Title
Ecology and Evolution of Diatom-Associated Cyanobacteria Through Genetic Analyses

Author
Hilton, Jason Andrew

Publication Date
2014

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ECOLOGY AND EVOLUTION OF DIATOM-ASSOCIATED CYANOBACTERIA THROUGH GENETIC ANALYSES

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

OCEAN SCIENCES

by

Jason A. Hilton

June 2014

The Dissertation of Jason A. Hilton
is approved:

Professor Jonathan Zehr, chair

Professor Raphael Kudela

Professor Jack Meeks

Professor Tracy Villareal

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Tables and Figures

Chapter 1: Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont

Chapter 2: The genomic sequence of the *Rhizosolenia clevei* symbiont and comparison with other diatom symbionts

Chapter 3: Metatranscriptomics of N_2-fixing cyanobacteria in the Amazon River plume

Chapter 4: Distribution and diversity of DNA elements within functional genes of heterocyst-forming cyanobacteria

Evolution and ecology of diatom-associated cyanobacteria through genetic analyses

Jason A. Hilton

Abstract

Primary production in a large fraction of the surface ocean communities is limited by the availability of nitrogen (N). Heterocyst-forming cyanobacteria associated with diatoms of the genera *Hemiaulus, Rhizosolenia,* and *Chaetoceros* can dominate surface communities throughout the global oceans, and are an important source of N to these communities. Additionally, these unique associations are directly linked to a highly efficient carbon export system. In order to study the nature and evolution of these symbioses, the genomes of representatives of symbionts from the three common diatom associations were sequenced. The genomes revealed the evolutionary pressure a symbiont undergoes as a result of long-term associations with unicellular diatoms. The implications of the genome streamlining, specifically in the N metabolism pathways, extend to mechanisms for maintenance of the association as well as possible symbiont regulation mechanisms by the diatom. The genomes could additionally be implemented in an unbiased examination of diatom symbiont transcription through an extensive environmental RNA sequence data set. This revealed significant reduction in photosystem II gene expression and possible regulation on the cyclic electron transport of the most abundant diatom symbiont population. The environmental sequences also showed very low diversity amongst the diatom symbiont populations. Lastly, the examination of interruption elements in heterocyst-forming cyanobacteria expanded the framework for which to study these genetic features. Although the majority of the element sequences proved to be highly dynamic, the recombinase sequences on each interruption element gave insights into the evolutionary paths of these unique genetic features. Additionally, the presence of elements in non-N_2 fixation genes and in cyanobacteria not capable of cell differentiation implicate the interruption elements in processes beyond heterocyst formation. The genomic sequences of oceanic diatom symbionts have opened a myriad of approaches for which to study the globally significant diatom-diazotroph associations. These analyses have revealed much about the ecology and evolution of diatom symbionts, and carry implications into global oceanic nutrient cycling, as well as plant-microbe associations, in general.

Acknowledgements

The following dissertation is a result of the efforts of many, and it is my hope that my sense of pride in these studies is shared by all of those who have contributed to it. Above all, I thank Jon Zehr for the support, guidance, and numerous opportunities provided to me that have allowed me to develop as a researcher. As a member of the Zehr Laboratory, I have had the privilege of working beside many intelligent and insightful people, and I am very thankful for all of the assistance I received from them. I also thank my entire thesis committee for their guidance through this dissertation and their help in viewing each scientific puzzle from a new perspective. Graduate students, faculty, and staff within the UCSC Ocean Sciences Department, as well as those from other institutes that I have had the great fortune of meeting and working with, have assisted me in every way ranging from technical support to casual discussions. Finally, I am eternally grateful for the ongoing support from my family and friends

Introduction

Oceanic N_2 fixation

Primary production, which was responsible for oxygenating our atmosphere, is limited by the availability of fixed inorganic nitrogen (N) in many oceanic environments. In the oceans, N limitation can be avoided by some primary producers, as well as some heterotrophic microorganisms, by their ability to convert N_2 gas into ammonia (NH_3). Additionally, the biologically-available N produced through N_2 fixation can fuel primary production throughout ecosystems (Eppley & Peterson, 1979; Vitousek *et al.*, 2002; LaRoche & Breitbarth, 2005). Thus, N_2-fixing populations are globally significant for N cycling within communities, as well as carbon cycling through the support of primary production.

The organisms capable of N_2 fixation, termed diazotrophs, avoid N limitation, but have several physiological challenges. Among them, the enzyme that catalyzes N_2 fixation, nitrogenase, is sensitive to inactivation by oxygen (O_2) (Wong & Burris, 1972). Many marine diazotrophs are cyanobacteria, which evolve O_2 through photosynthesis and use various mechanisms to separate N_2 fixation and photosynthesis activities. Some diazotrophs avoid nitrogenase inhibition by photosynthesizing during the day and fixing N_2 at night (Berman-Frank *et al.*, 2007). Other diazotrophic cyanobacteria physically separate the two processes by confining N_2 fixation to specialized cells, called heterocysts (Stewart *et al.*, 1969). The activity of O_2-evolving photosystem II is reduced or absent within heterocysts, and a thick envelope around the cell wall prevents gas diffusion, creating a low O_2 microenvironment (Wolk *et al.*, 2004). Through cellular differentiation, heterocyst-forming cyanobacteria are able to spatially separate two vital, yet potentially conflicting, metabolic processes.

In marine environments, observations of free-living heterocyst-forming cyanobacteria are very rare (Gomez *et al.*, 2005; White, Prahl, *et al.*, 2007; Zhao *et al.*, 2012). It has been hypothesized that the advantage of heterocysts is reduced at warmer temperatures where oceanic diazotrophy is commonly found, and thus, free-living heterocyst-forming cyanobacteria are outcompeted (Staal *et al.*, 2003). Instead, the heterocyst-forming cyanobacteria in the ocean are commonly observed in symbiotic relationships with diatoms, a diverse group of unicellular marine algae (Bowler *et al.*, 2008). The heterocyst-forming cyanobacteria capable of forming these unique associations are the focus of this dissertation.

Heterocyst-forming cyanobacteria in symbiosis

Associations between heterocyst-forming cyanobacteria and eukaryotic partners are not unusual. Heterocyst-forming cyanobacteria form associations with a wide range of taxonomically-diverse multicellular eukaryotes including fungi (Schüssler *et al.*, 1996), bryophytes (Enderlin & Meeks, 1983), gymnosperms (Lindblad *et al.*, 1985), and angiosperms (Johansson & Bergman, 1994). However, diatoms are the only known unicellular partners to form associations with heterocyst-forming cyanobacteria.

