concentrations and a larger contribution to primary production in the summer than in the winter. The seasonal pattern of picocyanobacteria observed in the CB are also seen in other estuarine systems (James E. Cloern & Jassby, 2010). Early studies could also differentiate pigmentation profiles among *Synechococcus* strains.

The ratio of PC-rich and PE-rich *Synechococcus* strains has been used to investigate the distribution of different pigment types of CB picocyanobacteria over time and space. It has been reported that the PC-rich type was 8 times more prevalent than the PE-rich type (Ray et al., 1989); a striking contrast to the open ocean where PE-rich *Synechococcus* are prominent (Campbell & Carpenter, 1987). Furthermore, PC-rich *Synechococcus* dominated in surface waters during the summer, they accounted for 73.8% of surface picophytoplankton. In the winter, benthic PE-rich *Synechococcus* were more productive and comprised 65.4% of benthic picophytoplankton (Affronti & Marshall, 1993). These early studies of CB picocyanobacteria mainly focused on the spatiotemporal distribution of *Synechococcus* based on their abundance, pigmentation, biomass, and productivity. Little was known about ecophysiology and genetic diversity of *Synechococcus* in the Bay, as no CB *Synechococcus* were isolated and characterized until 2004.

Unique and diverse Synechococcus isolated from the Chesapeake Bay

To better understand the taxonomy and physiology of CB picocyanobacteria, it is necessary to isolate and cultivate cyanobacteria from the Bay. Thirteen picocyanobacteria were isolated from various locations in the Chesapeake Bay, including Baltimore Inner Harbor, middle, and lower Bay, mostly during the summer months (F. Chen et al., 2004). Microscopic

identification of these isolates showed that they are unicellular cyanobacteria with coccoid or rod shape and cell size between 1-3 μm. Among these 13 isolates, seven strains are phycocyanin-rich and six are phycoerythrin-rich *Synechococcus*. Five motile strains were also identified in these *Synechococcus* cultures. During the course of isolation, it was estimated that 80-90% of colonies recovered from the Baltimore Inner Harbor and the upper Chesapeake Bay were green picocyanobacteria, while the percentile of green colonies decreased to 56-65% at the mouth of the Bay. The salinity is usually in the range of 5-10 ppt in the Inner Harbor, and 20-30 ppt in the lower bay. These CB *Synechococcus* isolates are able to grow in culture media with a wide range of salinity (0-30 ppt). In contrast, many coastal and open ocean *Synechococcus* strains do not grow at lower salinities (F. Chen et al., 2004). The trend of abundant PC-rich *Synechococcus* at lower salinity, or abundant PE-rich *Synechococcus* at higher salinity was also evident based on the enumeration *via* epifluorescence microscopy (Figure 1.2) (F. Chen et al., 2004). This was the first study focused on isolating and characterizing picocyanobacteria from estuarine environments. The availability of these CB *Synechococcus* enables the further phylogenetic and genomic studies. Many CB cyanophages were also isolated and characterized when these cultures became available (Wang and Chen 2008).

The phylogenetic position of these CB *Synechococcus* was first illustrated based on the ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) large subunit gene (*rbcL*) sequences (F. Chen et al., 2004). The *rbcL* phylogeny showed that the Chesapeake Bay contains diverse *Synechococcus* and the vast majority of CB *Synechococcus* do not cluster with marine cluster A *Synechococcus,* which contains coastal and oceanic *Synechococcus*. In the early 2000's, the taxonomy of marine *Synechococcus* mainly refers to the three major clusters (marine

cluster A, B and C) (Herdman et al., 2001; Waterbury, 1986). At the time, marine cluster A contained many *Synechococcus* strains isolated from coastal and open oceans, while marine cluster B and C only contained a few marine *Synechococcus* strains. Compared to marine cluster A and C, marine cluster B had not been well characterized. By including CB *Synechococcus* strains in the *rbcL* tree, it becomes clear that the majority of CB *Synechococcus* isolates belong to marine cluster B (F. Chen et al., 2004). The *rbcL* amino acid phylogeny showed that these estuarine *Synechococcus* are more closely related to marine *Synechococcus* than to freshwater *Synechococcus*, suggesting a close relationship between estuarine and marine *Synechococcus*. Several other interesting findings related to the clustering of *Synechococcus* include: 1) Marine cluster B can contain PE-rich *Synechococcus* (the previous systematics only included PC-rich *Synechococcus*); 2) motility does not necessarily cluster *Synechococcus* together; 3) members of marine cluster B including WH8007 belong to the Form IA *rbcL* type, not the Form IB *rbcL* type (Pichard et al., 1997); and 4) the *rbcL* genotypes varies dramatically from the upper to lower bay.

The study of Chen et al. 2004 suggests that certain phenotypic features like pigmentation and motility may no longer be valid for classifying or clustering *Synechococcus* due to their diverse and polyphyletic nature. Both PE and PC rich *Synechococcus* can be clustered closely within the same clade or subcluster, and there is no clear separation between PC and PE rich *Synechococcus* strains. This incongruence was also confirmed by later studies in the Baltic Sea (T. Haverkamp et al., 2008; T. H. A. Haverkamp et al., 2009). Characterization of CB picocyanobacterial cultures not only provides morphological and physiological information, but also sheds light on the diversity and genetic nature of estuarine *Synechococcus*. This study

first confirmed that the vast majority of picocyanobacteria in the Bay are *Synechococcus* and they most likely have a marine rather than freshwater origin based on the *rbcL* phylogeny. The Chesapeake Bay provides a unique environment for marine cluster B to thrive and adapt.

In a later study, the MC-B group of *Synechococcus* was re-defined as subcluster 5.2 *Synechococcus* based on the 16S-23S rRNA internal transcribed spacer (ITS) sequences (F. Chen et al., 2006b). The ITS is a non-coding RNA region and is less conserved than 16S or 23S rRNA. The less conserved nature of the ITS allows for a higher resolution between closely related strains. Based on the ITS phylogeny, many Chesapeake Bay *Synechococcus* isolates fell into two subgroups, CB4 and CB5, which form subcluster 5.2. *Synechococcus* subcluster 5.2 prevailed in the upper bay, while subcluster 5.1 or marine cluster A *Synechococcus* was prevalent in the lower bay. This study laid the foundation for the phylogenetic position of subcluster 5.2 *Synechococcus*. Many later studies in estuarine and coastal environments identified even more diverse members of subcluster 5.2. For example, the presence of subcluster 5.2 *Synechococcus* was also reported in different environments including the East China Sea (Choi & Noh, 2009), the coastal estuary of Hong Kong (X. Xia, Vidyarathna, et al., 2015), the Baltic Sea (Larsson et al., 2014), the Northern South China Sea (X. Xia, Guo, et al., 2015), the Bering Sea and Chukchi Sea (Huang et al., 2011), and the Massachusetts Coastal Observatory (Hunter-Cevera et al., 2016). It is striking to see that subcluster 5.2 *Synechococcus* (CB5 clade) comprise the vast majority (ca. 80%) of picocyanobacteria in the Chukchi Sea at the highest latitude subzero waters with temperatures around 0°C (Huang et al., 2011). The Chukchi Sea is a shallow shelf sea (ca. 50m deep) which can be influenced by freshwater from ice melt. Due the low picocyanobacterial abundance (<1000 cells per ml) in winter Chesapeake Bay or in a cold region like the polar

ocean, it is difficult to detect and obtain picocyanobacterial sequences using metagenomics or from a 16S rRNA gene clone library. The picocyanobacteria-specific PCR primers based on the ITS region should be considered under these conditions (Huang et al., 2011).

Later studies in the Chesapeake Bay found that *Synechococcus* populations present in winter are distinct from those in the summer (Cai et al., 2010). Environmental clones recovered from the upper, middle, and lower Bay in February 2005 showed that two new clades CB6 and CB7 are present in the winter. CB6 is more closely related to marine *Synechococcus*, while CB7 clusters with *Cyanobium,* which is commonly found in the freshwater system. It has been known that *Synechococcus* cell counts in winter are usually 2-3 orders of magnitude lower than those in summer. A good linear correlation between *Synechococcus* abundance and water temperature in the Bay was reported based on a five-year survey of picocyanobacteria (K. Wang et al., 2011). *Synechococcus* abundance in the Bay often exceeds 1 million cells per milliliter in the summer 'blooming' season and can be less than 100 cells per ml in winter. The seasonal pattern of CB *Synechococcus* is clear and annually recurring. Co-variation between *Synechococcus* and cyanophages were also observed suggesting that viral infection is an important factor that can influence the population dynamics of *Synechococcus*.

Winter Chesapeake Bay isolates

Early studies have shown that distinct genotypes of *Synechococcus* are present in the Chesapeake Bay during winter. The Chesapeake Bay *Synechococcus* isolates reported by Chen et al. in 2004 and 2006 were mainly recovered during warmer months. A total of 17 picocyanobacterial strains were isolated from the Baltimore during the winter season (Xu et al.,

2015). These winter cultures were isolated between December 2010 and February 2011, and the water temperature fluctuated between 2 and 8°C during this period. Seven PC-rich and ten PE-rich strains were recovered. The winter isolates are colorful and exhibit blue-green, yellowish, brown, and pink colors, suggesting a wide chromatic adaptation by winter picocyanobacteria. The winter *Synechococcus* strains belong to five distinct phylogenetic clusters, which differ from *Synechococcus* strains that are dominant during summer months (i.e. subcluster 5.2 *Synechococcus*) (Figure 1.3). These five lineages include the Bornholm Sea cluster (named after the 50 m halocline Bornholm Basin located in the Baltic Sea (Jakobsen, 1996)), subalpine cluster II, CB7 cluster, and two other novel clusters. Interestingly, many winter CB isolates are closely related to picocyanobacteria isolated from Baltic Sea, subalpine waters, and Arctic Sea, suggesting a common origin of cold-adapted *Synechococcus*. The Bornholm Sea cluster was first established when picocyanobacteria were isolated and identified from the Baltic Sea (Ernst et al., 2003a). This cluster only contains the strains isolated from the Baltic Sea at two sampling sites where salinity was 7 and 9 ppt, respectively. Noticeably, none of the Baltic Sea picocyanobacteria belong to *Synechococcus* subcluster 5.2 which contains many CB summer isolates. The lack of subcluster 5.2 in the Baltic Sea collection could be related to the difference in cultivation method. A culture medium comprised of 1 part ASN III and 3 parts BG11 with salinity of ~6 ppt was used for isolation of picocyanobacteria in the Baltic Sea (Ernst et al., 2003a), while SN15, a modified lower nutrient cyanobacterial medium, was used for isolation of picocyanobacteria in the Chesapeake Bay (F. Chen et al., 2004; Xu et al., 2015). A later study based on metagenomics showed the presence of many contigs related to subcluster 5.2 *Synechococcus* in the Baltic Sea (Larsson et al., 2014). The Baltic Sea is one of world's largest

estuaries with strong salinity gradient, but it differs from the Chesapeake Bay in many aspects. For example, the Baltic Sea has a larger water volume, deeper water column, much longer residence time (25 years), and is located in a colder climate (higher latitude) compared to the Chesapeake Bay. Despite this, picocyanobacteria in the Chesapeake Bay and Baltic Sea share many common lineages.

The effect of temperature (4, 10, 15, 23, 25, or 28°C) on the growth rate of winter and summer CB *Synechococcus* strains, coastal, and open ocean *Synechococcus* strains was compared. Save for CBW1108, all winter *Synechococcus* (8 representative strains) were able to grow slowly at 4 and 10°C, but none of the coastal and open ocean *Synechococcus* grew at these temperatures. The winter CB isolates were able to maintain slow growth or prolonged dormancy at 4°C and resume normal growth at room temperature. This phenomenon was not observed in open ocean *Synechococcus*. Interestingly, several *Synechococcus* strains in the Bornholm Sea cluster exhibited various cell lengths during exponential growth. For example, the cell length of CBW1112 can vary from 1.21 to 21 μm during exponential growth at 23°C. Many winter CB strains displayed a 2-3 fold cell enlargement during prolonged exposure to 4 °C. Cell enlargement or increase in cell volume (not elongation) under the cold condition has been reported in freshwater *Synechococcus* like cyanobacteria (Jezberová & Komárková, 2007). Cell elongation (up to 50 fold) was found later in freshwater *Synechococcus* PCC 7942 during the stationary phase at room temperature, while elongation was caused by phosphorus limitation (Goclaw-Binder et al., 2012).

Isolation and characterization of CB picocyanobacteria from both warm and cold seasons sheds light on taxonomy, physiology, and genetic diversity of estuarine *Synechococcus*.

The Bay *Synechococcus* spp. exhibit wider range of salt and thermal tolerance compared to open ocean counterparts. It appears that subcluster 5.2 *Synechococcus* dominates in summer, while the Bornholm Sea cluster *Synechococcus* prevails in winter, in the Chesapeake Bay. Early studies only included clade CB4 and CB5 into subcluster 5.2 (Ahlgren & Rocap, 2012; F. Chen et al., 2006b; Huang et al., 2011), while the number of clades in subcluster 5.2 is much lower compared to that in subcluster 5.1. Other picocyanobacterial clusters such as the Bornholm Sea cluster, subalpine clusters, CB7 cluster, *Cyanobium gracile* cluster are close to those in subcluster 5.2 (Huang et al., 2011; Xu et al., 2015). The broader question is raised: Should they be considered as clades in subcluster 5.2? Or, should subcluster 5.2 be extended to include other closely related picocyanobacteria?

A broader subcluster 5.2?

When more genomes of freshwater or non-marine picocyanobacteria were sequenced, phylogenomic trees showed that freshwater, brackish or estuarine, and some marine picocyanobacteria can form a monophyletic subcluster which is broader than previously defined subcluster 5.2 (Di Cesare et al., 2018; Patricia Sánchez-Baracaldo et al., 2019). Coutinho et al., (2016) proposed a new genus name *Parasynechococcus* which includes *Synechococcus* strains isolated from estuarine, coastal, and oceanic *Synechococcus* based on a phylogenomic reconstruction. The original genus *Synechococcus* only includes freshwater strains according to the classification system defined by Coutinho et al. in 2016. This system abandons the division of subcluster 5.1, 5.2 and 5.3, and can generate some confusion as subcluster 5.1 has been well defined for marine *Synechococcus*, especially for coastal and open ocean *Synechococcus*.

Recently, a freshwater cyanobacterium, *Vulcanococcus limneticus* sp. nov. (formerly *Synechococcus* LL) isolated from a volcanic lake in central Italy, was found closely related to CB4 and CB5 groups in subcluster 5.2 (Di Cesare et al., 2018). According to the recent phylogenomic study, it appears that subcluster 5.2 can now be extended to include even more cyanobacterial members (i.e. *Cyanobium*) from diverse habitats such as freshwater, brackish, and coastal water. The broader subcluster 5.2 was further confirmed by a most recent phylogenomic study based on 136 cyanobacterial proteins (Patricia Sánchez-Baracaldo et al., 2019). These two latest studies strongly support that subcluster 5.2 should be extended to include freshwater picocyanobacteria (i.e. *Cyanobium* spp.) and estuarine *Synechococcus*. While the broader subcluster 5.2 was proposed, no further studies were conducted to compare with the phylogeny based on the 16S rRNA gene or ITS region. These two gene markers have been widely used to study the phylogenetic relationship and genetic diversity of picocyanobacteria in various aquatic environments (Callieri et al., 2013; Ernst et al., 2003a; T. H. A. Haverkamp et al., 2009; Jing et al., 2009; A. Wilmotte et al., 2017). It will be interesting to see if the gene marker-based phylogeny also supports the broader subcluster 5.2 defined by core genomes of picocyanobacteria. We are currently evaluating this newly defined broader subcluster 5.2 based on the 16S rRNA gene and ITS sequences.

In summary, isolation of CB *Synechococcus* using culture medium with adjusted salinity recovered many indigenous estuarine species of *Synechococcus* from different locations in the Chesapeake Bay and from different seasons. Physiological studies showed that these estuarine *Synechococcus* spp. have a higher tolerance to environmental stressors compared to coastal and open ocean *Synechococcus* spp. Phylogenetic analysis demonstrated that the Chesapeake

Bay contains diverse and unique *Synechococcus* distinct from well-studied open ocean *Synechococcus*. It is believed that strong environmental gradients in the Chesapeake Bay select for estuarine *Synechococcus* with diverse pigment types, genotypes, and ecotypes. However, little is known about the mechanism used by estuarine *Synechococcus* to cope with such a dynamic ecosystem.

Genome sequencing of Synechococcus sp. CB0101

Synechococcus strain CB0101 has been used as a model strain for Chesapeake Bay picocyanobacteria as its genotype is commonly found in the Bay. Strain CB0101 has been used to understand ecophysiology of estuarine picocyanobacteria and to isolate cyanobacterial viruses from the Chesapeake Bay (K. Wang et al., 2011). In physiological growth rate evaluations, estuarine *Synechococcus* CB0101 was more resilient than the coastal *Synechococcus* WH7803 and open-ocean *Synechococcus* WH7805 in variable growth conditions with wide ranges of salinity, temperature, nutrient, and metal concentrations (D. W. Marsan, 2016). CB0101 had a higher growth rate than the coastal and open-ocean strains in a wider array of nutrient, salinity, and temperature conditions. Growth rates of CB0101 show that it is well adapted to grow in low light, nutrient replete conditions, exemplary of the turbid conditions found in its endemic Chesapeake Bay. These physiological responses suggest that *Synechococcus* CB0101 has a genetic capacity to tolerate more stressful conditions than its coastal and open ocean counterparts.

The draft genome of CB0101 was reported by Marsan et al. in 2014, representing the first genome sequence for subcluster 5.2 *Synechococcus* isolated from the Chesapeake Bay. The

genome of CB0101 contains many genes related to transport, store, utilize, and export metals, especially copper, nickel, cobalt, and magnesium, indicating its extensive capacity to sense and respond to changes in the Chesapeake Bay (D. Marsan et al., 2014). The complete genome of *Synechococcus* CB0101 was sequenced (Fucich et al., 2019). As a representative *Synechococcus* strain for the estuarine environment, the genome sequence of CB0101 can be compared to the genome sequences of *Synechococcus* isolated from coastal and oceanic waters to understand the genetic features involved in such a niche adaptation. Unique genetic features may be present in estuarine *Synechococcus* strains that are not found in marine, coastal, and freshwater strains. This may give estuarine strains a specialized ability to adapt and cope with environmental conditions unique to turbulent estuarine environments. More *Synechococcus* strains from estuarine environments would have to be isolated and sequenced to investigate the "pangenome" of estuary *Synechococcus* which could provide insight on genes that are most important for estuarine life.

Toxin-antitoxin genes are present in marine Synechococcus

Toxin-Antitoxin (TA) systems are small genetic elements that are traditionally comprised up of a two-component system: a toxin, which can act on cellular targets to arrest growth, and a cognate antitoxin which negates its toxin (Unterholzner et al., 2014a). TA systems are nearly ubiquitous in bacteria and archaea and are known to regulate cell growth in response to environmental stressors (Page and Peti 2016). Toxin-Antitoxin systems are classified into four main types I-IV (Harms et al., 2018a) which can be differentiated by their method of action, or cellular target, and molecular type of the antitoxin; either RNAs or amino acids (Figure 1.4).

Toxin-antitoxin types affect many central cellular actions including translation, replication, cellular membrane integrity and biosynthesis, among others (Unterholzner et al., 2014a). These outcomes can be permanent or recoverable depending on their method of action.

Type II TA systems are the most well studied and experimentally verified TA systems. These systems are characterized by a protein-protein toxin-antitoxin system where the antitoxin nullifies the action of the toxin through direct interaction (Pandey & Gerdes, 2005). Type II TA systems, such as RelE/RelB, can act as effectors of bacterial persister cells and are often associated with the activation of proteases, such as Lon, for antitoxin degradation. Type II toxins often interrupt translation by endonuclease activity, either in a ribosomal-dependent or ribosomal-independent manner. In this way, they temporarily disrupt translation without effecting cellular death. When the toxin and antitoxin are at equilibrium, the antitoxin negates the toxin, and the toxin is rendered inactive; cellular activity continues as normal (Figure 1.5). However, in a stressful environment where the secondary messenger alarmone (p)ppGpp is present and Lon protease is activated, the antitoxin is degraded and the toxin is free to act on its cellular targets (Maisonneuve & Gerdes, 2014). This process results in reversible growth arrest which is advantageous in conditions of interim cellular stress.

However, little is known about the presence of TA systems in cyanobacteria because current research is narrowly focused on few, scattered model organisms using methods lacking a repeatable and systematic approach. Prediction software can quickly become defunct as support may end abruptly. The source database may lie stagnant (Sevin & Barloy-Hubler, 2007) in contrast to the ever-increasing knowledge of TA families, their methods of action, and conserved domains (Figure 1.4) (Harms et al., 2018a).

TA systems have been predicted in freshwater cyanobacteria including *Microcystis aeruginosa* (Makarova et al., 2009), *Synechocystis* PCC6803 (Kaneko et al., 2003), and on the pANL plasmid in *Synechococcus* PCC7942 (Y. Chen et al., 2011). The scope of these TA genes was small with only a few cyanobacterial genomes being available or included. TA families like VapB, VapC and PemK were reported in these well studied freshwater cyanobacterial strains, however no uniform, systematic survey of TA genes was performed at the genomic level for all sequenced strains of marine *Synechococcus*.

TA systems in CB0101

The first 7 chromosomal TA systems in marine *Synechococcus* were described in the estuarine *Synechococcus* strain CB0101 (D. Marsan et al., 2017). These include some common TA families such as *relE/relB* and *vapC/vapB* (Figure 1.6). CB0101, isolated from the Chesapeake Bay, belongs to *Synechococcus* subcluster 5.2 (F. Chen et al., 2006b). *In vivo* transcriptomics of CB0101 reveals a tight coupling between the upregulation of particular toxins, such as relE, with simulated stress conditions, suggesting that TA systems could be an important genetic feature for estuarine *Synechococcus* to adapt to a highly variable environment like the Chesapeake Bay (D. Marsan et al., 2017).

More recently, the search for chromosomal TA systems in CB0101 and other picocyanobacteria continued, using improved methodology and ever-expanding subject databases (Shao et al., 2011; Xie et al., 2018). In CB0101, the original 7 predicted chromosomal TA pairs were confirmed, and 14 more potential TA pairs were predicted. Of the original 7 predicted TA pairs, a correction to a misidentified toxin and antitoxin pair at the locus:

gsyne_1326 and gsyne_1325 was made. These were originally annotated as the antitoxin yoeB and toxin yefM. Newer prediction methods identified gsyne_1326 as the toxin and gsyne_1325 as the antitoxin. Conserved domains in each of these open reading frames annotated as the toxin relE/parE family and PhD/yefM family, respectively. When coupled with available transcriptomic data, the statistically significant upregulation of the toxin relE (gsyne_1326) is apparent during oxidative stress simulated by nitrogen starvation and zinc toxicity (Fucich, Unpublished).

Further, new TA prediction and annotation coupled with access to available transcriptomic data suggest an active TA system at loci gsyne_2550-2551. This toxin is annotated with a conserved domain of DUF5616 which includes a PIN domain, which is the active RNase N-terminus of the vapC toxin (Rocker & Meinhart, 2016). This putative vapC toxin has a cognate open reading frame (ORF) with the conserved domain of unknown function (DUF) 433 that may act as an antidote antitoxin, as it is frequently predicted alongside PIN associated antitoxins (Makarova et al., 2009). Regardless of lack of conserved domains to known TA systems in some cases, this putative vapC at gsyne_2550 shows upregulation during all simulated stressors: nitrogen and phosphorous starvation as well as zinc toxicity (Fucich, Submitted).

TA systems in greater Synechococcus

We recently investigated complete genomes of *Synechococcus* and *Prochlorococcus* to understand the prevalence of TA systems in picocyanobacteria (Fucich and Chen, in press). Using the TAfinder software, Type II TA systems were predicted in 27 of 33 (81%)

Synechococcus strains, but none of the 38 *Prochlorococcus* strains contain TA genes. *Synechococcus* strains with larger genomes tend to contain more putative type II TA systems. The number of TA pairs varies from 0 to 42 in *Synechococcus* strains isolated from various environments (Fucich and Chen, in press). Linear correlations between the genome size and the number of putative TA systems in coastal and freshwater *Synechococcus* was established, respectively (r^2= 0.9152, p<0.00001 and r^2=0.8296, p<0.005). In general, open ocean *Synechococcus* contain no or few TA systems, while coastal and freshwater *Synechococcus* contain more TA systems. Type II TA systems inhibit microbial translation *via* ribonucleases and allow cells to enter the persister or "dormant" stage under adverse conditions. Our survey shows that TA systems are widely present in many freshwater, coastal, and estuarine *Synechococcus*. Inheritance of more TA genes in these strains could be an important mechanism for them to survive in their highly dynamic environments.

Diversity of putative TA systems in Synechococcus

Synechococcus toxin genes have more conserved protein domains than their cognate antitoxins. When considering the putative TA pairs of all 27 *Synechococcus* species with predicted TA pairs, the majority of those toxins were identified as VapC (41%), either from their direct annotation, or from their inclusion of a PilT N-terminus (PIN) domain (Fucich and Chen, 2020). This is consistent with other bacterial TA modules, as the VapBC family is the most abundant family (Robson et al., 2009). Putative antitoxin sequences were less conserved than toxins. Only 35% of antitoxin genes contained a conserved domain with a traditionally named TA system. Many conserved domains in putative antitoxin genes had generic names such as

"domains of unknown function" (DUF) or "cluster of orthologous groups" (COG). These COG's and DUF's are cryptic, so the identity of the putative TA system could be unknown. But in all observed cases, antitoxin sequences contain less conserved domains than toxins.

Interesting questions

The prevalence of TA genes in freshwater and estuarine environments and absence of TA genes in *Prochlorococcus* and open ocean *Synechococcus* imply an interesting environmental selection on the TA systems. Marine *Synechococcus* evolved from freshwater *Synechococcus* (Patricia Sánchez-Baracaldo et al., 2019). It is plausible that TA genes were lost when marine *Synechococcus* occupied the ocean. *Synechococcus* adapted to the estuarine environment could serve as excellent models to understand the evolution of TA systems in unicellular cyanobacteria. Cyanobacteria are ancient and diverse organisms, and widely distributed in nearly all aquatic habitats. Because of this nature, extensive studies have been done to understand the ecology, evolution, and molecular biology of cyanobacteria. However, little is known about the role of TA systems in the ecological adaption of cyanobacteria.

TA genes are not conserved and are involved in frequent horizontal gene transfer. These characteristics makes it difficult to study conserved domains or genes. Genomics and metagenomics are becoming extremely powerful to study the TA genes in microorganisms. Do certain network patterns exist for picocyanobacterial TA genes? Why do freshwater and estuarine *Synechococcus* carry more TA genes and marine *Synechococcus*? Which TA genes are functional or important? How do they respond to environmental stress? How do they function

in concert with other stress response systems in cyanobacteria? How do TA system coordinate with other stress response systems such as heat shock or cold shock proteins?

Recently, the full genomes of four winter CB *Synechococcus* (CBW1002, CBW1004, CBW1006, and CBW1108) have been sequenced. Phylogenomic analysis has placed these four strains into the broader subcluster 5.2 *Synechococcus* (Luo et al., unpublished data). Genomic analysis of these winter *Synechococcus* will deepen our understanding of cold adaptation of *Synechococcus*. Our preliminary analysis showed that one of winter strains contains 80 TA gene pairs, a number that is considered very high in all microorganisms. What biological features does this strain gain by carrying such a high number of TA genes?

Learning more about stress, and specifically cold stress, genes in CBW strains is important to understand their ability to survive in cold weather conditions. Cold stress response genes in the cyanobacterial model *Synechocystis* sp. PCC 6803 have been studied (Sinetova & Los, 2016). Homologs for these cold induced genes in Chesapeake Bay strains revealed that winter CB strains did not contain more cold induced genes than summer CB strains, or other marine *Synechococcus.* Similar numbers of cold induced homologs found between *Synechococystis* and other *Synechococcus* strains is surprising, as cold adapted strains were expected to have significantly more cold adapted genes than summer, or open ocean *Synechococcus*. It may be possible that the high number of TA systems in CBW strains could play a role in cold adaptation. This remains as an interesting hypothesis to test in the future. While there are many interesting questions related to TA systems in cyanobacteria, I plan to address a few important ones in my dissertation such as:

1. How common are TA systems in picocyanobacteria?
2. What are the ecological implications of the TA systems in picocyanobacteria; Is there a link between endemic habitat and TA system abundance?
3. Are TA systems important to the cold adaptation of winter *Synechococcus*?
4. Are TA systems conserved among *Synechococcus*?

Specifically, I raise three hypotheses as follows:

Hypothesis 1: Synechococcus spp. from highly variable environments, like the Chesapeake Bay, contain more toxin-antitoxin systems compared to Synechococcus spp. and Prochlorococcus spp. from relatively stable environments with streamlined genomes.

Hypothesis 2: Synechococcus strains isolated during the winter in the Chesapeake Bay harbor more stress response genes compared to the summer CB Synechococcus and other marine Synechococcus species.

Hypothesis 3: TA systems are not conserved in Synechococcus even among closely related Synechococcus strains making it impossible to identify conserved domains for the broader group of Synechococcus.

Brief outline of my dissertation chapters

Chapter I: Introduction

Chapter II: Presence of toxin-antitoxin systems in picocyanobacteria and their ecological implications

In this chapter, I searched for the presence of TA genes in all complete and publicly available *Synechococcus* and *Prochlorococcus* genomes, to gain an insight on the distribution of TA genes in picocyanobacterial genomes. The finding of TA systems in *Synechococcus* CB0101 (D. Marsan et al., 2017) is interesting and raises many new questions. There is no systematic survey to investigate how many TA genes could be found in picocyanobacteria. The strains included in this chapter represented picocyanobacteria from a wide array of environments, including freshwater, estuary, coastal and open ocean. We found that coastal, estuarine, and freshwater *Synechococcus* tend to have larger genomes and contain more TA genes compared to *Synechococcus* and *Prochlorococcus* living in the open ocean (Hypothesis 1). Interestingly, a novel correlation between genome size and putative TA systems were found. *Synechococcus* from highly variable environments contain more TA systems than *Synechococcus.* The prevalence of TA systems in *Synechococcus* from marine and non-marine environments was first reported in our study. The linear relationship between the number of TA genes and genome size has not been observed in other bacteria, at the domain or genera-specific level. We speculate that having more TA genes could be important for *Synechococcus* to adapt to more variable and even stressful conditions. This work has been published in the ISME Journal.

Chapter III: Genomic features for cold adaptation of winter Chesapeake Bay *Synechococcus*

In an earlier study, many *Synechococcus* strains were isolated from the Baltimore Inner Harbor during the winter time, and they exhibited impressive cold tolerance through

physiological testing (Xu et al., 2015). Recently, complete genomes of four representative CB winter *Synechococcus* were sequenced. To explain this phenomenon, homologs of known cold induced stress genes from *Synechocystis* PCC 6803 were queried in CB winter *Synechococcus* genomes in order to test hypothesis 2. Surprisingly, all four winter CB *Synechococcus* isolates did not contain more cold induced *Synechocystis* homologs than summer CB, and open ocean *Synechococcus.* This result suggests there could be other unique genetic elements involved in cold adaptation of winter *Synechococcus*. CBW strains contain an impressive amount of putative toxin-antitoxin systems. We hypothesize that TA systems are important for cold adaption of CB winter *Synechococcus*. Future experiments are needed to test this hypothesis.

Chapter IV: Abundance and complexity of toxin-antitoxin systems in *Synechococcus* from various aquatic environments

Chapter IV investigates the abundance, diversity, and activity of TA systems in *Synechococcus* strains isolated from the Chesapeake Bay, and other marine and freshwater environments. The winter CB *Synechococcus* genomes contain an unusually high frequency of putative TA pairs which are diverse and form complex association patterns compared to those from coastal and oceanic waters. Freshwater strains (*i.e.* PCC6307 and PCC6312) are comparable to winter CB TA systems in terms of abundance and complexity. CBW1002 and CBW1006 are in the Bornholm Sea cluster, making the similar TA profile to freshwater strains notable. Amino acid sequences from putative CB strains contain a wide variety of conserved domains. However, even sequences in the same TA family are not "conserved" at the level of marker genes where meaningful alignments could be constructed (hypothesis 3). This result suggests that unlike many house-keeping or core genes, TA genes are subject to frequent

horizontal gene transfer, which is seen in other bacteria (Leplae et al., 2011). Therefore, TA systems are not conserved at the genomic level. In most cases, it is not possible to use a conserved domain within a toxin or antitoxin gene as a genetic marker to investigate the genetic diversity of TA genes in the natural environment. One RelE toxin from CB0101 was confirmed to arrest the growth of *E. coli* through plasmid induction. However, another toxin, VapC, did not significantly arrest *E. coli* growth. These mixed results suggest that different TA pairs in *Synechococcus* may not have the same function or activity when exposed to certain stressors and environmental conditions. This was also evident when the expression of TA genes in CB0101 were examined under different stress conditions. Having a high number of putative TA genes in estuarine and freshwater *Synechococcus* is an interesting observation. We believe that TA genes play an important role in ecological adaptation of *Synechococcus*, but specific function and coordination under environmental stressors is a future focus of research.

Chapter V: Conclusion and future prospects

Figures

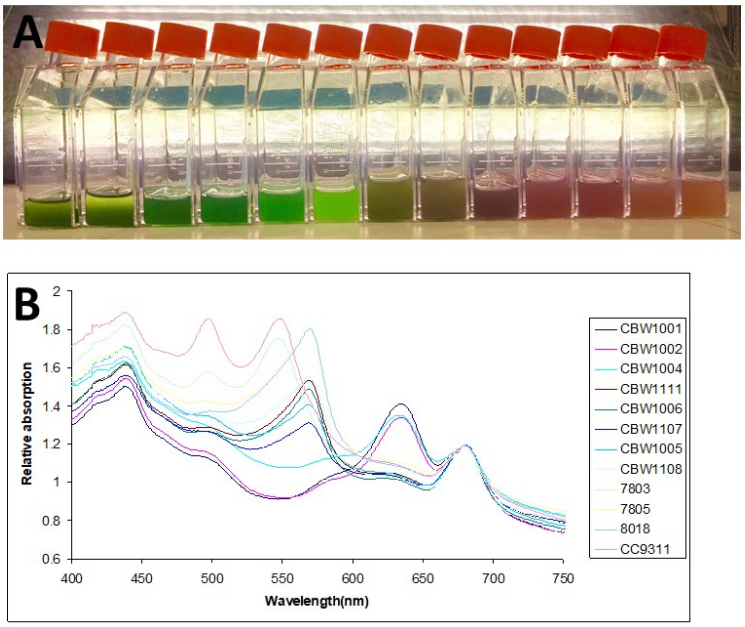


Figure 1.1. Chromatic adaptation of estuarine *Synechococcus* and their light absorption spectra. A) Accessory pigments such as phycoerythrin (PE) and phycocyanin (PC) in variable ratios result in unique phenotypes in many Chesapeake Bay *Synechococcus* strains. B) These phenotypic changes result in differential absorption spectra. When these absorption spectra are overlaid, several *Synechococcus* strains can maximize absorption between 400-700 nm (Xu et al., 2015).

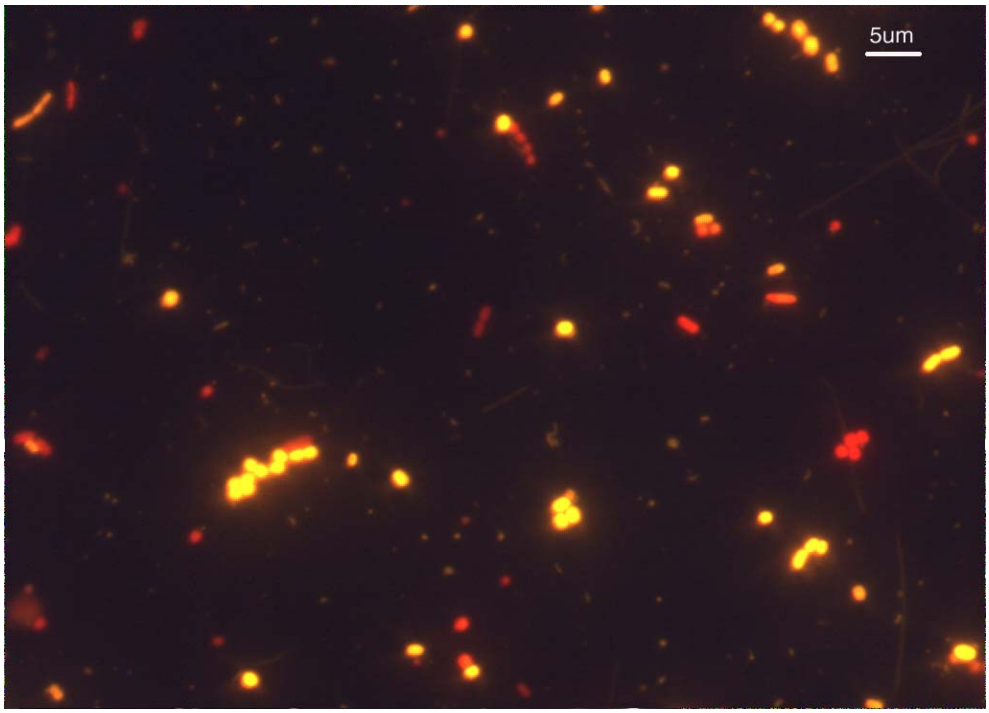


Figure 1.2. Picocyanobacterial community from the Baltimore Inner Harbor, viewed under epifluorescence microscopy. PC-rich and PE-rich strains can be differentiated as red and orange color, respectively (Courtesy of Feng Chen).

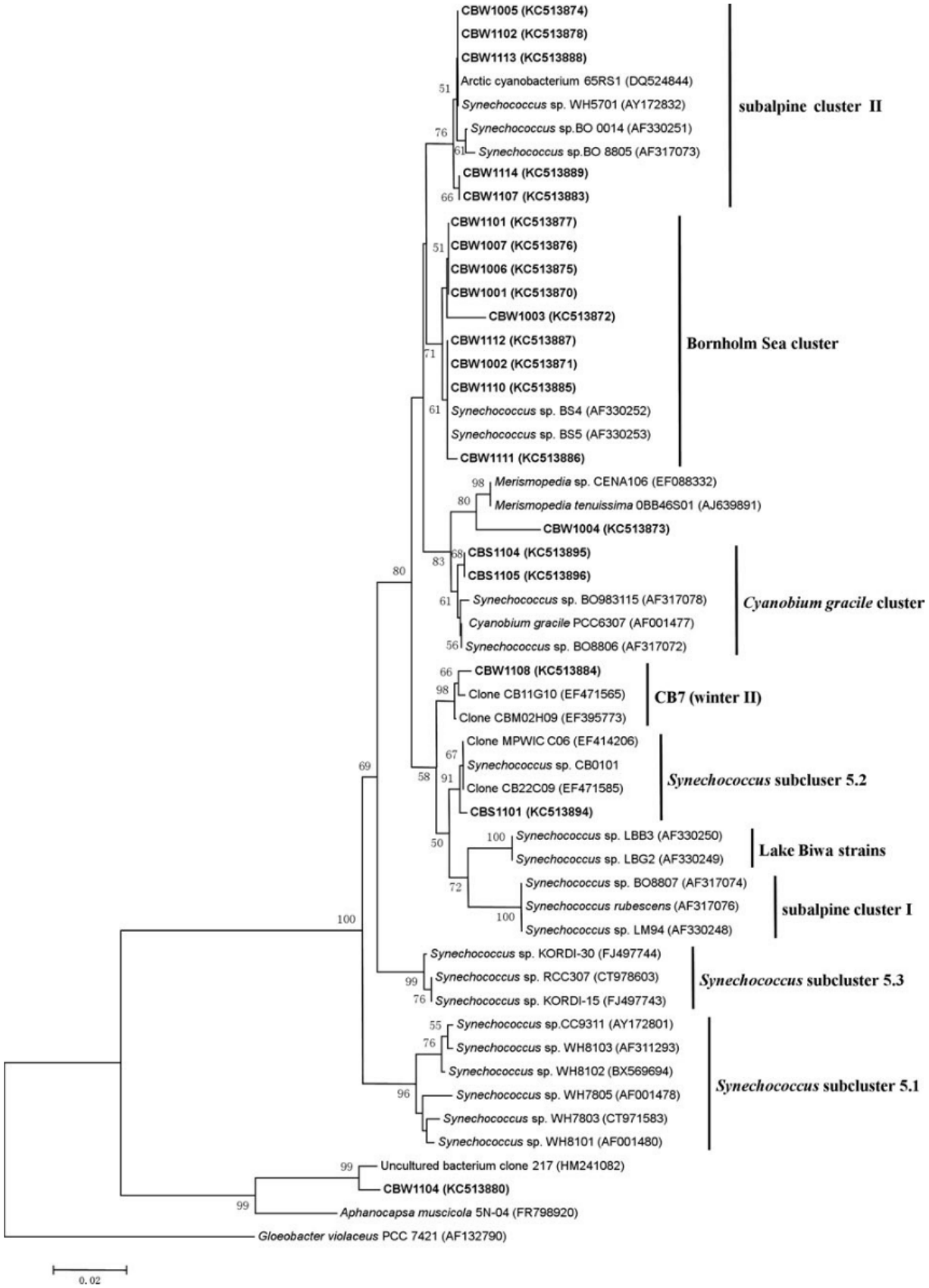

CBW1005 (KC513874)
CBW1102 (KC513878)
CBW1113 (KC513888)
Arctic cyanobacterium 65RS1 (DQ524844)
Synechococcus sp. WH5701 (AY172832)
Synechococcus sp.BO 0014 (AF330251)
Synechococcus sp.BO 8805 (AF317073)
CBW1114 (KC513889)
CBW1107 (KC513883)
subalpine cluster II
CBW1101 (KC513877)
CBW1007 (KC513876)
CBW1006 (KC513875)
CBW1001 (KC513870)
CBW1003 (KC513872)
CBW1112 (KC513887)
CBW1002 (KC513871)
CBW1110 (KC513885)
Synechococcus sp. BS4 (AF330252)
Synechococcus sp. BS5 (AF330253)
CBW1111 (KC513886)
Bornholm Sea cluster
Merismopedia sp. CENA106 (EF088332)
Merismopedia tenuissima 0BB46S01 (AJ639891)
CBW1004 (KC513873)
CBS1104 (KC513895)
CBS1105 (KC513896)
Synechococcus sp. BO983115 (AF317078)
Cyanobium gracile PCC6307 (AF001477)
Synechococcus sp. BO8806 (AF317072)
Cyanobium gracile cluster
CBW1108 (KC513884)
Clone CB11G10 (EF471565)
Clone CBM02H09 (EF395773)
CB7 (winter II)
Clone MPWIC C06 (EF414206)
Synechococcus sp. CB0101
Clone CB22C09 (EF471585)
CBS1101 (KC513894)
Synechococcus subcluser 5.2
Synechococcus sp. LBB3 (AF330250)
Synechococcus sp. LBG2 (AF330249)
Lake Biwa strains
Synechococcus sp. BO8807 (AF317074)
Synechococcus rubescens (AF317076)
Synechococcus sp. LM94 (AF330248)
subalpine cluster I
Synechococcus sp. KORDI-30 (FJ497744)
Synechococcus sp. RCC307 (CT978603)
Synechococcus sp. KORDI-15 (FJ497743)
Synechococcus subcluster 5.3
Synechococcus sp.CC9311 (AY172801)
Synechococcus sp. WH8103 (AF311293)
Synechococcus sp. WH8102 (BX569694)
Synechococcus sp. WH7805 (AF001478)
Synechococcus sp. WH7803 (CT971583)
Synechococcus sp. WH8101 (AF001480)
Synechococcus subcluster 5.1
Uncultured bacterium clone 217 (HM241082)
CBW1104 (KC513880)
Aphanocapsa muscicola 5N-04 (FR798920)
Gloeobacter violaceus PCC 7421 (AF132790)
0.02

Figure 1.3. Winter Chesapeake Bay picocyanobacteria neighbor joining tree based on partial 16S rRNA gene (Xu et al., 2015). Chesapeake Bay winter strains are diverse and unique, distinct from Chesapeake Bay strains isolated from summer months.