Relative to the other symbioses involving heterocyst-forming cyanobacteria, the associations formed between heterocyst-forming cyanobacteria and diatoms, termed diatom-diazotroph associations (DDAs), have been generally understudied. It is unclear if the diatom symbionts are obligate, and dependent on the partner for survival, or if they are facultative, and have a free-living state. The majority of heterocyst-forming cyanobacterial symbionts are facultative (Kneip *et al.*, 2007), but some are obligate (Peters & Meeks, 1989; Ran *et al.*, 2010). The common theme in all of the heterocyst-forming cyanobacteria associations is that the symbiont transfers fixed N to its partner (Adams *et al.*, 2006), and the transfer of N has recently been shown in DDAs (Foster *et al.*, 2011). However, it is not clear if the N is transferred to the diatoms in the form of ammonia, as is most commonly observed in heterocyst-forming cyanobacteria associations (Rai *et al.*, 1983; Meeks, Enderlin, *et al.*, 1985; Silvester *et al.*, 1996), or amino acids, as are likely transferred from symbiont to the gymnosperm *Cycas revoluta* (Lindblad *et al.*, 1987). The benefits to the diatom symbionts are unknown. In associations with *Cycas revoluta* and *Gunnera* spp., the cyanobacterial population cannot access sufficient light to photosynthesize from within the eukaryotic partner, and therefore rely on the partner for fixed carbon (Lindblad *et al.*, 1987; Söderbäck & Bergman, 1993). However, at least some of the diatom symbionts appear photosynthetically active (Weare *et al.*, 1974; Janson *et al.*, 1995), and likely do not require additional carbon from the diatom.

This dissertation aims to resolve many of the unanswered questions about the nature of the diatom associations through the sequencing of the diatom symbiont genomes. In addition to the metabolic capabilities of an organism, the genome of a symbiont also holds many clues as to whether the associations it forms is obligatory or facultative (Moran, 2003), and even if an obligatory partnership is ancient or has been more recently formed (Moran & Plague, 2004).

Three common DDAs

The DDAs observed in the oceans have several distinctions that separate them from each other, but it is unclear if these differences are indicative of varying ecological roles. Diatoms of the genera *Hemiaulus, Rhizosolenia*, and *Chaetoceros* are commonly observed in association with heterocyst-forming cyanobacterial symbionts (Heinbokel, 1986; Villareal, 1990; Janson *et al.*, 1999). *Guinardia* and *Bacteriastrum* have also been observed with symbionts, although less frequently than the other three diatoms (Villareal, 1992; Kulkarni *et al.*, 2010). Thus, it is mainly just the three commonly observed diatom-diazotroph associations that are addressed in this dissertation. Phylogenetic analysis of the diatom-associated heterocyst-forming cyanobacteria studied to-date revealed that each diatom harbors symbionts genetically distinct from those of other diatom species (Janson *et al.*, 1999; Foster & Zehr, 2006). Even the symbionts associated with *Hemiaulus hauckii* have sequence divergence from those associated with *H. membranaceus*, based on partial sequences of *hetR*, a heterocyst differentiation gene (~96% ID, DNA) (Janson *et al.*, 1999). The specificity between diatoms and associated cyanobacteria has been confirmed by partial sequences of the symbiont 16S rRNA gene and the *nifH* gene, which encodes the nitrogenase iron protein (Foster & Zehr, 2006).

The location of the symbiont within or on the diatom partner also differs depending on the diatom genus. Symbionts have been observed residing inside the siliceous cell wall, or frustule, of *Hemiaulus* spp. and *Rhizosolenia* spp. (Heinbokel, 1986; Villareal, 1990). The *Rhizosolenia* spp. symbiont has been shown to reside outside of the plasmalemma (Villareal, 1990), but it is unknown whether the *Hemiaulus* spp. symbiont is found in the same location or if it is truly an intracellular symbiont. The symbionts of the diatom genera *Rhizosolenia* and *Hemiaulus* are passed from one diatom generation to the next (Villareal, 1989; Zeev *et al.*, 2008). Meanwhile, heterocyst-forming cyanobacteria attach externally to the diatom *Chaetoceros* spp. (Janson *et al.*, 1999), but the mechanism of symbiont transmission in this association is unknown.

The frequency at which diatom cells harbor symbionts in the environment also differs between diatoms of the genera *Hemiaulus* and *Rhizosolenia* (there are no current data for *Chaetoceros*). When *Hemiaulus* spp. has been observed containing symbionts, it is typical for >95% of the diatom cells to possess at least one symbiont filament (Villareal, 1994; Vaillancourt *et al.*, 2003), with the lowest-reported fraction being 82% (Heinbokel, 1986). Similar numbers have been observed for symbiont-containing *Rhizosolenia* spp. (Heinbokel, 1986), but a much wider range has been documented for these diatom partners, even along a single transect. All *Rh. styliformis* cells in the tropical and subtropical North Pacific Ocean contained symbionts, but none of the diatom population contained symbionts in equatorial waters (Marumo & Asaoka, 1974). This contrast in frequency of symbiont-harboring diatoms has also been observed on smaller spatial scales. In waters near Hawaii, 91% of *Rh. styliformis* cells outside of an eddy harbored symbionts, but the diatom cells within an adjacent eddy, with elevated nitrate and nitrite concentrations, were devoid of any symbionts (Vaillancourt *et al.*, 2003). The ability for *Rhizosolenia* spp. to not only survive without the symbionts, but preferentially outgrow them in the presence of an N source, is supported by laboratory cultures of *Rh. clevei* growing without the symbionts upon addition of nitrate to the media (Villareal, 1990). Thus, there appears to be a distinction in the ability for each diatom to persist without a symbiont population.

Few N_2 fixation rate measurements of diatom symbionts have been documented to date, but initial results indicate the N_2 fixation rates of the symbionts in each association also differ. Separate studies that measured N_2 fixation by acetylene reduction found that the symbionts in association with *Rhizosolenia* spp. fixed N_2 at a rate two to five times higher than that of *Hemiaulus* spp. or *Chaetoceros compressum* symbionts (Villareal, 1991; Kitajima *et al.*, 2009).