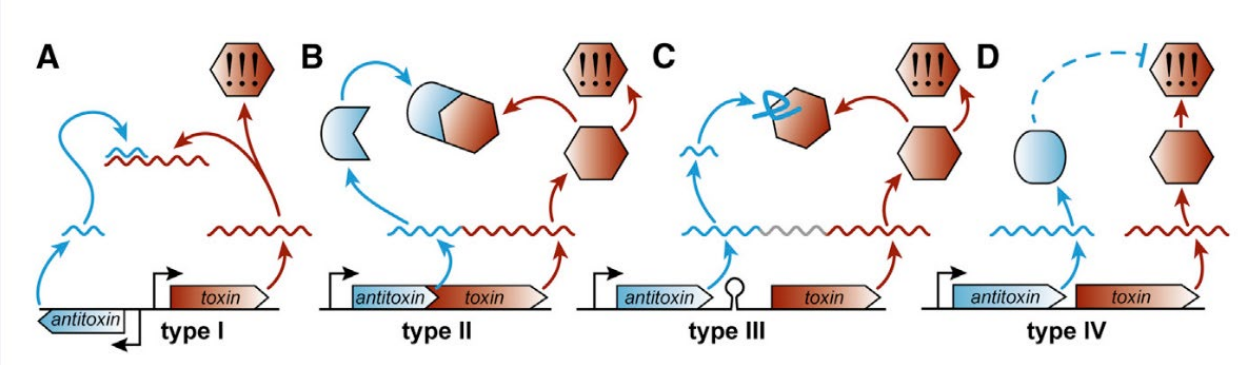


Figure 1.4. Four Main Toxin-Antitoxin Families (Harms et al., 2018a). "Different modes of toxins (red) are controlled by cognate antitoxins (blue) in type I–IV TA modules. Genetic loci and the positions of promoters are shown with colored and black arrows, respectively. RNAs are drawn as curly lines. Active toxin molecules that have been freed from antitoxin control are highlighted by exclamation marks. (A) Type I TA module. (B) Type II TA module. (C) Type III TA module. (D) Type IV TA module..."

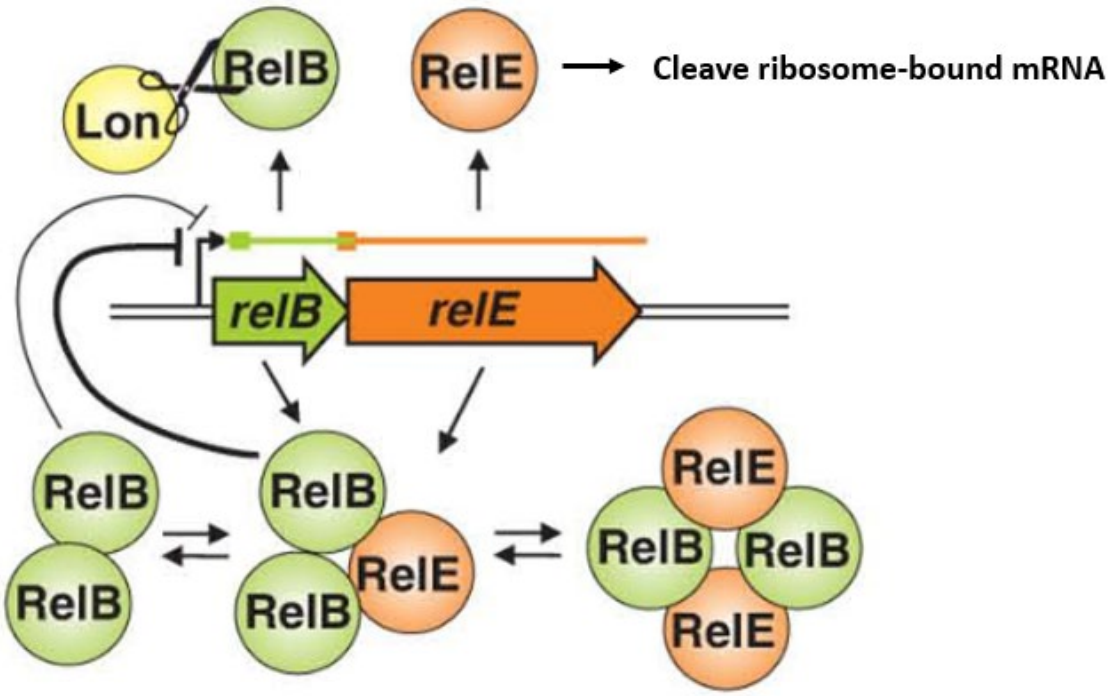


Figure 1.5. Method of action for RelE/RelB. Adapted from Unterholzner, Poppenberger, and Rozhon 2014: Free RelE toxin results in the inhibition of translation via the cleavage of ribosome-bound mRNA.

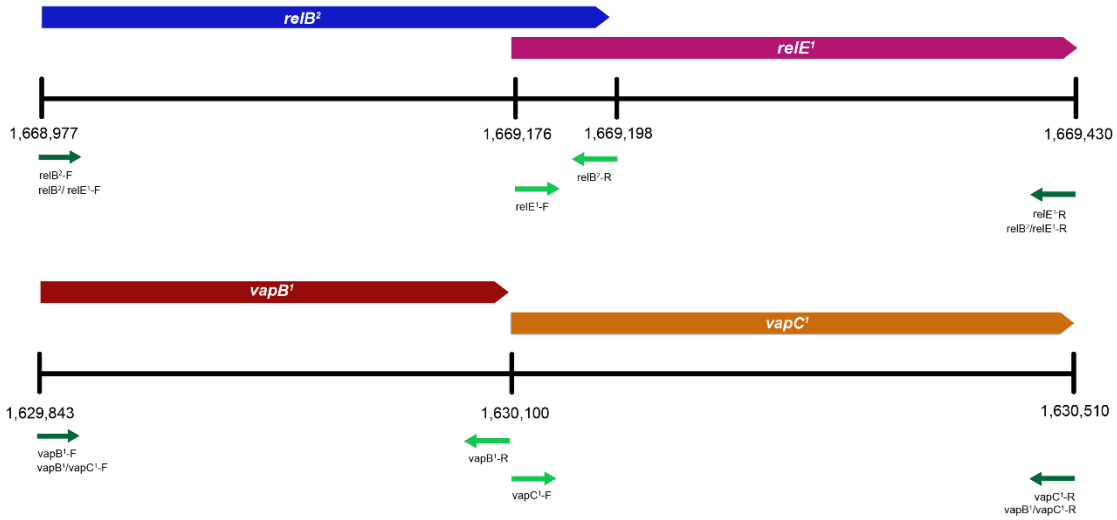


Figure 1.6. RelE/RelB and VapC/VapB Toxin Antitoxin system operons in *Synechococcus* CB0101. RelE/RelB has a 20 nucleotide overlap region. VapC/VapB does not overlap with ORFs that are immediately sequential. Green arrows indicate primer sites for gene amplification and heterologous expression.

Chapter II: Presence of toxin-antitoxin systems in picocyanobacteria and their ecological implications

Abstract

Picocyanobacteria (mainly *Synechococcus* and *Prochlorococcus*) contribute significantly to oceanic primary production. Toxin-Antitoxin (TA) systems present in bacteria and archaea are known to regulate cell growth in response to environmental stresses. However, little is known about the presence of TA systems in picocyanobacteria. This study investigated complete genomes of *Synechococcus* and *Prochlorococcus* to understand the prevalence of TA systems in picocyanobacteria. Using the TAfinder software, Type II TA systems were predicted in 27 of 33 (81%) *Synechococcus* strains, but none of 38 *Prochlorococcus* strains contain TA genes. *Synechococcus* strains with larger genomes tend to contain more putative type II TA systems. The number of TA pairs varies from 0 to 42 in *Synechococcus* strains isolated from various environments. A linear correlation between the genome size and the number of putative TA systems in both coastal and freshwater *Synechococcus* was established. In general, open ocean *Synechococcus* contain no or few TA systems, while coastal and freshwater *Synechococcus* contain more TA systems. The type II TA systems inhibit microbial translation via ribonucleases and allow cells to enter the "dormant" stage in adverse environments. TA systems are widely present in many freshwater and marine *Synechococcus*. Inheritance of more TA genes in freshwater and coastal *Synechococcus* could confer a recoverable persister state which would be an important mechanism to survive in highly dynamic environments.

Picocyanobacteria are small unicellular cyanobacteria, and they contribute greatly to carbon fixation in the aquatic ecosystem. Marine picocyanobacteria contain two major genera, *Prochlorococcus* and *Synechococcus*, which together can contribute about 25% of net primary production in the ocean (Flombaum et al., 2013; Li & Url, 1994). While *Prochlorococcus* is more restricted to warm oligotrophic water, *Synechococcus* is widely distributed in various aquatic environments ranging from open oceans to freshwater (Dvořák et al., 2014a). The average cell size of *Synechococcus* (0.9 µm) is larger than that of *Prochlorococcus* (0.6 µm) (Morel et al., 1993). In addition, the average genome size of *Prochlorococcus* (1.8 Mb) is also smaller than that of *Synechococcus* (2.9 Mb) (Frédéric Partensky et al., 1999). Genome streamlining provides less ecological flexibility to marine *Prochlorococcus*, on the other hand, the relatively large genome size of *Synechococcus* provides more genomic plasticity which enables them to adapt to more variable habitats (Biller et al., 2015; Dufresne et al., 2005; Larsson et al., 2011; Sun & Blanchard, 2014).

Diverse *Synechococcus* strains have been isolated from freshwater, estuarine, coastal, and oceanic water (Li & Url, 1994; P. Sánchez-Baracaldo et al., 2005; D J Scanlan et al., 2009; David J. Scanlan, 2012), suggesting that *Synechococcus* can adapt to distinct aquatic environments. In the estuarine environment, picocyanobacteria (mostly *Synechococcus*) can make up 20-40% of phytoplankton chlorophyll a and up to 60% of primary production in summer (K. Wang et al., 2011). Freshwater *Synechococcus* can also play an important role in carbon fixation and nutrient cycling in ponds, lakes, and rivers (Callieri, 2008; Callieri & Stockner, 2002; Stockner,

1988). Phylogenetic analyses of freshwater and marine *Synechococcus* show that *Synechococcus* is polyphyletic (Dvořák et al., 2014a; Honda et al., 1999; Rippka et al., 1979; A. M. R. Wilmotte & Stam, 1984). Molecular systematics has challenged the traditional taxonomy of *Synechococcus* in the past 20 years (Coutinho et al., 2016; Honda et al., 1999; Robertson et al., 2001). Genetic diversity of *Synechococcus* has been studied in various aquatic environments (Fuller et al., 2003; Rocap et al., 2002; Toledo & Palenik, 1997; Zwirglmaier et al., 2008). In marine waters, three subclusters of *Synechococcus* have been defined (Dufresne et al., 2008; D J Scanlan et al., 2009) and subdivided into 28 clades based on the ITS sequences (Huang et al., 2011). In the freshwater system, 6-8 clusters of *Synechococcus* have been identified based on the 16S rRNA gene or other genetic markers (Callieri et al., 2013; Crosbie et al., 2003; Ernst et al., 2003b; Huang et al., 2011; Jasser et al., 2011). Freshwater *Synechococcus* are deeply branched and are less congruent compared to marine *Synechococcus* (P. Sánchez-Baracaldo et al., 2005). Because of their ubiquity in aquatic systems, *Synechococcus* contain highly diverse phylotypes and ecotypes.

Comparative genomics of cyanobacteria has greatly advanced our understanding of molecular evolution, metabolic potential and ecological adaptation of different cyanobacterial types (Dufresne et al., 2008; D J Scanlan et al., 2009). Unicellular cyanobacteria with smaller genomes (<3.3 Mb) appear to have relatively more genes involved in amino acid metabolism, but fewer genes for environmental sensing (signal transduction) and cell motility compared to cyanobacteria with larger genomes (>3.3 Mb) (Larsson et al., 2011). In the marine environment, ecological adaptation of *Synechococcus* to different niches is evident at the genomic level. The first marine *Synechococcus* genome (strain WH8102) was sequenced in 2003 (B. Palenik et al.,

2003). By comparing the genomes of a coastal *Synechococcus* strain (CC9311) and the oceanic *Synechococcus* strain (WH8102), Palenik et al. showed that the coastal strain has a greater capacity to sense and respond to changes in their environment compared to the oceanic counterpart (Brian Palenik et al., 2006). Open ocean *Synechococcus* ('specialists') tend to have smaller genomes and less genome islands than coastal *Synechococcus* ('opportunists' or 'generalists') (Dufresne et al., 2008; D J Scanlan et al., 2009). Coastal *Synechococcus* strains have an increased tolerance for copper and oxidative stress through distinct transcriptional responses and genomic features (Stuart et al., 2009, 2013). Coastal *Synechococcus* genomes contain a large portion of accessory and unique genes which provide them considerable flexibility to adapt to diverse habitats (Dufresne et al., 2008). Novel genes in picocyanobacterial genome islands can provide selective advantage for niche adaptation (D J Scanlan et al., 2009). Recently, genome sequencing of a Chesapeake Bay *Synechococcus* strain CB0101 unveiled its increased capacity in environmental sensing, transportation, regulation, and stress response (Fucich et al., 2019). The presence of toxin-antitoxin (TA) genes and their functional assignment in *Synechococcus* CB0101 suggests that TA systems can be important to the high environmental endurance of estuarine *Synechococcus* (D. Marsan et al., 2017).

TA systems are known to be involved in stress responses in microbes, but little is known about TA systems in picocyanobacteria. TA systems are genetic modules comprised of a toxin, which often arrests translation and subsequently growth, and a cognate antitoxin which negates the interruption of the toxin (Page & Peti, 2016; Unterholzner et al., 2013). TA system activation often results in persister cell formation which can be advantageous for bacterial survival in highly variable environments. While TA systems have been broadly described as ubiquitous in

nearly all bacterial species, TA systems in cyanobacteria have only recently been described. TA systems have been predicted in freshwater cyanobacteria including *Microcystis aeruginosa* (Makarova et al., 2009), *Synechocystis* PCC6803 (Kaneko et al., 2003), and *Synechococcus* PCC7942 (Y. Chen et al., 2011). Only a few TA genes, i.e. VapB, VapC and PemK were reported in the freshwater cyanobacterial strains, no systematic survey on TA genes was performed at the genomic level on those strains. The first chromosomal TA system in marine *Synechococcus* was described in the estuarine *Synechococcus* strain CB0101 (D. Marsan et al., 2017). CB0101, isolated from the Chesapeake Bay, belongs to *Synechococcus* subcluster 5.2 (F. Chen et al., 2006a). Transcriptomic analysis of CB0101 reveals a tight coupling between the upregulation of particular toxins, such as relE, with environmental stressors like zinc heavy metal toxicity and high light intensity (D. Marsan et al., 2017). Marsan et al. showed that TA systems can be important to the environmental stress response in *Synechococcus* (D. Marsan et al., 2017). However, little is known about the occurrence, diversity, evolution, and ecological functions of type II TA systems in *Synechococcus* and other picocyanobacteria.

The goal of this study is to investigate the presence of TA genes in picocyanobacteria using the TAfinder software (Xie et al., 2018). Our search comprised of 71 complete picocyanobacterial genomes, including 33 *Synechococcus* and 38 *Prochlorococcus* genomes. An interesting linear relationship between the number of TA pairs and genome size was found in *Synechococcus*.

Methods

Complete *Synechococcus* (n=33) and *Prochlorococcus* (n=38) genomes were downloaded in September, 2019 from both the National Center for Biotechnology Information (NCBI) RefSeq database (Agarwala et al., 2017a; O'Leary et al., 2016) and the Joint Genome Institute (JGI) genome portal (Nordberg et al., 2014). To ensure quality, we omitted incomplete genomes from this study. The *Synechococcus* and *Prochlorococcus* genomes included in this study cover the majority of major known phylogenetic clades and subclusters (Table 2.1).

Toxin-antitoxin systems were predicted using the TAfinder software which utilizes the Toxin-Antitoxin Database (TADB) (Shao et al., 2011; Xie et al., 2018). Genomes that were not included in the TAfinder's available genome list were downloaded locally and manually uploaded to TAfinder. TAfinder was used to predict type II TA pairs in *Synechococcus* (freshwater, estuarine, coastal, and open ocean strains) and marine *Prochlorococcus* genomes using default settings (BLAST e-value=0.01, HMMer=1, Maximum length =300 aa, Distance=-20_150). *Synechococcus* strains were classified into habitats based on literature searches for original isolation information. Because estuarine strains, *Synechococcus* CB0101 and PCC7002, are underrepresented, they were categorized into the coastal habitat category for the purpose of linear regression data analysis.

To estimate relative diversity of the putative TA families, predicted amino acid sequences were searched against the NCBI conserved domain database (version CDD v3.18 - 55570 PSSMs) (Marchler-Bauer et al., 2017). Short names for conserved domains were manually reviewed and determined to be of a consensus of a major TA family. If the gene did not fall into one of the

traditional TA families, it was categorized as “Other” for the consensus. If the predicted amino acid sequence did not have a significant match to the conserved domain database, it was categorized as “Unknown”.

Linear regression and linear models were completed using Rstudio software (Core, 2017) and figures were made using ggplot2 (Wickham, 2009). Genome Island regions were predicted using IslandViewer 4 software (Bertelli et al., 2017).

Results

TAfinder predicted at least one TA pair in 27 of 33 *Synechococcus* genomes (81%). The number of TA systems in *Synechococcus* varies from 0 to 42 (A toxin antitoxin system is normally comprised of one toxin gene and one cognate antitoxin gene). Only five strains of *Synechococcus* did not contain putative TA systems. A total of 986 putative toxin and antitoxin genes were predicted, constituting 493 TA systems, in 27 complete *Synechococcus* genomes. The occurrence frequency of TA systems in *Synechococcus* is shown in Figure 2.1. The 27 TA-containing *Synechococcus* strains were isolated from various aquatic environments including freshwater, Antarctic (cold adapted), hot spring (thermophile), estuarine, coastal, and oceanic waters and belong to diverse phylogenetic lineages (Table 2.1).

TAfinder did not predict any TA systems in any of the *Prochlorococcus* genomes (n=38). These *Prochlorococcus* genomes were representative of many clades from both high light and low light adapted strains. These queried *Prochlorococcus* genomes ranged in size from 1.6 to 2.7 Mb.

Freshwater and coastal *Synechococcus* contained many putative TA systems. For example, freshwater *Synechococcus* strains PCC6312 and PCC6307 both contained 42 putative TA systems. These 84 genes accounted for ~1.2% of their total coding sequences (Table 2.1). Coastal strains PCC7003, and PCC7117 contained 38, and 37 TA pairs, respectively, accounting for 1.24% and 1.17% of their coding sequences.

In general, *Synechococcus* living in coastal, estuarine, and freshwater environment tend to have larger genomes compared to their counterparts living in the open ocean. It appears that *Synechococcus* with larger genomes contain more TA genes than *Synechococcus* with smaller genomes. Interestingly, a good linear correlation between the genome size and the number of putative TA pairs ($r^2=0.6235$, $p<0.0001$) (Figure. 2.2a) was found in *Synechococcus*, further confirming the above observation that larger *Synechococcus* genomes contain more TA genes. This apparent relationship between genome size and putative TA pairs becomes more clear in cases when endemic ecological conditions are considered; specifically, for coastal and freshwater *Synechococcus*. When analyzed separately, better linear regressions ($r^2= 0.9152$, $p<0.00001$ and $r^2=0.8296$, $p<0.005$) between genome size and putative TA pairs were found when coastal and freshwater *Synechococcus* were analyzed separately (Figure 2.2b). Conversely, the general correlative trend between genome size and the number of TA pairs in all *Synechococcus* strains was not found when only the open ocean strains were analyzed. *Synechococcus* toxin genes contained more known conserved domains than antitoxins (Figure 2.3). About 77% of toxin genes had a known conserved domain with an annotation of a traditionally named TA system. The most common toxin gene included a conserved PIN domain

which is characteristic of the VapC toxin which cleaves tRNAs or rRNAs (Winther & Gerdes, 2011). Nearly 41% of putative toxin genes contained the conserved domain for VapC.

Putative antitoxin sequences contained fewer NCBI conserved domains than toxins. Only 35% of antitoxin genes had a conserved domain with a traditionally annotated TA system. Many conserved domains in putative antitoxin genes had generic names such as “domains of unknown function” (DUF) or COG.

Discussion

A survey on picocyanobacterial TA systems leads to an interesting finding that *Synechococcus* strains with larger genome size contain more TA systems. Although many genetic features of picocyanobacterial genomes have been explored (Dufresne et al., 2005; D J Scanlan et al., 2009), little is known about the prevalence of TA genes in picocyanobacteria. *Synechococcus* has a remarkable adaptation capability, which is reflected by their occupancy in diverse environments ranging from lakes, rivers, estuaries, coastal and oceanic water. The presence of a specific group or genus over such a wide range of habitats makes *Synechococcus* an ideal model to explore the relationship between their ecological adaptation and genomic features. TA systems have been well studied in bacteria and archaea. One of well-known functions of TA systems is that it enables cells to go dormant or enter the persister stage under stressed conditions and recover when the adverse stresses are released (Harms et al., 2018b; Makarova et al., 2009; Unterholzner et al., 2013). While the actual functions of *Synechococcus* TA genes have not be tested, it is believed that inheritance of more TA genes may allow some *Synechococcus* strains to endure more variable environments which could confer a competitive

advantage against other less resilient picocyanobacteria. Coastal, estuarine, and freshwater environments are characterized by rapid changes in environmental conditions, the higher occurrence frequency of TA genes can provide adaptive advantages for *Synechococcus* living in these types of aquatic habitats.

TA systems have been shown to provide recoverable persister states when *Synechococcus* cells were exposed to conditions to induce oxidative stress (D. Marsan et al., 2017). These conditions are expected in rapidly changing environments such as estuaries, coastal, and some freshwater environments. In another cyanobacterial species, *Synechococystis* PCC 6803, TA systems have been found and were predicted to have RNase activity which could have a drastic effect on the transcriptomic remodeling (Kopfmann et al., 2016). Such a remodeling is possible through RNA degradation as a result of toxin overexpression, which can have a significant impact on slowing translation. Type II TA systems can have other methods of actions in other bacteria including post segregational killing and abortive infection (Harms et al., 2018b). Unfortunately, these remain poorly studied and understood in picocyanobacterial systems.

Interestingly, a linear correlation was found between the genome size and the number of TA genes in *Synechococcus*. Previous studies have had mixed results. One study found a similar linear correlation in prokaryotes (Makarova et al., 2009). In other similar work, such a linear correlation has not been found in bacteria (Leplae et al., 2011). When testing 2,181 genomes of prokaryotes (archaea and bacteria from both obligate intracellular species to free living species) (Leplae et al., 2011) and 65 genomes of *Acetobacter* (with sources ranging from fermented food, to fruits, to symbiotes in the fruit fly *Drosophila melanogaster*) (K. Xia et al., 2019), the number of TA gene pairs does not increase linearly with increased genome size. The clear linear

trend seen in *Synechococcus* is likely related to larger genome sizes having a wide array of CDS. For strains with expanded genetic capacity, it may be advantageous to retain a multitude of TA systems in aquatic habitats with highly variable chemical and physical features. While *Synechococcus* is ubiquitous in nearly all aquatic ecosystems, the presence of Synechococcal TA systems is not; this suggests that TA systems are advantageous in some, but not all, aquatic environments. The genome size of *Synechococcus* available for this in this study ranges from 2.1 to 3.7 Mb. *Synechococcus* genomes have previously been shown to correlate strongly with the length of hypervariable genome island regions (Dufresne et al., 2008). In *Synechococcus*, TA genes can be located on these genome islands, but the majority of the TA pairs are not located on hypervariable genome islands (Table 2.1).

The lack of TA systems in *Prochlorococcus* is likely related to their relatively stable habitats. The endemic habitat of *Prochlorococcus* is the pelagic ocean which is characterized by its stable, nutrient limiting environment coupled with a predictably high cellular density (Frédéric Partensky et al., 1999). The genus *Prochlorococcus* is a highly diverse group comprised of 12 specialized clades with genomic features uniquely adapted to specific conditions in oceanic ecosystems (Biller et al., 2015). High light adapted group II has some of the smallest genomes (~1.7 Mb) and lowest GC content (~33 %), which is indicative of genomic reduction (Dufresne et al., 2005). Some *Prochlorococcus* strains (such as low light adapted group IV) have relatively large genomes (2.4 to 2.6 Mb) and many unique genes (Biller et al., 2014). Regardless of their large genetic capacity, no TA genes were detected in the genomes of group IV *Prochlorococcus* strains. Despite the diversity of *Prochlorococcus* ecotypes, TA systems may not be needed due to *Prochlorococcus* specific adaptation to the oligotrophic ocean.

Among the 33 *Synechococcus* strains examined in this study, 11 are open ocean strains. Oceanic *Synechococcus* strains in general contain no TA genes or only a few TA genes. The five open ocean *Synechococcus* strains that are void of TA genes are WH8109, KORDI-52, KORDI-100, CC9605, and CC9902, while four oceanic *Synechococcus* strains (MIT9504, MIT9508, MIT9509, and KORDI-49) contain few (1 to 5) putative TA pairs. The exception of this is open ocean *Synechococcus* strain WH8102, which contains 15 putative systems. WH8102 was originally isolated from the Sargasso Sea, and its genome is more indicative of a 'generalist' with features acquired via horizontal gene transfer (B. Palenik et al., 2003). Like marine *Prochlorococcus*, oceanic *Synechococcus* may not need TA genes due to their acclimation to the stable oligotrophic environment.

Along with genome size, endemic ecological conditions and habitats are an important indicator of the prevalence of TA systems in *Synechococcus*, and more broadly picocyanobacteria. *Synechococcus* strains from more variable environments like coastal and freshwater locations tend to have more TA pairs than open ocean strains that are streamlined to a stable pelagic lifestyle. This phenomenon may also explain the broader pattern of TA system distribution in picocyanobacteria; the prevalence of TA pairs in picocyanobacteria living in the nutrient rich and dynamic habitats and the rareness and complete absence of TA in picocyanobacteria living in the oligotrophic open ocean. The presence of TA systems may be one of the many genetic features that allow *Synechococcus* to inhabit a wide array of aquatic ecosystems and achieve a cosmopolitan distribution. The lack of TA systems in *Prochlorococcus* is consistent with their reduced genomes and oligotrophic lifestyle (David J. Scanlan & West, 2002).

Originally, 7 TA pairs were predicted in CB0101 using BLASTCLUST (McWilliam et al., 2013) and confirmed using the RASTA-bacteria (Sevin & Barloy-Hubler, 2007) and TADB (Shao et al., 2011). More recently, the TAfinder search tool was used to search whole genomes (Xie et al., 2018), rather than specific gene pairs to predict type II TA systems. Due to the ever-expanding TADB and improved prediction methods like TAfinder, 22 TA systems, including the original 7 pairs, were found in CB0101. These new pairs were confirmed manually, and conserved domains were predicted using NCBI's conserved domain database and Interpro for protein functional analysis.

The scope of this study is constrained by the use of TAfinder. TAfinder is capable of predicting type II TA systems, which are the most well studied and characterized TA systems. Type II TA systems comprise 99% of the TA genes in the TADB (Shao et al., 2011). To ensure that other, less known TA families (I, II-VI) were not overlooked, a blast search for those few systems against all the genomes of *Synechococcus* and *Prochlorococcus* was completed. No significant matches were reported using default settings. Antitoxin sequences contained fewer conserved domains than toxins. Antitoxin sequences appear to be highly diverse and variable among *Synechococcus* strains. Multiple antitoxin structures may function to bind their cognate toxin. When the paired gene can sufficiently neutralize the toxin, it acts as an antitoxin and selection for highly conserved sequences may be relaxed. Although toxins contain more conserved domains than antitoxins, it is important to note that TA systems are not present in all *Synechococcus* and they are highly variable in terms of the number and type of TA systems. Even within the closely related *Synechococcus* strains, it is difficult to identify suitable genetic markers for phylogenetic analysis due to the overall poor gene conservation. VapC, and its

cognate VapB antidote, are the largest family of bacterial toxin-antitoxin modules (Robson et al., 2009). A wide variety of toxin functionality is represented in *Synechococcus* as both ribosomal-dependent mRNA endonucleases like RelE and ribosomal-independent mRNA endonucleases like HicA and MazF were predicted (Harms et al., 2018b).

Conclusion

The tight correlation of genome size and the number TA genes in coastal and freshwater *Synechococcus* suggest that the retention of TA systems could be advantageous for *Synechococcus* living in highly variable environments. All the tested *Prochlorococcus* genomes (n=38) do not contain any TA genes, given that their genome sizes range from 1.6 to 2.7 Mb. This result suggests that *Prochlorococcus* do not have a TA system mediated dormancy in response to changing environments. This also applies to some *Synechococcus* living in open oceans where chemical and hydrological conditions are relatively stable compared to coastal, estuarine, and freshwater environments. It is interesting that the number of TA genes is linearly correlated with increasing genome sizes of *Synechococcus*. It appears that the acquisition and retention of TA genes in *Synechococcus* is not only influenced by genome size, but also environmental stability. *Synechococcus* strains with large genomes, especially those that inhabit fluctuating ecosystems (coastal, estuarine, and freshwater) have more TA systems than strains with smaller genomes that are present in stable environments like the open ocean. Compared to *Prochlorococcus*, *Synechococcus* has a relatively large genome, with space for more coding sequences, ample TA systems, and a wide variety of environmental response genes that allow for their ubiquitous distribution in diverse aquatic environments. TA systems in *Synechococcus*

could confer an ability to enter persister states in the presence of stressful stimuli, which is advantageous in highly variable conditions which characterize coastal and freshwater environments.

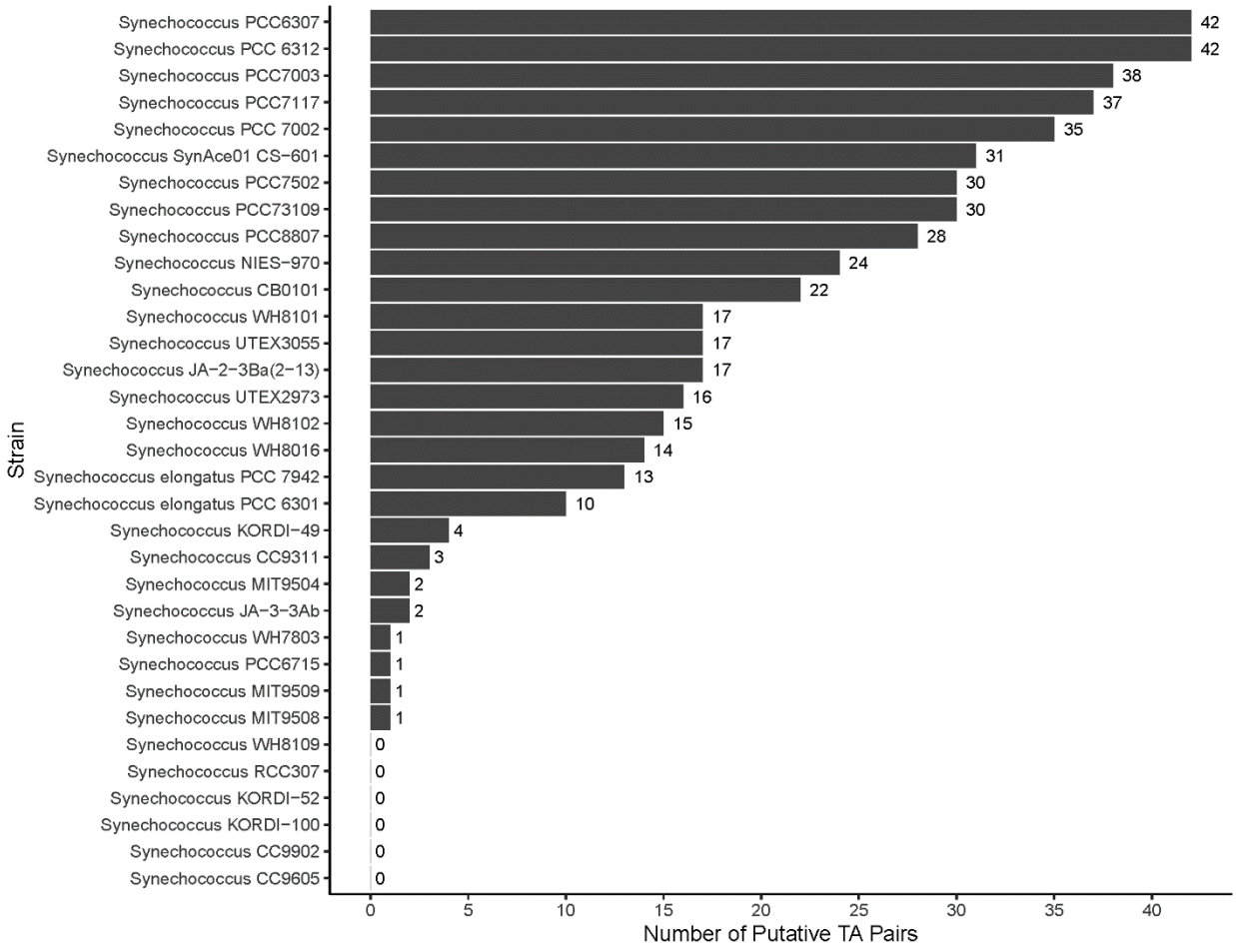


Figure 2.1. Occurrence frequency of putative TA systems in 33 strains of *Synechococcus* isolated from various aquatic environments

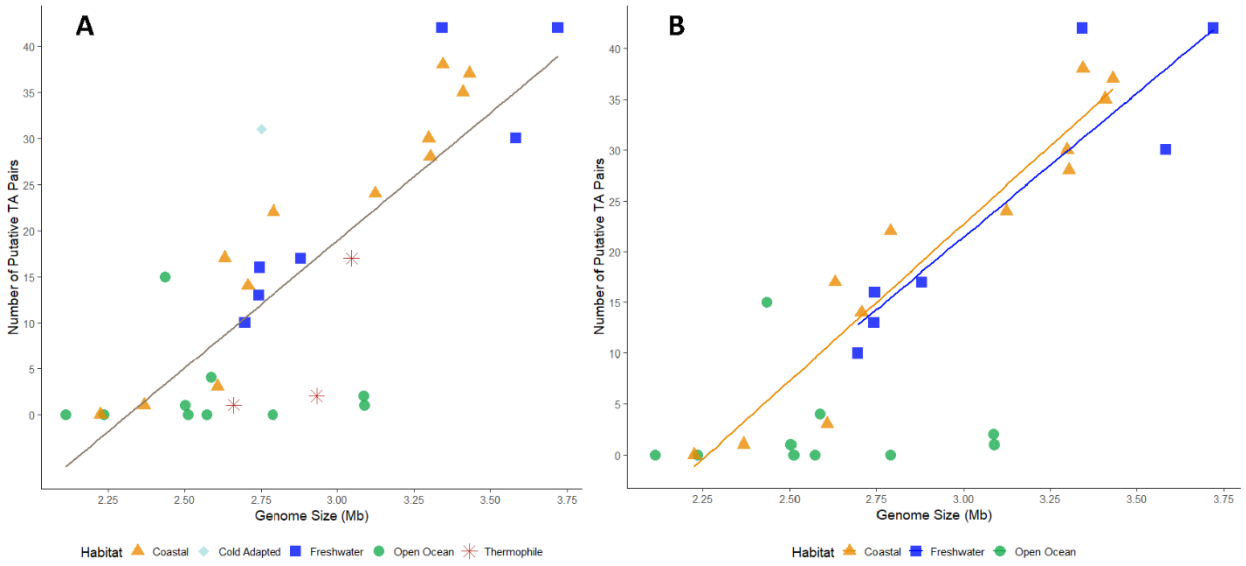


Figure 2.2. Relationship between genome size and the number of putative TA pairs in *Synechococcus*. A) Linear correlation between genome size and putative TA systems for all complete genomes of *Synechococcus* strains isolated from all habitats (r^2= 0.6235, $p<0.0001$); B) Linear correlation between genome size (Mb) and the number of putative TA pairs in coastal and freshwater *Synechococcus*, (r^2= 0.9152, $p<0.00001$ and r^2=0.8296, $p<0.005$ respectively). No such correlation was found in open ocean *Synechococcus*.

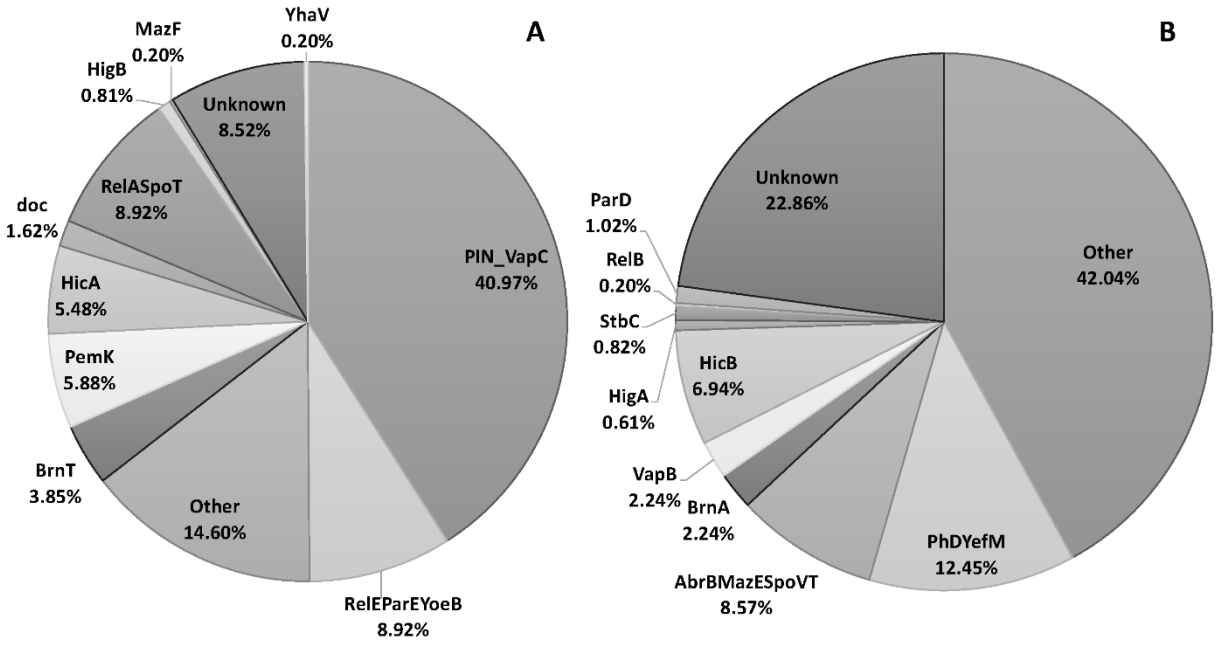


Figure 2.3. Conserved domain regions of putative A) Toxins and B) Antitoxins. Putative toxin and antitoxin sequences that contained a conserved domain that was not a traditional TA system were categorized as 'Other'. Sequences that did not contain a conserved domain were categorized as 'Unknown'.

Tables

Table 2.1. *Synechococcus* and *Prochlorococcus* genome accession numbers, strain names, classification, reference, putative TA pairs, genome size, coding sequences, habitat, TA as a function of total open reading frames (ORFs), and percent of TA pairs located on genomic islands.

Accession Number	Name	Classification	Reference	Number of Putative TA Pairs	Genome Size Mb	Coding Sequences	Habitat	TA Percent of Total ORFs	Percent of TA Systems located on Genome Islands
CP000576	*Prochlorococcus* marinus MIT 9301	HLII	Kettler et al. 2007	0	1.64	1785		0.00	0.00
NZ_CP018344	*Prochlorococcus* RS50	Unclassified	Sosa et al. 2019	0	1.66	1951		0.00	0.00
NZ_CP018346	*Prochlorococcus* RS04	Unclassified	Sosa et al. 2019	0	1.66	1952		0.00	0.00
NZ_CP018345	*Prochlorococcus* RS01	Unclassified	Sosa et al. 2019	0	1.66	1945		0.00	0.00
GCA_000011465.1	*Prochlorococcus* marinus pastoris CCMP1986	Unclassified		0	1.66	1790		0.00	0.00
CP000551	*Prochlorococcus* marinus AS9601	HLII	Kettler et al. 2007	0	1.69	1787		0.00	0.00
CP000878	*Prochlorococcus* marinus MIT 9211	LLIII	Kettler et al. 2007	0	1.69	1902		0.00	0.00
CP000552	*Prochlorococcus* marinus MIT 9515	HLI	Kettler et al. 2007	0	1.70	1792		0.00	0.00
2681813573	*Prochlorococcus* MIT1314	HLII	Becker et al. 2019	0	1.70	1982		0.00	0.00
CP000111	*Prochlorococcus* marinus MIT 9312	HLII	Kettler et al. 2007	0	1.71	1826		0.00	0.00
2681812903	*Prochlorococcus* MIT0919	Unclassified		0	1.72	1927		0.00	0.00
CP000825	*Prochlorococcus* marinus MIT 9215	HLII	Kettler et al. 2007	0	1.74	1983		0.00	0.00
GCF_000007925.1	*Prochlorococcus* marinus CCMP1375	Unclassified		0	1.75	1882		0.00	0.00
CP007753	*Prochlorococcus* MIT0604	HLII	Biller et al. 2014	0	1.78	2089		0.00	0.00
2681812902	*Prochlorococcus* MIT0918	Unclassified		0	1.79	1997		0.00	0.00

2681813568	*Prochlorococcus* MIT1223	LL.MIT1223	Berube et al. 2019	0	1.80	1991		0.00	0.00
CP000095	*Prochlorococcus* marinus NATL2A	LLI	Kettler et al. 2007	0	1.84	1953		0.00	0.00
2681813570	*Prochlorococcus* MIT1300	Unclassified		0	1.86	2020		0.00	0.00
CP000553	*Prochlorococcus* marinus NATL1A	LLI	Kettler et al. 2007	0	1.86	1976		0.00	0.00
2681812900	*Prochlorococcus* MIT0913	LLI	Berube et al. 2019	0	1.88	2203		0.00	0.00
2681812899	*Prochlorococcus* MIT0912	LLI	Berube et al. 2019	0	1.90	2206		0.00	0.00
2681812859	*Prochlorococcus* MIT0917	LLI	Berube et al. 2019	0	1.92	2224		0.00	0.00
CP007754	*Prochlorococcus* MIT0801	LLI	Biller et al. 2014	0	1.93	2218		0.00	0.00
2681813567	*Prochlorococcus* MIT1214	LLI	Berube et al. 2019	0	1.93	2206		0.00	0.00
2681813574	*Prochlorococcus* MIT1341	Unclassified		0	1.94	2090		0.00	0.00
2681812901	*Prochlorococcus* MIT0915	LLI	Berube et al. 2019	0	1.99	2252		0.00	0.00
2681813572	*Prochlorococcus* MIT1307	Unclassified		0	2.03	2198		0.00	0.00
2681812904	*Prochlorococcus* MIT1013	Unclassified		0	2.05	2428		0.00	0.00
ASM16179v2	*Synechococcus* WH8109	Subcluster 5.1 (Clade II)	Rocap et al. 2002	0	2.11	2696	Open Ocean	0.00	0.00
ASM6352	*Synechococcus* RCC307	Subcluster 5.3	Dufresne et al. 2008	0	2.22	2388	Open Ocean	0.00	0.00
ASM1250	*Synechococcus* CC9902	Subcluster 5.1 (Clade IV)	Dufresne et al. 2008	0	2.23	2337	Open Ocean	0.00	0.00
ASM6350	*Synechococcus* WH 7803	Subcluster 5.1 (Clade V)	Rocap et al. 2002	1	2.37	2456	Open Ocean	0.04	0.00
BX548175	*Prochlorococcus* marinus MIT 9313	LLIV	Kettler et al. 2007	0	2.41	2369		0.00	0.00

ASM19597	*Synechococcus* WH 8102	Subcluster 5.1 (Clade III)	Rocap et al. 2002	15	2.43	2513	Open Ocean	0.60	0.00
2681812948	*Prochlorococcus* marinus MIT1323	LLIV	Cubillos-Ruiz et al. 2017	0	2.44	2577		0.00	0.00
2681812928	*Prochlorococcus* marinus MIT1320	LLIV	Cubillos-Ruiz et al. 2017	0	2.50	2661		0.00	0.00
2681812924	*Prochlorococcus* MIT1306	LLIV	Cubillos-Ruiz et al. 2017	0	2.50	2699		0.00	0.00
NZ_LVHU01000000	*Synechococcus* MIT9508	Subcluster 5.1 (CRD1)	Cubillos-Ruiz et al. 2017	1	2.50	2983	Open Ocean	0.03	0.00
ASM1262	*Synechococcus* CC9605	Subcluster 5.1 (Clade II)	Dufresne et al. 2008	0	2.51	2665	Open Ocean	0.00	0.00
2681812950	*Prochlorococcus* marinus MIT1342	LLIV	Cubillos-Ruiz et al. 2017	0	2.54	2702		0.00	0.00
2681812925	*Prochlorococcus* marinus MIT1312	LLIV	Cubillos-Ruiz et al. 2017	0	2.55	2732		0.00	0.00
2681812923	*Prochlorococcus* MIT1303	LLIV	Cubillos-Ruiz et al. 2017	0	2.56	2849		0.00	0.00
ASM73759	*Synechococcus* KORDI-52	Subcluster 5.1 (WPC2)	Choi et al. 2009	0	2.57	2598	Open Ocean	0.00	0.00
2681812927	*Prochlorococcus* marinus MIT1318	LLIV	Cubillos-Ruiz et al. 2017	0	2.58	2710		0.00	0.00
ASM73757	*Synechococcus* KORDI-49	Subcluster 5.1 (WPC1)	Choi et al. 2009	4	2.59	2528	Open Ocean	0.16	0.00
2681812949	*Prochlorococcus* marinus MIT1327	LLIV	Cubillos-Ruiz et al. 2017	0	2.59	2701		0.00	0.00
2681812926	*Prochlorococcus* marinus MIT1313	LLIV	Cubillos-Ruiz et al. 2017	0	2.59	2707		0.00	0.00
ASM1458	*Synechococcus* CC9311	Subcluster 5.1 (Clade I)	Dufresne et al. 2008	3	2.61	2663	Coastal	0.11	0.00
ASM420977	*Synechococcus* WH8101	Subcluster 5.1 (Marine B)	Rocap et al. 2002	17	2.63	2693	Coastal	0.63	0.00
ASM275493	*Synechococcus* PCC6715	Unclassified		1	2.66	2227	Thermophile	0.04	0.00

ASM1006	*Synechococcus* elongatus PCC 6301	Unclassified	Sugita et al. 2007	10	2.70	2602	Freshwater	0.38	0.00
PRJNA61805	*Synechococcus* WH8016	Subcluster 5.1 (Clade I)	Rocap et al. 2002	14	2.71	3046	Coastal	0.46	0.00
ASM1252	*Synechococcus* elongatus PCC 7942	Unclassified		13	2.74	2685	Freshwater	0.48	0.00
ASM81732	*Synechococcus* UTEX2973	Unclassified		16	2.74	2678	Freshwater	0.60	0.00
ASM188521	*Synechococcus* SynAce01 CS-601	Antarctic	Tang et al. 2019	31	2.75	2701	Cold Adapted	1.15	0.13
ASM73753	*Synechococcus* KORDI-100	Subcluster 5.1 (UC-A)	Choi et al. 2009	0	2.79	2822	Open Ocean	0.00	0.00
CP039373	*Synechococcus* CB0101	Subcluster 5.2 (CB4)	Chen et al. 2006	22	2.79	3126	Coastal (Estuary)	0.70	0.27
ASM395780	*Synechococcus* UTEX3055	Unclassified		17	2.88	2815	Freshwater	0.60	0.00
ASM1320	*Synechococcus* JA-3-3Ab	Unclassified	Schirrmeister et al. 2005	2	2.93	2611	Thermophile	0.08	0.00
ASM1322	*Synechococcus* JA-2-3Ba(2-13)	Unclassified	Schirrmeister et al. 2005	17	3.05	2692	Thermophile	0.63	0.94
NZ_LVHT00000000	*Synechococcus* MIT9504	Subcluster 5.1 (CRD1)	Cubillos-Ruiz et al. 2017	2	3.09	3712	Open Ocean	0.05	0.00
NZ_LVHV01000000	*Synechococcus* MIT9509	Subcluster 5.1 (CRD1)	Cubillos-Ruiz et al. 2017	1	3.09	3752	Open Ocean	0.03	0.00
ASM235621	*Synechococcus* NIES-970	Unclassified	Shimura et al. 2017	24	3.12	2864	Coastal	0.84	0.08
ASM152185	*Synechococcus* PCC73109	Group 5	Robertson et al. 2001	30	3.30	3037	Coastal	0.99	0.07
ASM169329	*Synechococcus* PCC8807	Group B (Subalpine I)	Everroad et al. 2012	28	3.30	3083	Coastal	0.91	0.00
PRJNA158695	*Synechococcus* PCC6307	Cyanobium Group A	Havercamp et al. 2009	42	3.34	3439	Freshwater	1.22	0.24

ASM169325	*Synechococcus* PCC7003	Group 5	Robertson et al. 2001	38	3.35	3073	Coastal	1.24	0.13
ASM1948	*Synechococcus* PCC 7002	Group 5	Robertson et al. 2001	35	3.41	3148	Coastal	1.11	0.00
ASM169327	*Synechococcus* PCC7117	Group 5	Robertson et al. 2001	37	3.43	3162	Coastal	1.17	0.00
ASM31708	*Synechococcus* PCC 7502	Unclassified	Walter et al. 2017	30	3.58	3442	Freshwater	0.87	0.28
ASM31668	*Synechococcus* PCC 6312	Unclassified		42	3.72	3528	Freshwater	1.19	0.25

Chapter III: Genomic features for cold adaptation of winter Chesapeake Bay *Synechococcus*.

Abstract

Synechococcus are abundant and important to aquatic ecosystems. They contribute significantly to the world's oceans primary productivity and are endemic to freshwater, estuarine, coastal, and pelagic environments. Diverse and unique *Synechococcus* are present in the Chesapeake Bay during cold winter months, they differ from the summer *Synechococcus* populations in the Bay. Seventeen strains of *Synechococcus* were isolated from the Baltimore Inner Harbor in the winter months, and they belong to 5 different phylogenetic clusters. Five Chesapeake Bay winter (CBW) strains (CBW1002, CBW1004, CBW1006, 1107 and CBW1108) were selected for genome sequencing, and they represent each major phylogenetic lineage that lacks genome sequences. The complete genome sequences from these five CBW strains allow us to explore their genomic characteristics and compare them with *Synechococcus* from different aquatic habitats. The genome size of these five CBW strains range from 3.20 Mb to 3.86 Mb, with CBW1002 and CBW1006 among the largest genome size for picocyanobacteria (~3.8 Mb). The five CBW strains have relatively high GC content (64 To 67%) and share many homologs that are unique and not shared with pelagic *Synechococcus.* CBW strains contain relatively high numbers of fatty acid desaturase, lipid A biosynthesis, chaperone, and transposase genes compared to coastal and open ocean *Synechococcus*, and these genes are known to play key roles in maintaining membrane fluidity, proper metabolite folding, and genome plasticity.

Introduction

Estuarine Synechococcus in winter

Synechococcus is a major cyanobacterial genus that contributes significantly to global primary productivity because of their abundance (Flombaum et al., 2013). *Synechococcus* has a ubiquitous distribution and has the capability to adapt to nearly every aquatic environment (Dvořák et al., 2014b; Waterbury, 1986). Many genomes of freshwater and marine *Synechococcus* strains have been reported and used for comparative genomics and phylogenomic analysis (Coutinho et al., 2016; Dufresne et al., 2008; B Palenik et al., 2003; Brian Palenik et al., 2006; Salazar et al., 2020), but few genomes of estuarine *Synechococcus* been reported. The draft genome sequences of estuarine *Synechococcus* WH5701, CB0101 and CB0205 were deposited in GenBank in 2006, 2009, and 2009, respectively. The first complete or closed genome of estuarine *Synechococcus* CB0101 isolated from the Chesapeake Bay was reported recently (Fucich et al., 2019). CB0101 was isolated from the Chesapeake Bay during the summer along with a dozen of other *Synechococcus* strains (F. Chen et al., 2004). These summer estuarine *Synechococcus* isolates have fortified the subcluster 5.2 lineage (F. Chen et al., 2006b). In the Chesapeake Bay, the winter picocyanobacterial community is dominated by distinct subpopulations which are not present in the summer (Cai et al., 2010). The abundance of picocyanobacteria exhibits a strong seasonal pattern in the Bay, high in summer and low in winter (K. Wang et al., 2011). In the winter season, the surface water of upper Chesapeake Bay can be frozen. *Synechococcus* cells are still present at the subzero temperature, but little is known about how they survive the cold and even freezing conditions in winter. To learn more

about the physiology of winter *Synechococcus*, 17 *Synechococcus* strains were isolated during the winter season (December 2010 to February 2011) in Baltimore's Inner Harbor (Xu et al., 2015). The growth data suggest that these winter isolates can grow in low temperature and have wide salinity tolerance. Winter *Synechococcus* isolates are not affiliated with subcluster 5.2 but related to several phylogenetic lineages for freshwater and brackish water picocyanobacteria. Of these Chesapeake Bay winter isolates, five strains (CBW1002, CBW1004, CBW1006, CBW1107 and CBW1108) were chosen for genome sequencing. The five CBW strains represent four different phylogenetic clades (subalpine cluster II, Bornholm Sea cluster, CB7 cluster, and the novel CBW1004 cluster) based on their 16S rRNA gene phylogeny (Xu et al., 2015). Many non-marine picocyanobacteria belong to subalpine cluster II, Bornholm Sea cluster, CB7 cluster, but no complete picocyanobacterial genomes have been reported for these well-defined clusters.

Cold Adaptation in Picocyanobacteria, Bacteria, and Beyond

Full genome sequences allow for broad genomic comparisons among the CB winter strains as well as model strains of interest. Availability of these sequences also allows for in depth comparison of specific genes with relevant known function. Cold stress response in cyanobacteria has been explored, based on what is known about cold stress in bacteria (Gualerzi et al., 2003; Weber & Marahiel, 2003). Major categories have been found to be important in the bacteria cold stress response signal recognition and transduction, membrane fluidity, protein folding, translational regulation.

Signal transduction is mediated by sensors and reporters, most famously the transmembrane histidine kinase *hik33,* which perceives cold and triggers response of many proteins such as desaturases and pathways for lipid production (Mikami et al., 2002; Suzuki et al., 2001). Hik33 is the histidine kinase involved in cold sensing in bacteria. However, the cold response is not solely controlled by *hik33*. Gene expression of desaturases and other response genes can be controlled: 1) completely by *hik33*, 2) partially controlled by hik33, or 3) unaffected by the *hik33*$^{-1}$ mutant (Suzuki et al., 2001).

Cold stress response in *Synechocystis* is controlled by multiple interacting genes and regulatory systems (Los et al., 2008). *Synechocystis* can modify membrane fluidity with alternate desaturases, without delta-12 and delta-6 desaturases (Mironov et al., 2012). Therefore, cold response genes can fall into two categories: 1) Genes that are controlled by *hik33*, membrane fluidity, and light and 2) Genes that are not controlled by *hik33*, which are membrane fluidity and light independent.

Other membrane proteins necessary to retain membrane fluidity is phosphatidic acid formation, a precursor to phospholipid formation. This process is controlled by *plsX, plsY (yneS),* and *plsC (yhdO).* In *Bacillus subtilis,* a knockout of *plsX* ceases both phospholipid and fatty acid synthesis completely, *plsY* and *plsC* knockouts only arrests fatty, denoting the importance of *plsX* (Paoletti et al., 2007).

Retention of membrane fluidity is a universal response to low temperatures by modification of fatty acids, usually through desaturation or synthesis (Barria et al., 2013; Los & Murata, 1999). There are several desaturase genes, with different methods of action, pathway,

placement, and metabolic products. The 3 main lineages of loosely related desaturases are delta-9 *(desC and desE)*, delta-12 *(desA)* /omega-3 (*desB* and *pfaA)*, and "front-end" or delta-5, delta-6, or delta-8 *(desD)* desaturases (López Alonso et al., 2003; L. Wang et al., 2020). These are believed to share a common origin because of a conserved histidine kinase region. There are many orthologs in various species across the tree of life, showing the universal necessity of fatty acid desaturation to retain membrane fluidity. In cyanobacteria, cold shock membrane composition change is achieved through acyl-lipid desaturases (Murata & Wada, 1995; Phadtare, 2004).

Chaperone proteins are also vital in cold acclimation and adaptation to ensure properly folded metabolites. The main bacterial cold shock protein in *E. coli*, *cspA,* is an RNA chaperone (Jiang et al., 1997) that prevents secondary structures of RNA at low temperatures. Inactivation of chaperonins such as GroE/L/S resulting in protein refolding failure have suggested the importance of chaperones in cold stress response (Strocchi et al., 2006). In the cyanobacterial system *Synechocystis* PCC6803, *hik*33 regulates protein chaperones such as DnaJ, GroEL, DnaK, and more (Mikami et al., 2002).

In addition to chaperones and desaturases, there are other miscellaneous genes associated with cold acclimation in bacteria and cyanobacteria. These include but are not limited to, translation initiation factors (*e.g. infB*), proteases (*e.g. HtrA*), RNA helicases (*e.g. deaD)*, and proteases (*e.g. HtrA* which is also regulated by *hik33* (Mikami et al., 2002), among others.

Although not explicitly linked to the cold stress response in bacteria, transposases are mobile genetic elements that are involved in genome plasticity. Transposases are high in both abundance and expression in bacteria, specifically the *Synechococcus*, found in the Baltic Sea (Vigil-Stenman et al., 2017). This environment is highly dynamic, with strong gradients of salinity, nutrients, oxygen, and temperature. Temperature ranges from -3 to 20 ^{o}C in the Baltic Sea, and the low temperature range is reminiscent of the CB during winter months.

Here, we investigate the genome sequences of five winter CB *Synechococcus*, and compared them with reference genomes of *Synechococcus* spp. isolated from freshwater, coastal and marine environments, in order to gain insight into cold adaptation of winter *Synechococcus*.

Materials and Methods

Strain Selection for Genome Sequencing

A total of five winter *Synechococcus* strains (CBW1002, CB1004, CBW1006, CBW1107, and CBW1108) were chosen for genome sequencing. CBW1002 and CBW1006 were selected because of their membership to the Bornholm Sea cluster. More than half of our winter isolates (9 out of 17) are in the Bornholm Sea cluster and are closely related to several picocyanobacteria isolated from the Baltic Sea (Ernst et al., 2003a). CBW1004 represents a novel branch without closely related *Synechococcus* in the phylogeny based on 16S rRNA sequences. CBW1107 represents subalpine cluster II, a well-defined phylogenetic cluster which contains freshwater and estuarine picocyanobacteria. CBW1108 was chosen for genome sequencing because it is the only cultured member in the CB7 clade, which is a predominant group of picocyanobacteria in the Bay during the winter time (Cai et al., 2010).

Selection of reference genomes

A total of 13 reference genomes were chosen to represent *Synechococcus* (except for *Synechocystis* PCC6803) in various habitats. Strains CB0101 and CB0205 were isolated from the Chesapeake Bay in the summer months, they represent estuarine *Synechococcus* in marine subcluster 5.2. WH8101 and CC9311 were chosen to represent coastal *Synechococcus*. Open ocean strains are represented by WH8102, WH7803, and RCC307. The five coastal and open ocean strains are all the members of marine subcluster 5.1. Four freshwater strains were selected to represent different genome sizes. Strains PCC6301 and UTEX2973 represent two freshwater *Synechococcus* with relatively small genome size, and PCC6312 and PCC7502 with relatively large genome size. *Synechococcus* CS-601 (SynAce01), a *Synechococcus* strain isolated from the Antarctic Ocean is included. A freshwater *Synechocystis* strain, PCC6803, was also included as a reference because it has been well studied with respect to cold adaptation and other stress responses (Mironov et al., 2012; Suzuki et al., 2001).

Sequencing Methods

The five CBW strains were grown in SN15 media (Xu et al., 2015). Genomic DNA was extracted by using phenol-chloroform (Kan et al., 2006). The DNA samples were sent to the Beijing Genome Institute (BGI) for sequencing. CBW complete genome sequences were obtained using a combination of Illumina HiSeq and PacBio Sequel platforms. For the raw reads from Illumina sequencing, low quality (<=20), high N nucleotide percentage (>10%), adapter and duplication reads were removed to obtain clean reads. For PacBio raw sequences, adapters and poor-quality reads were cut from polymerase reads to generate multiple subreads.

Subreads with less than 1,000 nucleotides were filtered out, and remaining subreads were integrated into one Circular Consensus Sequencing (CCS) read of insert. Subreads were corrected and constructed using Celera and Falcon to yield optimal assemblies. The yielded assemblies were checked with 2nd Generation Illumina seq for single-nucleotide correction (Quiver, GATK, SOAPsnp/SOAPindel) for the final assemblies. All five CBW strains are circular and contain only one chromosome with no plasmids.

To learn more about the cold stress response in genomes of Chesapeake Bay strains, several bioinformatic methods were used. Originally, a list of cold induced genes of known (n=64) and unknown function (n=47) was adapted from *Synechocystis* PCC6803 (Sinetova & Los, 2016). From this list of 111 genes, *in silico* homologs were predicted using reciprocal best hits (RBH) with blastp at the stringency level of e-value < 1e-20 (Table 3.5). Homologs were determined using reciprocal best hits of open reading frames with blastp (e-value < 1×10^{-10}) from one group of CDS to the other. Shared homologs are represented as ribbons between genomes. Highly shared homologs are in color (top 50^{th} percentile) while the fewest shared homologs are in grey. Stringency value was based on a histogram analysis of total hits at several different test e-values. To gain a better understanding of the broader bacterial cold stress response, a more simplistic blastp approach was adapted from Tang et al., 2019. A shorter list of genes from *E. coli* was used to tabulate the amount of blastp hits at a stringency of e-value < 1e-5. This gene list was adapted from Barria et al., 2013 and the e-value was chosen to replicate the methods of Tang et al., 2019. Finally, a manual text search of gene functions from automatic annotation was completed and verified using three sources, the RAST server (Aziz et al., 2008; Overbeek et al., 2014), PATRIC, the bacterial bioinformatics resource center (Brettin

et al., 2015; Davis et al., 2020), and the Beijing Genome Institute (BGI) standard output. For reference sequences, feature tables were downloaded from NCBI and compared to feature table for CBW strains. Coupling these annotations with a literature search, genes implicated in cold adaptation *i.e.* fatty acid desaturases, chaperones, and transposases were compared.

Results

Genomic Comparisons

The genome sequencing of all five (CBW1002, CBW1004, CBW1006, CBW1107, and CBW1108) resulted in one circular contig per genome. Genome sizes of the five CBW strains ranges from 3.20 Mb (CBW1107) to 3.86 Mb (CBW1006). Their GC content ranges from 64.35% (CBW1108) to 67.35% (CBW1004). For *Synechococcus,* the genome sizes and GC contents of these five strains are both relatively high, averaging at ~2.5 Mb and ~58.9% respectively. Complete genome sequences have been submitted to NCBI under the accession number PRJNA657291. Individual accession numbers can be found in Table 3.1.

Interestingly, the two winter *Synechococcus* strains (CBW1002 and CBW1006) in the Bornholm Sea cluster contain very large genomes. The genome size of CBW1002 and CBW1006 is 3.85 Mb and 3.86 Mb, respectively. These genome sizes are among the largest genomes for *Synechococcus*. According to the publicly available JGI and NCBI databases, the largest complete *Synechococcus* genome was the freshwater *Synechococcus* PCC6312 (3.72 Mb) in current databases. The CBW core genome contains 1,295 ORFs shared between all five strains, and a pan-genome of 8,274 ORFs. Around 49% of the genes remain hypothetical proteins without any

functional prediction. CBW strains have large genomes, an expansive pan-genome, and largely unknown functional annotation.

Functional annotation of the CBW strains was compared using the RAST server (Aziz et al., 2008; Brettin et al., 2015; Overbeek et al., 2014). Annotated genes are categorized into subsystems based on predicted metabolic activity automatically. Subsystem coverage and subsystem breakdown give an idea of the proportion of the genome with a putative functional assignment (Figure 3.1). CBW strains are compared to marine WH8102 and coastal CC9311. It should be noted that the annotation for the latter two strains is much more complete than any CB strain. They have more subsystem coverage (45% and 40%, respectively) than any CB strain (only ~25%). CBW strains contained between 35-38 stress response genes, including genes involved in osmotic, oxidative, and detoxification stress. Chesapeake Bay strains contain a variable number of phage and prophage elements. CBW1107 had six phage elements, while CBW1002, CBW1004, and CBW1006 had two and CBW 1108 had just one. Annotation of CB0101 did not predict any phage elements. The summer strain CB0101 had more phosphorous metabolism genes (n=48) than any CBW strain which ranged between as few as 19 (CBW1107) and as many as 37 (CBW1002). The subsystem for fatty acid metabolism contains between 19-31 genes for CBW strains. Fatty acid desaturases are the subject of closer investigation in this work. For many CBW strains, subsystem totals are fewer than CB0101 (Fucich et al., 2019), a CB strain isolated from the Inner Harbor during the summer. This is the result of all CBW strains contain more genes that fall outside of subsystems than CB0101. These unique genes have unknown functions yet to be determined.

CBW strains share unique homologs among themselves. In congruence with the phylogeny described previously (Xu et al., 2015), all CBW strains share more homologs with each other than with CB0101 (subcluster 5.2), marine *Synechococcus* WH8102, or freshwater *Synechocystis* PCC6803 (Figure 3.2). The proportion of a strains shared homologs can be seen on the outer most ring, while the number of homologs shared between any two strains can be read on the inner ring and in supplemental Table 3.6. CBW1002 and CBW1006 share the most homologs among the CBW strains (n=3,023), have the largest genomes and are the most closely related with regard to the partial 16S rRNA gene marker (Xu et al., 2015).

Cold induced genes in Chesapeake Bay Winter Synechococcus strains

All CBW strains contain the *hik33* homolog, a histidine kinase involved in cold sensing and transcriptional regulation (blastp evalue=0, pident >50%) (data not shown). This two component module is responsible for recognition and cold response regulation in *Synechococystis* sp. PCC 6803 (Sinetova & Los, 2016). *Hik33* is also present in other *Synechococcus* strains.

CBW strains contain many genes implicated in the bacterial cold response. Originally, a list of 111 upregulated genes from *Synechocystis* PCC6803 from (Sinetova & Los, 2016) was compiled and queried against available CBW strains (Table S4.6) as RBH. Some of these amino acid sequences were involved in the general stress response, whole some were specific to cold stress (highlighted in blue). Many of these genes were upregulated during stress without known function (Sinetova & Los, 2016). CBW strains contained 43 to 47 of these 111 stress-related genes as reciprocal best hits (Table 3.5). The results of this RBH search were inconclusive, as

few CBW strains contained more *in silico* homologs of these stress genes than CB0101 (43) and none contained more than the marine WH8102 (54). This could be the result of a high threshold (e-value 1e-20) and the cross-genus nature of the blastp experiment. In any case, further methods, including further blastp and text annotation searches were used. Threshold for this experiment was determined using a histogram of hits to retain a high degree of specificity without missing potential *bona fide* homologs. This is a different method than used in Table 3.2, where the methods in Tang et al., 2019 were replicated.

Cold stress responses have been well studied in bacteria and cyanobacteria, and the genes involved in cold shock have been summarized (Los & Murata, 1999; Weber & Marahiel, 2003). These genes have been used to search for the presence of cold stress genes in bacteria and picocyanobacteria (Barria et al., 2013; Tang et al., 2019). To compare the presence of these cold stress genes in CBW strains and reference strains, 28 genes were queried against all 18 *Synechococcus* genomes in this study. Amino acid sequences were sourced from *E. coli* K-12 (Barria et al., 2013). A simple blastp search (e-value < 1e-05) was performed to count occurrences of these cold implicated amino acid sequences in five CBW *Synechococcus*, and 13 reference strains (Table 3.2).

CBW strains tended to have multiple copies of the *deaD* helicase (~6), *dnaJ* molecular chaperone (~8), and *infB* initiation factor 2 (~6). These tended to be higher than open ocean strains and freshwater strains with both large and small genomes (Table 3.2). Overall, CBW strains tended to have the same or a few more blastp hits to genes implicated in the bacterial cold stress response. Strain CBW1108 had surprising blastp hits for 2 genes, *lpxP* and *otsB*. CBW1108 was the only picocyanobacterial strain with significant sequence similarity to *lpxP,* a

palmitoleoyl transferase, and *otsB,* a cold induced trehalose phosphate phosphatase. These alignments resulted in good query coverage (84% and 100%, respectively), but lukewarm percent identities (21% and 31%, respectively).

lpxP is a palmitoleoyl transferase which is an enzyme involved in the lipid A pathway and is induced by cold shock at 12^{o}C in *E. coli* (Carty et al., 1999). Cyanobacteria are known to have the first four enzymes in the pathway: *lpxA, B, C,* and *D*, but not *lpxP*. As a result, the pathway is only capable of producing lipid A disaccharide, what is believed to be a 'primordial form' of lipid A (Opiyo et al., 2010). Further inspection of PARTIC annotation reveals that CBW1108 has 26 genes involved in the cellular envelope maintenance, many of them with subclasses involved in lipid A biosynthesis.

otsB encodes for a trehalose phosphate phosphatase which, together with *otsA*, plays a critical role in bacterial viability at low temperatures (Kandror et al., 2002). Most picocyanobacterial strains had a hit to *otsA*, with notable exceptions being CBW1004 and all freshwater strains PCC6301, UTEX2973, PCC6312, and PCC7502. CBW1107 and CBW1108 actually had two hits to *otsA.* Conversely, *otsB* was rare as it had only one significant hit to CBW1108.

CBW strains contain many annotated genes that could function to maintain membrane fluidity and promote proper metabolite folding. Several desaturase and chaperone amino acid sequences from CBW strains were used as queries. These include genes implicated in the cold stress response, namely chaperones (*Htp, hslO, GrpE, ComM, GroEL, GroES, HtrA)* and desaturases (ctrQ, *HopC, desE, desE2, fad, desE3, Slr1293, PfaA,* and *ERG3)*. These sequences

originated in multiple CBW strains (CBW1002, CBW1004, CBW1006, and CBW1108) and were queried against the CBW strains, as well as representative picocyanobacterial strains from estuarine, open ocean, coastal, and freshwater environments (Table 3.3). A blastp search based on the amino acid sequences of these key genes was performed with e value of 1e-05.

For certain desaturases such as the pro-zeta carotene desaturase *ctrQ and* squalene/phytoene desaturase *HopC*, CBW strains and freshwater *Synechococystis* PCC6803 seemed to have more hits than *Synechococcus* from freshwater, coastal, and marine sources. The sterol desaturase, *ERG3,* was sourced from CBW1108, and had hits in all freshwater strains. However, it only had hits to CBW1107 and CB0205, and the coastal WH8101. *PfaA* was sourced from CBW1004 and had 10 instances of sequence similarity, the most of any other genomes. Other CBW strains also had multiple hits to *PfaA*. Closer inspection found that these hits have poor query coverage as they were only gene fragments. The chaperone *HtrA* generally has more hits to CBW strains than freshwater, coastal, and open ocean strains.

Finally, the occurrence of fatty acid desaturase, transposase, and chaperone among these 18 *Synechococcus* genome annotations was compared (Table 3.4). This was completed with a simple text search of gene names against the feature annotation tables. Freshwater *Synechococcus* PCC7502 with a large genome has the highest number (223) of transposase genes, higher than *Synechocystis* PCC6803 which has 116 transposase genes. The number of transposase genes varied from 7 to 59 among the five CBW strains. CBW1002 and CBW1006 had 59 and 35 transposase genes (Table 3.4), and the high number of transposase genes in these two strains is consistent with their large genome sizes. The other three CBW strains (CBW1004, CBW1107 and CBW1108) had fewer transposase genes (7-15). The coastal and open

ocean *Synechococcus* strains appear to lack or have very few transposase genes except for WH8101. No clear trend could be seen when the term "desaturase" was searched. The number of desaturase genes varied from 6 to 16 among the 18 picocyanobacterial strains. PCC7502 has the highest number of desaturases. When searched against the term "fatty acid desaturase", the number of fatty acid desaturase genes varied from 1 to 7 in all the 18 picocyanobacterial strains. CBW strains had 3-5 fatty acid desaturase genes, similar to many reference strains. Coastal strain PCC9311 and open ocean strain RCC307, and freshwater strain UTEX2973 only contained one fatty acid desaturase gene. In general, searching with "fatty acid desaturase" resulted in nearly one third of the results searched by "desaturase" ("desaturase" avg=9.11, "fatty acid desaturase" avg=~3.61). This indicates that many other desaturases are present in picocyanobacteria, and their role in cold adaptation is not known. CBW strains contained 16 to 21 chaperone genes, higher than all the reference strains. Most of marine and freshwater *Synechococcus* contained 9-12 chaperone genes (Table 3.4).

Discussion

Desaturases

Three homologs of delta-9 TA desaturases were found in all CBW strains and at least 1 omega-3 desaturase, *pfaA*, only found in CBW1004. Delta-5 desaturases are not found in CBW strains which is expected as delta-5 desaturases are not present in higher plants and cyanobacteria (López Alonso et al., 2003). In *Synechococcus* PCC 7002, the delta-9 desaturase *desE* is responsible for the formation of the double bond of 1,14-nonadecadiene, a hydrocarbon that accumulated when cells are grown at low temperatures (Mendez-perez et al., 2014). *ΔdesE*

knock outs showed that at these low temperatures, this desaturase is necessary for growth. In cyanobacteria, the amount of desaturases present is likely related to habitat, given that thermophilic *Synechococcus* only have 1 delta-9 desaturase in contrast to mesophilic *Synechococcus,* which normally contain 2 delta-9 desaturase genes (Chi et al., 2008). CBW1004 is unique in that it also contains an omega-3 desaturase *pfaA*. Blastp shows that there are hits in other CBW strains (Table 3.3). However, most of these hits are incomplete with poor percent identity and low conservation (data not shown).

Polyunsaturated fatty acids (PUFAs) are directly related to the fluidity of biological membranes (Sakamoto & Murata, 2002). Together with PUFAs, pigments are believed to play a role in membrane fluidity. Carotenoids are believed to be an important component of the cold adaptive strategy in *Staphylococcus xylosus* (Seel et al., 2020) while chlorophyll-a and PUFAs were core components of the cellular membrane in the cyanobacteria *Nodularia spumigena* CHS1 (Hassan et al., 2020).

Other desaturases were annotated in CBW strains. For example, a zeta-carotene desaturase was annotated in the CBW strains and has between 3-5 copies in the strains endemic to the Chesapeake Bay. Other *Synechococcus* strains do not have any hits to this interesting gene besides WH8102, WH7803, and *Synechocystis* PCC6803. Such a carotenoid desaturase would be seemingly unrelated to cold response, despite some evidence in PCC7803 showing that *crtQ,* a 9 9-di-cis-zeta-carotene desaturase expression is constant in low temperature despite oscillating low and high light conditions (Guyet et al., 2020).

Chaperones

Protein folding, fatty acid and phospholipid synthesis are important for the bacterial adaptation of many bacteria to survive in cold environments. Various fatty acid desaturases, chaperone proteins, and other code associated genes are regularly predicted in CBW strains, often in higher abundance than open ocean *Synechococcus* strains (Table 3.2). Among another cold adapted *Synechococcus* strain, SynAce01 compared to other picocyanobacteria, no major difference in number of homologs was found between genomes (Tang et al., 2019). This is not true when expanding both the cold induced genes list and reference genome list in comparison to the CB winter strains. CBW strains tend to have more copies of select cold induced genes of interest.

In *E. coli*, DnaK, DnaJ, and GrpE are considered heat shock proteins which refold denatured proteins. Under temperature stress, these genes can arrest the refolding of such proteins, and recoverably resume refolding after the stress condition has been removed (Diamant & Goloubinoff, 1998). Various chaperones, such as *DnaJ* and DnaK have more blastp hits among the CBW strains than other *Synechococcus* representatives. Three of the CBW strains contain the most hits to *DnaJ* (8), while other strains such as WH8101 and CC9311 contain only half as many hits.

CBW1002 and CBW1006, two very closely related strains, with some of the most homology (Figure 3.1) contain 5 hits to *hscA,* a *DnaK*-like molecular chaperone (Table 3.2). In *Shewanella sp. Ac10,* DnaK increased ATPase activity at low temperatures than the DnaK in *E. coli*, which is characteristic of a cold active enzyme (Yoshimune et al., 2005). Determining the expression and ATPase activity of these copies of *hscA* under cold temperatures could indicate cold adaptation.

Surprisingly, all known *E. coli* cold shock protein are absent from CBW strains, and all other *Synechococcus* in the survey. This may suggest that CBW strains may have a differential cold shock mechanism, or that the *E. coli* sequences were not similar enough to the *Synechococcus* strains to transcend the e-value of 1e-5. Two copies of *GroEL* were found in all CBW, but this result is not different than other strains.

Transposases

These mobile genetic elements are not generally conserved among species and have loose relation to the bacterial cold response. Namely, for their apparent duplication in the psychrophilic *Methanococcoides burtonii.* Upon close inspection of the CBW strain annotations, there are 37 transposase genes shared among the strains with multiple duplications in each genome. The most interesting and abundant transposase in CBW strains is an IS5 family variant of transposases. This particular transposase (CBW1006GL001879 locus=Chromosome1:1864150:1865853:+) is frequently duplicated and highly conserved among CBW strains. IS5 transposase is present in CBW genomes between 27 and 33 times (Figure 3.3). In the Baltic Sea, cyanobacteria were responsible for ~40% of IS5 transposase metagenomic reads and ~50% of the metatranscriptomic reads for the appropriate size fraction (0.8-3.0 µm) (Vigil-Stenman et al., 2017). To search for this particular IS5 transposase in other *Synechococcus* genomes, a blastp was conducted against the coding sequences in a similar fashion to Table 3.3. Partial duplications were removed by applying a threshold of e-value < 1e-5, percent identity > 35%, and query coverage > 50%.

Other cold induced genes in CBW Synechococcus

Various elements such as helicases, initiation factors, and histidine kinases are involved in the bacterial cold stress response. CBW strains contain some of these elements in higher frequencies than their temperate *Synechococcus* representatives (Tables 3.2, 3.3, and 3.4).

First, CBW1107 contains the most hits (7) to the *deaD* helicase, while all other CBW strains contain 5, save CBW1004, with 4 hits. Compared to coastal, freshwater, and especially open ocean strains, which only have 1-3 copies (Table 3.2). This is significant as RNA helicases are involved in the cold acclimation in cyanobacteria (Chamot et al., 1999). Further, *deaD* helicase is essential for survival at low temperatures for *Caulobacter crescentus* (Aguirre et al., 2017).

Initiation factors *(infA, infB,* and *infC)* are often implicated in bacterial cold response (Barria et al., 2013). While initiation factor 1 *(infA)* and 3 *(infC)* contained 1 hit in all CBW and all other picocyanobacterial representatives, initiation factor 2 *(infB)* was highly variable among strains. CBW strains have among the most copies of *infB* (6), second only to the Antarctic SyneAce01 (7) (Table 3.2).

All CBW strains contain at least one copy of the histidine kinase *hik33.* Additionally, according to the BGI annotation (data not shown), CBW strains contain five core histidine kinases; 3 signal transduction histidine kinase and 2 K+ sensing *KdpD* (their function is unclear, but they could sense turgor pressure or osmolarity) (Mascher et al., 2006).

CBW strains contain important genes for phosphatidic acid formation, a precursor to membrane phospholipids (Paoletti et al., 2007). All CBW winter strains contain *plsX* (acyl-acyl carrier protein [ACP]: phosphate acyltransferase) and *plsY* (acyl-phosphate: glycerol-phosphate

acyltransferase). These genes were annotated using RAST annotation as well as annotation completed by the Beijing Genome Institute.

Interestingly, the presence of *plsC* (acyl-ACP:1-acylglycerol-phosphate acyltransferase) was not present in any CBW annotation or blastp search (hits had high [0.09] e-values and low percent identities [~23%]) (data not shown) The function of *plsC* was the least necessary when compared to the other *pls* genes during phosphatidic acid formation, as fatty acid formation continued at a high rate despite its deletion (Paoletti et al., 2007).

CBW strains contain relatively large genome sizes (between 3.2-3.8 Mb) and high GC content (between 64-67%). These features are comparable to the genomes of several freshwater picocyanobacteria, particularly freshwater *Synechococcus* spp. and *Cyanobium* spp. CBW1002 and CBW1006 have the largest genome size among known picocyanobacterial genomes. They are members of the Bornholm Sea cluster, which contain most the picocyanobacteria isolated from the Baltic Sea and Chesapeake Bay. It is possible that picocyanobacteria in this phylogenetic lineage share the characteristic trait of large genome size with high GC content. Given what is known about evolutionary genomics, this free living clade may have such features as a result of exposure to more complex and variable environments and therefore have a higher chance to exchange genes horizontally (Mann & Chen, 2010).

CB0101 shares many homologs with CBW strains compared to the marine WH8102 and freshwater *Synechococystis* PCC6803. These results suggest that homologs among the CB strains could harbor genes that are important for surviving in the Chesapeake Bay. The high

number of homologs shared by CBW1002 and CBW1006 displays that the phylogenetic relationship based on 16S rRNA is accurate and could be confirmed using further phylogenomic studies.

Except for a few genes *(deaD*, *dnaJ*, and *infB*) CBW strains are similar to the reference strains in terms of the occurrence of well-studied cold stress genes from *E. coli*. CBW have 1-2 more copy of these genes *(deaD*, *dnaJ*, and *infB*) compared to open ocean *Synechococcus* (Table 3.2). Although this comes as a surprise, a recent study also showed no clear indication of more cold stress genes in the genome of Antarctic *Synechococcus* sp. CS-601 (SynAce01) was sequenced (Tang et al., 2019). Previous analysis of reciprocal blast hits on stress and cold induced genes from *Synechocystis* PCC6803 (Sinetova & Los, 2016) had not provided a clear indication about how these CBW *Synechococcus* are capable of cold adaptation (Table 3.5). Early work on *Synechocystis* PCC6803 was focused on hik33 the histidine kinase which senses cold and osmotic stress (Mikami et al., 2002; Suzuki et al., 2001). It is possible that other searching methods and different genetic markers should be considered.

CBW strains contain more chaperone proteins than selected reference genomes, this is especially true of *HtrA* (Table 3.3) and *dnaJ* (Table 3.3). Chaperones are known to be vital for proper protein folding at cold temperatures (Strocchi et al., 2006). CBW strains may have more chaperones than *Synechococcus* in stable, warm climates with less seasonal temperature change. Cold Shock Proteins like CspA act as RNA chaperones to prevent mRNAs from forming secondary structures at low temperatures (Jiang et al., 1997; Kishor PB, 2019). Interestingly, no cold shock proteins from *E. coli* had significant amino acid sequence to any picocyanobacterial

reference strain. Perhaps a more closely related closer cyanobacteria would result in sequence homology specific enough to overcome the threshold.

Transposases are also enriched in the CBW strains compared to open ocean *Synechococcus*. It appears that some freshwater picocyanobacteria with large genome size also contain high amount of transposase genes. In general, the number of transposase genes in CBW strains is close to that of those freshwater *Synechococcus* with large genomes. It is notable that coastal strain WH8102 contains 52 transposase genes, while oceanic strains contain one or no transposase genes. Transposons are often vehicles for horizontal gene transfer and can be responsible for a substantial portion of the genome. By some annotations, the transposases are poorly characterized and have generic names or are simply a hypothetical protein, despite that they are highly conserved at the amino acid level with high query coverage. Ability to gain genetic function via horizontal gene transfer could be important for picocyanobacteria living in the stressed environment, including the cold winter. These could be implicated in the transfer of other stress related genes, such as toxin antitoxin systems which are encoded on transposons (Lima-Mendez et al., 2020). The abundance and the association pattern of TA systems in CBW and other reference strains are analyzed in detail in chapter 4.

Conclusion

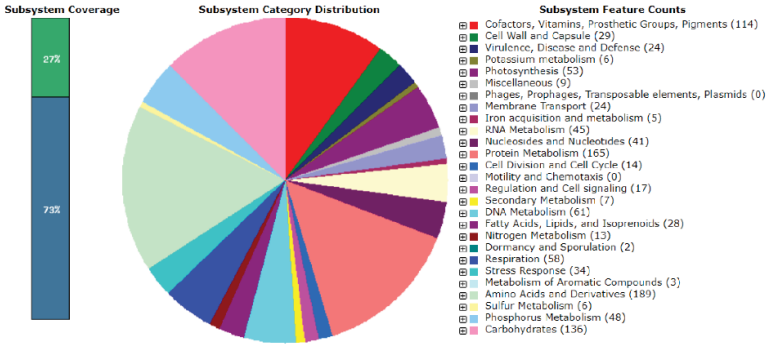
CBW genomes have some of the largest genomes among the *Synechococcus* with relatively high GC content. While they represent four distinct phylogenetic clades, they still share many homologs with cryptic gene function. Their genomes contain stress related,

phosphorous metabolism, and even phage genes, but between 50-70% of their coding sequences still fall outside of a subsystem function.

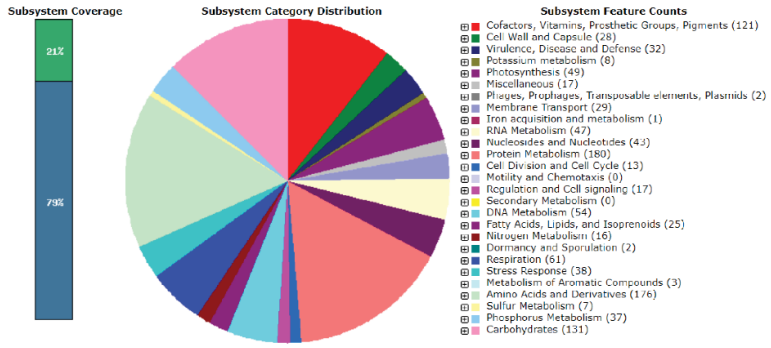
CBW estuarine strains generally tend to contain more cold induced and cold stress associated genes than freshwater, coastal, and open ocean *Synechococcus.* These genes were implicated using amino acid similarity and automatic annotation. Their genomes are equipped with desaturase genes and lipid A enzymes to maintain membrane fluidity, chaperone proteins for proper protein folding, and others for sensing a drop in temperature such as *hik33.*

A group of highly conserved transposases with around duplications in CBW strains, but only few in WH8101 and SyneAce01 was also found. It is unclear what the potential function of homologs for transposases in the CB winter strains is, and why there are over 20 copies in each strain, while there are very few in most other *Synechococcus.* Their highly conserved nature among genes within each genome suggests that they are the result of paralogous duplications, with few exceptions.