These observations serve as strong evidence that the nature of each partnership is unique and viewing DDAs as one general relationship would be incorrect and misleading. An inaccurate depiction of DDAs would alter global N budgets and ecosystem modeling. In the following dissertation, comparative genomics of a representative symbiont genome from each of the three associations was sequenced in order to distinguish their metabolic capabilities. Additionally, metatranscriptomic analysis of natural DDA populations was also conducted to compare their potential metabolic activities.

DDA global distribution and significance

The N supply from the symbionts allows the diatoms, a group normally associated with high-nutrient coastal waters (Goldman, 1993), the ability to form blooms in open ocean environments where low N concentrations commonly limit phytoplankton growth (Venrick, 1974; Brzezinski *et al.*, 1998; Foster & Zehr, 2006; White, Spitz, *et al.*, 2007). The distribution of the three common symbiont-harboring diatoms has been well-documented throughout global tropical and subtropical marine environments. Additionally, the significance of DDAs, as a whole, has been recognized in surface ocean nutrient cycling and export to the deep ocean (Carpenter *et al.*, 1999; Scharek *et al.*, 1999; Subramaniam *et al.*, 2008; Wilson *et al.*, 2008; Padmakumar *et al.*, 2010; Karl *et al.*, 2012; Yeung *et al.*, 2012).

DDAs have been extensively studied in the North Pacific Subtropical Gyre, and symbionts have been observed there in association with the diatoms *Hemiaulus hauckii, H. membranaceus, Chaetoceros compressus,* and many different *Rhizosolenia* species (Marumo & Asaoka, 1974; Heinbokel, 1986; Vaillancourt *et al.*, 2003; Gomez *et al.*, 2005; Foster & Zehr, 2006; Wilson *et al.*, 2008; Kitajima *et al.*, 2009). Blooms in the North Pacific are frequently dominated by symbiont-harboring *H. hauckii* (Brzezinski *et al.*, 1998), *Rhizosolenia cylindrus* (Venrick, 1974), and

DDAs in general (White, Spitz, *et al.*, 2007). It has also been estimated that the N added to the surface community by DDAs in this region has fueled the formation of blooms comprised of other phytoplankton (Wilson *et al.*, 2008). Additionally, DDAs in the North Pacific are largely responsible for export events, or pulses, to the deep (Scharek *et al.*, 1999; Karl *et al.*, 2012). DDAs have not been studied in the South Pacific Ocean at such extent, but *H. hauckii*, *H. membranaceus,* and *Rh. clevei* have all been observed there with symbionts (Janson *et al.*, 1999).

DDAs in the Atlantic Ocean have also been well-studied, specifically in the western side of the basin. In the western Atlantic Ocean and Caribbean Sea, associations with *Hemiaulus* spp. (*H. hauckii, H. membranaceus,* and *H . sinensis*) are typically more abundant that those with *Rhizosolenia* spp. (Villareal, 1991, 1994). In the tropical North Atlantic Ocean, symbiont densities reached nearly 10^4 heterocysts L^{-1} during a bloom that was dominated by symbiont-containing *H. hauckii* (Foster & Zehr, 2006). The lone detection of the *Chaetoceros* association in this ocean basin came from a molecular assay targeting the *nifH* sequence of the symbionts in the eastern equatorial Atlantic (Foster *et al.*, 2009).

The role of DDAs in the nutrient cycles in the Indian Ocean has only recently been documented. *Rh. hebetata* has been frequently observed in association with heterocyst-forming cyanobacteria in the Indian Ocean and connecting waters (Kulkarni *et al.*, 2010; Jabir *et al.*, 2013; Madhu *et al.*, 2013), and even form mats of nearly 10^5 diatom cells L^{-1} (Padmakumar *et al.*, 2010). The N_2 fixed by the cyanobacterial symbionts within these *Rhizosolenia* mats supports the primary production in the surface communities (Padmakumar *et al.*, 2010). Symbionts are also seen with *Rh. formosa, Hemiaulus* spp., and *Guinardia* spp. in this region (Kulkarni *et al.*, 2010; Jabir *et al.*, 2013).

The nutrients brought to oligotrophic waters from river plumes are known to increase primary production (Smith & Demaster, 1996; Lohrenz *et al.*, 1999), and tend to be DDA hot spots. The mixture of nutrient-rich river water with low nutrient oceanic water creates a nutrient gradient in the plume region. DDAs are especially prevalent between the river input and open ocean, in transitional waters where enough riverine fixed N has been assimilated by the community to provide an advantage to diazotrophs, but riverine P, Fe, and Si remain high enough to support DDA growth (Yeung *et al.*, 2012; Goes *et al.*, 2014). Most notably, *H. hauckii* and *Rh. clevei* with symbionts are abundant within a large diazotrophic community in and around the Amazon River plume (Foster *et al.*, 2007). The abundance of DDAs in this region has been directly linked to a highly efficient carbon export system (Subramaniam *et al.*, 2008; Yeung *et al.*, 2012). Furthermore, a bloom of symbiont-containing *H. hauckii* in this region reached densities of 10^6 heterocyst L^{-1} and was responsible for an estimated 0.45 Tg N added to the surface ocean (Carpenter *et al.*, 1999). Additionally, symbionts associated with *Hemiaulus* spp. and *Rhizosolenia* spp. have comprised a large portion of the N_2-fixing communities in the Mekong River plume (Bombar *et al.*, 2011) and the Mississippi River plume (Knapke, 2012).

DDAs are also commonly found in marginal seas, although their impact on the nutrient cycling in these environments has not been studied as deeply. Symbionts were found associated mainly with *Hemiaulus* spp. in the Red Sea (Kimor *et al.*, 1992) and the Mediterranean Sea (Zeev *et al.*, 2008). More than 10^3 heterocysts L^{-1} of symbionts have been observed in association with *Rhizosolenia* spp. in the Gulf of California (White, Prahl, *et al.*, 2007), and this association has also been observed in the Gulf of Carpentaria, Australia (Burford *et al.*, 1995). The presence of heterocyst-forming cyanobacteria in association with diatoms is not restricted to the major ocean basins, but are present in marine environments throughout the world.

The associations allow each partner to extend their habitat and thrive in regions that neither is typically found in as a free-living organism. The abundance of diatoms harboring heterocyst-forming cyanobacteria in global ocean environments and their influence on surface community primary production and carbon export make it imperative to more fully understand these unique partnerships.