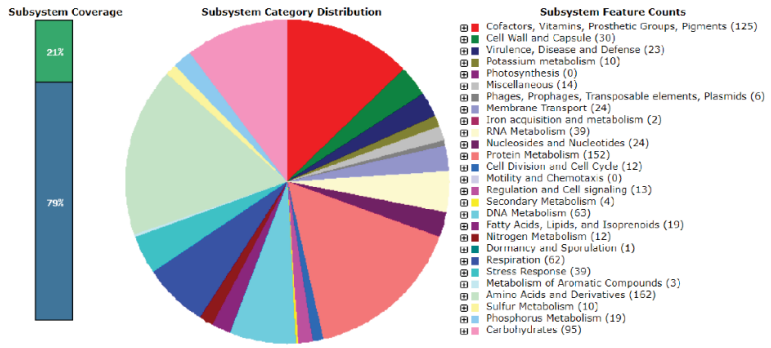
Figures



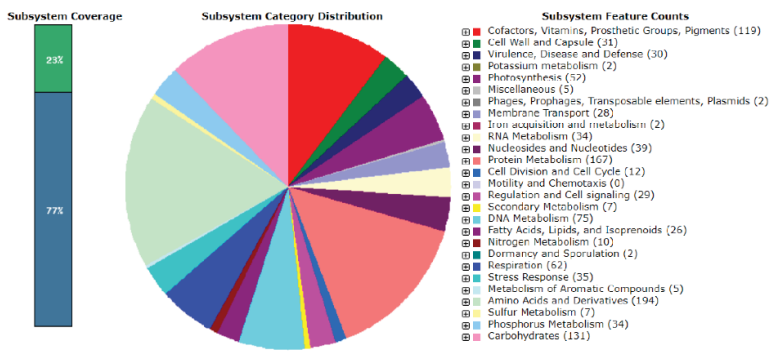
CB0101



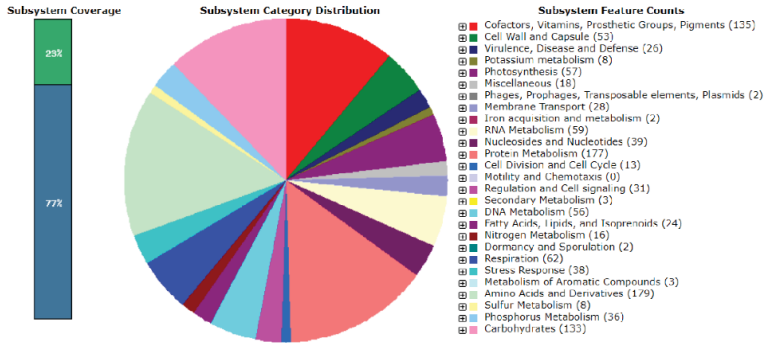
CBW1002



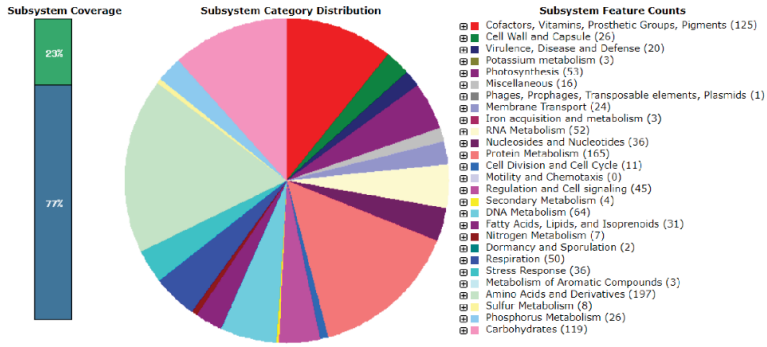
CBW1107



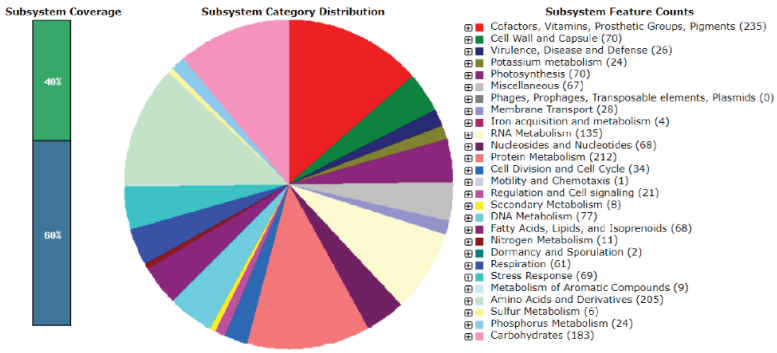
CBW1004



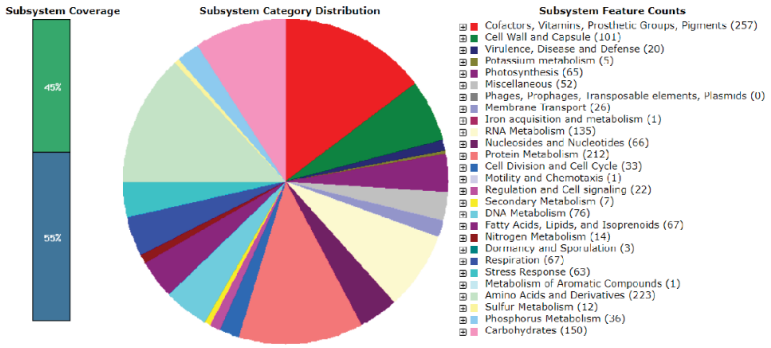
CBW1006



CBW1108



CC9311



WH8102

Figure 3.1. RAST annotation subsystem coverage and breakdown for each Chesapeake Bay strain. CB0101 represents a summer strain while CBW1002, CBW1004, CBW1006, CBW1107, and CBW1108 represent Chesapeake Bay winter strains. Many annotated genes fall outside of functional subsystems indicating the cryptic function of many CB *Synechococcus* coding sequences.

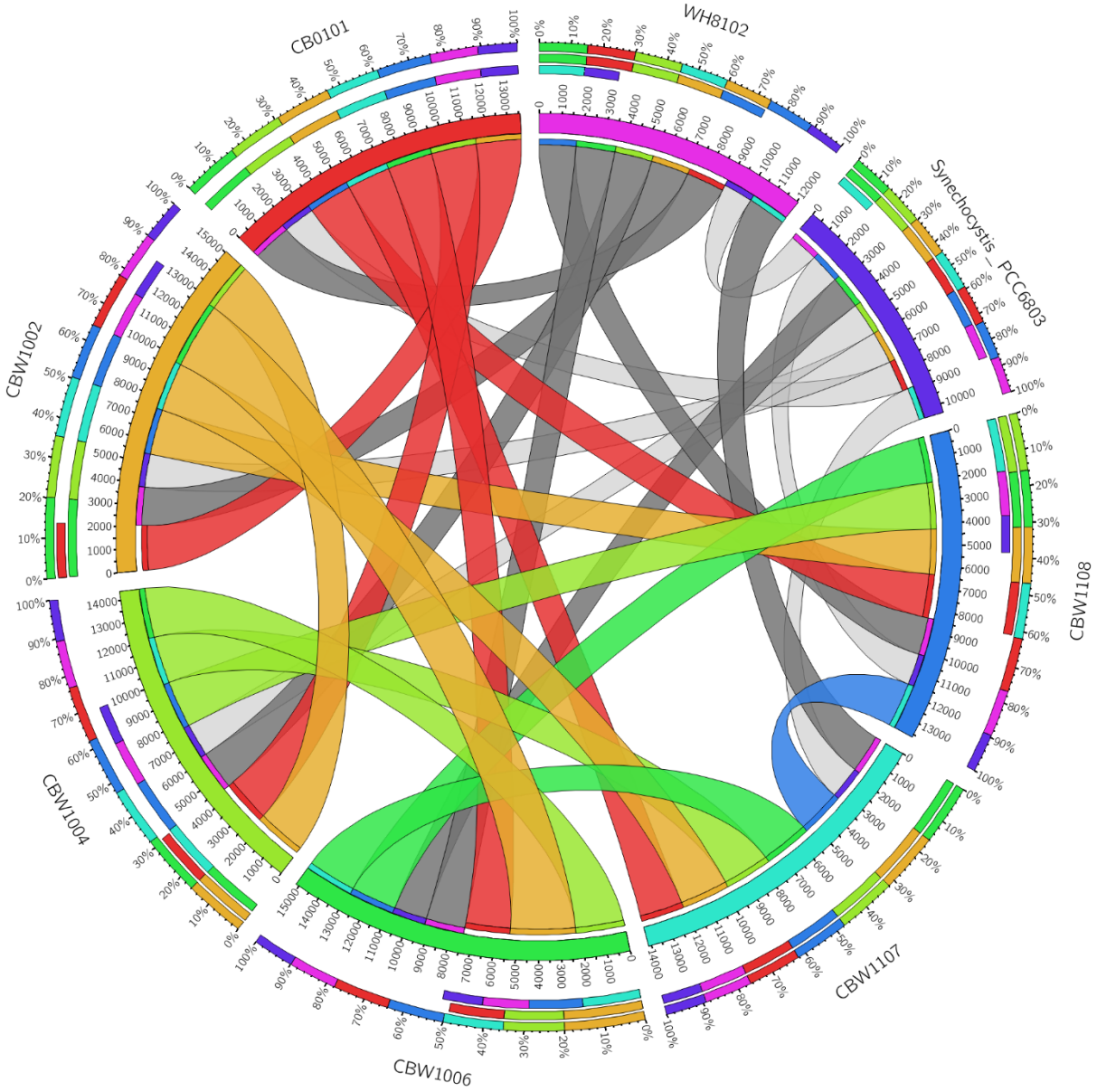


Figure 3.2. Homologs between selected picocyanobacteria genomes. Chesapeake Bay strains include 5 winter (CBW) and 1 summer (CB0101) strain. Model marine strain WH8102, and model freshwater strain *Synechocystis* PCC6803 were included as outgroups for comparison. Highly shared homologs are in color (top 50th percentile) while the fewest shared

homologs are in grey. Homologs were determined *in silico* using reciprocal best hits were using the blastp (e value < 1 x 10^{-10}) of amino acid sequences from open reading frames.

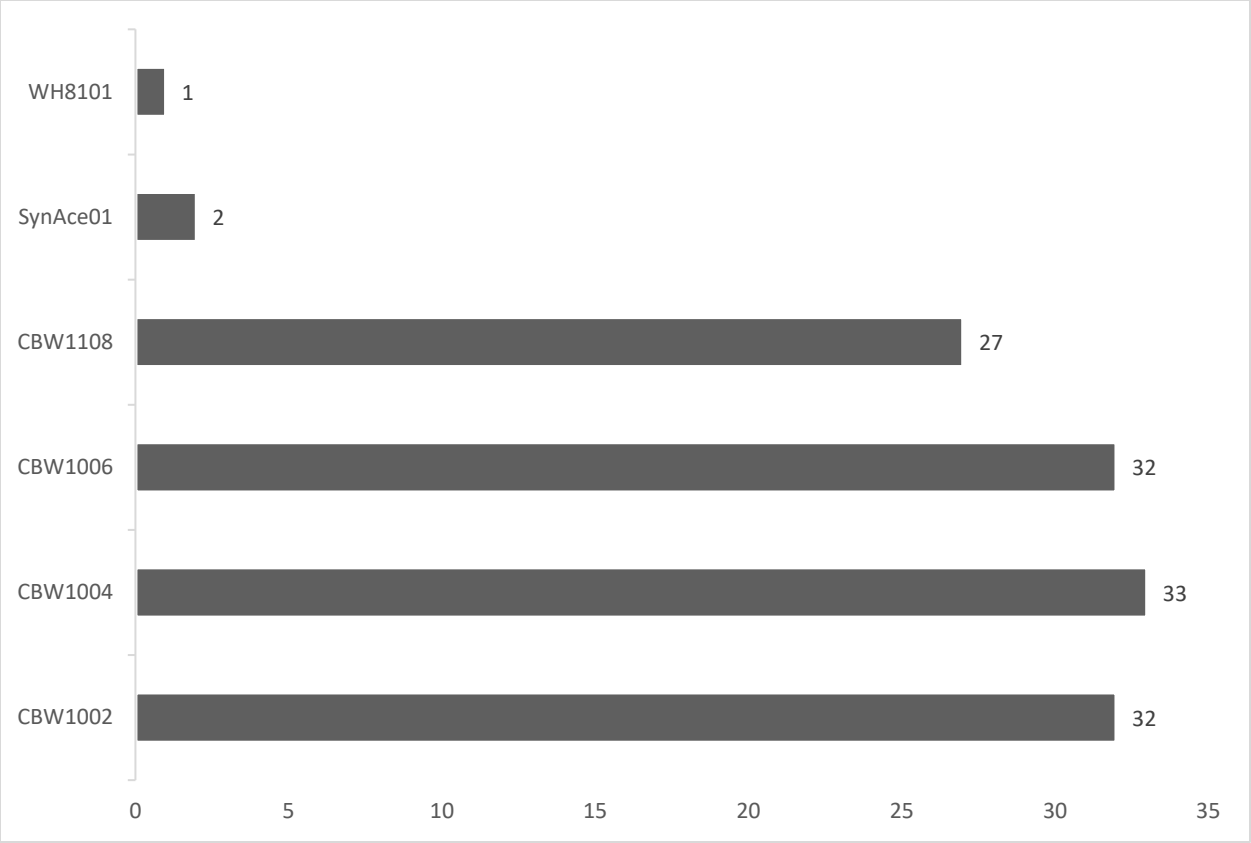


Figure 3.3. IS5 family transposase duplications in each *Synechococcus* genome. Genomes lacking significant similarity to this particular IS5 transposase were omitted. Blastp was performed using the CBW1006GL001879 locus=Chromosome1:1864150:1865853:+ CDS with thresholds of e-value < 1e-5, percent identity > 35%, and query coverage > 50%.

Table 3.1. Complete genome information for *Synechococcus* strains isolated from the Inner Harbor, Chesapeake Bay during the winter months (December 2010 to February 2011).

Synechococcus strain *name*	Length (bp)	GC % Content	Gene Number	% CDS	ncRNA	Accession Number
CBW1002	3,854,122	65.15	3,994	87.54	61	CP060398
CBW1004	3,672,318	67.35	3,668	87.41	83	CP060397
CBW1006	3,860,130	65.08	4,047	87.65	62	CP060396
CBW1107	3,202,093	66.86	3,446	88.90	54	CP064908
CBW1108	3,226,220	64.35	3,744	88.50	48	CP060395

Table 3.2. Numbers of blastp (threshold set to e-value of 1e-5) hits for genes implicated in cold stress response. Genes were derived from *E. coli* largely from lists in Barria et al., 2013 and Tang et al., 2019.

			Synechocystis	*Synechococcus*																	
			Freshwater	Freshwater Small Genome		Freshwater Large Genome		Artic	Summer		Winter					Coastal		Open Ocean		
Gene	Description	Reference	PCC6803	PCC6301	UTEX2973	PCC6312	PCC7502	SynAce01	CB0101	CB0205	CBW1002	CBW1004	CBW1006	CBW1107	CBW1108	WH8101	CC9311	WH8102	WH7803	RCC307
deaD	DEAD-like RNA helicase	Tang et al. 2019	4	1	1	1	2	5	4	4	5	4	5	7	5	4	3	2	3	2
desA	Fatty acid desaturase	Tang et al. 2019	2	0	0	0	2	2	2	1	1	1	1	1	2	0	0	1	2	1
dnaA	Replication initiation protein	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dnaJ	Molecular chaperone	Tang et al. 2019	7	7	7	7	7	6	5	6	8	6	8	8	5	4	4	6	5	6
gyrA	DNA gyrase subunit A	Tang et al. 2019	2	2	2	3	2	3	2	2	2	2	2	2	2	2	2	2	2	2
hscA	DnaK-like chaperone	Lelivelt & Kawula (1995)	4	5	5	5	6	3	4	5	5	2	5	4	4	2	4	4	4	2
hupB	Nucleoid protein, DNA supercoiling	Giangrossi et al. (2002)	1	2	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1
infA	Translation initiation factor IF-1	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
infB	Translation initiation factor IF-2	Tang et al. 2019	7	4	4	5	4	4	6	6	6	6	6	6	6	4	4	5	6	4
infC	Translation initiation factor IF-3	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
lpxP	Lipid A synthesis; cold-inducible	Vorachek-Warren et al. (2002)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
lpxA	Lipid A synthesis; cold-inducible	Opiyo et al.,2010	2	3	3	3	3	2	2	2	2	2	2	2	2	1	2	2	2	3
lpxB	Lipid A synthesis; cold-inducible	Opiyo et al.,2011	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1
lpxC	Lipid A synthesis; cold-inducible	Opiyo et al.,2012	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
lpxD	Lipid A synthesis; cold-inducible	Opiyo et al.,2013	2	4	5	3	3	2	2	2	2	2	2	2	2	2	3	3	2	3
nusA	Transcription termination/antitermination/elongation L factor	Bae et al. (2000)	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1
otsA	Trehalose phosphate synthase; cold- and heat-induced, critical for viability at low temperatures	Kandror et al. (2002)	1	0	0	0	0	1	1	1	1	0	1	2	2	1	1	1	1	1

otsB	Trehalose phosphate phosphatase; cold- and heat-induced, critical for viability at low temperatures	Kandror et al. (2002)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
pnp	3′–5′ exoribonuclease; component of RNA degradosome; cold shock protein required for growth at low temperatures	Yamanaka & Inouye (2001)	3	1	1	1	1	2	2	2	2	2	2	2	2	2	2	1	1	2
rnr	3′–5′ exonucleases; increases 10-fold in cold shock	Cairrão et al. (2003)	2	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
mtnA	Translation initiation factor IF-2B subunit alpha	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
pdhA	Pyruvate dehydrogenase E1 subunit alpha	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
pdhB	Pyruvate dehydrogenase E1 subunit beta	Tang et al. 2019	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
rbfA	Ribosome-binding factor A	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
recA	Recombination and DNA repair	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tig	Protein-folding chaperone	Tang et al. 2019	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
yfiA	Protein Y, associated with 30S ribosomal subunit, inhibits translation	Di Pietro et al. (2013)	0	0	0	0	1	1	0	1	1	0	1	0	0	0	0	0	0	0

Table 3.3 Numbers of blastp (threshold set to e-value of 1e-5) hits for desaturase and chaperone genes with functions likely involved in the bacterial cold response. Query sequences originated in CBW strains and were queried to other reference picocyanobacterial strains.

		Synechocystis	*Synechococcus*																	
		Freshwater	Freshwater Small Genome		Freshwater Large Genome		Artic	Summer		Winter					Coastal		Open Ocean			
Gene	Description	PCC6803	PCC6301	UTEX2973	PCC6312	PCC7502	SynAce01	CB0101	CB0205	CBW1002	CBW1004	CBW1006	CBW1107	CBW1108	WH8101	CC9311	WH8102	WH7803	RCC307
ctrQ	Pro-zeta-carotene_desaturase	5	3	3	2	2	4	3	3	4	3	5	4	4	4	4	4	4	2
HopC	Squalene/phytoene_desaturase	3	2	2	2	2	2	3	3	3	2	2	3	3	3	3	3	3	2
desE	Delta-9_fatty_acid_desaturase	1	1	1	2	3	3	3	2	3	3	3	4	3	2	2	1	2	2
desE2	Delta-9_fatty_acid_desaturase	1	1	1	2	3	3	3	2	3	3	3	4	3	2	2	1	2	2
fad	Generic FA des	2				1	3	3	3	1	3	1	3	2		1	2	2	1
Slr1293	Neurosporene_C-3',4'_desaturase	2	2	2	1	1	2	2	2	2	4	2	2	3	2	1	3	3	2
PfaA	omega-3 polyunsaturated fatty acid synthase subunit, PfaA	3	2	2	3	3	2	2	2	5	10	5	3	6	1	4	3	3	2
ERG3	sterol desaturase family protein	1	1	1	1	1	1		1				1	1	1				
Htp	Chaperone protein, has ATPase activity	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
hslO	Redox regulated molecular chaperone. Protects both thermally unfolding and oxidatively damaged proteins from irreversible aggregation.	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GrpE	hyperosmotic and heat shock by preventing the aggregation of stress-denatured proteins, in association with DnaK	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ComM	Induced during competence development. Not needed for DNA uptake.	2	2	2	2	1	2	2	2	2	3	2	2	2	2	3	2	2	2
GroEL	Prevents misfolding and promotes the refolding and proper assembly of unfolded polypeptides generated under stress conditions.	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
GroES	Binds to Cpn60 in the presence of Mg-ATP and suppresses the ATPase activity of the latter.	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1
HtrA	protease/chaperone protein	3	3	3	3	3	3	3	4	4	3	5	5	3	3	4	3	3	3

Table 3.3. Genes of interest to cold adaptation found in CBW strain annotation. Amino acid sequences found in CBW strains were queried against coding sequences of other *Synechococcus* strains with a stringency e-value of 1e-5.

		Synechococcus																	*Synechocystis*
		Summer		Winter					Coastal		Open Ocean			Freshwater Small Genome		Freshwater Small Genome		Artic	Freshwater
Gene	Description	CB0101	CB0205	CBW1002	CBW1004	CBW1006	CBW1107	CBW1108	WH8101	CC9311	WH8102	WH7803	RCC307	PCC6301	UTEX2973	PCC6312	PCC7502	SynAce01	PCC6803
GroEL	Prevents misfolding and refolds polypeptides under stress	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
GroES	Chaperone; Binds to Cpn60 in the presence of Mg-ATP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1
HtrA	protease/chaperone protein	3	4	4	3	5	5	3	3	4	3	3	3	3	3	3	3	3	3
ctrQ	Pro-zeta-carotene_desaturase	3	3	4	3	5	4	4			4	4							5
HopC	Squalene/phytoene_desaturase	3	3	3	2	2	3	3			3	3							3
desE	Delta-9_fatty_acid_desaturase	3	2	3	3	3	4	3			1	2							1
FA3	Generic FA des	3	3	1	3	1	3	2			2	2							2
Slr1293	Neurosporene_C-3',4'_desaturase	2	2	2	3	2	2	3			3	3							2
PfaA	omega-3 polyunsaturated fatty acid synthase subunit, PfaA	2	2	5	10	5	3	6	1	4	3	3	2	2	2	3	3	2	3
ERG3	sterol desaturase family protein		1				1	1	1					1	1	1	1	1	1

Table 3.4. Annotation feature table keyword search for CBW *Synechococcus* and reference strains. Keywords were chosen based on relationship to cold adaptation.

	Synechocystis	*Synechococcus*																
	Freshwater	Freshwater Small Genome		Freshwater Large Genome		Arctic	Summer		Winter					Coastal		Open Ocean		
Gene	PCC6803	PCC6301	UTEX2973	PCC6312	PCC7502	SynAce01	CB0101	CB0205	CBW1002	CBW1004	CBW1006	CBW1107	CBW1108	WH8101	CC9311	WH8102	WH7803	RCC307
Desaturase	12	8	7	9	16	11	11	9	8	8	9	10	8	11	6	8	7	6
Fatty Acid Desaturase	4	3	1	5	6	4	5	3	4	4	4	5	3	5	1	5	2	1
Transposase	116	4	4	44	223	54	52	8	59	15	35	7	55	52	0	0	1	0
Chaperone	15	12	11	10	9	12	11	12	19	17	21	17	16	11	12	11	9	11

Table 3.5. Reciprocal best hits (RBH) of genes implicated in the cyanobacterial cold response adapted from (Sinetova & Los, 2016). RBH threshold was set at the -value of 1e-20 and queried against the coding sequence of Chesapeake Bay strains. Genes highlighted in blue are induced specifically by cold stress in *Synechocystis sp.* PCC 6803.

CBW1002	CBW1004	CBW1006	CBW1108	CB0101	WH8102	**ORF**	**Gene**	**Function**	**Category**
					syn:sll0790	**sll0790**	hik31	Two-component sensor histidine kinase	Signal perception and transduction
syn:sll2014	syn:sll2014	syn:sll2014	syn:sll2014	syn:sll2014	syn:sll2014	**sll2014**	sfsA	Transcription factor: sugar fermentation stimulation protein	Signal perception and transduction
					syn:sll2012	**sll2012**	sigD	Group2 RNA polymerase sigma factor SigD	Transcription and RNA maintenance
syn:sll1742	syn:sll1742	syn:sll1742	syn:sll1742	syn:sll1742	syn:sll1742	**sll1742**	nusG	Transcription antitermination protein NusG	Transcription and RNA maintenance
syn:sll0517		syn:sll0517	syn:sll0517	syn:sll0517	syn:sll0517	**sll0517**	rbpA1	RNA binding protein A1	Transcription and RNA maintenance
syn:slr0083	syn:slr0083	syn:slr0083	syn:slr0083	syn:slr0083	syn:slr0083	**slr0083**	crhR	RNA helicase	Transcription and RNA maintenance
syn:sll1818	syn:sll1818	syn:sll1818	syn:sll1818	syn:sll1818	syn:sll1818	**sll1818**	rpoA	RNA polymerase alpha subunit	Transcription and RNA maintenance
syn:slr1639	syn:slr1639	syn:slr1639	syn:slr1639	syn:slr1639	syn:slr1639	**slr1639c**	smpBc	SsrA-binding protein	Translation
syn:sll0767	syn:sll0767	syn:sll0767	syn:sll0767	syn:sll0767	syn:sll0767	**sll0767c**	rplTc	50S ribosomal protein L20	Translation
syn:sll1743	syn:sll1743	syn:sll1743	syn:sll1743	syn:sll1743	syn:sll1743	**sll1743**	rplK	50S ribosomal protein L11	Translation
syn:sll1096	syn:sll1096	syn:sll1096	syn:sll1096	syn:sll1096	syn:sll1096	**sll1096**	rpsL	30S ribosomal protein S12	Translation
syn:slr0082	syn:slr0082	syn:slr0082	syn:slr0082	syn:slr0082	syn:slr0082	**slr0082**	rimO	Ribosomal protein S12 methylthiotransferase	Translation
syn:slr1105	syn:slr1105	syn:slr1105	syn:slr1105	syn:slr1105	syn:slr1105	**slr1105**	fus	GTP-binding protein TypA/BipA homolog	Translation
					syn:sll1865	**sll1865c**	pfbBc	Peptide chain release factor 2	Translation
syn:sll0533	syn:sll0533	syn:sll0533	syn:sll0533	syn:sll0533	syn:sll0533	**sll0533**	tig	Ribosome trigger factor	Translation
syn:slr0649	syn:slr0649	syn:slr0649	syn:slr0649	syn:slr0649	syn:slr0649	**slr0649**	metS	Methionyl-tRNA synthetase	Translation
syn:slr0955	syn:slr0955	syn:slr0955	syn:slr0955	syn:slr0955	syn:slr0955	**slr0955**	slr0955	tRNA/rRNA methyltransferase	Translation
syn:slr0399	syn:slr0399	syn:slr0399	syn:slr0399	syn:slr0399	syn:slr0399	**slr0399**	ycf39	Chaperon-like protein for quinone binding in photosystem II	Photosynthesis and respiration
					syn:slr1291	**slr1291**	ndhD2	NADH dehydrogenase subunit 4	Photosynthesis and respiration

			syn:sll1441	syn:sll1441	syn:sll1441	**sll1441a**	desBa	ω3 fatty acid desaturase	Lipid and fatty acid metabolism
	syn:slr1992		syn:slr1992		syn:slr1992	**slr1992**	gpx2	Hydroperoxy fatty acid reductase	Lipid and fatty acid metabolism
syn:slr0321	syn:slr0321	syn:slr0321	syn:slr0321	syn:slr0321	syn:slr0321	**slr0321c**	erac	GTP-binding protein, ERA homolog (cell growth and elongation)	Nucleotide binding and modification
syn:slr0426	syn:slr0426	syn:slr0426	syn:slr0426	syn:slr0426	syn:slr0426	**slr0426**	folE	GTP cyclohydrolase I (riboflavin synthesis)	Nucleotide binding and modification
syn:sll1258	syn:sll1258	syn:sll1258	syn:sll1258	syn:sll1258	syn:sll1258	**sll1258c**	dcdc	dCTP deaminase	Nucleotide binding and modification
syn:sll1854	syn:sll1854	syn:sll1854	syn:sll1854	syn:sll1854	syn:sll1854	**sll1854**	xthA	Exodeoxyribonuclease III	Nucleotide binding and modification
syn:slr1392	syn:slr1392	syn:slr1392		syn:slr1392		**slr1392**	feoB	Ferrous iron transport protein	Transport and binding proteins
					syn:slr0796	**slr0798**	ziaA	Zn exporter	Transport and binding proteins
syn:slr1512	syn:slr1512	syn:slr1512				**slr1512**	sbtA	Na-dependent bicarbonate transporter	Transport and binding proteins
					syn:sll0385	**sll0385**	cbiO	ATP-binding protein of ABC transporter	Transport and binding proteins
syn:slr1238	syn:slr1238	syn:slr1238	syn:slr1238	syn:slr1238	syn:slr1238	**slr1238c**	gshBc	Glutathione synthetase	Other functions
syn:sll1541	syn:sll1541	syn:sll1541	syn:sll1541	syn:sll1541	syn:sll1541	**sll1541**	syc2	Carotene oxygenase	Other functions
syn:slr0239	syn:slr0239	syn:slr0239	syn:slr0239	syn:slr0239	syn:slr0239	**slr0239c**	cbiFc	Precorrin-4 C11-methyltransferase (cobalamin biosynthesis)	Other functions
syn:slr0901	syn:slr0901	syn:slr0901	syn:slr0901	syn:slr0901	syn:slr0901	**slr0901**	moaA	Molybdopterin biosynthesis protein A	Other functions
syn:slr0323	syn:slr0323	syn:slr0323	syn:slr0323	syn:slr0323	syn:slr0323	**slr0323c**	ams1c	α-Mannosidase	Other functions
syn:slr0017	syn:slr0017	syn:slr0017	syn:slr0017	syn:slr0017	syn:slr0017	**slr0017c**	murAc	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	Other functions
syn:slr1072		syn:slr1072	syn:slr1072		syn:slr1072	**slr1072a**	yefAa	GDP-D-mannose dehydratase	Other functions
syn:slr0550	syn:slr0550	syn:slr0550	syn:slr0550	syn:slr0550	syn:slr0550	**slr0550**	dapA	4-Hydroxy-tetrahydrodipicolinate synthase	Other functions
syn:sll0322	syn:sll0322	syn:sll0322	syn:sll0322	syn:sll0322		**sll0322**	hypF	Putative hydrogenase expression/formation protein HypF	Other functions
					syn:sll1029	**sll1029**	ccmK1	Carbon dioxide concentrating mechanism protein CcmK	Other functions
syn:sll1383	syn:sll1383	syn:sll1383	syn:sll1383	syn:sll1383	syn:sll1383	**sll1383**	suhB	Probable myo-inositol-1(or 4)-monophosphatase	Other functions
syn:slr0077	syn:slr0077	syn:slr0077	syn:slr0077	syn:slr0077	syn:slr0077	**slr0077**	nifS	Cysteine desulfurase	Other functions
syn:slr0427	syn:slr0427	syn:slr0427	syn:slr0427	syn:slr0427	syn:slr0427	**slr0427**	slr0427	Putative competence-damage protein	Other functions

					syn:slr0549	**slr0549**	asd	Aspartate beta-semialdehyde dehydrogenese	Other functions
syn:ssl3044	syn:ssl3044	syn:ssl3044	syn:ssl3044	syn:ssl3044	syn:ssl3044	**ssl3044**	ssl3044	Probable ferredoxin, hydrogenase component	Other functions
syn:sll0157	syn:sll0157	syn:sll0157		syn:sll0157	syn:sll0157	**ssl3335**	secE	Preprotein translocase SecE subunit	Other functions
syn:sll0355	syn:sll0355	syn:sll0355	syn:sll0355	syn:sll0355	syn:sll0355	**sll0355c**			Proteins of unknown function
					syn:sll0462	**sll0462c**			Proteins of unknown function
syn:sll0556	syn:sll0556	syn:sll0556			syn:sll0556	**sll0556**			Proteins of unknown function
syn:sll1411		syn:sll1411	syn:sll1411			**sll1411c**			Proteins of unknown function
			syn:sll2013			**sll2013**			Proteins of unknown function
syn:slr0320	syn:slr0320	syn:slr0320	syn:slr0320	syn:slr0320	syn:slr0320	**slr0320**			Proteins of unknown function
syn:slr0400	syn:slr0400	syn:slr0400	syn:slr0400	syn:slr0400	syn:slr0400	**slr0400**			Proteins of unknown function
syn:slr0551	syn:slr0551	syn:slr0551	syn:slr0551	syn:slr0551	syn:slr0551	**slr0551**			Proteins of unknown function
syn:slr0553	syn:slr0553	syn:slr0553	syn:slr0553	syn:slr0553	syn:slr0553	**slr0553c**			Proteins of unknown function
					syn:slr0612	**slr0612**			Proteins of unknown function
syn:slr0755		syn:slr0755		syn:slr0755		**slr0755c**			Proteins of unknown function
syn:slr0959	syn:slr0959	syn:slr0959	syn:slr0959	syn:slr0959	syn:slr0959	**slr0959**			Proteins of unknown function
syn:slr1077	syn:slr1077	syn:slr1077				**slr1077b**			Proteins of unknown function
syn:slr1599	syn:slr1599	syn:slr1599	syn:slr1599	syn:slr1599	syn:slr1599	**slr1599c**			Proteins of unknown function
syn:slr1974	syn:slr1974	syn:slr1974	syn:slr1974	syn:slr1974	syn:slr1974	**slr1974c**			Proteins of unknown function
					syn:slr2123	**slr2123c**			Proteins of unknown function
47	44	47	44	43	54	**111**	**Total**		

Table 3.6. Homologs shared between CBW and reference strains. Homologs were determined using a reciprocal best hit blastp strategy (RBH) with a threshold e-value of 1e-10.

	CB0101	CBW1002	CBW1004	CBW1006	CBW1108	WH8102	*Synechocystis*_PCC6803	CBW1107
CB0101	-	2083	2100	2102	2013	1793	1470	2018
CBW1002	-	-	2301	3023	2097	1779	1529	2253
CBW1004	-	-	-	2283	2166	1785	1536	2202
CBW1006	-	-	-	-	2112	1815	1545	2260
CBW1108	-	-	-	-	-	1738	1433	2078
WH8102	-	-	-	-	-	-	1356	1783
*Synechocystis*_PCC6803	-	-	-	-	-	-	-	1526
CBW1107	-	-	-	-	-	-	-	-

Chapter IV: Abundance and complexity of toxin-antitoxin systems in *Synechococcus* from various aquatic environments

Abstract

Synechococcus spp. are abundant and important to aquatic ecosystems. They contribute significantly to the world's oceans primary productivity and are endemic to coastal, freshwater, pelagic, and estuarine environments. This distribution includes the Chesapeake Bay, Maryland during winter months where they can be counted and isolated even in near freezing brackish water. These *Synechococcus* strains contain several genetic elements that may help them survive in such a variable environment, they also contain genes not traditionally involved in the cold shock response. Toxin-antitoxin systems are small genetic elements that are activated by the bacterial stringent response and can result in a persister state. Chesapeake Bay winter (CBW) strains contain a particularly high abundance of these TA pairs with complex association networks. They feature promiscuous toxins which support the mix and match hypothesis as well as some more monogamous toxins which tend to pair with their traditionally named antitoxin. Activity of select TA systems in a foreign host is consistent with transcriptomic data in CB0101. Further investigation is necessary to understand why CBW strains have such ample TA suites, and what proportion of these TA pairs are in fact functional.

Introduction

Small unicellular picocyanobacteria are widespread, abundant, and contribute significantly to primary production in the world's oceans (Garcia-Pichel et al., 2009; P. W. Johnson & Sieburth, 1979; Li & Url, 1994; Waterbury et al., 1979). Picocyanobacteria, mainly *Synechococcus* and *Prochlorococcus* have differential genomic, physiological, and morphological characteristics, which equip them for differential ecological conditions (Dufresne et al., 2008; D J Scanlan et al., 2009; David J. Scanlan, 2012). *Synechococcus,* in relative comparison to *Prochlorococcus,* have larger genomes (Dufresne et al., 2008; D J Scanlan et al., 2009) and a ubiquitous distribution (Olson, Chisholm, & Zettler, 1990; Frédéric Partensky et al., 1999). *Prochlorococcus* by contrast is tailored to an oligotrophic environment (Dufresne et al., 2003) and constrained to a limited latitudinal distribution between 40 ^{o}N and 40 ^{o}S (Z. I. Johnson, Zinser, Coe, McNulty, et al., 2006; Olson, Chisholm, Zettler, et al., 1990; Frédéric Partensky et al., 1999; Shalapyonok et al., 2001).

In ecological theory, the concepts of generalists and specialists have been used to categorize organisms based on their ecological strategies: Generalists have broad environmental tolerances, while specialists have specific and narrow habitat tolerances (Pandit et al., 2009). This ecological theory is also true in bacterial communities (Fierer et al., 2007; Lindstrom & Langenheder, 2012), including those in coastal oceans (Mou et al., 2008). On a broad scale for picocyanobacterial species, generalists and specialists are best exemplified by the genera *Synechococcus* and *Prochlorococcus*, respectively. This is apparent when comparing their distribution, mentioned above, and their genetic components. *Synechococcus* has clearly

defined phyletic subgroups categorized as coastal/opportunists (Dufresne et al., 2008) and are capable of surviving nearly all aquatic environments. *Prochlorococcus* is confined to the world's oligotrophic oceans and their genomes have undergone significant specialization and reduction, save for the low light clade IV (Dufresne et al., 2005; Kettler et al., 2007). When comparing *Synechococcus* and *Prochlorococcus* in a broad sense, the former tends to be generalists while the latter are specialists. On a granular level, this dichotomy of generalists and specialists can also exist more subtly at the genus-specific level among *Synechococcus.* Coastal and estuarine *Synechococcus* are considered as generalists, while open ocean *Synechococcus* is believed to have a specialist lifestyle (Dufresne et al., 2008; Brian Palenik et al., 2006). Generalists tend to have an expanded genetic capacity and therefore, a greater ability to sense and respond to environmental stimuli, or stressors. This advantage is reflected in genetic systems that have allowed organisms to adapt to a range of environmental conditions.

Toxin-Antitoxin (TA) systems are small intracellular elements that can regulate bacterial and archaeal cell growth (Unterholzner et al., 2013). They are comprised of a protein toxin and a cognate antitoxin, which can be either protein, or non-coding RNA (ncRNA). Depending on the genetic material of the antitoxin and method of action of the toxin, TA systems are classified into 5 (Unterholzner et al., 2014b) sometimes, 4 main families (Harms et al., 2018a). By far, the most well studied category of TA systems is Type II, in which both the toxin and antitoxin components of Type II TA systems are proteins. Functions of type II toxins like relE and yoeB often exhibit RNA degradation and a similar structure to RNases (Kamada & Hanaoka, 2005). Their method of action can alter gene expression and ultimately induce a persister state in cells. Among cyanobacteria, 81% of type II toxin-antitoxin systems in *Synechocystis* 6803 displayed

RNase activity during fluorescence assays (Kopfmann et al., 2016). Picocyanobacteria like *Synechococcus* may regulate gene expression *via* RNase activity of their toxin-antitoxin systems.

To date, more than three hundred of picocyanobacterial genomes (including draft and complete) have been sequenced. Comparative genomics often consider genome size, GC content, core and variable genomes, interesting functional genes, etc. in picocyanobacteria. However, TA genes have previously been overlooked during traditional genomic analysis, thus the knowledge about TA systems in picocyanobacteria is very limited. TA systems have been predicted in WH1802 and freshwater cyanobacteria including *Microcystis aeruginosa* (Makarova et al., 2009), *Synechocystis* PCC6803 (Kaneko et al., 2003), and on the pANL plasmid in *Synechococcus* PCC7942 (Y. Chen et al., 2011).

The first chromosomal TA systems in estuarine *Synechococcus* were described in the *Synechococcus* strain CB0101 endemic to the Chesapeake Bay (D. Marsan et al., 2017). *In vivo* transcriptomics of CB0101 reveals a tight coupling between the upregulation of particular toxin genes, such as *relE*^{1}, with simulated oxidative stress conditions (*i.e.* high zinc or high light exposure). In CB0101, growth arrest co-occurred with a four-fold increase in *relE*^{1} expression. When the stressor was removed, growth rate returned to normal. This recoverable growth arrest coincides with the upregulation of *relE*^{1} and downregulation of the corresponding antitoxin gene *relB*^{2} in CB0101. These phenomena were also observed in nitrogen starvation and zinc toxicity experiments in the same study.

CB0101 has been used as the representative strain for the Chesapeake Bay since isolation in 2004 and was isolated in the summer (F. Chen et al., 2004). *Synechococcus* cell

abundance increases with temperature, and *Synechococcus* cells are still counted in frigid waters in the Chesapeake Bay (K. Wang et al., 2011). In other frigid waters such as the Bornholm Sea and subalpine waters (Ernst et al., 2003a) *Synechococcus* is still detected. More recently, interesting and unique Chesapeake Bay *Synechococcus* strains have been isolated during winter months (Xu et al., 2015). These strains display an incredible ability to grow at very low temperatures (4^{o}C) and recover normal growth at 23^{o}C after exposure to such cold shock. Unfortunately, little is known about their genetic capacity which result in such incredible physiological capabilities.

Recent work has revealed an interesting correlation (r^2= 0.6235, p<0.0001) between genome size and the occurrence of TA systems in *Synechococcus* (Fucich & Chen, 2020), but such a pattern is not exhibited in other bacteria or archaea (Leplae et al., 2011). More TA systems were predicted in *Synechococcus* strains with larger genomes, but tighter linear correlations were observed (r^2=0.9152, p<0.00001 and r2=0.8296, p<0.005) when strains were grouped into unique habitat types, specifically coastal and freshwater respectively (Fucich & Chen, 2020). This suggests that habitat may be a principal contributing factor in the retention of TA pairs in *Synechococcus*, rather than genome size. In the broader picocyanobacteria, TA prevalence appears to follow the same distribution pattern. TA systems are abundant in *Synechococcus* endemic to nutrient rich, dynamic habitats, but rare or completely absent in strains (*Synechococcus* and *Prochlorococcus*) that are specialized to the pelagic.

Our recent study suggests that TA systems could be an important mechanism for stress response and niche partitioning for picocyanobacteria. Despite the fact that a high number of TA genes are present in certain *Synechococcus* spp., little is known about their diversity,

association, and activity. Toxin-antitoxin diversity and evolutionary history is difficult to study given their lack of amino acid conservation and frequent horizontal gene transfer (Chandra et al., 2016; Van Melderen, 2010).

In this study, fourteen *Synechococcus* strains were chosen to represent a variety of habitats, including estuarine (n=6), coastal (n=1), open ocean (n=4), and freshwater (n=2) bodies. A dinoflagellate ectosymbiont (n=1) was also included in the study. The aim of this study is threefold: (1) predict TA systems in *Synechococcus* strains endemic to the Chesapeake Bay during winter and compare them to strains from freshwater, coastal, and marine habitats; (2) assess the diversity of predicted TA genes and note patterns of complexity in association networks; and finally (3) verify toxin-antitoxin functionality and growth regulation of at least two pairs (*relB*^{2}*/relE*^{1} and *vapB*^{1}*/vapC*^{1}) found in *Synechococcus* strain CB0101.

Results

Synechococcus ecotypes from diverse habitats

Genomes from 14 *Synechococcus* strains isolated from open ocean, coastal, freshwater, ectosymbiont, and estuarine environments were selected as representatives (Table 4.1). Their genome sizes range from 1.8 to 3.8Mb. The genome size of 3.8 Mb is among the largest known *Synechococcus* genomes. The smallest genome (1.8 Mb) belongs to *Synechococcus* OmCyn01, the ectosymbiont of *Ornithocercus magnificus*. This is not a surprise as a symbiont, genome reduction, loss of unnecessary coding sequences, and lack of TA pairs are expected. Chesapeake Bay winter strains CBW1002 and CBW1006 have some of the largest genomes among *Synechococcus* with 3.85 Mb and 3.86 Mb, respectively. Freshwater strains, such as PCC6312

and PCC7502 have similarly large genomes at 3.7 Mb and 3.6 Mb respectively (Table 1.1) . Large genomes are seemingly more common among freshwater, coastal, and estuarine *Synechococcus* rather than open ocean strains.

Chesapeake Bay Synechococcus strains have ample TA suites

Synechococcus strains isolated from the Chesapeake Bay (CB) tend to have more TA pairs (TA pairs $\bar{x}$=46) compared to strains from freshwater ($\bar{x}$=26), open ocean ($\bar{x}$=6.6), and coastal environments (n=3). Freshwater strain PCC6307 was comparable to CB strains with 42 putative TA pairs, but PCC6301 had far fewer with only 10 putative TA pairs.

relE/parE *toxins are common in Chesapeake Bay strains*

The toxin gene *vapC* was the most abundant toxin in all *Synechococcus*. A notable difference between the Chesapeake Bay TA association network maps compared to the marine and freshwater strains is the discrepancy in abundance between *relE* and *NT_KNT* between estuarine and freshwater/marine strains. In Chesapeake Bay strains, the *relE/parE* family is often the second most abundant toxin. In marine and freshwater *Synechococcus*, conserved nucleotidyltransferase domains from the superfamily cl11966 (*NT_KNT*) were the second most predicted toxin. Superfamily *cl11966* contains nucleotidyltransferase conserved domains which are also found in DNA polymerase and kanamycin resistance genes. It appears that *Synechococcus* in different habitats may acquire different toxin families.

Unique Chesapeake Bay Synechococcus

CBW1108 contained the most putative TA pairs (n=80) and connections (c=34), which included the most traditionally named TA families and many additional conserved domains that do not fall into known TA families. Having many TA families, or nodes, resulted in the association map with the most edges, or connections, and therefore the most complex network organization. CBW1108 contained as many as 6 predicted toxins are connected to the main network (Figure 4.1). This pattern displays that multiple diverse putative toxin families can associate with antitoxins that have similar conserved amino acid sequences. This is exhibited in CBW1108, CBW1006, CB1004, and CB0101 where phd/yefM acts as the antidote to vapC and relE/parE.

CBW1108 contains both of the putative antitoxins (n=2) that could not be assigned a conserved domain, traditional AT family, or even gene fragment name (Figure 4.1). Although these blastp results showed that these families are found in other cyanobacteria (data not shown), they are not adequately annotated. These cryptic genes could be functioning to negate the effect of a toxin. CBW1004 contained a hypothetical AT 1 that is associated with a predicted toxin with a *NT_KNT* conserved domain. The only non-Chesapeake Bay *Synechococcus* to contain a hypothetical antitoxin was the marine strain WH8102.

Association complexity of TA systems in Synechococcus

Network association maps of toxin-antitoxin systems in Chesapeake Bay *Synechococcus* show an increase in complexity with an increase in putative TA system abundance. CBW1108 had the most complex association network with 34 connections, or edges, between 80 putative TA pairs, while CB0101 had the least complex network with only 12 connections between 22 TA

pairs (Figure 4.1 and Table 4.1). Expanding this to other habitats, this pattern continues until the logical conclusion, bottoming out at the open ocean strain WH7803 which has only one connection shared between its only putative TA pair *PIN_vapC/phdyefM*.

The TA association network maps are organized relative to complexity. The simplest association networks are in the upper left (WH7803, open ocean), and the most complex networks are in the bottom right (CBW1108 Estuary). The trend of complex association network maps roughly follows genome size increase (Table 4.1), with a few notable exceptions. CBW1108 has the most putative TA pairs, but only has the 3rd largest genome (3.2 Mb). This results in a dense proportion of toxin antitoxins as a function of coding sequences at 4.27%, the highest among any genome surveyed.

The discrepancy of predicted TA pairs between CBW1002 and CBW1006 is notable. These strains have the largest genomes, both around 3.8 Mb, but CBW1006 has 54 putative TA pairs while CBW1002 has only 29 predicted pairs. Why CBW1002 contains nearly half as many TA systems as CBW1006 is still unknown, especially when considering that they are both belong to the Bornholm Sea Cluster (Figure 1.3) and they share the most *in silico* homologs of any CBW strain pair (Figure 3.2). Habitat trends tend to explain the trends in association network complexity more so than genome size. While CBW strains can have some of the largest genomes of the *Synechococcus* surveyed, their putative TA pairs are atypically high, especially CBW1108, CBW1006, and CBW1004. These 3 strains have the most TA system dense genomes as a function of coding sequences. CB strains (besides CB0205) all have the highest TA % of coding sequences. This trend is interrupted by PCC6307, which was selected to represent a

freshwater genome with particularly abundant TA systems. Habitat, particularly the turbulent estuary is the best indicator of an abundant TA suite with rich association networks (Figure 4.1).

Some toxin and antitoxin families were abundant, and often promiscuously paired with many different toxin or antitoxin families. *PIN_vapC* was the most abundant toxin family and was the association network "hub" for all Chesapeake Bay *Synechococcus* strains (Figure 4.1). This was also true of marine, coastal, and freshwater *Synechococcus* strains. *PIN_vapC* was the center of as many as ten connections to antitoxins in Chesapeake Bay strains CBW1004 and CBW1108 and was never connected to fewer than four antitoxin (AT) families in CB0101 and CBW1108.

The next most promiscuous toxin family is the *relE/parE* family. *relE/parE* connects to five different antitoxin families in CBW1004 and CBW1006. In cases where the *relE/parE* toxin is predicted, it is often paired with the antitoxin family *phd/yefM*. For many of the amino acid sequences, especially abundant and promiscuous toxins like *vapC*, multiple sequence alignment resulted in poor alignments with very few conserved residues (Figure 4.2).

Other toxin-antitoxin families are much more exclusive in their pairing. The *brnT/brnA* family only associate with their cognate protein and are never paired with any other known toxin or antitoxin family. The *brnT/brnA* TA family is predicted to be in *Synechococcus* strains CBW1004, CBW1006, CBW1108, CB0101, and freshwater strain PCC6307. In CBW1006, one *brnA* AT is predicted normally with the *brnT* toxin. The only exception to this monogamous link between *brnT/brnA* is in CBW1006 where one *brnA* antitoxin is paired with a putative toxin

which contains the gene fragment DUF4258, which is likely associated with the *brnT* toxin according to entries in Pfam (El-Gebali et al., 2019).

The *hicA/hicB* family is predicted in freshwater strains PCC6301 and PCC6307 and in Chesapeake Bay strains CBW1004, CBW1006, and CBW1108. Interestingly, the *hicA/hicB* family displays monogamous pairing in some strains and promiscuous pairing in others. In CBW1004 and PCC6307, the *hicA/hicB* are only predicted together in a monogamous fashion. In other strains, there can be more than one occurrence of this TA family, pairing with different antidote proteins. In CBW1006, *hicA* pairs with *hicB*, *copG*, and a 2-oxo acid dehydrogenase complex (OADH). *hicA* in CBW1108 is paired with *hicB* and a similar OADH. This hypothetical antitoxin labeled OADH was identified through protein BLAST. While in these Chesapeake Bay strains, the toxin *hicA* showed selective promiscuity, in the freshwater PCC6301, antitoxin *hicB* was paired as an antidote to *hicA*, and *PIN_vapC*.

The *hipA* toxin is paired with a Helix-Turn-Helix (*HTH*) motif as a predicted antitoxin. The *hipA/HTH* pairing are exclusive to each other in CBW1002 and CBW1006, but pairing is not exclusive in CBW1004 and CBW1108. In both of these winter CB strains, *hipA* pairs with an XRE domain (cl22854) which includes a *HTH* motif as well.

Hypothetical Toxins and Antitoxins in Chesapeake Bay Synechococcus

Some putative toxins and antitoxins predicted in Chesapeake Bay *Synechococcus* strains are not well annotated. Putative toxins may exhibit a conserved domain but are not able to be categorized into a traditional toxin family. These include, but are not limited to, the kanamycin nucleotidyltransferase (*NT_KNT*) from the superfamily cl11966, the N-Acyltransferase (*NAT_SF*)

from the superfamily cl17182, and the aptly named *RES* domain which include the conserved residues arginine, glutamine, and serine throughout the cl02411 superfamily.

In comparison to toxins, fewer antitoxins contain a conserved domain associated with a known antitoxin family. Many putative antitoxins in Chesapeake Bay *Synechococcus* strains contain a domain of unknown function (DUF) such as: DUF86 (superfamily cl01031), DUF2191 (superfamily unknown), DUF433 (superfamily cl01030), and DUF1778 (superfamily unknown). DUF1778 has suspected helix structure and sequence similarity with the hicB antitoxin, however these DUF are often cryptic in function.

Another predicted antitoxin without a traditional family name includes the nucleotide binding domain HEPN from the superfamily cl00824 which, in bacteria, is accompanied by a nucleotidyltransferase (Grynberg et al., 2003). Another nucleotidyltransferase (EC 2.7.7) (NTase_sub_bind) is part of a superfamily cl23885 with roles in polynucleotide modification (Lehmann et al., 2003).

Growth arrest of E. coli by Synechococcus CB0101 TA pairs

The growth of transformed strain *K12*:*relE*^{1} (the toxin gene) was significantly inhibited with addition of IPTG. The transformed *E. coli* cells with both genes *relB*^{2}/*relE*^{1} (toxin and antitoxin) grew similarly as non-transformed strain *K12* when induced with IPTG. The growth of *K12:relE*^{1} in the cultures with added IPTG decreased significantly compared to the control strain *K12* containing the empty vector, both with and without added IPTG (p-value=0.02) (Figure 4.4A). There was some growth inhibition on *K12:relE*^{1} without IPTG compared to the control

(K12 without *relE*[1]), but growth arrest was not significant compared to the control. The K12 culture grew similarly with or without IPTG inducer (Figure 4.4A).

All of the strains containing genes from the *vapB*[1]*/vapC*[1] TA system showed statistically similar growth. The growth curves for *K12:vapC*[1]*, K12:vapB*[1] and *K12:vapB*[1]*/vapC*[1] in cultures with and without added IPTG are all comparable to the non-induced and induced cultures of the control strain *K12* containing the empty plasmid (Figure 4.5A, 4.5B, 4.5C). When the *vapC*[1] toxin was expressed in the *K12:vapC*[1] strain there was no significant change in the measured optical density.

Discussion

Winter Chesapeake Bay Synechococcus strains contain abundant TA pairs

Winter *Synechococcus* strains contain more putative TA pairs than summer strains. While the sample size is quite small, the draft genome of CB0205 contained one putative TA system: *vapC/phd*, while the complete CB0101 genome contained 22 putative TA pairs. The four winter strains ranged from as many as 80 putative TA pairs in CBW1108 to as few as 29 putative TA pairs in CBW1002. The high abundance of TA systems in Chesapeake Bay winter strains (CBW strains) is striking due to their physiological resilience to low temperatures. In laboratory tests, CBW strains are able to resume growth at low temperatures, and in all cases resume growth after exposed to near-freezing (4^{o}C) water temperatures (Xu et al., 2015). TA system upregulation in CB0101 has been linked to simulated oxidative stress conditions *via* Zn toxicity, high light conditions, and nutrient starvation (D. Marsan et al., 2017). The increased abundance of TA systems in Chesapeake Bay winter strains may be a cellular regulatory

mechanism activated by the general stress response. This may allow cold adapted Chesapeake Bay strains to survive freeze and thaw cycles in the temperate bay. This hypothesis remains to be tested.

Further, Chesapeake Bay strains tend to have more putative TA pairs than coastal and open ocean marine strains. This finding has been seen with a larger sample size of all available complete *Synechococcus* genomes and appears to be linked to endemic habitat and a statistically significant correlation between genome size and putative TA pairs (Fucich & Chen, 2020).

Some *Synechococcus* strains (CB0205, WH7805, and OmCyn01) did not contain any putative TA systems. This may be the result of their status as draft sequences as any putative toxin antitoxin systems could have been missed in sequencing. However, this is highly unlikely, as they both have large sequences >2.4 Mb, a genome size in which *Synechococcal* toxin antitoxin systems have been predicted previously (Fucich & Chen, 2020). Environmentally, CB0205 and WH7805 could be *Synechococcus* strains that are less adapted to estuarine environments and more so for open ocean environments. Though not expected in CB0205, as it was originally isolated from the Chesapeake Bay, this strain could be better adapted for pelagic environments. WH7805 was isolated from open ocean. It is common for TA systems to be absent from open ocean marine *Synechococcus* strain genomes.

Synechococcus strains support the 'mix and match' hypothesis

TA systems are frequently found in non-traditional pairing arrangements, known as 'Mix and Match' pairing (Arbing et al., 2010; Fasani & Savageau, 2015; Guglielmini & Van Melderen,

2011). This results in most toxins and antitoxins having poor sequence alignment. Even in the most conserved toxin PIN domain, with the exception of 3 conserved residues distributed along the sequence, there is poor sequence alignment across the family (Arcus et al., 2011). The three-dimensional structure places these particular residues at the putative active site (Arcus et al., 2004; Bunker et al., 2008).

A putative toxin or antitoxin, one without a match to the NCBI CDD for a traditional toxin or antitoxin family, could be considered cognate "guilt by association" protein (Leplae et al., 2011). In our survey, this type of prediction most often occurred with a known toxin and non-traditional antitoxin. Most of these 'hypothetical proteins' (93%) that needed to be identified with blastp were antitoxins rather than toxins. The non-conserved nature of antitoxins in *Synechococcus* strains is consistent with toxin antitoxin systems in other bacteria (Mittenhuber, 1999). The promiscuous nature of toxins matching with many different antitoxins results in different families of toxins and antitoxins acting as an antidote for another family. Such non-traditional toxin-antitoxin pairing is referred to as the "mix and match hypothesis" (Arbing et al., 2010; Guglielmini & Van Melderen, 2011). The nontraditional pairing of different toxin and antitoxin families in *Synechococcus*, like *PIN_vapC* and *relE*/*parE* with various antitoxins, is evidence of this idea. These instances are most easily illustrated by the central 'hubs' in the association network maps (Figure 4.1).

Although toxins and antitoxins contained 'conserved' domains, these domains are not sufficient to be treated as genetic markers. Among some of the most frequently predicted toxins, *PIN_vapC*, the alignment of eight versions was poor with only three conserved residues (Figure 4.2A). In the case of the least promiscuous toxin, brnT, amino acid alignments were also

poor (Figure 4.2B). The alignment consisted five sequences from three strains, and only two conserved residues. For these reasons, toxin-antitoxin systems are poor genetic markers.

Cryptic TA genes in Synechococcus suggest regulatory functions

Functions of toxin and antitoxin genes with no known family often have a cryptic function or unknown function. Many amino acid sequences resulted in domains of unknown functions (DUF) or clusters of orthologous groups (COGs). Further investigation of these amino acid sequences often results in the prediction of structures related to nucleotide binding, nucleotide modification, and regulatory mechanisms. These various RNase, nucleotidyltransferase, helix-turn-helix motifs, *etc.* have predicted functions that are very similar to the known methods of action in type II TA systems (Makarova et al., 2009; Robson et al., 2009; Winther & Gerdes, 2011).

Hypothetical proteins with no immediate connection to a conserved domain were clarified by blastp. It was possible to glean some indication of functionality from these blastp results. Many of the hypothetical antitoxins associated with *brnT* had sequence similarity to *brnA*. This clarified the largely monogamous nature of the *brnT/brnA* family. Further analysis of many DUF and clusters of orthologous groups (COG) domains were linked to known toxin or antitoxin sequences. For example, DUF2191, which was predicted by blastp several times on hypothetical proteins, is often found in *vapB* domains in *Mycobacterium tuberculosis* (Ramage et al., 2009). In all CBW strains, the HipA toxin is paired with an HTH or XRE domain as a predicted antitoxin. Both of these domains incorporate an HTH, suggesting that such a structure could act to negate the activity of HipA. HTH domains are capable of binding to DNA

and are involved with many proteins that regulate gene expression (Brennan & Matthews, 1989).

Putative toxins and antitoxins that do not belong to a known family often include a conserved nucleotide binding domain that may be accompanied by a nucleotidyltransferase. This is plausible given that one of the functions of type II TA pairs is translation arrest by RNA degradation. These nucleotide binding domains may be responsible for recognizing RNA transcripts to disrupt translation.

Scattered toxin-antitoxin loci

Toxin-antitoxin systems in *Synechococcus* display differential distributions along genomes. In many cases, TA systems appear concentrated at certain locations along the genome, as seen in CBW1006, or CB0101 (Figure 4.3). Other genomes have a more even distribution, like CBW1004. In either case, it is still being determined if the TA systems are located on genome islands. This, along with the poor conservation of TA system amino acid sequence suggests they are transferred horizontally, as is a known transmission vector for TA systems in other bacteria (Guglielmini & Van Melderen, 2011; Leplae et al., 2011).

Confirmation of relE^1 activity from Synechococcus CB0101

It is important to note that only two putative toxin antitoxin systems from CB0101 were tested in this study. The apparent activity of the relE toxin in *E. coli* suggests that the expression of *relE^1* can strongly inhibit the growth of *E. coli* and the *relE^1* toxin most likely plays an active role on arresting the growth of CB0101 under stressful conditions. The slight variation in the

growth of *K12:relE*^{1} in relation to the control, even when the toxin gene was not being expressed could be due to the added metabolic load brought on by the chloramphenicol antibiotic solution by utilizing the resistance component of the pCA24N plasmid.

The growth inhibition in the transformed strain *K12*:*relE*^{1} and no growth inhibition in cells with *relB*^{2}/*relE*^{1} upon IPTG induction suggest that this TA gene pair regulates cell growth as a typical type II TA system. This behavior is consistent with the transcriptional data from Marsan et al. 2017 and the growth regulation under stress hypothesis has been confirmed in similar type II TA systems found in other picocyanobacteria like *Synechocystis* sp. PCC 6803 (Kopfmann et al., 2016).

The apparent non-activity of vapC in *E. coli* is consistent with the previously described transcriptional data, where there was no significant upregulation in the vapC toxin when *Synechococcus* CB0101 was exposed to a variety of stressors. This could be due to type I error of the TA prediction software of the gene sequence. More likely, this operon of *vapB*^{1}/*vapC*^{1} could be non-functional. Considering that there are more than two versions of the *vapC* toxin gene at different loci in CB0101, it is likely this pair of TA genes is silent or has unknown function.

By cloning cyanobacterial TA genes into *E. coli*, we demonstrate that the TA genes *relB*^{2}/*relE*^{1} from *Synechococcus* CB0101 are functional and are able to inhibit or arrest the cell growth. We conclude that the *relE*^{1}gene acts as a type II bacterial toxin and because the growth inhibition can be released with the expression of its cognate antitoxin *relB*^{2}. The functionality test performed in this study further confirms the integral role of *relB*^{2}/*relE*^{1} in the stress

response of *Synechococcus* CB0101 (D. Marsan et al., 2017). The TA gene pair *vapB*^{1}*/ vapC*^{1} did not exhibit the characteristic cell growth regulation as *relB*^{2}/*relE*^{1}.

Putative TA presence does not equal activity

Abundance of TA genes may not be related to toxin activity; in fact, the opposite may be true. If there are several copies of a toxin, many may not be active or functional in times of stress or normalcy. The toxin relE showed growth suppression in a foreign host and the greatest transcriptomic response to an environmental stressor (D. Marsan et al., 2017), despite having few copies predicted in its genome. Conversely, vapC is the most frequently predicted toxin in *Synechococcus*, but the loci tested *vapC*^{1} fails to show any significant transcriptomic response or response in a foreign host. In these two cases, transcriptomic data matches growth in a foreign host. To decide which TA genes are most likely to be functional, cloning and expression of these TA genes in a foreign host is ultimately necessary. However, the process is time and resource consuming. To determine TA pairs that are most likely functional in the natural host, gene expression *via* transcriptomics may first be used to generate a short list of likely functional candidates.

The activity of toxin antitoxin systems has been investigated. Among others, MqsR, a type II toxin can be inactivated by a mutation in the toxin promotor or by a chromosomal mutation in *mhpR* (Fernandez-Garcia et al., 2019). While this type II toxin is not predicted in any of the CBW strains, and confirmed with no significant similarity by blastp, a similar method of inactivation may be possible by adjacent or distant genes.

relE was not predicted as frequently as *vapC*, but it showed much greater growth retardation in the foreign *E. coli* host. This difference in confirmed activities, in transcriptomic response and activity in a foreign host, suggests that many of these putative TA pairs may not be active in the stringent response. The type II toxin antitoxin systems predicted may not have any regulatory effect in some, or all stressful situations. Future work is necessary to test the *in vivo* activity in a foreign host before determining the functionality of toxin antitoxin systems in *Synechococcus* and to investigate their potential role in environmental tolerance.

Methods

Genome Sequence Acquisition

Genome sequences were obtained from multiple sources. CB0101, PCC6307, PCC6301, KORDI-49, and WH8102, PCC7803, PCC7805, WH8102, and CC9311 were downloaded from NCBI, as they are publicly available (accession numbers found in Table 4.1). CB0205 is a draft genome and was obtained from Illumina reads. CB0205 is a draft genome and consists of 78 contigs, is 2,427,308 bp in length, with an N50 of 63,410. CBW strains, (CBW1002, CBW1004, CBW1006, and CBW1108) are complete sequences which were completed by the Beijing Genome Institute (BGI). These sequences have been deposited into NCBI under the bioproject: PRJNA657291.

Toxin Antitoxin Prediction

Genomes were annotated using the RAST server using standard bacterial settings (Overbeek et al., 2014). TA systems were predicted using the TAfinder tool (Xie et al., 2018)

which uses experimentally validated and *in silico* predicted TA systems in the Toxin Antitoxin Database (TADB) (Shao et al., 2011). In order to predict putative TA pairs, the original FASTA sequence file and the exported GenBank annotation file from the RAST server was uploaded to the TAfinder software.

Conserved Domain Prediction

Putative toxin and antitoxin amino acid sequences were queried against the NCBI conserved domain database (CDD) (Marchler-Bauer et al., 2017) for the presence of a known toxin or antitoxin family (Figure 4.6). Annotations were manually verified and categorized in to uniform TA families to correct discrepancies in annotation in the CDD (*i.e.* "*PIN_vapC*-like" and "*PIN_vapC4-5*_FitB-like" are condensed into the same superfamily: *cl28905*). If no conserved domain was found for a query sequence, the amino acid sequence was queried with protein BLAST (blastp) using default settings (Agarwala et al., 2017b). In the case of these hypothetical toxins or antitoxins, the result with the highest bit score and lowest e-value was selected; given the result had a known toxin or antitoxin name, other gene, or gene element name. In only three cases was a "hypothetical protein" name assigned. Hypothetical AT 1 was assigned for CBW1004_38_AT and CBW1108_14_AT. Hypothetical AT 2 was assigned to CBW1108_24_AT. These names were only assigned after there were no matches to NCBI's CDD and blastp results contained all hypothetical proteins. To check for amino acid conservation, Clustal Omega (McWilliam et al., 2013) was used for multiple sequence alignment of like-conserved domains.

Association figures were subsequently derived from the putative TA systems and their conserved domains or for a subsection of putative sequences, blastp results. Networks were

based on the putative toxins or antitoxins, assigned as elements, and their frequency of connection, depicted by line thickness in Kumu (https://kumu.io/).

Confirmation of Toxin Antitoxin Activity

Preparation of CB0101 culture and DNA extraction

Synechococcus CB0101 culture (30 ml) was grown in liquid SN medium with a 15 percent salinity (Waterbury, 1986) in a 25 cm^2 culture flask (Corning Inc.). Cultures were shaken once daily for 30 seconds. DNA was extracted from 2 ml of centrifuged culture using the MOBIO UltraClean Microbial DNA Isolation Kit (MOBIO). Extracted DNA was recovered in 30 µl of the included elution buffer.

PCR amplification of TA genes

The primers for each TA genes were shown in Table 4.3. The primers were designed for blunt-end cloning by digesting pCA24N with the restriction enzyme *StuI* (Thermo Fisher Scientific). Genes in Table 4.1 were amplified with high fidelity PrimeStar GXL DNA polymerase in 30 cycles with an annealing temperature of 58ºC and an elongation temperature of 72ºC. After separating the PCR products through gel electrophoresis, the fragments of expected size were extracted with the GeneJET Gel Extraction and PCR Purification Kits (Thermo Fisher Scientific).

Cloning the TA genes

To confirm the functionality of *relB*^{2}/*relE*^{1} and *vapB*^{1}/*vapC*^{1} in *Synechococcus* CB0101, these genes were cloned into vector pCA24N. The vector *pCA24N* (constructed by (Kitagawa et al., 2005), GenBank accession number AB052891) has Chloramphenicol resistance and a high copy number and its expression is induced with Isopropyl β-D-1-thiogalactopyranoside (IPTG). Six separate expression vectors were designed and used for transformation resulting in the strains in Table 4.3. The toxin and antitoxin genes were cloned individually into the *pCA24N* plasmid for expression resulting in plasmids *relE*^{1}*:pCA24N*, *relB*^{2}*:pCA24N, vapB*^{1}*:pCA24N* and *vapC*^{1}*:pCA24N.* The TA genes were also cloned together as a construct to form *relB*^{2}*:relE*^{1}*:pCA24N* and *vapB*^{1}*:vapC*^{1}*:pCA24N*. The plasmids were transformed by 42ºC heat shock into *E. coli* K12 ER2738 obtained from a frozen stock kindly provided by Dr. Xiaoxue Wang at the South China Institute of Oceanology. The cells were made chemically competent through a $MgCl_2$ and $CaCl_2$ wash protocol. Positive clones were confirmed by Illumina sequencing and gel electrophoresis (Figure 4.7).

The growth of transformed clones

All six transformed strains were streaked separately in LB agar plates with chloramphenicol at a 25 μg/ml concentration and grown overnight at 37ºC, as well as the control strain containing the empty plasmid (denominated *K12* for this study). Single colonies were transferred into 1 ml of LB media and grown overnight at 37ºC with shaking. The cultures were diluted at a hundred-fold in 40 ml of LB media with chloramphenicol in 6 different flasks per each one of the seven strains. Flasks were kept at 37ºC in a shaking incubator and the optical density (OD600) was measured to assess growth every 30 minutes for a total of 10

hours. IPTG in a final concentration of 0.5 mM to induce gene expression was added to three of the six flasks per strain when the OD600 was measured at approximately at 0.15.

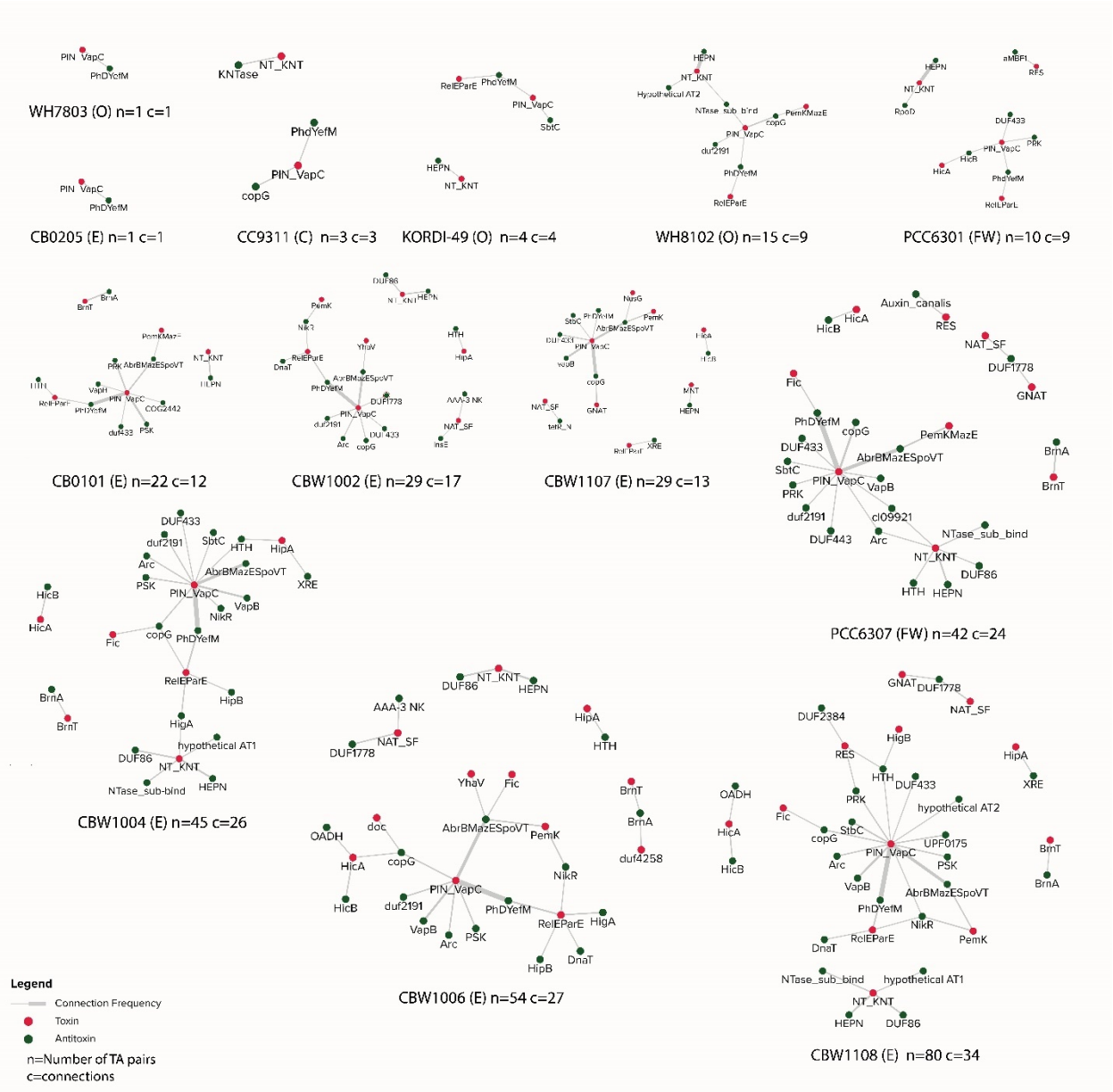


Figure 4.1. Association network maps showing predicted TA systems in *Synechococcus* strains isolated from the Chesapeake Bay estuary (E), coastal water (C), freshwater (FW), and open ocean (O). Putative toxins (red) and antitoxins (green) are represented by nodes with their frequency of association denoted by the stroke of connection.

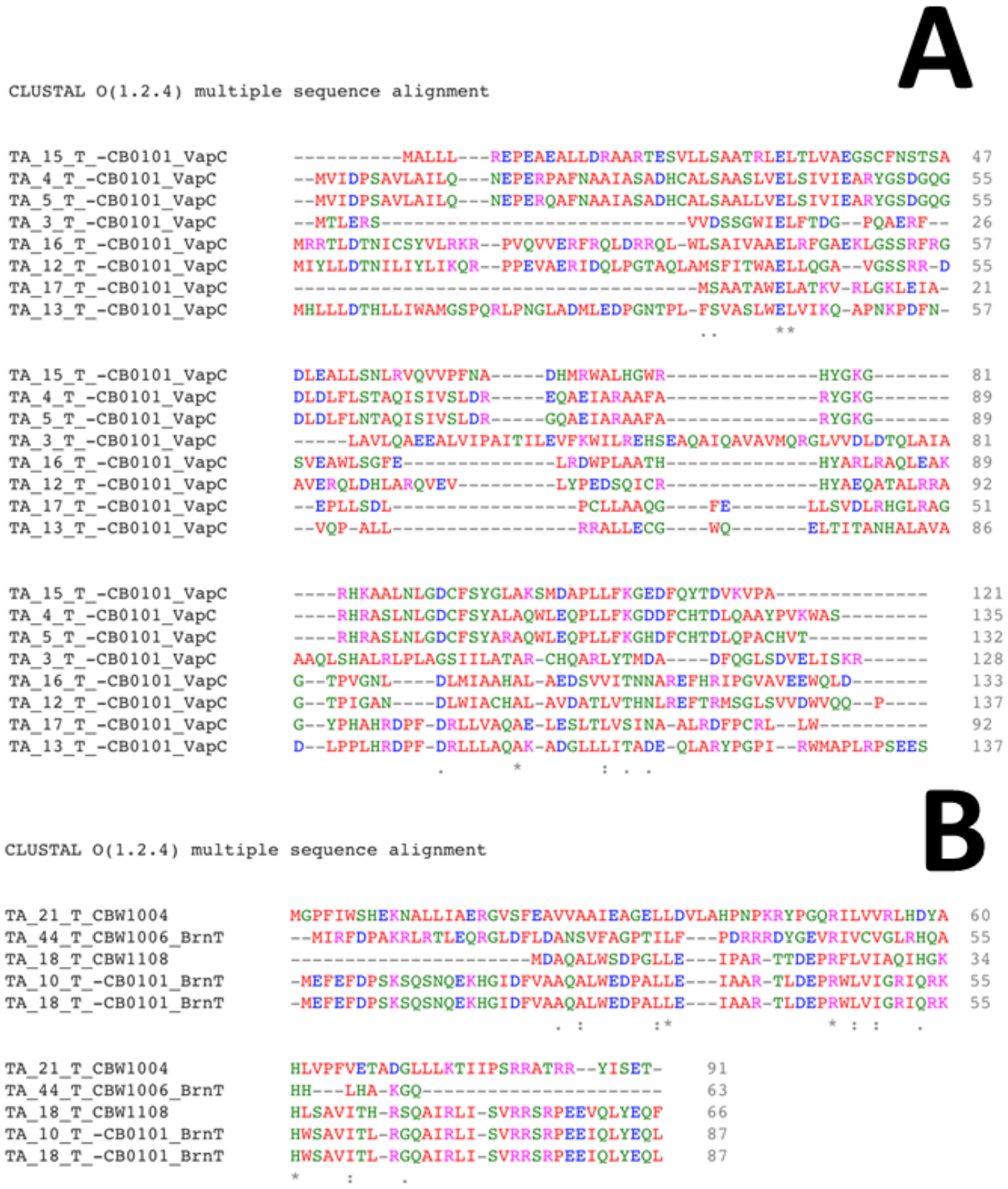
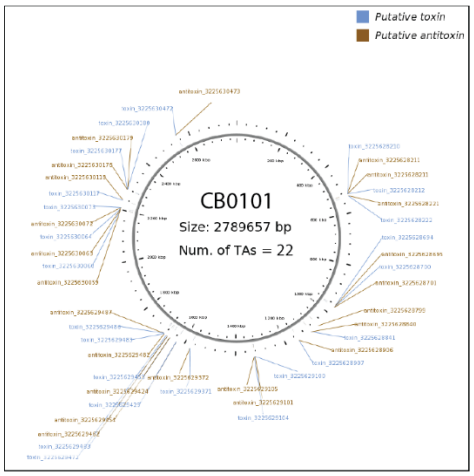
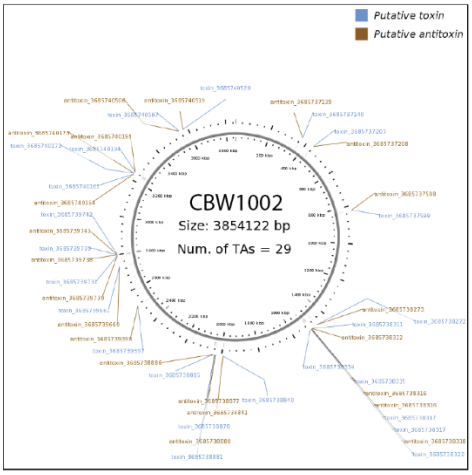


Figure 4.2. Amino acid alignments of select toxin-antitoxin systems in Chesapeake Bay

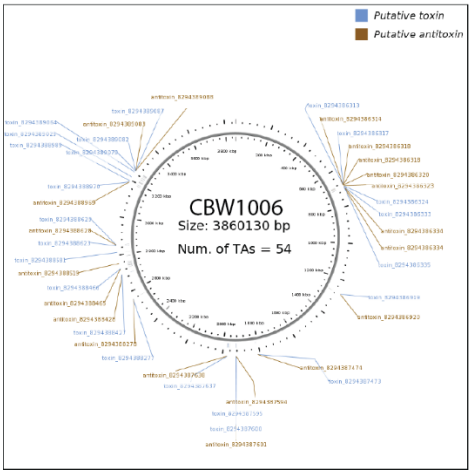
Synechococcus strains. A) Eight vapC toxin amino acid alignments sourced from CB0101 resulting in three conserved residues (E, L, and A denoted by *). B) Five BrnT toxin amino acid alignments from CB0101, CBW1108, CBW1006, and CBW1004 resulting in only two conserved residues (R and H denoted by *).


Putative toxin
Putative antitoxin
CB0101
Size: 2789657 bp
Num. of TAs = 22


Putative toxin
Putative antitoxin
CBW1107
Size: 3202093 bp
Num. of TAs = 29


Putative toxin
Putative antitoxin
CBW1002
Size: 3854122 bp
Num. of TAs = 29


Putative toxin
Putative antitoxin
CBW1004
Size: 3672318 bp
Num. of TAs = 46


Putative toxin
Putative antitoxin
CBW1006
Size: 3860130 bp
Num. of TAs = 54

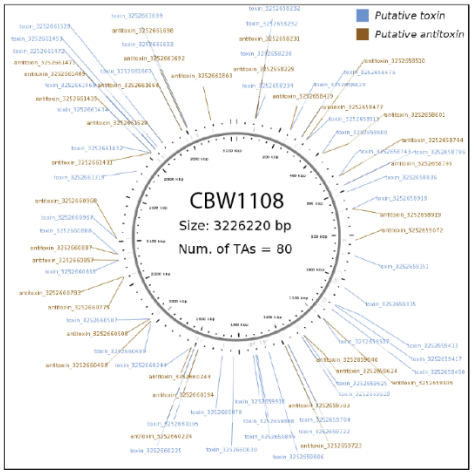
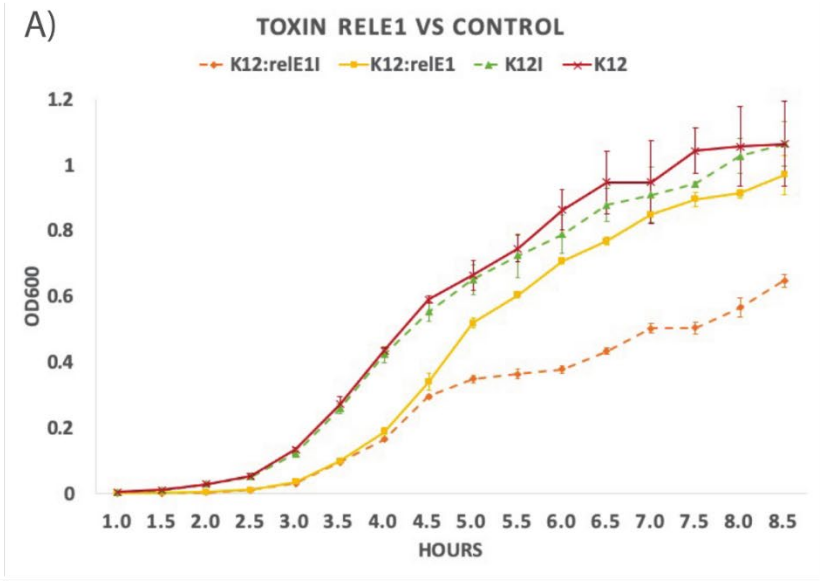
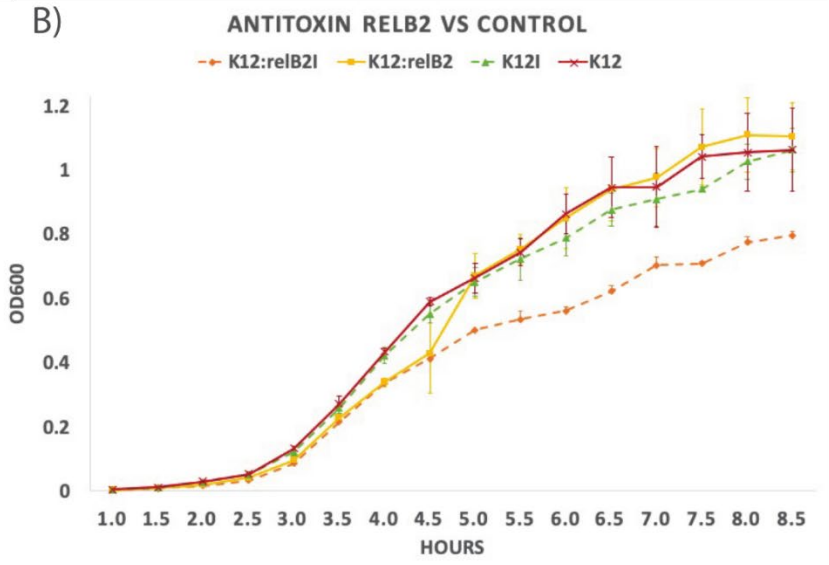

Putative toxin
Putative antitoxin
CBW1108
Size: 3226220 bp
Num. of TAs = 80

Figure 4.3. Circular genome maps of Chesapeake Bay *Synechococcus* strains and reference strains with toxin antitoxin systems highlighted.


A)
TOXIN RELE1 VS CONTROL
K12:relE1I
K12:relE1
K12I
K12
OD600
HOURS


B)
ANTITOXIN RELB2 VS CONTROL
K12:relB2I
K12:relB2
K12I
K12
OD600
HOURS

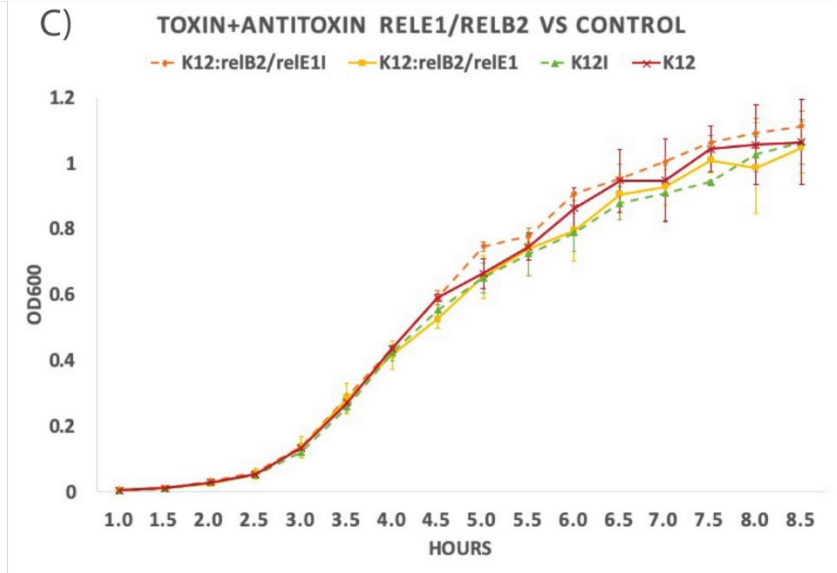
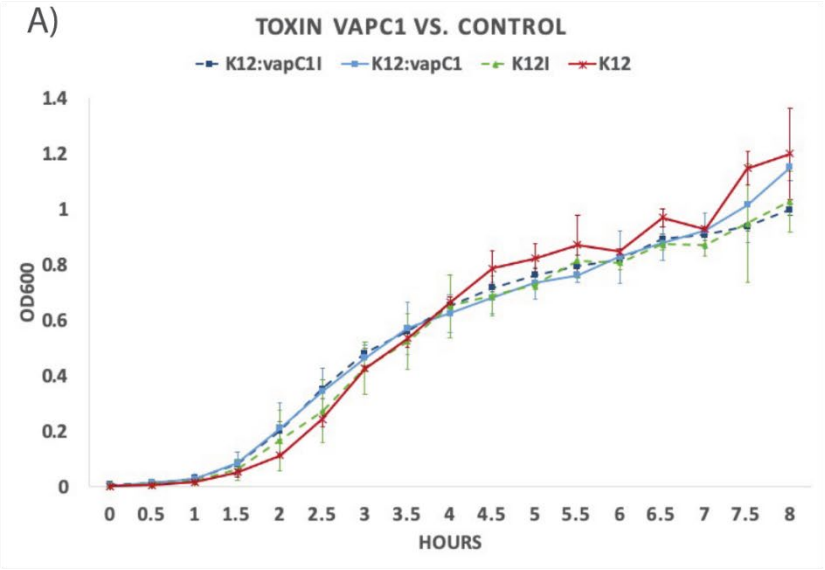
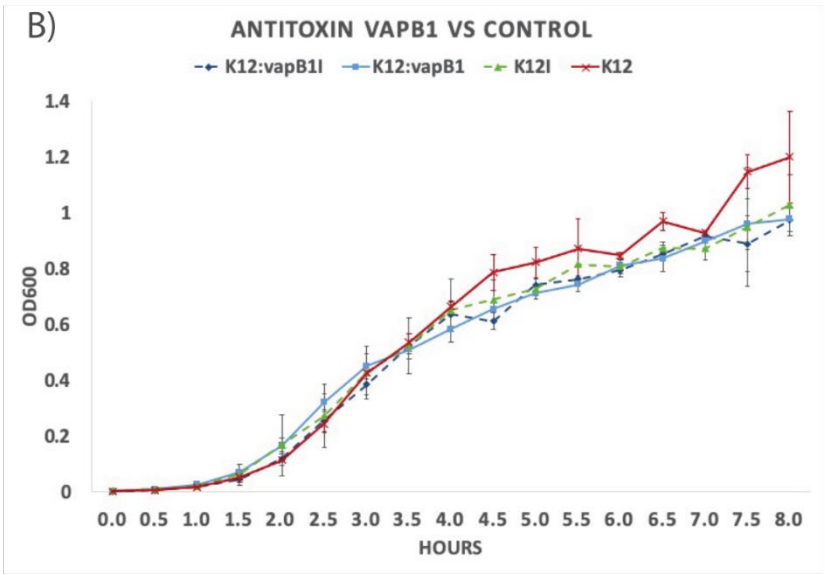

C)
TOXIN+ANTITOXIN RELE1/RELB2 VS CONTROL
K12:relB2/relE1I
K12:relB2/relE1
K12I
K12
OD600
HOURS

Figure 4.4. Growth of *E. coli* K12 which contained toxin *relE*[1], antitoxin *relB*[2], toxin-antitoxin (*relE*[1]*/relB*[2]) or control vector *pCA24N* (no gene inserted) with IPTG (dash line) or without IPTG (solid line). (A) Induced toxin strain represented by K12:*relE1I*, (B) induced antitoxin strain represented by *K12:relB2I*. (C) Induced strain expressing toxin-antitoxin complex represented by *K12:relB2/relE1I*. K12 strains represent the control strain E coli K12 ER2738 containing the empty vector *pCA24N*, induced control strain represented by K12I.