The metabolism of diatom symbionts

The first study of this dissertation describes the genomic sequence and comparative genomics of diatom symbionts representative of each of the three common DDAs. Identification of metabolic deficiencies of each diatom symbionts may elucidate the benefits of forming associations with diatoms, and reveal the dependency of each symbiont on the association. The genomes of the diatom symbionts also enabled comparison of their metabolic capabilities with each other, free-living heterocyst-forming cyanobacteria, and those found in association with multicellular eukaryotes. The genome sequence of an organism provides information about how that organism survives in the environment by revealing which metabolic capabilities are present and which are lacking. The recently assembled UCYN-A genome revealed this unicellular diazotrophic cyanobacterium lacks genes required for photosystem II and other metabolic pathways common to cyanobacteria, meaning its requirements differ greatly from other similar organisms in the same open ocean environment (Tripp *et al.*, 2010; Zehr *et al.*, 2008). The genomes of many symbiotic rhizobia lack a *nifV* gene, coding a homocitrate synthase, which provides homocitrate for the iron-molybdenum cofactor in nitrogenase. However, the rhizobia are able to fix significant amounts of N_2 when in symbiosis with legume *Lotus japonicus*, because of the homocitrate synthase encoded by the *L. japonicus* gene *FEN1* (Hakoyama *et al.*, 2009). Genomes of symbionts are especially insightful as they hold many clues as to whether the associations it forms are obligatory or facultative (Moran, 2003), and even if an obligatory partnership has been ancient or more recently formed (Moran & Plague, 2004). Shared metabolic traits of the three symbiont genomes revealed metabolic implications for all DDAs, while the distinctions amongst the genomes confirmed the nature of the association differs with each diatom genera.

Once the effects of associations with diatoms on the evolution of the symbiotic cyanobacterial metabolic capabilities were determined, the genomes could then be utilized to explore how those alterations were reflected in the gene expression of natural symbiont populations. To accomplish this, diatom symbiont RNA sequences from a large environmental data set were examined, using the previously-constructed genomes as references. Metatranscriptomics provides a full transcription snapshot of a community while avoiding potential bias stemming from targeting predetermined organisms or processes. Studying metatranscriptomes of oceanic microbial communities, in general, have revealed the abundance of novel transcripts and small RNAs (sRNAs) (Gilbert *et al.*, 2008; Shi *et al.*, 2009), the intricacies of diatom population response to iron limitation (Marchetti *et al.*, 2012), and the synchronicity of diel transcription amongst bacterial and archaeal populations (Ottesen *et al.*, 2013). Although more community-based research is enabled through the use of metatranscriptomes, only a few studies have utilized this tool to elucidate the physiological state of cells of diazotrophic populations. Important information such as the expression of key nutrient limitation response genes, as well as highly-expressed genes of unknown function, were obtained from metatranscriptomic analyses of *Crocosphaera* (Hewson, Poretsky, Beinart, *et al.*, 2009) and *Trichodesmium* populations (Hewson, Poretsky, Dyhrman, *et al.*, 2009). The use of metatranscriptomic analysis in the Amazon River plume revealed distinctions within key metabolic processes between diatom-associated diazotrophic populations and free-living diazotrophs within the same environment.

Interruption elements in heterocyst-forming cyanobacteria

The final study of this dissertation analyzes the origin and evolution of unique genetic features found in the diatom symbiont genomes, and many other heterocyst-forming cyanobacteria. During heterocyst differentiation, DNA sequences, termed elements, are excised from within key N_2 fixation genes, which are then contiguous and functional in the heterocyst genome (Golden *et al.*, 1985; Golden & Wiest, 1988; Vintila *et al.*, 2011). Although they have been studied for several decades, the origin of these interruption elements, how they have evolved, and what possible advantages or disadvantages they might provide to organisms are largely unknown. In order to examine the evolutionary paths of the elements, possible functions of these unique interruption elements, and to test their application as a possible marker for lineage, 101 elements were identified within genes from 28 heterocyst-forming cyanobacterial genomes. This included a total of seven elements within the three diatom symbiont genomes. The comprehensive analysis provided the framework to study the history and behavior of these mysterious sequences in all heterocyst-forming cyanobacteria.

The findings from the studies of this dissertation provide advancements for the field of marine microbial ecology as well as the study of plant-microbe associations. This has given much-needed framework for which to study the evolution of cyanobacteria capable of forming heterocysts, in addition to showing how associations formed by these organisms influence the evolution of their metabolism. The metabolic capabilities and activities of the diatom-associated cyanobacteria allow an improved illustration of the interaction between the symbiosis partners and their role in oceanic nutrient cycling. As a result of the studies compiled in this dissertation, there is now a more complete picture of the biology of these globally-significant microorganisms.

Chapter 1

Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont[1]

[1] This chapter has been published as: Hilton, J. A., Foster, R.A., Tripp, H.J., Carter, B.J., Zehr, J.P., and Villareal, T.V. (2013). Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. Nature Communications 4:1767 doi: 10.1038/ncomms2748.

Abstract

Diatoms with symbiotic N_2-fixing cyanobacteria are often abundant in the oligotrophic open ocean gyres. The most abundant cyanobacterial symbionts form heterocysts (specialized cells for N_2 fixation) and provide nitrogen (N) to their hosts, but their morphology, cellular locations and abundances differ depending on the host. Here, the location of the symbiont and its dependency on the host are shown to be linked to the evolution of the symbiont genome. The genome of *Richelia* (found inside the siliceous frustule of *Hemiaulus*) is reduced and lacks ammonium transporters, nitrate/nitrite reductases and glutamine:2-oxoglutarate aminotransferase. In contrast, the genome of the closely related *Calothrix* (found outside the frustule of *Chaetoceros*) is more similar to those of free-living heterocyst-forming cyanobacteria. The genome of *Richelia* is an example of metabolic streamlining that has implications for the evolution of N_2-fixing symbiosis and potentially for manipulating plant–cyanobacterial interactions.

Introduction

Cyanobacteria form partnerships with taxonomically diverse hosts that are usually multicellular, and these symbioses are ubiquitous in terrestrial and aquatic environments (Usher *et al.*, 2007). Cyanobacteria are autotrophic microorganisms and some can convert dinitrogen (N_2) gas to ammonium. Two groups of understudied planktonic symbioses are the partnerships between marine diatoms and the heterocyst-forming cyanobacteria, *Richelia intracellularis* and *Calothrix rhizosoleniae* (Figure 1a-c).

Richelia and *Calothrix* species fix N_2 and transfer the fixed N to their host (Foster *et al.,* 2011). *Richelia* and *Calothrix* associate with different hosts and also differ in cellular location (internal versus external), implying different life histories and mechanisms for nutrient exchanges with their partners. The *Richelia* symbionts of the diatom genera *Rhizosolenia* and *Hemiaulus* reside inside the diatom cell wall and are passed on to the next generation of the host (Villareal, 1992). The *Rhizosolenia* symbiont is outside the plasmalemma in the periplasmic space (Villareal, 1992); the *Hemiaulus* symbiont's location is unknown. In contrast, *Calothrix* attaches externally to *Chaetoceros* spp. and can be cultured without the host diatom (Foster *et al.*, 2010). Reports of free-living *Richelia* may be a result of broken diatoms (Lyimo, 2011; Zhang *et al.*, 2011), whereas *Calothrix* have been observed as individual trichomes in the plankton (Gomez *et al.*, 2005; White, Prahl, *et al.*, 2007). The mechanism of formation of a *Calothrix-Chaetoceros* association and whether the symbiont is transmitted to the next generation is unknown.

I compared the genomes of two of the *Richelia* internal symbiont strains (*Ri. intracellularis* HH01, RintHH, symbiont of *Hemiaulus hauckii* and *Ri. intracellularis* HM01, RintHM, symbiont of *H. membranaceus*) with that of the external symbiont *Calothrix rhizosoleniae* SC01 (CalSC). Genome size and content, especially N metabolism genes, differed substantially, suggesting the cellular location (intracellular versus extracellular) has dictated varying evolutionary paths and resulted in different mechanisms involved in maintaining the symbiosis (Table 1).

Methods

H. hauckii and *H. membranaceus* symbiont DNA preparation

Stable *Hemiaulus–Richelia* cultures, isolated from the western Gulf of Mexico, were grown in N-free YBC-II medium at 25 °C (Pyle, 2011), filtered on to a 3-µm pore size, 25-mm diameter polyester filter (Sterlitech) and frozen for storage. TE buffer (1 ×) was added, and once the filter thawed the cells were resuspended by vortexing for 1 min. The majority of diatoms were broken at this stage, releasing the symbionts in the process. Samples were then analyzed on the Influx flow cytometer and cell sorter (BD Biosciences), and cyanobacteria cells were distinguished from other cells by their phycoerythrin pigmentation (Figure 2). For the *H. hauckii* symbionts, the vegetative trichomes and heterocysts had separated during the sample preparation and the cells formed separate populations on the flow cytometer based on slightly different chlorophyll and phycoerythrin signatures (Figure 2). Sorting gates, defined by relative pigment values, allowed for the isolation of vegetative cells from the rest of the sample. Two replicate sorts of 5,000 symbiont vegetative trichomes (3–5 cells per trichome) were sorted. Genomic DNA in each sample was amplified by multiple displacement amplification using the Repli-g Midi kit (Qiagen). The manufacturer's protocol for 0.5 µl of cell material was followed with one exception: after buffer D2 was added, the samples were incubated for 5 min at 65 °C and then put on ice for 1 min, instead of 10 min on ice without a 65 °C incubation.

To ensure uncontaminated samples, each amplified DNA sample was PCR-amplified using universal 16S rRNA primers 27F (5′-

AGAGTTTGATCMTGGCTCAG-3′) and 1492R (5′-GGTTACCTTGTTACGACTT-3′) (Lane, 1991). The PCR was carried out in 50 μl reactions consisting of 1× PCR buffer, 2 mM $MgCl_2$, 200 μM dNTPs, 0.2 μM of each primer and 1.5 U of Platinum Taq DNA polymerase (Invitrogen). A touchdown PCR was performed as follows: an initial denaturing step at 94 °C for 5 min, followed by 30 cycles of three 1-min steps (denaturation at 94 °C, annealing at 53–41 °C and elongation at 72 °C) and a final elongation step at 72 °C for 10 min. The first cycle annealing took place at 53 °C and was lowered by 0.4 °C for each cycle to reach 41 °C for the final cycle. Resulting products were run on a 1.2% agarose gel, the distinct bands of ~1,500 bp were excised and then recovered using the Zymoclean DNA Recovery Kit (Zymo Research). The recovered DNA was then ligated and plated for blue/white screening using the pGem-T and pGem-T Easy Vectors Systems (Promega). Twenty-four colonies per sample were picked and grown overnight at 37 °C in 2 × LB media with carbenicillin (200 μg ml^{-1}). The Montáge Plasmid $Miniprep_{HTS}$ Kit (Millipore) was used following the manufacturer’s instructions for the full lysate protocol for plasmid DNA miniprep. Samples were sequenced at UC-Berkeley DNA Sequencing Facility and each sequence was subject to BLAST analysis against the nr/nt database (blastn). All sequences were identical and had top hits to 16S rRNA sequences of heterocyst-forming cyanobacteria, supporting no contaminant genomes were present in the samples.

DNA concentration and quality were checked (Agilent 2100 Bioanalyzer, Agilent Technologies) before submission for 454 Titanium sequencing (Roche) at the UCSC Genome Technology Center.

The symbionts of *H. membranaceus* were processed in the same manner, but heterocysts and vegetative cells did not separate during sample preparation, and both cell types were present in the sorted samples. Moreover, we were confident from flow cytometry that the cell preparation was pure enough to determine the comparative features of interest, and that the closely related symbiont could be distinguished from the few bacteria that could be carried through by flow cytometry. Therefore, no contamination or DNA quality checks were performed in preparation of the RintHM samples.

H. hauckii and *H. membranaceus* symbiont genome assembly

A total of 433,028 reads were sequenced from RintHM samples. The reads assembled to nearly 8 Mb and the assembly contained four 16S rRNA sequences with low similarity to each other (<83% ID), indicating multiple DNA sources in the data. RintHM contigs were defined as those which had a better BLAST hit to RintHH than to any other organism in the nr/nt database. The resulting 2,212,909 bp (941 contigs, coverage depth 13 ×) were made up of 77,324 reads averaging 380 bp each. An additional 31 contigs, totaling 97,821 bp, had a top hit in the nr/nt database to a cyanobacterium other than RintHH, but none of those contigs contained any of the N metabolism genes of interest.