A)
TOXIN VAPC1 VS. CONTROL
K12:vapC1I
K12:vapC1
K12I
K12
OD600
HOURS


B)
ANTITOXIN VAPB1 VS CONTROL
K12:vapB1I
K12:vapB1
K12I
K12
OD600
HOURS

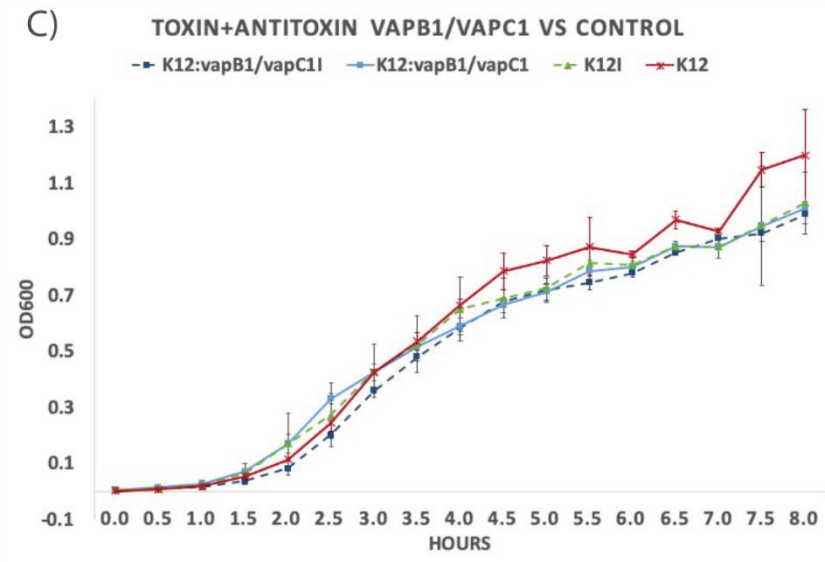

C)
TOXIN+ANTITOXIN VAPB1/VAPC1 VS CONTROL
K12:vapB1/vapC1I
K12:vapB1/vapC1
K12I
K12
OD600
HOURS

Figure 4.5. Growth of *E. coli* K12 which contained toxin *vapC*^{1}, antitoxin *vapB*^{1}, toxin-antitoxin (*vapB*^{1}/*vapC*^{1}) or control vector *pCA24N* (no gene inserted) with IPTG (dash line) or without IPTG (solid line). (A) Induced toxin strain represented by K12:*vapC1*, (B) induced antitoxin strain represented by *K12:vapB1*. (C) Induced strain expressing toxin-antitoxin complex represented by *K12:vapB1/vapC1*. K12 strains represent the control strain E coli K12 ER2738 containing the empty vector *pCA24N*, induced control strain represented by K12I.

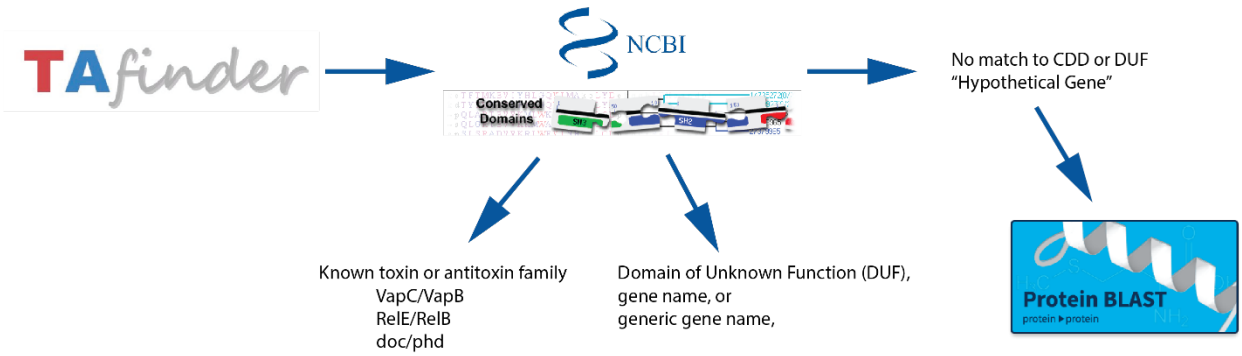


Figure 4.6. Decision algorithm to determine identity of putative toxin antitoxin systems in *Synechococcus.* Type 2 toxin-antitoxins were predicted using TAfinder software. To determine their identity these predicted amino acid sequences were queried against NCBI's Conserved Domain Database (CDD). If a known toxin or antitoxin family was predicted, such as *vapC/vapB, relE/relB*, or *doc/phd*, that traditional name was used. If a less specific, Domain of Unknown Function (DUF) or generic name was found, it was used for categorization. 97% of TA pairs predicted were able to be categorized this way. However, for the few hypothetical genes, a blastp search was used to determine the best category for the putative protein.

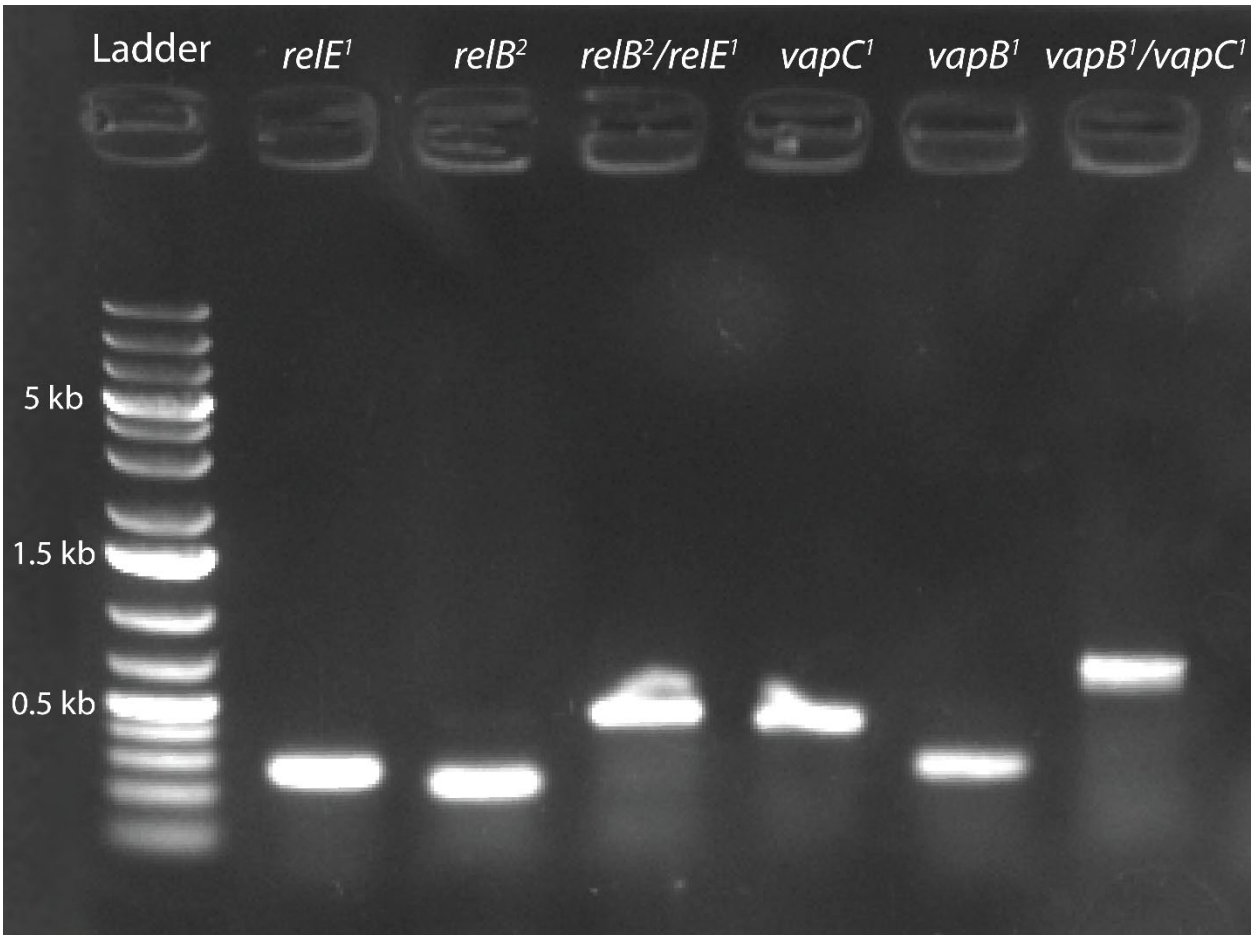


Figure 4.7. Various toxin-antitoxin genes amplified from corresponding transformed *E. coli* strains to confirm positive cloning.

Table 4.1. Genomic statistics for Chesapeake Bay estuary, coastal, freshwater, and open ocean *Synechococcus*.

Synechococcus strain *name*	Status	Isolation Location	Habitat	Reference	Contigs	Length (bp)	GC % Content	Gene Number	ncRNA	Accession Number	Putative TA Pairs	TA % of CDS	Network Map Association Connections
CB0101	Complete	Chesapeake Bay, Maryland	Estuary	Marsan et al. 2017; Fucich et al. 2019	1	2,789,657	64.1	3,128	76	CP039373	22	1.41	12
CB0205	Draft	Chesapeake Bay, Maryland	Estuary	Marsan et al. 2017	78	2,427,308	63	2,788	47	GCA_000179255.1	1	0.07	1
CBW1002	Complete	Chesapeake Bay, Maryland	Estuary	This study	1	3,854,122	65.15	3,994	61	CP060398	29	1.45	17
CBW1004	Complete	Chesapeake Bay, Maryland	Estuary	This study	1	3,672,318	67.35	3,668	83	CP060397	45	2.45	26
CBW1006	Complete	Chesapeake Bay, Maryland	Estuary	This study	1	3,860,130	65.08	4,047	62	CP060396	54	2.67	27
CBW1107	Complete	Chesapeake Bay, Maryland	Estuary	This study	1	3,202,093	66.3	3,456	50	CP064908	29	1.68	13
CBW1108	Complete	Chesapeake Bay, Maryland	Estuary	This study	1	3,226,220	64.35	3,744	48	CP060395	80	4.27	34
CC9311	Complete	Edge of California Current, Coast, Pacific Ocean USA	Coastal	Palenik et al., 2006	1	2,606,748	52.5	3,065	46	ASM1458	3	0.20	3
KORDI-49	Complete	Marine, South China Sea	Open Ocean	Choi et al. 2009	1	2,585,813	57.6	2,528	54	ASM73757	4	0.32	4
OmCyn01	Complete	Ectosymbiont, *Ornithocercus magnificus*	Open Ocean Symbiote	Nakayama et al. 2019	16	1,878,918	48.5	2,099	44	GCA_007996965.1	0	0.00	0
PCC6301	Complete	Freshwater Texas, USA 1952	Freshwater	Sugita et al. 2007	1	2,696,255	55.5	2,525	47	ASM1006	10	0.79	9
PCC6307	Complete	Freshwater lake Wisconsin, USA 1949	Freshwater	Havercamp et al. 2009	1	3,342,364	68.71	3,439	42	PRJNA158695	42	2.44	24
WH7803	Complete	Marine, Sargasso Sea	Open Ocean	Six et al., 2007	1	2,366,980	60.1	2,533	54	ASM6350	1	0.08	1
WH7805	Permanent Draft	Marine, Sargasso Sea	Open Ocean	Six et al. 2007	13	2,627,046	57.6	2,937	47	AAOK01000000	0	0.00	0
WH8102	Complete	Marine, Sargasso Sea	Open Ocean	Rocap et al. 2002	1	2,434,428	59.4	2,513	55	ASM19597	15	1.19	9

Table 4.2. Significant upregulation (red) or down-regulation (blue) of toxin-antitoxin systems of CB0101 under various stress conditions. Transcriptomic data were obtained from Marsan et al. 2017. Toxins are represented by italics; antitoxins are represented in bold. Significance is based on the threshold of $p<0.01$ and a minimum foldchange of two.

	locus	TA family	Toxin Activity	Nitrogen Deplete RNA-Seq	Phosphate Deplete RNA-Seq	Zinc Toxicity RNA-Seq	Original Publication
T	gsyne_1326	*relEparEyoeB*	Cleavage of ribosome-bound mRNA	2.2	1.5	2.1	Marsan et al., 2017
AT	gsyne_1325	**phdyefM**		-1.6	-1.4	-1.3	Marsan et al., 2017
T	gsyne_1792	*PIN_vapC*	Cleavage of tRNA	1.8	1.2	2.7	Marsan et al., 2017
AT	gsyne_1794	**phdyefM**		2.5	1.3	-1.2	Marsan et al., 2017
T	gsyne_1883	*relEparEyoeB*	Cleavage of ribosome-bound mRNA	2.3	3.1	10.6	Marsan et al., 2017
AT	gsyne_1882	**relB**		-2.5	1.2	1.1	Marsan et al., 2017
T	gsyne_2550	*PIN_vapC*	Cleavage of tRNA	3.2	2.3	2.0	This study
AT	gsyne_2551	**DUF433**		3.3	2.0	-2.7	This study

Table 4.3: Primer pairs designed for the amplification and cloning of the toxin and antitoxin genes. Strain name was determined for each one of the plasmids after transformation into E. coli K12 ER2738.

Genes	*Forward primer*	*Reverse primer*	*Type*	*Strain name*
relB[2]	GCCGCTCAGGTGACGGCCCGACT	CCATCCGCAGCGAGCAGCACA	Antitoxin	*K12:relB2*
relE[1]	GCCTGCTGCTCGCTGCGGATTAA	CCGCGATACACCTCCTTGCGAT	Toxin	*K12:relE1*
relB[2]**/ relE**[1]	GCCGCTCAGGTGACGGCCCGACT	CCGCGATACACCTCCTTGCGAT	Antitoxin-toxin complex	*K12:relB2/relE1*
vapB[1]	GCCACAGCGTCATTGCCTAGCCG	CCCAACCCCTCACGGGTCTGGA	Antitoxin	*K12:vapB1*
vapC[1]	GCCATCTATCTGCTCGACACCAA	CCTGGTTGCTGCACCCAATCCA	Toxin	*K12:vapC1*
vapB[1]**/vapC**[1]	GCCACAGCGTCATTGCCTAGCCG	CCTGGTTGCTGCACCCAATCCA	Antitoxin-toxin complex	*K12:vapB1/vapC1*

Chapter V: Conclusions and Future Directions

Synechococcus spp. are a vital component in nearly all aquatic environments. Because of their abundance and ubiquity, they contribute significantly to global carbon sequestration and are vital for primary productivity in aquatic ecosystems (Dvořák et al., 2014b; Li & Url, 1994). Their presence in open oceans is well studied because of their cohabitation with their related genus, the *Prochlorococcus*. The vast nature of the open ocean pelagic makes its global ecological significance clear (Flombaum et al., 2013). Therefore, the majority of picocyanobacterial research was focused on oceanic strains. The open ocean, although vast, is only one of the seemingly countless habitats of the *Synechococcus*.

Unlike *Prochlorococcus, Synechococcus* has a ubiquitous distribution, they can be found in the open ocean, on the coast, in freshwater and estuarine systems, frigid polar waters, and even in hot springs. *Synechococcus* thrives in the hot and the cold; the saline and the fresh; and every condition in between. *Synechococcus* has the apparent capability to adapt and tolerate nearly any environmental condition. This ability to conform to the conditions of almost any aquatic environment may be possible through vast genome plasticity (Callieri, 2017), a character that is not present in the *Prochlorococcus* (Dufresne et al., 2005).

Genetic elements for plasticity are often mobile genome islands, transposons, plasmids, *etc.* Other frequent members of the mobilome are toxin-antitoxin (TA) systems which are often exchanged through horizontal gene transfer (Guglielmini & Van Melderen, 2011; Van Melderen, 2010). These elements are observed in *Synechococcus*, but never in *Prochlorococcus*. This is of particular interest given the vastly different global distributions of the genera. The striking

difference of type II TA system presence and absence among the picocyanobacteria was the first concept introduced in Chapter 2.

TA systems may be helpful for environmental tolerance in *Synechococcus*. First, TA systems have a strong link to the general bacterial stringent response (Habib et al., 2018; Maisonneuve & Gerdes, 2014). Second, several TA systems have been shown to be upregulated during experimental high light exposure and simulated oxidative stress *in situ* in *Synechococcus* CB0101 (D. Marsan et al., 2017; D. W. Marsan, 2016). Lastly, putative TA systems are more common in strains that are endemic to more volatile coastal and estuarine environments (Fucich & Chen, 2020, Chapter 2). For these reasons, the abundance of TA systems in *Synechococcus* and conversely the absence from *Prochlorococcus* genomes is almost intuitive. The theme of generalists vs. specialists has been explored among the coastal and pelagic *Synechococcus* (Stuart et al., 2009), and on a more broad scale of the picocyanobacteria, exemplified by *Synechococcus* and *Prochlorococcus* (Dufresne et al., 2008). Generalists like *Synechococcus* are even capable of growing in frigid waters of the Chesapeake Bay and the Bornholm Sea. Several strains have been isolated from the Baltimore Inner Harbor in the Chesapeake Bay during winter months which cluster together with *Cyanobium* spp. (Xu et al., 2015).

Chesapeake Bay winter (CBW) *Synechococcus* strains represent several unique clades including the Bornholm Sea cluster, subalpine cluster II, CB7 (winter II) clade, and the novel CBW1004 clade. These representative strains share many genomic features described in Chapter 3. Among them are homologs including desaturases, chaperones, and transposases implicated in cold adaptation and general stress response. The Bornholm Sea cluster strains

CBW1002 and CBW1006 contain some of the largest genomes among *Synechococcus* (~3.8 Mb), and they share most homologs between them. Overall, CBW strains share more homologs among themselves than with subcluster 5.2 representative CB0101 and WH8102 from subcluster 5.1. Comparative genomics of CBW strains with other reference strains was presented in Chapter 3. CBW strains contain high number of fatty acid desaturase, transposase, and chaperone genes as examples, and these genomic features may allow them to survive cold or other stress conditions. Homologs shared between all Chesapeake Bay strains (CBW1002, CBW1004, CBW1006, CBW1107, CBW1108, and CB0101), but not pelagic strains from subcluster 5.1 could indicate genes implicated in survival in turbulent estuaries like the Chesapeake Bay.

To better understand cold adaptation by CBW strains, multiple methods were used to differentiate them from freshwater, coastal, and open ocean strains in Chapter 3. Cold stress response genes were queried using amino acid similarity and automatic annotation from the RAST and PATRIC servers (Aziz et al., 2008; Brettin et al., 2015; Davis et al., 2020; Overbeek et al., 2014). Amino acid sequences of interest were sourced from genes induced by cold shock in *Synechocystis* PCC6803 (Sinetova & Los, 2016) and canonical bacterial cold response genes in *E. coli* (Barria et al., 2013). In some cases, methods were repeated from work searching for cold response genes in the Antarctic *Synechococcus* strain SynAce01 (CS-601) (Tang et al., 2019). CBW strains tend to have high amount or cold implicated genes similar to freshwater strains, these include desaturases, chaperones, and transposases. In general, coastal and open ocean *Synechococcus* tend to have fewer these cold or stress response genes. In many cases, CBW strains contained duplicates of these genes. In Chapter 3, a highly conserved transposase of

interest (IS5 Family transposase) has nearly 20 duplications according to amino acid sequence homology (Figure 3.3) in CBW strains, but only 1 and 2 copies in WH8101 and SynAce01, respectively.

The high abundance of TA systems predicted in CBW strains is described in Chapter 4. CBW strains contain an unusually high amount of TA pairs compared to other *Synechococcus* strains. In the case of CBW1108, it contains nearly twice as many (n=80) putative TA pairs found in the next highest *Synechococcus,* freshwater strains PCC6307 and PCC6312 (n=42). CBW TA systems increase complexity with abundance. Strains with the most putative TA pairs, namely CBW1108 and CBW1006, have the most intricate association networks with several 'hub' toxins connecting to several different antitoxins. Strains with fewer TA pairs have more simple association network diagrams with fewer connections. Toxins often contain a conserved domain, and pair in a promiscuous manner with multiple antitoxins containing different conserved domains. These promiscuous toxins act as 'hubs' and in nearly all association maps are VapC and RelE, in few cases the nucleotidyl transferase NT_KNT. The nature of different antitoxins acting as the antidote to these promiscuous toxins supports the mix and match hypothesis (Fasani & Savageau, 2015; Guglielmini & Van Melderen, 2011).

In summary, the discovery of TA systems in *Synechococcus* is still relatively new (~10 years); no systematic approach was applied to their study in picocyanobacteria until this work. The stark contrast of the near ubiquity of TA systems in *Synechococcus* and the complete absence in *Prochlorococcus* is of note due to their differential roles in ecological theory. The complete genome sequences of five Chesapeake Bay winter *Synechococcus* show that estuarine *Synechococcus* contain rich cold and stress response genes. The large genome size and high

number of transposase genes enable more genomic plasticity for the Chesapeake Bay winter *Synechococcus*. The incorporation and retention of TA systems and transposases may be a contributing genetic factor that allows *Synechococcus* to fill the role of a generalist. *Prochlorococcus,* as a specialist that is dominant in stable, pelagic environments, has no use for persister cell formation. Further, the tight coupling of abundant and diverse TA pairs in strains inhibiting highly variable environments suggests the importance of TA systems to expanded environmental tolerance.

Future Directions

Current data suggests that some TA systems in *Synechococcus* are active when exposed to stressful conditions. In CB0101, significant upregulation of relE occurred during experimental high light exposure (D. Marsan et al., 2017). Similarly, VapC and other toxins were upregulated during simulated oxidative stress. These results suggest these toxins are active and may play a role in growth retardation and the formation of a subsequent persister state in *Synechococcus* CB0101. Confirmation of TA systems in foreign hosts is a common method to verify functionality. Few TA pairs from CB0101 have been tested and verified in an *E. coli* host, and it is necessary to confirm what proportion of these putative TA systems are functional. Quantitative PCR of putative TA systems under simulated stress conditions could be used to find likely candidates. Once identified, these select TA pairs could be transformed into a foreign *E. coli* host to verify activity, as demonstrated in supplemental material of Chapter 4.

Upon prediction and careful examination, several peculiar toxin antitoxin systems with unique organizations were found in CB0101 and other Chesapeake Bay *Synechococcus* strains. Interesting TA loci were organized in non-traditional fashion, like the gsyne_618, gsyne_619, gsyne_620 cassette in CB0101. This apparently contains a putative antitoxin flanked by two toxin genes. This is atypical for the traditional two gene toxin-antitoxin. Gene activity could be confirmed in several ways, including transformation to a foreign host as described above. Confirmation of activity could elucidate, listed in order of likelihood, if A) one toxin is active and the other is vestigial, B) both toxins are inactive and the TA system is not functional, or C) this is a truly unique TA pair with two toxins negated by one antitoxin.

A more in-depth genome wide comparisons of Chesapeake Bay strains would give insight into estuarine *Synechococcus* genomic adaptation. By identifying and assessing the pan genome, or the collection of genes shared by all strains of a defined group, the most important genes needed to survive in estuarine systems could be described. Given the size and unique seasonal conditions of the Chesapeake Bay, this would be an excellent opportunity to see what genes are important for survival in estuaries and in cold water temperatures. Further analysis and identification of cold or stress response genes are still needed to fully understand the diversity and phylogenetic relationship of these genes between CBW strains and other cyanobacteria. Whole genome comparison can give insight to the genetic evolution, potential horizontal gene transfer and synteny of Chesapeake Bay genomes.

In a similar way that a short list of potentially active TA systems can be made by inducing oxidative stress and measuring gene expression, qPCR can be used to measure the activity of cold response genes. The list of ~30 cold induced genes in CBW and their frequency in each genome (Table 3.2, 3.3, and 3.4) can be used to identify transcripts of interest. For example, it will be interesting to learn more about the role and interaction of fatty acid desaturase, chaperone, and transposase genes in cold adaptation of Chesapeake Bay winter picocyanobacteria. CBW strains can be exposed to cold temperatures (~4^{o}C) for temporary and extended periods of time and RNA expression of the cold induced genes of interest can be used to infer activity. Once potentially active genes are identified, antibodies for the few candidate strains can be produced to confirm protein production.

Appendix I

Supplementary Code

##TAfinderCommandRecord

```
scp -r PredictedSynTASystems dfucich@130.85.169.195:/data1/dfucich/


ssh dfucich@130.85.169.195


#I am getting a permission denied error in the server


I retried the copy on a Linux machine, and I have no issues with
permissions

scp -r PredictedSynTAPairs dfucich@130.85.169.195:/data1/dfucich/


cat *all_TA_proteins.fas > AllPredictedSynTAPairs.fas


nice -n 14 nohup blastp -query
PredictedSynTAPairs/AllPredictedSynTAPairs.fas -subject
PredictedSynTAPairs/AllPredictedSynTAPairs.fas -out
PredictedSynTAPairs/Results -outfmt "6" &
```

```
cat *all_TA_proteins.fas > AllPredictedSynTAPairs.fas

blastp -query PredictedSynTAPairs/AllPredictedSynTAPairs.fas -subject
ncnr -out PredictedSynTAPairs/Results -outfmt "6"

blastp -db <nr> -query
<PredictedSynTAPairs/AllPredictedSynTAPairs.fas> -out <outfile> -
outfmt "6 qseqid sseqid"



blastn -query transcripts.fa -out transcripts.blast.txt -task
megablast -db refseq_rna -num_threads 12 -evalue 1e-10 -
```

```
best_hit_score_edge 0.05 -best_hit_overhang 0.25 -outfmt 7 -
perc_identity 50 -max_target_seqs 1 &
```

```
scp dfucich@130.85.169.195:/data1/dfucich/PredictedSynTAPairs/Results
~/
```

```
grep TA_1_T_* PredictedSynTAPairs/AllPredictedSynTAPairs.fas
```

Remove Redundant results

```
awk '$1!=$2' Results > ResultsNonRedundant
```

or if you want to just print Redundant results

```
awk '$1==$2' Results > ResultsRedundant
```

```
scp dfucich@130.85.169.195:/data1/dfucich/PredictedSynTAPairs/Result*
~/
```

Uploaded CB0101 to Island viewer

To Do To make sense of my data

plot evalues as a histogram to find a reasonable cutoff

plot all evalues against themselves

Future, to get better data

1. Separate all T and all ATs
2. Blast these against each other

Ernest plan

3. pick a few indicative ones or interesting TAs and blast them against your SynTADB, find what is interesting, plot evalue pident and learn what you could begin to consider ‘families’

CB0101 relE and relB

but also, the HipA in all of the CBW strains

upload all of the TAs of interest from the local machine to the

scp -r TADB dfucich@130.85.169.195:/data1/dfucich/

so, let's get a look at our e values and particularly when our evalues fall off when we look at just a few of our "TA's of interest", with regard to the TADB.

I will also need to separate the "TADB" into really a toxin database: "TDB" and an antitoxin database "ATDB"

grep -A1 "_T_" AllPredictedSynTAPairsNoBreaks.fas > TDB

to remove line breaks

#this worked well except I have to remove the "enter" from the fasta file, as it only writes 1 line

#this works removes the break

```
awk '!/^>/ { printf "%s", $0; n = "\n" } /^>/ { print n $0; n = "" }
END { printf "%s", n }' input.fasta > Output.fasta
```

#this also

```
awk '/^>/{print s? s"\n"$0:$0;s="";next}{s=s
sprintf("%s",$0)}END{if(s)print s}' file > out
```

#now....SEE UPDATE BELOW!!!!!

```
grep -A1 "_T_" AllPredictedSynTAPairsNoBreaks.fas > TDB

#ugh this prints a -- between some of the lines SEE UPDATE BELOW!!!!!


#sweet so this is an undocumented case where you can modify grep to
not print out a group separator

grep -A1 --no-group-separator "_T_" AllPredictedSynTAPairsNoBreaks.fas
> TDB


#now for ATDB

grep -A1 --no-group-separator "_AT_"
AllPredictedSynTAPairsNoBreaks.fas > ATDB


#now let's go ahead and blastp each of the "interesting' TA systems
against out TDB and ATDB


#I would like: -outfmt "6 qseqid sseqid pident length evalue bitscore
sseq"


blastp -query TAsOfInterest/TOXINS/TA_14_T_-CB0101_relE.txt -subject
PredictedSynTAPairs/TDB -out TAsOfInterest/TOXINS/CB0101relEInTDB -
outfmt "6"


blastp -query TAsOfInterest/Antitoxins/TA_12_AT_-
CBW1002_Cognate_hipA.fsa -subject PredictedSynTAPairs/ATDB -out
TAsOfInterest/Antitoxins/CBW1002_TA_12_AT_-CBW1002_Cognate_hipAInTDB -
outfmt "6 qseqid sseqid pident length evalue bitscore sseq"

blastp -query TAsOfInterest/Antitoxins/TA_12_AT_-
CBW1002_Cognate_hipA.fsa -subject PredictedSynTAPairs/ATDB -out
TAsOfInterest/Antitoxins/CBW1002_TA_12_AT_-CBW1002_Cognate_hipAInTDB -
outfmt "6 qseqid sseqid pident length evalue bitscore sseq"


#I went through and manually blastp all of CB0101 putative TA pairs
and annotated them in the .fas file to aid in my blastp against the
TDB and ATDB


#can I use a multifasta file to blastp and get a reasonable output?
```

```
blastp -query PredictedSynTAPairs/ -subject PredictedSynTAPairs/TDB -
out TAsOfInterest/TOXINS/CB0101relEInTDB -outfmt "6"


#copied "annotated' fasta to server to replace old fasta

scp IMETServerFiles/Synechococcus\ sp.\ CB0101_all_TA_proteins.fas
dfucich@130.85.169.195:/data1/dfucich/PredictedSynTAPairs


#linearize

awk '!/^>/ { printf "%s", $0; n = "\n" } /^>/ { print n $0; n = "" }
END { printf "%s", n }' PredictedSynTAPairs/Synechococcus\ sp.\
CB0101_all_TA_proteins.fas > PredictedSynTAPairs/Synechococcus\ sp.\
CB0101_all_TA_proteinsNoBreaks.fas


#separate CB0101 toxins

grep -A1 --no-group-separator "_T_" PredictedSynTAPairs/Synechococcus\
sp.\ CB0101_all_TA_proteinsNoBreaks.fas >
TAsOfInterest/TOXINS/SynCB0101Toxins


#multifasta to check all

blastp -query TAsOfInterest/TOXINS/SynCB0101Toxins -subject
PredictedSynTAPairs/TDB -out
TAsOfInterest/TOXINS/CB0101LikeToxinsSynTDB -outfmt "6 qseqid sseqid
pident length evalue bitscore sseq"


#now let's do the same for ATs

grep -A1 --no-group-separator "_AT_"
PredictedSynTAPairs/Synechococcus\ sp.\
CB0101_all_TA_proteinsNoBreaks.fas >
TAsOfInterest/Antitoxins/SynCB0101Antitoxins


blastp -query TAsOfInterest/Antitoxins/SynCB0101Antitoxins -subject
PredictedSynTAPairs/ATDB -out
TAsOfInterest/Antitoxins/CB0101LikeAntitoxinsSynATDB -outfmt "6 qseqid
sseqid pident length evalue bitscore sseq"
```

```
#let’s download these

scp
dfucich@130.85.169.195:/data1/dfucich/TAsOfInterest/TOXINS/CB0101LikeT
oxinsSynTDB ~/IMETServerFiles


scp
dfucich@130.85.169.195:/data1/dfucich/TAsOfInterest/Antitoxins/CB0101L
ikeAntitoxinsSynATDB ~/IMETServerFiles


#i used R to plot the e values for these and now I want to just
extract ones with an acceptable e value

#now let's set that evalue to something highly specific let’s go .0001

blastp -query TAsOfInterest/TOXINS/SynCB0101Toxins -subject
PredictedSynTAPairs/TDB -out
TAsOfInterest/TOXINS/CB0101LikeToxinsSynTDBStringent -outfmt "6 qseqid
sseqid pident length evalue bitscore sseq" -evalue .0001


blastp -query TAsOfInterest/Antitoxins/SynCB0101Antitoxins -subject
PredictedSynTAPairs/ATDB -out
TAsOfInterest/Antitoxins/CB0101LikeAntitoxinsSynATDBStringent -outfmt
"6 qseqid sseqid pident length evalue bitscore sseq" -evalue .0001


#let's find how many of each we have

awk '{A[$1]++}END{for(i in A)print i,A[i]}'
TAsOfInterest/TOXINS/CB0101LikeToxinsSynTDBStringent


#obviously the nucleotidyltransferase is most popular, VapC being a
close second

awk '{A[$1]++}END{for(i in A)print i,A[i]}'
TAsOfInterest/TOXINS/CB0101LikeToxinsSynTDBStringent | sort -rn -k 2 >
TAsOfInterest/TOXINS/CB0101LikeToxinsSynTDBStringentCount


awk '{A[$1]++}END{for(i in A)print i,A[i]}'
TAsOfInterest/Antitoxins/CB0101LikeAntitoxinsSynATDBStringent | sort -
rn -k 2 >
TAsOfInterest/Antitoxins/CB0101LikeAntitoxinsSynATDBStringentCount
```

```
#convert a blastp output to a fasta using awk, take column 2 and 7 and
make a fasta from this

awk '{print ">"$2"\n"$7}'
TAsOfInterest/TOXINS/CB0101LikeToxinsSynTDBStringent  >
tabtofastaseqs.fa


#but let's be specific about which we want, so let's pipe it up to
specify

grep "TA_13_T_-CB0101_VapC"
TAsOfInterest/TOXINS/CB0101LikeToxinsSynTDBStringent | awk '{print
">"$2"\n"$7}'  >
TAsOfInterest/TOXINS/TA_13_T_CB0101_VapCtabtofastaseqs.fa

#so, with this fasta file I can do an alignment, find conserved areas,
and make a phylogenetic tree


grep "TA_17_AT_-CB0101_PhD/YefM"
TAsOfInterest/Antitoxins/CB0101LikeAntitoxinsSynATDBStringent | awk
'{print ">"$2"\n"$7}'  >
TAsOfInterest/Antitoxins/TA_17_AT_CB0101_PhDYefMtabtofastaseqs.fa


#so, I need to be much more stringent because I feel i am getting many
many false positives.  I have a plan to reduce these by maximizing
stringency with e values, and by making the minimum sequence length be
65 amino acids as per (Brown, 2003 A Novel Family of Escherichia coli
Toxin-Antitoxin Gene Pairs

)

#when I was aligning these using phylogeny.fr and mega7, I had some
poor alignments and therefore trees.

If i am stringent in prediction, I can be less stringent when
comparing the predicted Toxins and Antitoxins in Synechococcus

awk '{print $7}'
TAsOfInterest/Antitoxins/CB0101LikeAntitoxinsSynATDBStringent | wc -m


awk '{print $7}'
TAsOfInterest/Antitoxins/CB0101LikeAntitoxinsSynATDBStringent | awk
'{print length}' | sort -rn
```

#11.25.2019

#compiled the definitive list of NCBI available Syn genomes with TA putative pairs. allSyn33 is the folder

cat *all_TA_proteins.fas > AllPredictedSynTAPairsNCBI.fas

#separate toxins

grep -A1 --no-group-separator "_T_" AllPredictedSynTAPairsNCBI.fas > AllSynToxins.fas

#now let's do the same for ATs

grep -A1 --no-group-separator "_AT_" AllPredictedSynTAPairsNCBI.fas > AllSynAntiToxins.fas

#error, pAQ6 did not have _T_ or _AT_ it only had _T and _AT so this modification makes it correct see below

dfucich@bioinfo:~/all33Syn/AllPredictedSynTAPairs$ grep -A1 --no-group-separator "_T" AllPredictedSynTAPairsNCBI.fas > AllSynToxins.fas

dfucich@bioinfo:~/all33Syn/AllPredictedSynTAPairs$ grep -A1 --no-group-separator "_AT" AllPredictedSynTAPairsNCBI.fas > AllSynAntiToxins.fas

#I am onto comparing several of the CBW, CB0101 and CB0205 to other strains more closely. Namely, Marine Syn WH8102 and KORDI-49 and Freshwater strains PCC6307 and PCC6301. I digress. I am going to redo these "other" *Synechococcus* strains so that I have the most up to date NCBI CDD.