The two RintHH samples yielded a total 409,035 reads, averaging 344 bp each. The read data were pooled and assembled into 3,243,759 bp in 90 contigs (coverage depth 43 ×) and appeared to be non-contaminated. There are seven contigs longer than 100 kbp, an additional 32 contigs longer than 25 kbp (Figure 3) and 91% of bases with 15 × coverage or greater.

CalSC DNA preparation

CalSC genomic DNA was extracted from pelleted cells using a sucrose lysis protocol, including the optional back extraction (Cuvelier *et al.*, 2010). The exceptions to this protocol were the use of 10% SDS in the lysate for Fraction B instead of 20% SDS and the 1-h incubation of Fraction B after adding the lysate was at 37 °C rather than 55 °C. The genomic DNAs from Fraction B were pooled and divided into three equal volume samples. The three genomic extracts were checked for purity and quantity (Agilent 2100 Bioanalyzer, Agilent Technologies), and the DNA concentrations ranged between 38.37 and 66.38 ng μl^{-1}. Samples were then submitted to JCVI for 454 sequencing.

CalSC genome assembly

Once the read data from JCVI (2,477,040 reads, 968 MB) were assembled, the number of contigs (69,919) and size of the assembly (81.4 Mb) immediately suggested that more than one organism was in the sequencing samples. The longest contig of 1.2 MB in length contained a full-length rRNA operon predicted by RDP (Ribosomal Database Project) to be a Planctomycete, confirming the presence of organisms other than CalSC. A plot of the number of reads on each contig against the length of the contig showed strong linear relationships (Figure 4), representing defined clusters of coverage depth, based on the relative abundance of the each organism's genome in the sample. Spot-checking the phylogeny of BLAST (blastn) results for long open reading frames (ORFs) on long contigs revealed that the contigs lying along the line marked in red (representing a coverage depth of 30×) were those that came from CalSC (Figure 4). Each predicted ORF >450 bp on contigs with depth of coverage 15–45× was subject to BLAST analysis against the nr/nt database. A contig was considered to be part of the CalSC genome if at least one of these ORFs on the contig had a top hit to a cyanobacterial sequence, and 471 contigs met this criterion (5,967,587 bp). One additional contig (5,416 bp) containing the rRNA operon was added. It had been overlooked initially due to its lack of ORFs and its relatively higher coverage depth (71 × , indicating it is present in two copies in the genome). The end result was a 5,973,003 bp genome composed of 472 contigs (30 × coverage depth).

Genomic analysis

After assembly and contamination screening, the genomes were submitted to RAST (Rapid Annotation using Subsystem Technology) (Aziz *et al.*, 2008) for annotation.

The nitrogen metabolism genes not found in the RintHH genome were pulled from each Nostocales genome, and each gene was subject to BLAST analysis against a database of all 409,035 reads (tblastn, e-value <10). Two thousand seven hundred and twenty reads had hits at least 25% identical (AA) across at least 50% the length of the read or gene, whichever was shorter. A BLAST analysis of each of these reads against the nr database was performed (blastx, e-value <10). Twenty-one reads had a top hit to a GOGAT-encoding gene, and each of these reads is assembled into the intergenic region discussed below as likely GOGAT remnants in the RintHH genome. No other reads had a top hit in the nr database of a nitrogen metabolism-related gene.

Predicted ORFs in each genome with a BLAST hit in the Transporter Classification database (Saier *et al.*, 2009) (blastp, e-value $<10^{-19}$) were counted as transporter genes.

For the 16S rRNA and the *ntcA* phylogenetic trees, nucleotide sequences were acquired from DOE Joint Genome Institute for each of the seven previously sequenced Nostocales genomes and *Trichodesmium erythraeum* IMS101, and were aligned with the sequences from the three diatom symbiont genomes using Clustal W (Thompson *et al.*, 1994) (1,421 bp, 16S rRNA; 646 bp, *ntcA*). Phylogenetic analyses were rendered in Mega5 ((Tamura *et al.*, 2011)) using the Neighbor-Joining method (Saitou & Nei, 1987). The Tamura–Nei test was run to detect the best models. Statistical support for nodes was based on 1,000 bootstrap replicates (Felsenstein, 1985).

Results

General features of the diatom symbiont genomes

On the basis of the 16S rRNA and *ntcA* gene sequences, the diatom symbionts cluster within the cyanobacterial Order Nostocales (Figure 5), but their genome sizes vary greatly (RintHH, 3.2 Mb; CalSC, 6.0 Mb; Table 1). The percent coding information of the CalSC genome is only slightly lower than the free-living Nostocales members, whereas the RintHH genome percent coding is further reduced, similar to '*Nostoc azollae*' 0708 (Table 1). Similarly, the RintHH genome GC content and transporter count are lower than any other genome in the Order, whereas the CalSC genome is a more characteristic Nostocales genome in each respect (Table 1).

The genome of RintHM, the symbiont of *H. membranaceus*, a diatom that is closely related to *H. hauckii*, is only 2.2 Mb and is lacking a number of sequences expected of a full genome (including transfer RNAs for four amino acids and several nitrogenase genes). Therefore, it is likely a partial genome due to low-sequencing coverage (average depth of coverage 13×). However, 16S rRNA and *ntcA* sequences confirm the morphologically similar symbionts are also related genetically (Figure 5), as previously demonstrated by *nifH* and *hetR* sequences (Janson *et al.*, 1999; Foster & Zehr, 2006). In addition, analysis of the contigs showed that there are no evident gene insertions/deletions or genome rearrangements between the two *Hemiaulus* sp. symbiont genomes. The 1,671 shared genes of the symbionts average 97.5% sequence identity (DNA) (Figure 6) and show no significant difference in the GC content of the genes sequenced.

Nitrogen metabolism of the diatom symbionts

Given its small size, the RintHH genome is highlighted by many gene deletions, including numerous genes important in N metabolism, such as the transporters for ammonium and nitrate, and the genes encoding nitrate and nitrite reductases (Figure 7). The diatom symbiont genomes are each missing genes that encode urea transporters and urease, which are functional in all previously sequenced Nostocales genomes, except for the genome of '*N. azollae*' 0708 (Ran *et al.*, 2010).