#So, I will once again be parse out the toxins from the antitoxins and make a "otherTDB.fas" and "otherATDB.fas" so to speak.

grep -A1 --no-group-separator "_T_" OtherSynTAsystemsMarineWH8102_KORDI49_FW_PCC6307_PCC6301.fas > otherTDB.fas

```
grep -A1 --no-group-separator "_AT_"
OtherSynTAsystemsMarineWH8102_KORDI49_FW_PCC6307_PCC6301.fas >
otherATDB.fas
```

##CB0101andCBWOrthologCommandRecord

#File of results compiled in RBH*Synechocystis*PCC6803CIDin*Synechococcus*

#quick look at general orthologs between CB0101 and CBW

#blastp proteomes using reciprocal blast hits

#1108

```
blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
FilesForDucTape/CBW1108.Gene.pep.fasta -soft_masking "false" -out
CB0101-CBW1108 -evalue 1e-10 -outfmt "6"
```

```
blastp -query FilesForDucTape/CBW1108.Gene.pep.fasta -subject
FilesForDucTape/CB0101_6666666.413450.faa  -soft_masking "false" -out
CBW1108-CB0101 -evalue 1e-10 -outfmt "6"
```

```
python3.7 RBH-v1.py CB0101-CBW1108 CBW1108-CB0101 RBHCB0101-
CBW1108.csv
```

```
wc -l RBHCB0101-CBW1108.csv
```

#1002

```
blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
FilesForDucTape/CBW1002.Gene.pep.fasta -soft_masking "false" -out
CB0101-CBW1002 -evalue 1e-10 -outfmt "6"
```

```
blastp -query FilesForDucTape/CBW1002.Gene.pep.fasta -subject
FilesForDucTape/CB0101_6666666.413450.faa  -soft_masking "false" -out
CBW1002-CB0101 -evalue 1e-10 -outfmt "6"
```

```
python3.7 RBH-v1.py CB0101-CBW1002 CBW1002-CB0101 RBHCB0101-
CBW1002.csv
```

```
wc -l RBHCB0101-CBW1002.csv
```

```
#1004
```

```
blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
FilesForDucTape/CBW1004.Gene.pep.fasta -soft_masking "false" -out
CB0101-CBW1004 -evalue 1e-10 -outfmt "6"
```

```
blastp -query FilesForDucTape/CBW1004.Gene.pep.fasta -subject
FilesForDucTape/CB0101_6666666.413450.faa  -soft_masking "false" -out
CBW1004-CB0101 -evalue 1e-10 -outfmt "6"
```

```
python3.7 RBH-v1.py CB0101-CBW1004 CBW1004-CB0101 RBHCB0101-
CBW1004.csv
```

```
wc -l RBHCB0101-CBW1004.csv
```

```
#1006
```

```
blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
FilesForDucTape/CBW1006.Gene.pep.fasta -soft_masking "false" -out
CB0101-CBW1006 -evalue 1e-10 -outfmt "6"
```

```
blastp -query FilesForDucTape/CBW1006.Gene.pep.fasta -subject
FilesForDucTape/CB0101_6666666.413450.faa  -soft_masking "false" -out
CBW1006-CB0101 -evalue 1e-10 -outfmt "6"
```

```
python3.7 RBH-v1.py CB0101-CBW1006 CBW1006-CB0101 RBHCB0101-
CBW1006.csv
```

```
wc -l RBHCB0101-CBW1006.csv
```

```
#how about cb0101 to cb0101

#CB0101

blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
FilesForDucTape/CB0101_6666666.413450.faa -soft_masking "false" -out
CB0101-CB0101 -evalue 1e-10 -outfmt "6"


blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
FilesForDucTape/CB0101_6666666.413450.faa  -soft_masking "false" -out
CB0101-CB0101 -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py CB0101-CB0101 CB0101-CB0101 RBHCB0101-CB0101.csv



wc -l RBHCB0101-CB0101.csv


#how about WH8102 to cb0101

#WH8102

blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
WH8102/8102ProteinsGCF_000195975.1_ASM19597v1_protein.faa  -
soft_masking "false" -out CB0101-WH8102 -evalue 1e-10 -outfmt "6"


blastp -query
WH8102/8102ProteinsGCF_000195975.1_ASM19597v1_protein.faa -subject
FilesForDucTape/CB0101_6666666.413450.faa  -soft_masking "false" -out
WH8102-CB0101 -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py CB0101-WH8102 WH8102-CB0101 RBHWH8102-CB0101.csv


wc -l RBHWH8102-CB0101.csv
```

```
#cystisPCC6803

blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
cystis6803GCF_000009725.1_ASM972v1_protein.faa  -soft_masking "false"
-out CBProteomeRBH/CB0101-cystis -evalue 1e-10 -outfmt "6"


blastp -query cystis6803GCF_000009725.1_ASM972v1_protein.faa -subject
FilesForDucTape/CB0101_6666666.413450.faa   -soft_masking "false" -out
CBProteomeRBH/cystis-CB0101 -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py CBProteomeRBH/CB0101-cystis CBProteomeRBH/cystis-
CB0101 CBProteomeRBH/RBHcystis-CB0101.csv


wc -l RBHcystis-CB0101.csv


#After doing some organization, I put all of the Gene.pep.fasta's in
"Proteomes" and all of the Ortholog files will go into "OrthologsRBH"

# so, as a result the above command is a bit clunky but hopefully that
gets better below


#let’s get started with cystisPCC6803

#cystisPCC6803-WH8102


blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-WH8102.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/cystisPCC6803-WH8102.txt
OrthologsRBH/WH8102-cystisPCC6803.txt OrthologsRBH/RBHWH8102-
cystisPCC6803.csv


wc -l OrthologsRBH/RBHWH8102-cystisPCC6803.csv
```

```
#cystisPCC6803-CBW1108


blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-CBW1108.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/cystisPCC6803-CBW1108.txt
OrthologsRBH/CBW1108-cystisPCC6803.txt OrthologsRBH/RBHCBW1108-
cystisPCC6803.csv


wc -l OrthologsRBH/RBHCBW1108-cystisPCC6803.csv


#cystisPCC6803-CBW1002


blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-CBW1002.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/cystisPCC6803-CBW1002.txt
OrthologsRBH/CBW1002-cystisPCC6803.txt OrthologsRBH/RBHCBW1002-
cystisPCC6803.csv


wc -l OrthologsRBH/RBHCBW1002-cystisPCC6803.csv


#cystisPCC6803-CBW1004
```

```
blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-CBW1004.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/cystisPCC6803-CBW1004.txt
OrthologsRBH/CBW1004-cystisPCC6803.txt OrthologsRBH/RBHCBW1004-
cystisPCC6803.csv


wc -l OrthologsRBH/RBHCBW1004-cystisPCC6803.csv


#cystisPCC6803-CBW1006


blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-CBW1006.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/cystisPCC6803-CBW1006.txt
OrthologsRBH/CBW1006-cystisPCC6803.txt OrthologsRBH/RBHCBW1006-
cystisPCC6803.csv


wc -l OrthologsRBH/RBHCBW1006-cystisPCC6803.csv


#cystisPCC6803-cystisPCC6803


blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"
```

```
blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/cystisPCC6803-cystisPCC6803.txt
OrthologsRBH/cystisPCC6803-cystisPCC6803.txt
OrthologsRBH/RBHcystisPCC6803-cystisPCC6803.csv


wc -l OrthologsRBH/RBHcystisPCC6803-cystisPCC6803.csv


#now 1002

#CBW1002-CBW1004


blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-CBW1004.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-CBW1002.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1002-CBW1004.txt
OrthologsRBH/CBW1004-CBW1002.txt OrthologsRBH/RBHCBW1002-CBW1004.csv


wc -l OrthologsRBH/RBHCBW1002-CBW1004.csv


#CBW1002-CBW1006


blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-CBW1006.txt -evalue 1e-10 -outfmt "6"
```

```
blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-CBW1002.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1002-CBW1006.txt
OrthologsRBH/CBW1006-CBW1002.txt OrthologsRBH/RBHCBW1002-CBW1006.csv


wc -l OrthologsRBH/RBHCBW1002-CBW1006.csv



#CBW1002-CBW1108


blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-CBW1108.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-CBW1002.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1002-CBW1108.txt
OrthologsRBH/CBW1108-CBW1002.txt OrthologsRBH/RBHCBW1002-CBW1108.csv


wc -l OrthologsRBH/RBHCBW1002-CBW1108.csv


##CBW1002-WH8102


blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-WH8102.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-CBW1002.txt -evalue 1e-10 -outfmt "6"
```

```
python3.7 RBH-v1.py OrthologsRBH/CBW1002-WH8102.txt
OrthologsRBH/WH8102-CBW1002.txt OrthologsRBH/RBHCBW1002-WH8102.csv


wc -l OrthologsRBH/RBHCBW1002-WH8102.csv


##CBW1002-CBW1002

blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-CBW1002.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-CBW1002.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1002-CBW1002.txt
OrthologsRBH/CBW1002-CBW1002.txt OrthologsRBH/RBHCBW1002-CBW1002.csv


wc -l OrthologsRBH/RBHCBW1002-CBW1002.csv



#now 1006 with others that are missing


#CBW1004-CBW1006


blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-CBW1006.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-CBW1004.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1004-CBW1006.txt
OrthologsRBH/CBW1006-CBW1004.txt OrthologsRBH/RBHCBW1004-CBW1006.csv
```

```
wc -l OrthologsRBH/RBHCBW1004-CBW1006.csv



#CBW1004-CBW1108


blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-CBW1108.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-CBW1004.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1004-CBW1108.txt
OrthologsRBH/CBW1108-CBW1004.txt OrthologsRBH/RBHCBW1004-CBW1108.csv


wc -l OrthologsRBH/RBHCBW1004-CBW1108.csv


##CBW1004-WH8102


blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-WH8102.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-CBW1004.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1004-WH8102.txt
OrthologsRBH/WH8102-CBW1004.txt OrthologsRBH/RBHCBW1004-WH8102.csv


wc -l OrthologsRBH/RBHCBW1004-WH8102.csv
```

```
##CBW1004-CBW1004

blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-CBW1004.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-CBW1004.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1004-CBW1004.txt
OrthologsRBH/CBW1004-CBW1004.txt OrthologsRBH/RBHCBW1004-CBW1004.csv


wc -l OrthologsRBH/RBHCBW1004-CBW1004.csv


#now the rest of to 1006


#CBW1006-CBW1108


blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-CBW1108.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-CBW1006.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1006-CBW1108.txt
OrthologsRBH/CBW1108-CBW1006.txt OrthologsRBH/RBHCBW1006-CBW1108.csv


wc -l OrthologsRBH/RBHCBW1006-CBW1108.csv


##CBW1006-WH8102
```

```
blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-WH8102.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-CBW1006.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1006-WH8102.txt
OrthologsRBH/WH8102-CBW1006.txt OrthologsRBH/RBHCBW1006-WH8102.csv


wc -l OrthologsRBH/RBHCBW1006-WH8102.csv


##CBW1006-CBW1006

blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-CBW1006.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-CBW1006.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1006-CBW1006.txt
OrthologsRBH/CBW1006-CBW1006.txt OrthologsRBH/RBHCBW1006-CBW1006.csv


wc -l OrthologsRBH/RBHCBW1006-CBW1006.csv


#1108 and WH8102, that's it then also WH102-WH8102


##CBW1108-WH8102


blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-WH8102.txt -evalue 1e-10 -outfmt "6"
```

```
blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-CBW1108.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1108-WH8102.txt
OrthologsRBH/WH8102-CBW1108.txt OrthologsRBH/RBHCBW1108-WH8102.csv


wc -l OrthologsRBH/RBHCBW1108-WH8102.csv


##CBW1108-CBW1108

blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-CBW1108.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-CBW1108.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1108-CBW1108.txt
OrthologsRBH/CBW1108-CBW1108.txt OrthologsRBH/RBHCBW1108-CBW1108.csv


wc -l OrthologsRBH/RBHCBW1108-CBW1108.csv


#and now WH8102-WH8102


##WH8102-WH8102

blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-WH8102.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-WH8102.txt -evalue 1e-10 -outfmt "6"
```

```
python3.7 RBH-v1.py OrthologsRBH/WH8102-WH8102.txt
OrthologsRBH/WH8102-WH8102.txt OrthologsRBH/RBHWH8102-WH8102.csv


wc -l OrthologsRBH/RBHWH8102-WH8102.csv




#2020, we are back in the command line doing data analysis.


#add CBW1107

#itself

blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-CBW1107.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1107-CBW1107.txt
OrthologsRBH/CBW1107-CBW1107.txt OrthologsRBH/RBHCBW1107-CBW1107.csv


wc -l OrthologsRBH/RBHCBW1107-CBW1107.csv


#1107-1002


blastp -query Proteomes/CBW1002.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1002-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/CBW1002.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-CBW1002.txt -evalue 1e-10 -outfmt "6"
```

```
python3.7 RBH-v1.py OrthologsRBH/CBW1002-CBW1107.txt
OrthologsRBH/CBW1107-CBW1002.txt OrthologsRBH/RBHCBW1107-CBW1002.csv


wc -l OrthologsRBH/RBHCBW1107-CBW1002.csv


#1107-1004


blastp -query Proteomes/CBW1004.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1004-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/CBW1004.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-CBW1004.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1004-CBW1107.txt
OrthologsRBH/CBW1107-CBW1004.txt OrthologsRBH/RBHCBW1107-CBW1004.csv


wc -l OrthologsRBH/RBHCBW1107-CBW1004.csv


#1107-1006


blastp -query Proteomes/CBW1006.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1006-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/CBW1006.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-CBW1006.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1006-CBW1107.txt
OrthologsRBH/CBW1107-CBW1006.txt OrthologsRBH/RBHCBW1107-CBW1006.csv


wc -l OrthologsRBH/RBHCBW1107-CBW1006.csv
```

```
#1107-1108


blastp -query Proteomes/CBW1108.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1108-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/CBW1108.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-CBW1108.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CBW1108-CBW1107.txt
OrthologsRBH/CBW1107-CBW1108.txt OrthologsRBH/RBHCBW1107-CBW1108.csv


wc -l OrthologsRBH/RBHCBW1107-CBW1108.csv


CB0101_6666666.413450.faa

#1107-0101


blastp -query Proteomes/CB0101.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CB0101-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/CB0101.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-CB0101.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/CB0101-CBW1107.txt
OrthologsRBH/CBW1107-CB0101.txt OrthologsRBH/RBHCBW1107-CB0101.csv


wc -l OrthologsRBH/RBHCBW1107-CB0101.csv


#1107-cystisPCC6803.Gene.pep.fasta
```

```
blastp -query Proteomes/cystisPCC6803.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/cystisPCC6803-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/cystisPCC6803.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-cystisPCC6803.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/cystisPCC6803-CBW1107.txt
OrthologsRBH/CBW1107-cystisPCC6803.txt OrthologsRBH/RBHCBW1107-
cystisPCC6803.csv


wc -l OrthologsRBH/RBHCBW1107-wh8102.csv


#okay and lastly 1107-WH8102.Gene.pep.fasta


blastp -query Proteomes/WH8102.Gene.pep.fasta -subject
Proteomes/CBW1107.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/WH8102-CBW1107.txt -evalue 1e-10 -outfmt "6"


blastp -query Proteomes/CBW1107.Gene.pep.fasta -subject
Proteomes/WH8102.Gene.pep.fasta -soft_masking "false" -out
OrthologsRBH/CBW1107-WH8102.txt -evalue 1e-10 -outfmt "6"


python3.7 RBH-v1.py OrthologsRBH/WH8102-CBW1107.txt
OrthologsRBH/CBW1107-WH8102.txt OrthologsRBH/RBHCBW1107-WH8102.csv


wc -l OrthologsRBH/RBHCBW1107-WH8102.csv
```

##*ColdStressResponseCommandRecord*

#02182020

tblastn -query ColdStressResponse/*Synechocystis*\ PCC\ Cold\ Induced\ Genes\ ALL.fasta -subject CBWStrains/CBW1002.Chromosome.fasta

#for presence or absence....output to a tab delimited format

tblastn -query ColdStressResponse/*Synechocystis*\ PCC\ Cold\ Induced\ Genes\ ALL.fasta -subject CBWStrains/CBW1002.Chromosome.fasta -out CBWStrains/CBW1002ColdInducedGenes -evalue 0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"

tblastn -query ColdStressResponse/*Synechocystis*\ PCC\ Cold\ Induced\ Genes\ ALL.fasta -subject CBWStrains/CBW1004.Chromosome.fasta -out CBWStrains/CBW1004ColdInducedGenes -evalue 0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"

tblastn -query ColdStressResponse/*Synechocystis*\ PCC\ Cold\ Induced\ Genes\ ALL.fasta -subject CBWStrains/CBW1006.Chromosome.fasta -out CBWStrains/CBW1006ColdInducedGenes -evalue 0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"

tblastn -query ColdStressResponse/*Synechocystis*\ PCC\ Cold\ Induced\ Genes\ ALL.fasta -subject CBWStrains/CBW1108.Chromosome.fasta -out CBWStrains/CBW1108ColdInducedGenes -evalue 0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"

#02192020

#so, I think an issue I could have is false positives with this method. I want to ensure there is an ORF upstream of the similar sequences to limit false positives. So, if I blast the genes of interest against the putative proteome, I should be able to make sure the similarity is in the orfs

#these files are conveniently located in /FilesForDuctApe how convenient

#they are also peptides, so I will be changing to blastp

```
#also, I am pretty sure I had duplicated the Unknown genes in the All
file so I must change that


 blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1002.faa  -out
CBWStrains/ColdInducedGenesInCBW/CBW1002ColdInducedGenes.csv  -evalue
0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"


#probably good to check the line count

wc -l CBWStrains/ColdInducedGenesInCBW/CBW1002ColdInducedGenes.csv

#actually, quite similar to previous results wc - = 247 so I suppose I
reduced 6 potential false positives


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1004.faa  -out
CBWStrains/ColdInducedGenesInCBW/CBW1004ColdInducedGenes.csv  -evalue
0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1006.faa  -out
CBWStrains/ColdInducedGenesInCBW/CBW1006ColdInducedGenes.csv  -evalue
0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1108.faa  -out
CBWStrains/ColdInducedGenesInCBW/CBW1108ColdInducedGenes.csv  -evalue
0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CB0101.faa  -out
CBWStrains/ColdInducedGenesInCBW/CB0101ColdInducedGenes.csv  -evalue
0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"


 #what about with an open ocean strain?

 tblastn -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject WH8102/81 -out WH8102/8102ColdInducedGenes
-evalue 0.00001 -outfmt "6 qseqid sseqid pident evalue length qlen"
```

```
 blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject WH8102/81  -out
WH8102/8102ColdInducedGenes.csv  -evalue 0.00001 -outfmt "6 qseqid
sseqid pident evalue length qlen"



 #there is not much difference between CBW, CB, and even WH8102
strains.  This is very surprising to me, I expected CBW strains to
show the most


#After talking with Ernest and Tsetso, I've learned a few things, and
need to do a lot more to narrow down these orthologs.

#my e value is much too low, should be 10^-50 much better....

#also, I should try to do reciprocal blasts to verify that they are
the best hits to each genome.  Additionally I can turn off soft
masking to not eliminate blastp from not aligning "Low-complexity
regions and interspersed repeats typically match many sequences" these
are not normally biologically important, but for my purposes of trying
to find an identical gene this is very important

#also, it is advantageous to print both the query and subject
alignments so that I can work with them later, good to print out
regularly

#I am still not sure about the possibility of using soft masking when
using a -subject.  That is to say when not using a database.
Currently the protein fasta files are not in db format.



TO BE COMPLETED......
```

```
blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject WH8102/81 -out WH8102/test  -evalue 1e-50 -
outfmt "6 qseqid sseqid pident evalue length qlen qseq sseq"


ColdStressResponse/Stringentevalue


#this will complete the blastp with high stringency and then
immediately do a word count line to quickly tell you how many seqs
have aligned. I removed the printing of the seqs, that can be done
later quickly.


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1002.Gene.pep.fasta -out
ColdStressResponse/StringentEvalue/CBW1002cold.csv  -evalue 1e-50 -
outfmt "6 qseqid sseqid pident evalue length qlen" | wc -l
ColdStressResponse/StringentEvalue/CBW1002cold.csv


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1004.Gene.pep.fasta -out
ColdStressResponse/StringentEvalue/CBW1004cold.csv  -evalue 1e-50 -
outfmt "6 qseqid sseqid pident evalue length qlen" | wc -l
ColdStressResponse/StringentEvalue/CBW1004cold.csv


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1006.Gene.pep.fasta -out
ColdStressResponse/StringentEvalue/CBW1006cold.csv  -evalue 1e-50 -
outfmt "6 qseqid sseqid pident evalue length qlen"


wc -l ColdStressResponse/StringentEvalue/CBW1006cold.csv


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1108.Gene.pep.fasta -out
ColdStressResponse/StringentEvalue/CBW1108cold.csv  -evalue 1e-50 -
outfmt "6 qseqid sseqid pident evalue length qlen"


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1108.Gene.pep.fasta -
soft_masking "true" -out
```

```
ColdStressResponse/StringentEvalue/CBW1108cold.csv  -evalue 1e-50 -
outfmt "6 qseqid sseqid pident evalue length qlen"


#soft masking off....I believe it is false by default.

blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1108.Gene.pep.fasta -
soft_masking "false" -out
ColdStressResponse/StringentEvalue/CBW1108cold.csv  -evalue 1e-50 -
outfmt "6 qseqid sseqid pident evalue length qlen"


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CB0101.faa -soft_masking
"false" -out ColdStressResponse/StringentEvalue/CB0101cold.csv  -
evalue 1e-50 -outfmt "6 qseqid sseqid pident evalue length qlen"


#so simply doing the blastp is not good enough.  I must employ a
reciprocal blast hit strategy RBH to have a greater confidence in the
orthologs


#I found a script to parse through them

#this is an example with Bacillus anthracis and Bacillus subtilis168


#the inputs are blast outfmt 6, so you may determine your evalue.

#the output is simply a list of the RBH

#this protocol could be repeated for all of the CBW strains


python3.7 RBH-v1.py B.anthracis.Ames-B.subtilis168.FF.txt
B.subtilis168-B.anthracis.Ames.FF.txt outRBH.txt




#CBW1002


#first you need to make the blast table to show org1->org2
```

```
blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1002.Gene.pep.fasta -
soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/6803CID-CBW1002.txt -evalue 1e-50 -outfmt "6"


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1002.Gene.pep.fasta -
soft_masking "false" -out RBH\6803CID-CBW1002.txt -evalue 1e-20 -
outfmt "6"


#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1002.Gene.pep.fasta -subject
ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\ Genes\ ALL.fasta
-soft_masking "false" -out RBH\Simple-reciprocal-best-blast-hit-pairs-
master/CBW1002-6803CID.txt -evalue 1e-20 -outfmt "6"




blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1002.Gene.pep.fasta -
soft_masking "false" -out RBH\6803CID-CBW1002.txt -evalue 1e-20 -
outfmt "6"


#now run the python script to show potential orthologs




#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py 6803CID-CBW1002.txt CBW1002-6803CID.txt
outCBW1002CID.csv


#CBW1004


#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1004.Gene.pep.fasta -subject
ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\ Genes\ ALL.fasta
```

```
-soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/CBW1004-6803CID.txt -evalue 1e-20 -outfmt "6"


#now run the python script to show potential orthologs


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1004.Gene.pep.fasta -
soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/6803CID-CBW1004.txt -evalue 1e-20 -outfmt "6"


#move the directory

cd RBH/Simple-reciprocal-best-blast-hit-pairs-master


#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py 6803CID-CBW1004.txt CBW1004-6803CID.txt
outCBW1004CID.csv


 #CBW1006

#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1006.Gene.pep.fasta -subject
ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\ Genes\ ALL.fasta
-soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/CBW1006-6803CID.txt -evalue 1e-20 -outfmt "6"


#now run the python script to show potential orthologs


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1006.Gene.pep.fasta -
soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/6803CID-CBW1006.txt -evalue 1e-20 -outfmt "6"


#move the directory

cd RBH/Simple-reciprocal-best-blast-hit-pairs-master
```

```
#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py 6803CID-CBW1006.txt CBW1006-6803CID.txt
outCBW1006CID.csv



  #CBW1108

#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1108.Gene.pep.fasta -subject
ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\ Genes\ ALL.fasta
-soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/CBW1108-6803CID.txt -evalue 1e-20 -outfmt "6"


#now run the python script to show potential orthologs


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1108.Gene.pep.fasta -
soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/6803CID-CBW1108.txt -evalue 1e-20 -outfmt "6"


#move the directory

cd RBH/Simple-reciprocal-best-blast-hit-pairs-master


#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py 6803CID-CBW1108.txt CBW1108-6803CID.txt
outCBW1108CID.csv



  #CBW1108

#then the same thing with org2->org1
```

```
blastp -query FilesForDucTape/CBW1108.Gene.pep.fasta -subject
ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\ Genes\ ALL.fasta
-soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/CBW1108-6803CID.txt -evalue 1e-20 -outfmt "6"


#now run the python script to show potential orthologs


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1108.Gene.pep.fasta -
soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/6803CID-CBW1108.txt -evalue 1e-20 -outfmt "6"


#move the directory

cd RBH/Simple-reciprocal-best-blast-hit-pairs-master


#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py 6803CID-CBW1108.txt CBW1108-6803CID.txt
outCBW1108CID.csv




#WH8102

blastp -query
WH8102/8102ProteinsGCF_000195975.1_ASM19597v1_protein.faa -subject
ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\ Genes\ ALL.fasta
-soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/WH8102-6803CID.txt -evalue 1e-20 -outfmt "6"

#then the same thing with org2->org1


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject
WH8102/8102ProteinsGCF_000195975.1_ASM19597v1_protein.faa -
soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/6803CID-WH8102.txt -evalue 1e-20 -outfmt "6"

#now run the python script to show potential orthologs
```

```
#move the directory

cd RBH/Simple-reciprocal-best-blast-hit-pairs-master


#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py 6803CID-WH8102.txt WH8102-6803CID.txt
outWH8102CID.csv


#CB0101


#then the same thing with org2->org1

blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\ Genes\ ALL.fasta
-soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/CB0101-6803CID.txt -evalue 1e-20 -outfmt "6"


#now run the python script to show potential orthologs


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CB0101_6666666.413450.faa -
soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/6803CID-CB0101.txt -evalue 1e-20 -outfmt "6"


#move the directory

cd RBH/Simple-reciprocal-best-blast-hit-pairs-master | python3.7 RBH-
v1.py 6803CID-CB0101.txt CB0101-6803CID.txt outCB0101CID.csv


#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py 6803CID-CB0101.txt CB0101-6803CID.txt
outCB0101CID.csv


#My new worry is that the CID genes are too small of a database to
weed out false negatives.  The whole 6803 proteome may be necessary to
```

eliminate potential false positives that could be highly similar to other genes in the 6803 genome.

#so, what I need to do is add the CID to the proteome with the same fasta call entry so that it will show as a rbh for the python program

##

#CBW1002 RBH with full Synechococystis PCC6803

#first you need to make the blast table to show org1->org2

blastp -query ColdStressResponse/*Synechocystis*\ PCC\ Cold\ Induced\ Genes\ ALL.fasta -subject FilesForDucTape/CBW1002.Gene.pep.fasta -soft_masking "false" -out RBH\Simple-reciprocal-best-blast-hit-pairs-master\6803CID-CBW1002.txt -evalue 1e-20 -outfmt "6"

#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1002.Gene.pep.fasta -subject ColdStressResponse/*Synechocystis*\ PCC\ 6803\ Peptides\ CIDReplaced.fasta -soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-master/AgainstFUll6803Proteome/CBW1002-6803CID.txt -evalue 1e-20 -outfmt "6"

#this step will take a bit longer

#now run the python script to show potential orthologs

#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py RBH/Simple-reciprocal-best-blast-hit-pairs-master/oldsmallCID/6803CID-CBW1002.txt RBH/Simple-reciprocal-best-blast-hit-pairs-master/AgainstFUll6803Proteome/CBW1002-6803CID.txt

RBH/Simple-reciprocal-best-blast-hit-pairs-master/AgainstFUll6803Proteome/outCBW1002CID.csv

#CBW1004 RBH with full Synechococystis PCC6803

#first you need to make the blast table to show org1->org2

#these have been done previously, so they are copied for reference

blastp -query ColdStressResponse/*Synechocystis*\ PCC\ Cold\ Induced\ Genes\ ALL.fasta -subject FilesForDucTape/CBW1004.Gene.pep.fasta -soft_masking "false" -out RBH\Simple-reciprocal-best-blast-hit-pairs-master\6803CID-CBW1004.txt -evalue 1e-20 -outfmt "6"

#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1004.Gene.pep.fasta -subject ColdStressResponse/*Synechocystis*\ PCC\ 6803\ Peptides\ CIDReplaced.fasta -soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-master/AgainstFUll6803Proteome/CBW1004-6803CID.txt -evalue 1e-20 -outfmt "6"

#this step will take a bit longer

#now run the python script to show potential orthologs

#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py RBH/Simple-reciprocal-best-blast-hit-pairs-master/oldsmallCID/6803CID-CBW1004.txt RBH/Simple-reciprocal-best-blast-hit-pairs-master/AgainstFUll6803Proteome/CBW1004-6803CID.txt RBH/Simple-reciprocal-best-blast-hit-pairs-master/AgainstFUll6803Proteome/outCBW1004CID.csv

#CBW1006 RBH with full Synechococystis PCC6803

```
#first you need to make the blast table to show org1->org2

#these have been done previously, so they are copied for reference

blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1006.Gene.pep.fasta -
soft_masking "false" -out RBH\Simple-reciprocal-best-blast-hit-pairs-
master\6803CID-CBW1006.txt -evalue 1e-20 -outfmt "6"


#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1006.Gene.pep.fasta -subject
ColdStressResponse/Synechocystis\ PCC\ 6803\ Peptides\
CIDReplaced.fasta  -soft_masking "false" -out RBH/Simple-reciprocal-
best-blast-hit-pairs-master/AgainstFUll6803Proteome/CBW1006-
6803CID.txt -evalue 1e-20 -outfmt "6"

#this step will take a bit longer


#now run the python script to show potential orthologs



#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py RBH/Simple-reciprocal-best-blast-hit-pairs-
master/oldsmallCID/6803CID-CBW1006.txt RBH/Simple-reciprocal-best-
blast-hit-pairs-master/AgainstFUll6803Proteome/CBW1006-6803CID.txt
RBH/Simple-reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outCBW1006CID.csv


wc -l RBH/Simple-reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outCBW1006CID.csv




#CBW1108 RBH with full Synechococystis PCC6803


#first you need to make the blast table to show org1->org2
```

```
#these have been done previously, so they are copied for reference

blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CBW1108.Gene.pep.fasta -
soft_masking "false" -out RBH\Simple-reciprocal-best-blast-hit-pairs-
master\6803CID-CBW1108.txt -evalue 1e-20 -outfmt "6"


#then the same thing with org2->org1

blastp -query FilesForDucTape/CBW1108.Gene.pep.fasta -subject
ColdStressResponse/Synechocystis\ PCC\ 6803\
Peptides\CIDReplaced.fasta  -soft_masking "false" -out RBH/Simple-
reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/CBW1108-6803CID.txt -evalue 1e-20 -
outfmt "6"

#this step will take a bit longer


#now run the python script to show potential orthologs



#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py RBH/Simple-reciprocal-best-blast-hit-pairs-
master/oldsmallCID/6803CID-CBW1108.txt RBH/Simple-reciprocal-best-
blast-hit-pairs-master/AgainstFUll6803Proteome/CBW1108-6803CID.txt
RBH/Simple-reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outCBW1108CID.csv


wc -l RBH/Simple-reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outCBW1108CID.csv


#WH8102



#then the same thing with org2->org1
```

```
blastp -query
WH8102/8102ProteinsGCF_000195975.1_ASM19597v1_protein.faa -subject
ColdStressResponse/Synechocystis\ PCC\ 6803\ Peptides\
CIDReplaced.fasta  -soft_masking "false" -out RBH/Simple-reciprocal-
best-blast-hit-pairs-master/AgainstFUll6803Proteome/WH8102-6803CID.txt
-evalue 1e-20 -outfmt "6"



#now run the python script to show potential orthologs


#move the directory


#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py RBH/Simple-reciprocal-best-blast-hit-pairs-
master/oldsmallCID/6803CID-WH8102.txt RBH/Simple-reciprocal-best-
blast-hit-pairs-master/oldsmallCID/WH8102-6803CID.txt RBH/Simple-
reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outWH8102CID.csv


wc -l RBH/Simple-reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outWH8102CID.csv



#CB0101


#org1->org2


blastp -query ColdStressResponse/Synechocystis\ PCC\ Cold\ Induced\
Genes\ ALL.fasta -subject FilesForDucTape/CB0101_6666666.413450.faa
-soft_masking "false" -out RBH/Simple-reciprocal-best-blast-hit-pairs-
master/oldsmallCID/6803CID-CB0101.txt -evalue 1e-20 -outfmt "6"


#then the same thing with org2->org1
```

```
blastp -query FilesForDucTape/CB0101_6666666.413450.faa -subject
ColdStressResponse/Synechocystis\ PCC\ 6803\ Peptides\
CIDReplaced.fasta    -soft_masking "false" -out RBH/Simple-reciprocal-
best-blast-hit-pairs-master/AgainstFUll6803Proteome/CB0101-6803CID.txt
-evalue 1e-20 -outfmt "6"


#now run the python script to show potential orthologs


#move the directory

cd RBH/Simple-reciprocal-best-blast-hit-pairs-master | python3.7 RBH-
v1.py 6803CID-CB0101.txt CB0101-6803CID.txt outCB0101CID.csv


#usage of the python code

python3.7 RBH-v1.py BLASTOUTPUT1 BLASTOUTPUT2 RBH-list-outfile

python3.7 RBH-v1.py RBH/Simple-reciprocal-best-blast-hit-pairs-
master/oldsmallCID/6803CID-CB0101.txt RBH/Simple-reciprocal-best-
blast-hit-pairs-master/AgainstFUll6803Proteome/CB0101-6803CID.txt
RBH/Simple-reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outCB0101CID.csv


wc -l RBH/Simple-reciprocal-best-blast-hit-pairs-
master/AgainstFUll6803Proteome/outCB0101CID.csv


######################useful option for blastp

blastp -max_target_seqs 1

#this only prints the top hit for each query, that is helpful if you
are only looking for the top hit and not interested in the partial
sequence alignments.




#09/01/2020

#Working on my chapter 3, which is cold inducted genes.   I have
compiled a new list of cold induced genes from Barria 2013. I will be
doing something similar with the 111 genes from Synechococystis. this
method is simpler, just a blastp against the proteome See Tang, 2019.
```

They simply tallied up how many hits against these genes. I have contacted the author to find out more details, but they are waiting on who conducted the research. I think they had an evalue of 1e-05 based on their paper. I am looking at 34 genes in 17 genomes, so a bit broader with genomes, more focused with genes.

#master blastp, will be replacing the proteome component

```
blastp -query ColdStressResponse/CIDfromBarria.fasta -subject
Proteomes\ -soft_masking "false" -out
ColdStressResponse/CIDfromBarria/ -evalue 1e-5 -outfmt "6"
```

#so not so great....none of them hit at all. even with default parameters....so I can't even replicate the Tang, 2019 paper......bad stuff. SEE UPDATE BELOW

```
blastp -query ColdStressResponse/CIDfromBarria.fasta -subject
Proteomes/SynAce01_GCF_001885215.1_ASM188521v1_cds_from_genomic.fna -
soft_masking "false" -out ColdStressResponse/Barria/Barria-SynAce01
```

#oops, the cds files from NCBI, are not amino acid seqs. Kind of hard to use blastp with AA vs ntds...

#all that means is that I will be using blastp for some tblastn for others. I am worried about comparing tblastn to blastp because tblastn suggests frame shifts, while blastp does not. IT ENDS UP BEING OKAY.

#blastp is CBW1* and CB0101. [AND CBW1107] Also CB0205 and WH7803 these are all from RAST others will be tblastn

```
blastp -query ColdStressResponse/CIDfromBarria.fasta -subject
Proteomes/CBW1 -soft_masking "false" -out
ColdStressResponse/Barria/Barria-CBW1 -evalue 1e-5 -outfmt "6"
```

#tblastn

```
tblastn -query ColdStressResponse/CIDfromBarria.fasta -subject
Proteomes/SynAce01_GCF_001885215.1_ASM188521v1_cds_from_genomic.fna -
```

```
soft_masking "false" -out ColdStressResponse/Barria/Barria-SynAce01 -
evalue 1e-5 -outfmt "6"


tblastn -query ColdStressResponse/CIDfromBarria.fasta -subject
Proteomes/SynAce01_GCF_001885215.1_ASM188521v1_cds_from_genomic.fna -
soft_masking "false" -out ColdStressResponse/Barria/Barria-SynAce01 -
evalue 1e-5 -outfmt "6"


We've added CBW1107 to the genome list and several (13) other genes
from CBW annotation to the list. So now our matrix is 18*46=828
fantastic!


#All of the query genes should be in AA fasta format. CBW1107 is a
gene.pep.fasta file Feng handed me that file. so, I will be using
blastp. Let's go. #see edits


#I am going to overwrite the original files and then keep the db file
from excel. I too like to live on the edge.



blastp -query ColdStressResponse/CIDfromBarria.fasta -subject
Proteomes/CBW1* -soft_masking "false" -out
ColdStressResponse/Barria/Barria-CBW1* -evalue 1e-5 -outfmt "6"


tblastn -query ColdStressResponse/CIDfromBarria.fasta -subject
Proteomes/SynAce01_GCF_001885215.1_ASM188521v1_cds_from_genomic.fna -
soft_masking "false" -out ColdStressResponse/Barria/Barria-SynAce01 -
evalue 1e-5 -outfmt "6"



#WH8102!!!!!!!!!!!!!! also must use blastp, not tblastn

#cystisPCC6803 too? yes that is correct.

#see the spreadsheet with checkboxes "Cold Induced Genes for Blastp in
CBW and other strains" sheet genome list
```

#Okay so this could turn into a never-ending search. I have added all of the CBW strain desaturases to the search. FA4 could be novel and in a novel pathway of hydrocarbon processing. So, I am including them all in the blast, so that I can gather all of the data and take a bird's eye view.

#So, I have an updated Barria2 and CIDfromBarria2 so we cannot overwrite data and keep things organized.

#also changing the output to -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qseq sseq"

########templates:

#template: blastp -query ColdStressResponse/CIDfromBarria2.fasta -subject Proteomes/CBW1* -soft_masking "false" -out ColdStressResponse/Barria2/Barria-CBW1* -evalue 1e-5 -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qseq sseq"

#actual

blastp -query ColdStressResponse/CIDfromBarria2.fasta -subject Proteomes/CB0101.Gene.pep.fasta -soft_masking "false" -out ColdStressResponse/Barria2/Barria-CB0101 -evalue 1e-5 -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qseq sseq"

tblastn -query ColdStressResponse/CIDfromBarria.fasta -subject Proteomes/SynAce01_GCF_001885215.1_ASM188521v1_cds_from_genomic.fna -soft_masking "false" -out ColdStressResponse/Barria2/Barria-SynAce01 -evalue 1e-5 -outfmt "6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qseq sseq"

#####Note 1107 is in the Proteome subfolder!!!!!!!!!! CBWFromBGI

###10/8/2020 I am adding four lpx genes lpxA-D to the Barria file, same as above go to templates

```
blastp -query ColdStressResponse/CIDfromBarria2.fasta -subject
Proteomes/CB0101.Gene.pep.fasta -soft_masking "false" -out
ColdStressResponse/Barria3/Barria-CB0101 -evalue 1e-5 -outfmt "6
qseqid sseqid pident length mismatch gapopen qstart qend sstart send
evalue bitscore qseq sseq"
```

```
tblastn -query ColdStressResponse/CIDfromBarria2.fasta -subject
Proteomes/SynAce01_GCF_001885215.1_ASM188521v1_cds_from_genomic.fna -
soft_masking "false" -out ColdStressResponse/Barria3/Barria-SynAce01 -
evalue 1e-5 -outfmt "6 qseqid sseqid pident length mismatch gapopen
qstart qend sstart send evalue bitscore qseq sseq"
```

References

Affronti, L. F. (1990). *Seasonal and diel patterns of abundance and productivity of phototrophic picoplankton in the lower Chesapeake Bay*. https://doi.org/10.25777/q77g-a234

Affronti, L. F., & Marshall, H. G. (1993). Diel abundance and productivity patterns of autotrophic picoplankton in the lower Chesapeake Bay. *Journal of Plankton Research*, *15*(1), 1–8. https://doi.org/10.1093/plankt/15.1.1

Affronti, L. F., & Marshall, H. G. (1994). Using frequency of dividing cells in estimating autotrophic picoplankton growth and productivity in the Chesapeake Bay. *Hydrobiologia*, *284*, 193–203.

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bourexis, D., Brister, J. R., Bryant, S. H., Canese, K., Charowhas, C., Clark, K., DiCuccio, M., Dondoshansky, I., Feolo, M., Funk, K., Geer, L. Y., Gorelenkov, V., Hlavina, W., … Zbicz, K. (2017a). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *45*(D1), D12–D17. https://doi.org/10.1093/nar/gkw1071

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bourexis, D., Brister, J. R., Bryant, S. H., Canese, K., Charowhas, C., Clark, K., DiCuccio, M., Dondoshansky, I., Feolo, M., Funk, K., Geer, L. Y., Gorelenkov, V., Hlavina, W., … Zbicz, K. (2017b). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *45*(D1), D12–D17. https://doi.org/10.1093/nar/gkw1071

Aguirre, A. A., Vicente, A. M., Hardwick, S. W., Alvelos, D. M., Mazzon, R. R., Luisi, B. F., & Marques, M. V. (2017). Association of the cold shock DEAD-box RNA helicase RhlE to the

RNA degradosome in *Caulobacter crescentus*. *Journal of Bacteriology*, *199*(13), 1–13. https://doi.org/10.1128/JB.00135-17

Ahlgren, N. A., & Rocap, G. (2006). Culture isolation and culture-independent clone libraries reveal new marine *Synechococcus* ecotypes with distinctive light and N physiologies. *Applied and Environmental Microbiology*, *72*(11), 7193–7204. https://doi.org/10.1128/AEM.00358-06

Ahlgren, N. A., & Rocap, G. (2012). Diversity and distribution of marine *Synechococcus*: Multiple gene phylogenies for consensus classification and development of qPCR assays for sensitive measurement of clades in the ocean. *Frontiers in Microbiology*, *3*(JUN), 1–24. https://doi.org/10.3389/fmicb.2012.00213

Arbing, M. A., Handelman, S. K., Kuzin, A. P., Verdon, G., Wang, C., Su, M., Rothenbacher, F. P., Abashidze, M., Liu, M., Hurley, J. M., Xiao, R., Action, T., Inouye, M., Montelione, G. T., Woychik, N. A., & Hunt, J. F. (2010). Crystal structures of Phd-Doc , HigA , and YeeU establish multiple evolutionary links between microbial growth-regulating toxin-antitoxin systems. *Structure/Folding and Design*, *18*(8), 996–1010. https://doi.org/10.1016/j.str.2010.04.018

Arcus, V. L., Bäckbro, K., Roos, A., Daniel, E. L., & Baker, E. N. (2004). Distant structural homology leads to the functional characterization of an Archaeal PIN domain as an exonuclease. *Journal of Biological Chemistry*, *279*(16), 16471–16478. https://doi.org/10.1074/jbc.M313833200

Arcus, V. L., Mckenzie, J. L., Robson, J., & Cook, G. M. (2011). The PIN-domain ribonucleases and

the prokaryotic VapBC toxin-antitoxin array. *Protein Engineering, Design and Selection*, *24*(1–2), 33–40. https://doi.org/10.1093/protein/gzq081

Aziz, R. K., Bartels, D., Best, A. A., Dejongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., Mcneil, L. K., Paarmann, D., Paczian, T., Parrello, B., … Zagnitko, O. (2008). The RAST server : Rapid annotations using subsystems technology. *BMC Genomics*, *9:75*(75). https://doi.org/10.1186/1471-2164-9-75

Barria, C., Malecki, M., & Arraiano, C. M. (2013). Bacterial adaptation to cold. *Microbiology (United Kingdom)*, *159*(PART 12), 2437–2443. https://doi.org/10.1099/mic.0.052209-0

Bertelli, C., Laird, M. R., Williams, K. P., Lau, B. Y., Hoad, G., Winsor, G. L., & Brinkman, F. S. L. (2017). IslandViewer 4: Expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*, *45*(W1), W30–W35. https://doi.org/10.1093/nar/gkx343

Biller, S. J., Berube, P. M., Berta-Thompson, J. W., Kelly, L., Roggensack, S. E., Awad, L., Roache-Johnson, K. H., Ding, H., Giovannoni, S. J., Rocap, G., Moore, L. R., & Chisholm, S. W. (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data*, *1*, 1–11. https://doi.org/10.1038/sdata.2014.34

Biller, S. J., Berube, P. M., Lindell, D., & Chisholm, S. W. (2015). *Prochlorococcus*: The structure and function of collective diversity. *Nature Reviews Microbiology*, *13*(1), 13–27. https://doi.org/10.1038/nrmicro3378

Brennan, R. G., & Matthews, B. W. (1989). The helix-turn-helix DNA binding motif. *Journal of*

Biological Chemistry, *264*(4), 1903–1906.

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Iii, J. A. T., Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk : A modular and extensible implementation of the RAST algorithm for annotating batches of genomes. *Scientific Reports*, *5:8365*(8365), 1–6. https://doi.org/10.1038/srep08365

Brunet, C., & Lizon, F. (2003). Tidal and diel periodicities of size-fractionated phytoplankton pigment signatures at an offshore station in the southeastern English Channel. *Estuarine, Coastal and Shelf Science (Print)*, *56*(3–4), 833–843.

Buck, K. R., Chavez, F. P., & Campbell, L. (1996). Basin-wide distributions of living carbon components and the inverted trophic pyramid of the central gyre of the North Atlantic Ocean, summer 1993. *Aquatic Microbial Ecology*, *10*(3), 283–298. https://doi.org/10.3354/ame010283

Bunker, R. D., McKenzie, J. L., Baker, E. N., & Arcus, V. L. (2008). Crystal structure of PAEO151 from *Pyrobaculum aerophilum*, a PIN-domain (VapC) protein from a toxin-antitoxin operon. *Proteins: Structure, Function and Genetics*, *72*(1), 510–518. https://doi.org/10.1002/prot.22048

Cai, H., Wang, K., Huang, S., Jiao, N., & Chen, F. (2010). Distinct patterns of picocyanobacterial communities in winter and summer in the Chesapeake Bay. *Applied and Environmental Microbiology*, *76*(9), 2955–2960. https://doi.org/10.1128/AEM.02868-09

Callieri, C. (2008). Picophytoplankton in freshwater ecosystems: The importance of small-sized phototrophs. *Freshwater Reviews*, *1*(1), 1–28. https://doi.org/10.1608/frj-1.1.1

Callieri, C. (2017). *Synechococcus* plasticity under environmental changes. *FEMS Microbiology Letters*, *364*(23), 1–8. https://doi.org/10.1093/femsle/fnx229

Callieri, C., Coci, M., Corno, G., Macek, M., Modenutti, B., Balseiro, E., & Bertoni, R. (2013). Phylogenetic diversity of nonmarine picocyanobacteria. *FEMS Microbiology Ecology*, *85*(2), 293–301. https://doi.org/10.1111/1574-6941.12118

Callieri, C., & Stockner, J. G. (2002). Freshwater autotrophic picoplankton: A review. *Journal of Limnology*, *61*(1), 1–14. https://doi.org/10.4081/jlimnol.2002.1

Campbell, L., & Carpenter, E. J. (1987). Characterization of phycoerythrin-containing *Synechococcus* spp. populations by immunofluorescence. *Journal of Plankton Research*, *9*(6), 1167–1181.

Carpenter, E. J., Campbell, L., Carpenter, E. J., & Campbell, L. (1988). *Diel patterns of cell division and growth rates of Synechococcus spp. in Long Island Sound*. *47*(2), 179–183.

Carty, S. M., Sreekumar, K. R., & Raetz, C. R. H. (1999). Effect of cold shock on lipid a biosynthesis in *Escherichia coli*: Induction at 12 °C of an acyltransferase specific for palmitoleoyl-acyl carrier protein. *Journal of Biological Chemistry*, *274*(14), 9677–9685. https://doi.org/10.1074/jbc.274.14.9677

Chamot, D., Magee, W. C., Yu, E., & Owttrim, G. W. (1999). A cold shock-induced cyanobacterial RNA helicase. *Journal of Bacteriology*, *181*(6), 1728–1732.

https://doi.org/10.1128/jb.181.6.1728-1732.1999

Chandra, B., Ramisetty, M., & Santhosh, R. S. (2016). *Horizontal gene transfer of chromosomal Type II toxin – antitoxin systems of Escherichia coli*. *October 2015*, 1–7. https://doi.org/10.1093/femsle/fnv238

Chen, F., Wang, K., Kan, J., Bachoon, D. S., Lu, J., Lau, S., & Campbell, L. (2004). Phylogenetic diversity of *Synechococcus* in the Chesapeake Bay revealed by Ribulose-1,5-bisphosphate carboxylase-oxygenase (RuBisCO) large subunit gene (rbcL) sequences. *Aquatic Microbial Ecology*, *36*, 153–164. https://doi.org/10.3354/ame036153

Chen, F., Wang, K., Kan, J., Suzuki, M. T., & Wommack, K. E. (2006a). Diverse and unique picocyanobacteria in Chesapeake Bay, revealed by 16S-23S rRNA internal transcribed spacer sequences. *Applied and Environmental Microbiology*, *72*(3), 2239–2243. https://doi.org/10.1128/AEM.72.3.2239-2243.2006

Chen, F., Wang, K., Kan, J., Suzuki, M. T., & Wommack, K. E. (2006b). Diverse and unique picocyanobacteria in Chesapeake Bay, revealed by 16S-23S rRNA internal transcribed spacer sequences. *Applied and Environmental Microbiology*, *72*(3), 2239–2243. https://doi.org/10.1128/AEM.72.3.2239-2243.2006

Chen, Y., Holtman, C. K., Magnuseon, R. D., Youderian, P. A., & Golden, S. S. (2011). The complete sequence and functional analysis of pANL, the large plasmid of the unicellular freshwater cyanobacterium *Synechococcus* elongatus PCC 7942. *Plasmid*, *23*(1), 1–7. https://doi.org/10.1038/jid.2014.371

Chi, X., Qingli, Y., Fangqing, Z., Song, Q., Yu, Y., Junjun, S., & Hanzhi, L. (2008). Comparative analysis of fatty acid desaturases in cyanobacterial genomes. *Comparative and Functional Genomics*, *2008*. https://doi.org/10.1155/2008/284508

Choi, D. H., & Noh, J. H. (2009). Phylogenetic diversity of *Synechococcus* strains isolated from the East China Sea and the East Sea. *FEMS Microbiology Ecology*, *69*(3), 439–448. https://doi.org/10.1111/j.1574-6941.2009.00729.x

Cloern, J. E., Foster, S. Q., & Kleckner, A. E. (2014). Phytoplankton primary production in the world's estuarine-coastal ecosystems. *Biogeosciences*, *11*(9), 2477–2501. https://doi.org/10.5194/bg-11-2477-2014

Cloern, J E, Foster, S. Q., & Kleckner, A. E. (2016). Phytoplankton primary production in the world's estuarine-coastal ecosystems. *Biogeosciences*, *11*(May 2014), 2477–2501. https://doi.org/10.5194/bg-11-2477-2014

Cloern, James E., & Dufford, R. (2005). Phytoplankton community ecology: Principles applied in San Francisco Bay. *Marine Ecology Progress Series*, *285*(Cembella 2003), 11–28. https://doi.org/10.3354/meps285011

Cloern, James E., & Jassby, A. D. (2010). Patterns and scales of phytoplankton variability in estuarine-coastal ecosystems. *Estuaries Coast.*, *33*(2), 230–241. https://doi.org/10.1007/s12237-009-9195-3

Coleman, M. L., & Chisholm, S. W. (2007). Code and context : *Prochlorococcus* as a model for cross-scale biology. *Trends in Microbiology*, *15*(9).

https://doi.org/10.1016/j.tim.2007.07.001

Core, T. R. (2017). *R: A language and environment for statistical computing* (3.6.2). https://www.r-project.org

Coutinho, F., Tschoeke, D. A., Thompson, F., & Thompson, C. (2016). Comparative genomics of *Synechococcus* and proposal of the new genus *Parasynechococcus*. *PeerJ*, *4*, e1522. https://doi.org/10.7717/peerj.1522

Crosbie, N. D., Po, M., & Weisse, T. (2003). Dispersal and phylogenetic diversity of nonmarine picocyanobacteria, inferred from 16S rRNA gene and cpcBA-intergenic spacer sequence analyses. *Society*, *69*(9), 5716–5721. https://doi.org/10.1128/AEM.69.9.5716

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., Chlenski, P., Conrad, N., Dickerman, A., Dietrich, E. M., Gabbard, J. L., Gerdes, S., Guard, A., Kenyon, R. W., MacHi, D., Mao, C., Murphy-Olson, D., Nguyen, M., Nordberg, E. K., … Stevens, R. (2020). The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Research*, *48*(D1), D606–D612. https://doi.org/10.1093/nar/gkz943

Di Cesare, A., Cabello-Yeves, P. J., Chrismas, N. A. M., Sánchez-Baracaldo, P., Salcher, M. M., & Callieri, C. (2018). Genome analysis of the freshwater planktonic *Vulcanococcus limneticus* sp. nov. reveals horizontal transfer of nitrogenase operon and alternative pathways of nitrogen utilization. *BMC Genomics*, *19*(1), 1–12. https://doi.org/10.1186/s12864-018-4648-3

Diamant, S., & Goloubinoff, P. (1998). Temperature-controlled activity of DnaK-DnaJ-GrpE

chaperones: Protein- folding arrest and recovery during and after heat shock depends on the substrate protein and the GrpE concentration. *Biochemistry*, *37*(27), 9688–9694. https://doi.org/10.1021/bi980338u

Dufresne, A., Garczarek, L., & Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology*, *6*(2).

Dufresne, A., Ostrowski, M., Scanlan, D. J., Garczarek, L., Mazard, S., Palenik, B. P., Paulsen, I. T., de Marsac, N. T., Wincker, P., Dossat, C., Ferriera, S., Johnson, J., Post, A. F., Hess, W. R., & Partensky, F. (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biology*, *9*(5). https://doi.org/10.1186/gb-2008-9-5-r90

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., Makarova, K. S., Ostrowski, M., Oztas, S., Robert, C., Rogozin, I. B., Scanlan, D. J., De Marsac, N. T., Weissenbach, J., Wincker, P., … Hess, W. R. (2003). Genome sequence of the cyanobacterium *Prochlorococcus* marinus SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(17), 10020–10025. https://doi.org/10.1073/pnas.1733211100

Dvořák, P., Casamatta, D. A., Poulíčková, A., Hašler, P., Ondřej, V., & Sanges, R. (2014a). *Synechococcus*: 3 billion years of global dominance. *Molecular Ecology*. https://doi.org/10.1111/mec.12948

Dvořák, P., Casamatta, D. A., Poulíčková, A., Hašler, P., Ondřej, V., & Sanges, R. (2014b). *Synechococcus*: 3 billion years of global dominance. *Molecular Ecology*, *23*(22), 5538–5551.

https://doi.org/10.1111/mec.12948

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Ernst, A., Becker, S., Wollenzien, U. I. A., & Postius, C. (2003a). Ecosystem-dependent adaptive radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis. *Microbiol.*, *149*(1), 217–228. https://doi.org/10.1099/mic.0.25475-0

Ernst, A., Becker, S., Wollenzien, U. I. A., & Postius, C. (2003b). Ecosystem-dependent adaptive radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis. *Microbiol.*, *149*(1), 217–228. https://doi.org/10.1099/mic.0.25475-0

Fasani, R. A., & Savageau, M. A. (2015). Unrelated toxin-antitoxin systems cooperate to induce persistence. *Journal of the Royal Society Interface*, *12*(108). https://doi.org/10.1098/rsif.2015.0130

Fernandez-Garcia, L., Kim, J. S., Tomas, M., & Wood, T. K. (2019). Toxins of toxin/antitoxin systems are inactivated primarily through promoter mutations. *Journal of Applied Microbiology*, *127*(6), 1859–1868. https://doi.org/10.1111/jam.14414

Fierer, N., Bradford, M. A., & Jackson, R. B. (2007). Toward an ecological classification of soil bacteria. *Ecology*, *88*(6), 1354–1364.

Flombaum, P., Gallegos, J. L., Gordillo, R. a, Rincón, J., Zabala, L. L., Jiao, N., Karl, D., Li, W.,

Lomas, M., Veneziano, D., Vera, C., Vrugt, J. a, & Martiny, a C. (2013). Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *PNAS*, *110*(24), 9824–9829. https://doi.org/10.1073/pnas.1307701110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1307701110

Fortunato, C. S., & Crump, B. C. (2011). Bacterioplankton community variation across river to ocean environmental gradients. *Microbial Ecology*, *62*(2), 374–382. https://doi.org/10.1007/s00248-011-9805-z

Fucich, D., & Chen, F. (2020). Presence of toxin-antitoxin systems in picocyanobacteria and their ecological implications. *ISME Journal*, 17–19. https://doi.org/10.1038/s41396-020-00746-4

Fucich, D., Marsan, D., Sosa, A., & Chen, F. (2019). Complete genome sequence of subcluster 5.2 *Synechococcus* sp. strain CB0101, isolated from the Chesapeake Bay. *Microbiology Resource Announcements*, *August*, 6–8.

Fuller, N. J., Marie, D., Vaulot, D., Post, A. F., & Scanlan, D. J. (2003). Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Applied and Environmental Microbiology*, *69*(5), 2430–2443. https://doi.org/10.1128/AEM.69.5.2430

Garcia-Pichel, F., Belnap, J., Neuer, S., & Schanz, F. (2009). Estimates of global cyanobacterial biomass and its distribution. *Algological Studies*, *109*(1), 213–227. https://doi.org/10.1127/1864-1318/2003/0109-0213

Goclaw-Binder, H., Sendersky, E., Shimoni, E., Kiss, V., Reich, Z., Perelman, A., & Schwarz, R.

(2012). Nutrient-associated elongation and asymmetric division of the cyanobacterium *Synechococcus* PCC 7942. *Environmental Microbiology*, *14*(3), 680–690. https://doi.org/doi:10.1111/j.1462-2920.2011.02620.x

Goericke, R., & Repeta, D. J. (1992). The pigments of *Prochlorococcus* marinus: The presence of divinylchlorophyll a and b in a marine procaryote. *Limnology and Oceanography*, *37*(2), 425–433. https://doi.org/10.4319/lo.1992.37.2.0425

Grynberg, M., Erlandsen, H., & Godzik, A. (2003). HEPN: A common domain in bacterial drug resistance and human neurodegenerative proteins. *Trends in Biochemical Sciences*, *28*(5), 224–226. https://doi.org/10.1016/S0968-0004(03)00063-X

Gualerzi, C. O., Giuliodori, A. M., & Pon, C. L. (2003). Transcriptional and post-transcriptional control of cold-shock genes. *Journal of Molecular Biology*, *331*(3), 527–539. https://doi.org/10.1016/S0022-2836(03)00732-0

Guglielmini, J., & Van Melderen, L. (2011). Bacterial toxin-antitoxin systems: Translation inhibitors everywhere. *Mobile Genetic Elements*, *1*(4), 283–290.

Guyet, U., Nguyen, N. A., Doré, H., Haguait, J., Pittera, J., Conan, M., Ratin, M., Corre, E., Le Corguillé, G., Brillet-Guéguen, L., Hoebeke, M., Six, C., Steglich, C., Siegel, A., Eveillard, D., Partensky, F., & Garczarek, L. (2020). Synergic effects of temperature and irradiance on the physiology of the marine *Synechococcus* strain WH7803. *Frontiers in Microbiology*, *11*(July), 1–22. https://doi.org/10.3389/fmicb.2020.01707

Habib, G., Zhu, Q., & Sun, B. (2018). Bioinformatics and functional assessment of toxin-antitoxin

systems in *Staphylococcus aureus*. *Toxins*, *10*(437). https://doi.org/10.3390/toxins10110473

Harding, L. W., Mallonee, M. E., Perry, E. S., Miller, W. D., Adolf, J. E., Gallegos, C. L., & Paerl, H. W. (2016). Variable climatic conditions dominate recent phytoplankton dynamics in Chesapeake Bay. *Scientific Reports*, *6*, 1–16. https://doi.org/10.1038/srep23773

Harms, A., Brodersen, D. E., Mitarai, N., & Gerdes, K. (2018a). Toxins, targets, and triggers: An overview of toxin-antitoxin biology. *Molecular Cell*, *70*(June 7), 1–17. https://doi.org/10.1016/j.molcel.2018.01.003

Harms, A., Brodersen, D. E., Mitarai, N., & Gerdes, K. (2018b). Toxins, targets, and triggers: An overview of toxin-antitoxin biology. *Molecular Cell*, *70*(5), 768–784. https://doi.org/10.1016/j.molcel.2018.01.003

Hassan, N., Anesio, A. M., Rafiq, M., Holtvoeth, J., Bull, I., Williamson, C. J., & Hasan, F. (2020). Cell membrane fatty acid and pigment composition of the psychrotolerant cyanobacterium *Nodularia spumigena* CHS1 isolated from Hopar glacier, Pakistan. *Extremophiles*, *24*(1), 135–145. https://doi.org/10.1007/s00792-019-01141-4

Haverkamp, T., Acinas, S. G., Doeleman, M., Stomp, M., Huisman, J., & Stal, L. J. (2008). Diversity and phylogeny of Baltic Sea picocyanobacteria inferred from their ITS and phycobiliprotein operons. *Environmental Microbiology*, *10*(1), 174–188. https://doi.org/10.1111/j.1462-2920.2007.01442.x

Haverkamp, T. H. A., Schouten, D., Doeleman, M., Wollenzien, U., Huisman, J., & Stal, L. J.

(2009). Colorful microdiversity of *Synechococcus* strains (picocyanobacteria) isolated from the Baltic Sea. *ISME Journal*, *3*(4), 397–408. https://doi.org/10.1038/ismej.2008.118

Herbert, R. A. (1999). Nitrogen cycling in coastal marine ecosystems. *FEMS Microbiology Reviews*, *23*(5), 563–590. https://doi.org/10.1016/S0168-6445(99)00022-4

Herdman, M., Castenholz, R. W., Iteman, I., Waterbury, J. B., & Rippka, R. (2001). The Archaea and the deeply branching and phototrophic bacteria. In D. Boone & R. W. Castenholz (Eds.), *Bergey's Manual of Systematic Bacteriology,* (2nd ed., pp. 493–514). Springer Verlag.

Hess, W. R. (2004). Genome analysis of marine photosynthetic microbes and their global role. *Current Opinion in Biotechnology*, *15*, 191–198. https://doi.org/10.1016/j.copbio.2004.03.007

Holt, J., Krieg, N., Sneath, P., Staley, J., & Williams, S. (1994). Bergey's manual of determinative microbiology, 9th edn. *Lippincot, Williams and Wilkins, Baltimore*, 1710–1728.

Honda, D., Yokota, A., & Sugiyama, J. (1999). Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine *Synechococcus* strains. *Journal of Molecular Evolution*, *48*(6), 723–739. https://doi.org/10.1007/PL00006517

Huang, S., Wilhelm, S. W., Harvey, H. R., Taylor, K., Jiao, N., & Chen, F. (2011). Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. *The ISME Journal*, *6*(2), 285–297. https://doi.org/10.1038/ismej.2011.106

Hunter-Cevera, K. R., Post, A. F., Peacock, E. E., & Sosik, H. M. (2016). Diversity of *Synechococcus* at the Martha's Vineyard Coastal Observatory: Insights from culture isolations, clone libraries, and flow cytometry. *Microbial Ecology*, *71*(2), 276–289. https://doi.org/10.1007/s00248-015-0644-1

Jakobsen, F. (1996). The dense water exchange of the bornholm basin in the Baltic Sea. *Deutsche Hydrografische Zeitschrift*, *48*(2), 133–145. https://doi.org/10.1007/BF02799383

Jardillier, L., Zubkov, M. V., Pearman, J., & Scanlan, D. J. (2010). Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME Journal*, *4*(9), 1180–1192. https://doi.org/10.1038/ismej.2010.36

Jasser, I., Królicka, A., & Karnkowska-Ishikawa, A. (2011). A novel phylogenetic clade of picocyanobacteria from the Mazurian lakes (Poland) reflects the early ontogeny of glacial lakes. *FEMS Microbiology Ecology*, *75*(1), 89–98. https://doi.org/10.1111/j.1574-6941.2010.00990.x

Jezberová, J., & Komárková, J. (2007). Morphometry and growth of three *Synechococcus*-like picoplanktic cyanobacteria at different culture conditions. *Hydrobiologia*, *578*(1), 17–27. https://doi.org/10.1007/s10750-006-0429-0

Jiang, W., Hou, Y., & Inouye, M. (1997). CspA, the major cold-shock protein of *Escherichia coli*, is an RNA chaperone. *Journal of Biological Chemistry*, *272*(1), 196–202. https://doi.org/10.1074/jbc.272.1.196

Jing, H., Zhang, R., Pointing, S. B., Liu, H., & Qian, P. (2009). Genetic diversity and temporal

variation of the marine *Synechococcus* community in the subtropical coastal waters of Hong Kong. *Canadian Journal of Microbiology*, *55*(3), 311–318. https://doi.org/10.1139/W08-138

Jochem, F. (1988). On the distribution and importance of picocyanobacteria in a boreal inshore area (Kiel Bight , Western Baltic). *Journal of Plankton Research*, *10*(5), 1009–1022. https://doi.org/10.1093/plankt/10.5.1009

Johnson, P. W., & Sieburth, J. M. N. (1979). Chroococcoid cyanobacteria in the sea: A ubiquitous and diverse phototrophic biomass. *Limnology and Oceanography*, *24*(5), 928–935. https://doi.org/10.4319/lo.1979.24.5.0928

Johnson, Z. I., Zinser, E. R., Coe, A., Mcnulty, N. P., Malcolm, E. S., Chisholm, S. W., Woodward, E. M. S., & Chisholm, S. W. (2006). Niche Partitioning Among *Prochlorococcus* Ecotypes Along Ocean-Scale Environmental Gradients. *Science*, *311*(5768), 1737–1740. https://doi.org/10.1126/science.1118052

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., & Chisholm, S. W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science*, *311*(5768), 1737–1740. https://doi.org/10.1126/science.1118052

Kamada, K., & Hanaoka, F. (2005). Conformational change in the catalytic site of the ribonuclease YoeB toxin by YefM antitoxin. *Molecular Cell*, *19*, 497–509. https://doi.org/10.1016/j.molcel.2005.07.004

Kan, J., Crump, B. C., Wang, K., & Chen, F. (2006). Bacterioplankton community in Chesapeake Bay: Predictable or random assemblages. *Limnology and Oceanography*, *51*(5), 2157–2169. https://doi.org/10.4319/lo.2006.51.5.2157

Kandror, O., DeLeon, A., & Goldberg, A. L. (2002). Trehalose synthesis is induced upon exposure of *Escherichia coli* to cold and is essential for viability at low temperatures. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(15), 9727–9732. https://doi.org/10.1073/pnas.142314099

Kaneko, T., Nakamura, Y., Sasamoto, S., Watanabe, A., Kohara, M., Matsumoto, M., Shimpo, S., Yamada, M., & Tabata, S. (2003). Structural analysis of four large plasmids harboring in a unicellular cyanobacterium, *Synechocystis* sp. PCC 6803. *DNA Research*, *10*(5), 221–228. https://doi.org/10.1093/dnares/10.5.221

Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., & Chisholm, S. W. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genetics*, *3*(12), 2515–2528. https://doi.org/10.1371/journal.pgen.0030231

Kishor PB, K. (2019). Bacterial cold shock proteins - the molecular chaperones for multiple stress tolerance. *Advances in Biotechnology & Microbiology*, *12*(3). https://doi.org/10.19080/aibm.2019.12.555837

Kitagawa, M., Ara, T., Arifuzzaman, M., Ioka-Nakamichi, T., Inamoto, E., Toyonaga, H., & Mori, H. (2005). Complete set of ORF clones of *Escherichia coli* ASKA library (A complete set of E.

coli K-12 ORF archive): unique resources for biological research. *DNA Research*, *12*(5), 291–299. https://doi.org/10.1093/dnares/dsi012

Kopfmann, S., Roesch, S., & Hess, W. (2016). Type II toxin–antitoxin systems in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803. *Toxins*, *8*(7), 228. https://doi.org/10.3390/toxins8070228

Larsson, J., Celepli, N., Ininbergs, K., Dupont, C. L., Yooseph, S., Bergman, B., & Ekman, M. (2014). Picocyanobacteria containing a novel pigment gene cluster dominate the brackish water Baltic Sea. *ISME Journal*, *8*(9), 1892–1903. https://doi.org/10.1038/ismej.2014.35

Larsson, J., Nylander, J. A. A., & Bergman, B. (2011). Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evolutionary Biology*, *11*(187). https://doi.org/10.1186/1471-2148-11-187

Lehmann, C., Lim, K., Chalamasetty, V. R., Krajewski, W., Melamud, E., Galkin, A., Howard, A., Kelman, Z., Reddy, P. T., Murzin, A. G., & Herzberg, O. (2003). The HI0073/HI0074 protein pair from *Haemophilus influenzae* is a member of a new nucleotidyltransferase family: Structure, sequence analyses, and solution studies. *Proteins: Structure, Function and Genetics*, *260*, 249–260.

Leplae, R., Geeraerts, D., Hallez, R., Guglielmini, J., Drze, P., & Van Melderen, L. (2011). Diversity of bacterial type II toxin-antitoxin systems: A comprehensive search and functional analysis of novel families. *Nucleic Acids Research*, *39*(13), 5513–5525. https://doi.org/10.1093/nar/gkr131

Li, W. K. W., & Url, S. (1994). Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnology and Oceanography*, *39*(1), 169–175.

Lima-Mendez, G., Oliveira Alvarenga, D., Ross, K., Hallet, B., Van Melderen, L., Varani, A. M., & Chandler, M. (2020). Toxin-antitoxin gene pairs found in Tn 3 family transposons appear to be an integral part of the transposition module. *MBio*, *11*(2). https://doi.org/10.1128/mBio.00452-20

Lindstrom, E. S., & Langenheder, S. (2012). Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, *4*(1), 1–9. https://doi.org/10.1111/j.1758-2229.2011.00257.x

López Alonso, D., García-Maroto, F., Rodríguez-Ruiz, J., Garrido, J. A., & Vilches, M. A. (2003). Evolution of the membrane-bound fatty acid desaturases. *Biochemical Systematics and Ecology*, *31*(10), 1111–1124. https://doi.org/10.1016/S0305-1978(03)00041-3

Los, D. A., & Murata, N. (1999). Responses to cold shock in cyanobacteria. *Journal of Molecular Microbiology and Biotechnology*, *1*(2), 221–230.

Los, D. A., Suzuki, I., Zinchenko, V. V, & Murata, N. (2008). Stress responses in *Synechocystis*: regulated genes and regulatory systems. *The Cyanobacteria. Molecular Biology, Genomics and Evolution*, *January*, 117–157.

Mackey, K. R. M., Hunter-Cevera, K., Britten, G. L., Murphy, L. G., Sogin, M. L., & Huber, J. A. (2017). Seasonal succession and spatial patterns of *Synechococcus* microdiversity in a salt

marsh estuary revealed through 16s rRNA gene oligotyping. *Frontiers in Microbiology*, *8*(AUG), 1–11. https://doi.org/10.3389/fmicb.2017.01496

Maisonneuve, E., & Gerdes, K. (2014). Molecular mechanisms underlying bacterial persisters. *Cell*, *157*(3), 539–548. https://doi.org/10.1016/j.cell.2014.02.050

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2009). Comprehensive comparative-genomic analysis of Type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biology Direct*, *4*(1), 19. https://doi.org/10.1186/1745-6150-4-19

Mann, S., & Chen, Y. P. P. (2010). Bacterial genomic G + C composition-eliciting environmental adaptation. *Genomics*, *95*(1), 7–15. https://doi.org/10.1016/j.ygeno.2009.09.002

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., … Bryant, S. H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, *45*(D1), D200–D203. https://doi.org/10.1093/nar/gkw1129

Marsan, D., Place, A., Fucich, D., & Chen, F. (2017). Toxin-antitoxin systems in estuarine *Synechococcus* strain CB0101 and their transcriptomic responses to environmental stressors. *Frontiers in Microbiology*, *8*(JUL), 1–11. https://doi.org/10.3389/fmicb.2017.01213

Marsan, D. W. (2016). *Adaptive mechanisms of an estuarine Synechococcus based on genomics, transcriptomics, and proteomics*. University of Maryland.

Marsan, D., Wommack, K. E., Ravel, J., & Chen, F. (2014). Draft genome sequence of *Synechococcus* sp. strain CB0101, isolated from the Chesapeake Bay estuary. *Genome*, *2*(1), 2011–2012. https://doi.org/10.1128/genomeA.01111-13.Copyright

Mascher, T., Helmann, J. D., & Unden, G. (2006). Stimulus Perception in Bacterial Signal-Transducing Histidine Kinases. *Microbiology and Molecular Biology Reviews*, *70*(4), 910–938. https://doi.org/10.1128/mmbr.00020-06

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P., & Lopez, R. (2013). Analysis tool web services from the EMBL-EBI. *Nucleic Acids Research*, *41*(Web Server issue), 8934. https://doi.org/10.1093/nar/gkt376

Mendez-perez, D., Herman, N. A., & Pfleger, F. (2014). A desaturase gene involved in the formation of 1,14-nonadecadiene in *Synechococcus* sp. strain PCC 7002. *Applied and Environmental Microbiology*, *80*(19), 6073–6079. https://doi.org/10.1128/AEM.01615-14

Mikami, K., Kanesaki, Y., Suzuki, I., & Murata, N. (2002). The histidine kinase Hik33 perceives osmotic stress and cold stress in *Synechocystis* sp. PCC 6803. *Molecular Microbiology*, *46*(4), 905–915. https://doi.org/10.1046/j.1365-2958.2002.03202.x

Mironov, K. S., Sidorov, R. A., Trofimova, M. S., Bedbenov, V. S., Tsydendambaev, V. D., Allakhverdiev, S. I., & Los, D. A. (2012). Light-dependent cold-induced fatty acid unsaturation, changes in membrane fluidity, and alterations in gene expression in *Synechocystis*. *Biochimica et Biophysica Acta - Bioenergetics*, *1817*(8), 1352–1359. https://doi.org/10.1016/j.bbabio.2011.12.011

Mittenhuber, G. (1999). Occurence of MazEF-like antitoxin/toxin systems in bacteria. *J. Mol. Microbiol. Biotechnol*, *1*(2), 295–302. www.caister.com/bacteria-plant

Moore, L. R., Coe, A., Zinser, E. R., Saito, M. A., Sullivan, M. B., Lindell, D., Frois-Moniz, K., Waterbury, J., & Chisholm, S. W. (2007). Culturing the marine cyanobacterium *Prochlorococcus*. *Limnology and Oceanography: Methods*, *5*(10), 353–362. https://doi.org/10.4319/lom.2007.5.353

Morel, A., Yu-Hwan Ahn, Partensky, F., Vaulot, D., & Claustre, H. (1993). *Prochlorococcus* and *Synechococcus*: a comparative study of their optical properties in relation to their size and pigmentation. *Journal of Marine Research*, *51*(3), 617–649. https://doi.org/10.1357/0022240933223963

Mou, X., Sun, S., Edwards, R. A., Hodson, R. E., & Moran, M. A. (2008). *Bacterial carbon processing by generalist species in the coastal ocean*. *451*(February). https://doi.org/10.1038/nature06513

Murata, N., & Wada, H. (1995). Acyl-lipid desaturases and their importance in the tolerance and acclimatization to cold of cyanobacteria. *Biochemical Journal*, *308*(1), 1–8. https://doi.org/10.1042/bj3080001

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, *42*(D1), 26–31. https://doi.org/10.1093/nar/gkt1069

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Olson, R. J., Chisholm, S. W., & Zettler, E. R. (1990). Pigments, size, and distribution of *Synechococcus* in the North Atlantic and Pacific Oceans. *Limnology and Oceanography*, *35*(l).

Olson, R. J., Chisholm, S. W., Zettler, E. R., Altabet, M. A., & Dusenberry, J. A. (1990). Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. *Deep Sea Research Part A. Oceanographic Research Papers*, *37*(6), 1033–1051.

Opiyo, S. O., Pardy, R. L., Moriyama, H., & Moriyama, E. N. (2010). Evolution of the Kdo2-lipid A biosynthesis in bacteria. *BMC Evolutionary Biology*, *10*(362). https://doi.org/10.1186/1471-2148-10-362

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, *42*(November 2013), 206–214. https://doi.org/10.1093/nar/gkt1226

Page, R., & Peti, W. (2016). Toxin-antitoxin systems in bacterial growth arrest and persistence.

Nature Chemical Biology, *12*(4), 208–214. https://doi.org/10.1038/nchembio.2044

Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J., Paulsen, I., Dufresne, A., Partensky, F., Webb, E. A., & Waterbury, J. (2003). The genome of a motile marine *Synechococcus*. *Nature*, *424*(6952), 1037–1042. https://doi.org/10.1038/nature01943

Palenik, B, Barahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., J., M., Paulsen, I., Dufresne, A., Partensky, F., Webb, E. A., & Waterbury, J. (2003). The genome of a motile marine *Synechococcus*. *Nature*, *424*(August), 1037–1042. https://doi.org/10.1038/nature01883.1.

Palenik, Brian, Ren, Q., Dupont, C. L., Myers, G. S., Heidelberg, J. F., Badger, J. H., Madupu, R., Nelson, W. C., Brinkac, L. M., Dodson, R. J., Durkin, A. S., Daugherty, S. C., Sullivan, S. A., Khouri, H., Mohamoud, Y., Halpin, R., & Paulsen, I. T. (2006). Genome sequence of *Synechococcus* CC9311 : Insights into adaptation to a coastal environment. *PNAS*, *103*(36), 13555–13559.

Pandey, D. P., & Gerdes, K. (2005). Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Research*, *33*(3), 966–976. https://doi.org/10.1093/nar/gki201

Pandit, S. N., Kolasa, J., & Cottenie, K. (2009). Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology*, *90*(8), 2253–2262.

Paoletti, L., Lu, Y. J., Schujman, G. E., De Mendoza, D., & Rock, C. O. (2007). Coupling of fatty acid and phospholipid synthesis in *Bacillus subtilis*. *Journal of Bacteriology*, *189*(16), 5816–5824. https://doi.org/10.1128/JB.00602-07

Partensky, F, Hess, W. R., & Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol.Mol Biol.Rev.*, *63*(1), 106–127. https://doi.org/doi:1092-2172/99/$04.00

Partensky, Frédéric, Blanchot, J., & Vaulot, D. (1999). Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters : a review. *Bulletin de l'Institut Océanographique*, *19*(19), 457–475.

Phadtare, S. (2004). Recent developments in bacterial cold-shock response. *Current Issues in Molecular Biology*, *6*(2), 125–136. https://doi.org/10.21775/cimb.006.125

Pichard, S. L., Campbell, L., & Paul, J. H. (1997). Diversity of the ribulose bisphosphate carboxylase/oxygenase form I gene (rbcL) in natural phytoplankton communities. *Applied and Environmental Microbiology*, *63*(9), 3600–3606. https://doi.org/10.1128/aem.63.9.3600-3606.1997

Ramage, H. R., Connolly, L. E., & Cox, J. S. (2009). Comprehensive functional analysis of *Mycobacterium tuberculosis* toxin-antitoxin systems: Implications for pathogenesis, stress responses, and evolution. *PLoS Genetics*, *5*(12). https://doi.org/10.1371/journal.pgen.1000767

Ray, R., Haas, L., & Sieracki, M. (1989). Autotrophic picoplankton dynamics in a Chesapeake Bay

sub-estuary. *Marine Ecology Progress Series*, *52*, 273–285. https://doi.org/10.3354/meps052273

Rippka, R., Deruelles, J., & Waterbury, J. B. (1979). Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *Journal of General Microbiology*, *111*(1), 1–61. https://doi.org/10.1099/00221287-111-1-1

Robertson, B. R., Tezuka, N., & Watanabe, M. M. (2001). Phylogenetic analyses of *Synechococcus* strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. *International Journal of Systematic and Evolutionary Microbiology*, *51*(3), 861–871. https://doi.org/10.1099/00207713-51-3-861

Robson, J., McKenzie, J. L., Cursons, R., Cook, G. M., & Arcus, V. L. (2009). The vapBC operon from *Mycobacterium smegmatis* is an autoregulated toxin-antitoxin module that controls growth via inhibition of translation. *Journal of Molecular Biology*, *390*(3), 353–367. https://doi.org/10.1016/j.jmb.2009.05.006

Rocap, G., Distel, D. L., Waterbury, J. B., & Chisholm, S. W. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA Internal Transcribed Spacer sequences. *Applied and Environmental Microbiology*, *68*(3), 1180–1191. https://doi.org/10.1128/AEM.68.3.1180

Rocker, A., & Meinhart, A. (2016). Type II toxin: antitoxin systems. More than small selfish entities? *Current Genetics*, *62*(2), 287–290. https://doi.org/10.1007/s00294-015-0541-7

Sakamoto, T., & Murata, N. (2002). Regulation of the desaturation of fatty acids and its role in tolerance to cold and salt stress. *Current Opinion in Microbiology*, *5*(2), 206–210. https://doi.org/10.1016/S1369-5274(02)00306-5

Salazar, V. W., Tschoeke, D. A., Swings, J., Cosenza, C. A., Mattoso, M., Thompson, C. C., & Thompson, F. L. (2020). A new genomic taxonomy system for the *Synechococcus* collective . *Environmental Microbiology*, *00*, 1–14. https://doi.org/10.1111/1462-2920.15173

Sánchez-Baracaldo, P., Hayes, P. K., & Blank, C. E. (2005). Morphological and habitat evolution in the cyanobacteria using a compartmentalization approach. *Geobiology*, *3*, 145–165. https://doi.org/10.1111/j.1472-4669.2005.00050.x

Sánchez-Baracaldo, Patricia, Bianchini, G., Di Cesare, A., Callieri, C., & Chrismas, N. A. M. (2019). Insights into the evolution of picocyanobacteria and phycoerythrin genes (mpeBA and cpeBA). *Frontiers in Microbiology*, *10*(JAN), 1–17. https://doi.org/10.3389/fmicb.2019.00045

Scanlan, D J, Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W. R., Post, A. F., Hagemann, M., Paulsen, I., & Partensky, F. (2009). Ecological genomics of marine picocyanobacteria. *Microbiology and Molecular Biology Reviews*, *73*(2), 249–299. https://doi.org/10.1128/MMBR.00035-08

Scanlan, David J. (2012). Marine Picocyanobacteria. In *Ecology of Cyanobacteria II: Their Diversity in Space and Time* (Vol. 9789400738, pp. 503–533). https://doi.org/10.1007/978-94-007-3855-3

Scanlan, David J., & West, N. J. (2002). Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiology Ecology*, *40*(1), 1–12. https://doi.org/10.1016/S0168-6496(01)00217-3

Seel, W., Baust, D., Sons, D., Albers, M., Etzbach, L., Fuss, J., & Lipski, A. (2020). Carotenoids are used as regulators for membrane fluidity by *Staphylococcus xylosus*. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-019-57006-5

Sevin, E. W., & Barloy-Hubler, F. (2007). RASTA-Bacteria: A web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biology*, *8*(8). https://doi.org/10.1186/gb-2007-8-8-r155

Shalapyonok, A., Olson, R. J., & Shalapyonok, L. S. (2001). Arabian Sea phytoplankton during Southwest and Northeast Monsoons 1995: Composition, size structure and biomass from individual cell properties measured by flow cytometry. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *48*(6–7), 1231–1261. https://doi.org/10.1016/S0967-0645(00)00137-5

Shao, Y., Harrison, E. M., Bi, D., Tai, C., He, X., Ou, H. Y., Rajakumar, K., & Deng, Z. (2011). TADB: A web-based resource for Type 2 toxin-antitoxin loci in bacteria and archaea. *Nucleic Acids Research*, *39*(SUPPL. 1), 606–611. https://doi.org/10.1093/nar/gkq908

Sinetova, M. A., & Los, D. A. (2016). New insights in cyanobacterial cold stress responses: Genes, sensors, and molecular triggers. *Biochimica et Biophysica Acta - General Subjects*, *1860*(11), 2391–2403. https://doi.org/10.1016/j.bbagen.2016.07.006

Six, C., Thomas, J. C., Garczarek, L., Ostrowski, M., Dufresne, A., Blot, N., Scanlan, D. J., & Partensky, F. (2007). Diversity and evolution of phycobilisomes in marine *Synechococcus* spp.: A comparative genomics study. *Genome Biology*, *8*(12). https://doi.org/10.1186/gb-2007-8-12-r259

Stockner, J. G. (1988). Phototrophic picoplankton: An overview from marine and freshwater ecosystems. *Limnology and Oceanography*, *33*(4part2), 765–775. https://doi.org/10.4319/lo.1988.33.4part2.0765

Strocchi, M., Ferrer, M., Timmis, K. N., & Golyshin, P. N. (2006). Low temperature-induced systems failure in *Escherichia coli*: Insights from rescue by cold-adapted chaperones. *Proteomics*, *6*(1), 193–206. https://doi.org/10.1002/pmic.200500031

Stuart, R. K., Brahamsha, B., Busby, K., & Palenik, B. (2013). Genomic island genes in a coastal marine *Synechococcus* strain confer enhanced tolerance to copper and oxidative stress. *The ISME Journal*, *7*(6), 1139–1149. https://doi.org/10.1038/ismej.2012.175

Stuart, R. K., Dupont, C. L., Johnson, D. A., Paulsen, I. T., & Palenik, B. (2009). Coastal strains of marine *Synechococcus* species exhibit increased tolerance to copper shock and a distinctive transcriptional response relative to those of open-ocean strains. *Applied and Environmental Microbiology*, *75*(15), 5047–5057. https://doi.org/10.1128/AEM.00271-09

Sun, Z., & Blanchard, J. L. (2014). Strong genome-wide selection early in the evolution of *Prochlorococcus* resulted in a reduced genome through the loss of a large number of small effect genes. *PLoS ONE*, *9*(3), e88837. https://doi.org/10.1371/journal.pone.0088837

Suzuki, I., Kanesaki, Y., Mikami, K., Kanehisa, M., & Murata, N. (2001). Cold-regulated genes under control of the cold sensor Hik33 in *Synechocystis*. *Molecular Microbiology*, *40*(1), 235–244. https://doi.org/10.1046/j.1365-2958.2001.02379.x

Tang, J., Du, L. M., Liang, Y. M., & Daroch, M. (2019). Complete genome sequence and comparative analysis of *Synechococcus* sp. CS-601 (SynAce01), a cold-adapted cyanobacterium from an oligotrophic antarctic habitat. *International Journal of Molecular Sciences*, *20*(1), 1–17. https://doi.org/10.3390/ijms20010152

Toledo, G., & Palenik, B. (1997). *Synechococcus* diversity in the California Current as seen by RNA polymerase (rpoC1) gene sequences of isolated strains. *Applied and Environmental Microbiology*, *63*(11), 4298–4303.

Unterholzner, S. J., Poppenberger, B., & Rozhon, W. (2013). Toxin-antitoxin systems. *Bioengineered*, *5*(3), 1–13. https://doi.org/10.4161/mge.26219

Unterholzner, S. J., Poppenberger, B., & Rozhon, W. (2014a). Toxin-antitoxin systems. *Bioengineered*, *5*(3). https://doi.org/10.4161/mge.26219

Unterholzner, S. J., Poppenberger, B., & Rozhon, W. (2014b). Toxin-antitoxin systems. *Bioengineered*, *5*(3), 1–13. https://doi.org/10.4161/mge.26219

Van Melderen, L. (2010). Toxin-antitoxin systems: Why so many, what for? *Current Opinion in Microbiology*, *13*(6), 781–785. https://doi.org/10.1016/j.mib.2010.10.006

Vigil-Stenman, T., Ininbergs, K., Bergman, B., & Ekman, M. (2017). High abundance and expression of transposases in bacteria from the Baltic Sea. *ISME Journal*, *11*(11), 2611–

2623. https://doi.org/10.1038/ismej.2017.114

Wang, K., Wommack, K. E., & Chen, F. (2011). Abundance and distribution of *Synechococcus* spp. and cyanophages in the Chesapeake Bay. *Applied and Environmental Microbiology*, *77*(21), 7459–7468. https://doi.org/10.1128/AEM.00267-11

Wang, L., Chen, L., Yang, S., & Tan, X. (2020). Photosynthetic conversion of carbon dioxide to oleochemicals by cyanobacteria: Recent advances and future perspectives. *Frontiers in Microbiology*, *11*(April), 1–14. https://doi.org/10.3389/fmicb.2020.00634

Waterbury, J. B. (1986). Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Photosynthetic Picoplankton Canadian Bulletin of Fisheries Aquatic Sciences*, *214*, 71–120.

Waterbury, J. B., Watson, S. W., Guillard, R. R. L., & Brand, L. E. (1979). *Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium*. *277*(January), 293–294.

Weber, M. H., & Marahiel, M. A. (2003). Bacterial cold shock responses. *Science Progress*, *86*(Pt 1-2), 9–75. https://doi.org/10.3184/003685003783238707

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag. https://ggplot2.tidyverse.org

Wilmotte, A., Dail Laughinghouse IV, H., Capelli, C., Rippka, R., & Salmaso, N. (2017). Taxonomic Identification of Cyanobacteria by a Polyphasic Approach. *Molecular Tools for the Detection and Quantification of Toxigenic Cyanobacteria*, 79–134. https://doi.org/10.1002/9781119332169.ch4

Wilmotte, A. M. R., & Stam, W. T. (1984). Genetic relationships among cyanobacterial strains originally designated as "*Anacystis nidulans*" and some other *Synechococcus* strains. *Journal of General Microbiology*, *130*(10), 2737–2740. https://doi.org/10.1099/00221287-130-10-2737

Winther, K. S., & Gerdes, K. (2011). Enteric virulence associated protein VapC inhibits translation by cleavage of initiator tRNA. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(18), 7403–7407. https://doi.org/10.1073/pnas.1019587108

Wood, A. M., Horan, P. K., Muirhead, K., Phinney, D. A., Yentsch, C. M., & Waterbury, J. B. (1985). Discrimination between types of pigments in marine *Synechococcus* spp. by scanning spectroscopy, epifluorescence microscopy, and flow cytometry. *Limnology and Oceanography*, *30*(6), 1303–1315. https://doi.org/10.4319/lo.1985.30.6.1303

Xia, K., Bao, H., Zhang, F., Linhardt, R. J., & Liang, X. (2019). Characterization and comparative analysis of toxin–antitoxin systems in Acetobacter pasteurianus. *Journal of Industrial Microbiology & Biotechnology*, *46*(6), 869–882. https://doi.org/10.1007/s10295-019-02144-y

Xia, X., Guo, W., & Liu, H. (2015). Dynamics of the bacterial and archaeal communities in the Northern South China Sea revealed by 454 pyrosequencing of the 16S rRNA gene. *Deep-Sea Research Part II*, *117*(Complete), 97–107. https://doi.org/10.1016/j.dsr2.2015.05.016

Xia, X., Vidyarathna, N. K., Palenik, B., Lee, P., & Liu, H. (2015). Comparison of the seasonal variations of *Synechococcus* assemblage structures in estuarine waters and coastal waters

of Hong Kong. *Applied and Environmental Microbiology*, *81*(21), 7644–7655. https://doi.org/10.1128/AEM.01895-15

Xie, Y., Wei, Y., Shen, Y., Li, X., Zhou, H., Tai, C., Deng, Z., & Ou, H. Y. (2018). TADB 2.0: An updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Research*, *46*(D1), D749–D753. https://doi.org/10.1093/nar/gkx1033

Xu, Y., Jiao, N., & Chen, F. (2015). Novel psychrotolerant picocyanobacteria isolated from Chesapeake Bay in the winter. *Journal of Phycology*, *51*(4), 782–790. https://doi.org/10.1111/jpy.12318

Yoshimune, K., Galkin, A., Kulakova, L., Yoshimura, T., & Esaki, N. (2005). Cold-active DnaK of an Antarctic psychrotroph Shewanella sp. Ac10 supporting the growth of dnaK-null mutant of *Escherichia coli* at cold temperatures. *Extremophiles*, *9*(2), 145–150. https://doi.org/10.1007/s00792-004-0429-9

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F., Massana, R., Ulloa, O., & Scanlan, D. J. (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environmental Microbiology*, *10*(1), 147–161. https://doi.org/10.1111/j.1462-2920.2007.01440.x