The most unusual gene deletion in RintHH is the gene for an important enzyme in C and N metabolism, glutamate synthase, also known as glutamine:2-oxoglutarate amidotransferase (GOGAT). This enzyme is part of glutamine synthetase (GS)-GOGAT (GS-GOGAT), a generally universal pathway for high-affinity N assimilation (found in all other sequenced cyanobacterial genomes (Larsson *et al.*, 2011), including CalSC and '*N. azollae*' 0708), which uses glutamine, synthesized by GS, and a C skeleton, 2-oxoglutarate, to produce two glutamate molecules. The glutamate produced by GOGAT is then recycled for further ammonium assimilation by GS. The gene encoding GS is present and functional in each symbiont genome; however, they are each lacking a gene that encodes a GS-inactivating factor that is found in all previously sequenced Nostocales genomes (asl2329 in *Nostoc* sp. PCC 7120).

The multiple N metabolism genes missing from the RintHH genome are common to, and widely dispersed across, the genomes of all closed Nostocales genomes (Figure 8). Given this, and the high-sequencing coverage of the RintHH draft genome (average depth of coverage >40×), it is unlikely that the missing genes are actually present in the RintHH genome. The RintHH genome does contain a tRNA for each of the 20 amino acids, as expected from a complete genome. Other sequences expected to be present are also in the assembly, such as the previously studied genes *hetR* and *nifH* (Janson *et al.*, 1999; Foster & Zehr, 2006), and genes responsible for known characteristics of RintHH, such as nitrogen fixation, heterocyst formation, and phycoerythrin and chlorophyll pigments. Moreover, to decrease any possible bias during the process, two RintHH samples were separately sorted, amplified and sequenced.

The majority of the N metabolism genes found to be lacking in the RintHH genome show no similarity to any sequence in the genome. However, an intergenic sequence, which contains a small predicted hypothetical protein, has a top hit to the *Raphidiopsis brookii* GOGAT-encoding gene in the nr database (blastx, e-value=$3e^{-14}$; Figure 9). The intergenic sequence covers less than 20% of the GOGAT gene and aligns in all three unidirectional frames. This intergenic region is found downstream of two genes that are part of a conserved region downstream of the GOGAT gene in Nostocales genomes. A single contig from the RintHM genome also shows similarity to the GOGAT gene in the same manner (Figure 9).

Gene interruptions on the diatom symbiont *nif* operon

The similarities with other heterocyst-forming cyanobacteria include the presence of insertion sequences in the middle of RintHH and CalSC N_2-fixation genes (Golden *et al.*, 1985; Carrasco *et al.*, 1995). The RintHH *nifH* gene is interrupted in this manner by a 9.1-kb sequence (Figure 10). The CalSC *nifH* and *nifK* genes are each interrupted in the same manner by longer sequences (each >20 kb).

The *nifH* interruptions in RintHH and CalSC appear to occur at the same location within the *nifH* gene; however, the CalSC *nifH* element is at least twice as long as that in the RintHH genome. Recombination genes found on each *nifH* elements show high similarity to each other (71% ID, protein) and are presumably the mechanisms for excision of the element during heterocyst formation.

Discussion

To date, the intracellular RintHH genome is the smallest N_2-fixing, heterocyst-forming cyanobacteria genome sequenced. Within Nostocales, the *Ra. brookii* D9 genome is slightly smaller than that of RintHH, but *Ra. brookii* D9 is unable to form heterocysts or fix N_2 (Stucken *et al.*, 2010). In contrast, the CalSC genome is similar in size and content to the genomes of free-living organisms in this Order and *N. punctiforme,* a facultative symbiont.

The genome reduction in RintHH, marked by its size, percent coding and GC content, is similar to that of '*N. azollae*' 0708, the obligate, or host-dependent, symbiont of the water fern *Azolla filiculoides* (Ran *et al.*, 2010). These features are commonly exhibited by genomes of obligate symbionts, indicating that RintHH is also dependent on its host. Obligate symbionts have more unnecessary genes than free-living or facultative symbiotic organisms due to metabolic redundancy encoded by the host genome and the lack of full exposure to the environment (Moran, 2003). Examples of genes dispensable to obligate symbionts may be those absent or non-functional in both RintHH and '*N. azollae*' 0708, but present in other heterocyst-forming cyanobacteria genomes (Table 2). Decreased evolutionary pressure to keep functional genes leads to a lower percent coding and eventually to genome size reduction as non-functioning genes are deleted. The smaller genome leads to accelerated sequence evolution, increasing AT bias (Moran, 2003). The lack of CalSC genome reduction may be taken as evidence that this organism is a facultative symbiont. This is consistent with the external location of CalSC on the diatom setae (spine-like projections) and the ability to maintain it in culture independent of the host diatom in filtered seawater-based media (Foster *et al.*, 2010). In contrast, RintHH lives inside the host diatom cell wall, and possibly even within the cytoplasm, with little or no exposure to the external environment, and thus the genome reduction is consistent with that of an obligate symbiont.

The numerous absent N metabolism genes appear to have been selectively deleted from multiple regions throughout the RintHH genome. The lack of ammonium transporters and enzymes required to take up and assimilate urea or nitrate limits the possible N sources for RintHH to amino acids, N_2, and passive diffusion of ammonia in oceanic environments, where concentrations of amino acids and ammonium are extremely low. Therefore, deletions in N metabolism genes ensure N_2 fixation within the partner diatom persists, and is likely important for maintaining the symbiotic partnership.

The lack of GOGAT, on the other hand, likely streamlines host–symbiont interactions and seems to be a more recent deletion than the other N metabolism genes, given the similarity between GOGAT genes and intergenic space in the RintHH genome. Without GOGAT, RintHH must use an alternate pathway for assimilation of N_2-derived ammonium with glutamate dehydrogenase (GDH; Figure 7), unless the host diatom provides glutamate for the symbiont. In contrast, GS-GOGAT is the main N assimilation pathway used by *Anabaena azollae* in obligate symbiosis with host *Azolla caroliniana*, and very little N is assimilated through GDH (Meeks, Steinberg, *et al.*, 1985). Given the high N_2 fixation rates by the cyanobacterial symbiont when associated with the host diatom (Foster *et al.*, 2011), it is feasible that intracellular ammonium concentrations are elevated and facilitate assimilation by the low-affinity GDH enzyme (Florencio *et al.*, 1987). However, an adequate concentration of 2-oxoglutarate would also be needed to support ammonium assimilation. If these C skeletons are provided by the host, as in the *Nostoc–Gunnera* symbiosis (Söderbäck & Bergman, 1993), the symbiont may perceive the increase of intracellular C:N as N starvation (Muro-Pastor *et al.*, 2001), causing continued N_2 fixation by the cyanobacterium. Thus, the lack of GOGAT eliminates a common metabolic pathway and could create an N exchange pathway between host and symbiont that provides the host with a way to regulate the symbiont’s growth and activity.

The lack of a GS-inactivating factor streamlines N metabolism further in RintHH. GS catalyses the conversion of glutamate to glutamine, and without an inactivating factor it will maintain low intracellular glutamate concentrations. The subsequent increasing glutamine pool may indicate this amino acid is the form of N passed to the host. The absence of this regulator shows parallels between the *Richelia–Hemiaulus* and *Calothrix–Chaetoceros* associations, and separates the diatom symbionts from other heterocyst-forming cyanobacteria.

However, with regard to N metabolism, the similarities are minimal and the fundamental differences between the RintHH and CalSC genomes reflect the evolutionary selection of their metabolic interactions and cellular locations with the partner diatom. The extracellular CalSC symbiont is exposed to the open ocean environment at all times, and can therefore use a suite of dissolved inorganic nitrogen sources, albeit at low concentrations. Furthermore, the CalSC genome possesses a gene to encode GOGAT and, thus, the symbiont is capable of assimilating N through the high-affinity GS-GOGAT, in addition to GDH. However, a scenario for enhancing N_2 fixation by C transfer from the diatom to the external symbiont CalSC, as hypothesized for the *Richelia–Hemiaulus* association, would require a direct host–symbiont transport system. Otherwise, the extracellular C would likely be diluted immediately and available to other microorganisms. Thus, the extracellular location of CalSC on *Chaetoceros* spp. likely requires different mechanisms for N metabolism and exchange than intracellular RintHH. The differences in genome content and metabolic potential reflect the differences between obligate and facultative symbionts.

Many heterocyst-forming cyanobacteria have DNA sequences interrupting N_2 fixation-related genes in vegetative cells, which are excised during genome rearrangements coincident with heterocyst development (Carrasco *et al.*, 1994), but the functional significance and evolutionary origin of these elements are unknown. These interrupting sequences have been seen previously in several genes (Golden *et al.*, 1985; Carrasco *et al.*, 1995; Vintila *et al.*, 2011), but the CalSC genome is the first example of a *nifK* element. The location of elements within *nifH* and high similarity between the genes likely responsible for the excision of the interrupting sequence are the only apparent similarities between the two *nifH* elements in these closely related cyanobacteria. Although their similarities indicate the *nifH* elements in each organism have the same evolutionary origin, there seems to be little evolutionary pressure on the contents and length of the element.

Conclusions

The characteristics of the genomes of symbiotic heterocyst-forming cyanobacteria reflect the differences in cellular location and host dependency. The absence of basic N metabolism enzymes and transporters in the RintHH genome streamline it, while maintaining the association and providing a mechanism for host regulation of the symbiont. In contrast, the genome of CalSC has few deletions relative to free-living heterocyst-forming cyanobacteria. The differences between genomes suggest mechanisms that may be important in defining facultative or obligate symbioses, with implications for the biology and ecology of these widespread symbiotic associations in the sea. Furthermore, differences in the genomic composition of morphologically and taxonomically similar microorganisms provides an important example of how one partner's metabolic capabilities can evolve with a symbiosis. Finally, the genomes reported in this study, in addition to other recent discoveries of extensive metabolic streamlining in N_2-fixing cyanobacteria (Thompson *et al.*, 2012), yield the possibility of yet undiscovered plants or algae containing N_2-fixing organelles.

Acknowledgements

This work was sponsored in part by the Gordon and Betty Moore Foundation, NSF and the NSF Center for Microbial Oceanography: Research and Education. The

CalSC genome was sequenced through the Microbial Genome Sequencing Project managed by JCVI and was sponsored by the Gordon and Betty Moore Foundation. We thank J. Meeks for comments on the manuscript; M. Hogan, S. Bench and K. Turk for technical help and discussions; K. Karplus and J. Long for their effort on the CalSC genome assembly; and A. Worden and H. Wilcox for suggestions on CalSC culture DNA extraction methods.

Tables and Figures

Table 1. Nostocales genomes. Statistics of genomes from cyanobacteria of the order Nostocales available at the time of this report.

Cyanobacterium	**Accession number**	**Symbiotic state**	**Size (Mb)**	**Percent GC**	**Percent coding**	**TCs**
Anabaena variabilis ATCC 29413	PRJNA10642	Free-living	7.1	41	82	570
Nostoc punctiforme PCC 73102	PRJNA216	Facultative	9.1	41	77	575
Nostoc sp. PCC 7120	PRJNA244	Free-living	7.2	41	82	559
‘*Nostoc azollae’* 0708	PRJNA30807	Obligate	5.5	38	52	286
Raphidiopsis brookii D9 [a]	PRJNA40111	Free-living	3.2	40	86	300
Cylindrospermopsis raciborskii CS-505[a]	PRJNA40109	Free-living	3.9	40	85	344
Nodularia spumigena CCY 9414 [a]	PRJNA13447	Free-living	5.3	41	82	428
Richelia intracellularis HH01 [a]	PRJEA104979	Obligate	3.2	34	56	190
Calothrix rhizosoleniae SC01 [a]	PRJNA19291	Facultative	6.0	39	76	400

TC, transporter classification

[a] Genome is in a draft state

Table 2. Genes of obligate symbionts. Gene presence/absence in the *Ri. intracellularis* HH01 genome of genes uniquely present or uniquely absent in the *N. azollae* genome relative to other heterocyst-forming cyanobacterial genomes (Ran et al. 2010). *N. punctiforme* sequences are provided as representative functional gene sequences for genes not found in either obligate symbiont genome.

Gene Product	**Gene symbol**	***N. azollae***	***Ri. intracellularis***	***N. punctiforme***
uracil-DNA glycosylase	*ung*	Aazo_0465	RintHH_15290	np
dihydrofolate reductase	*folA*	Aazo_1890	np	np
thymidylate synthase	*thyA*	Aazo_5011	np	np
conserved hypothetical protein		*Aazo_3194	RintHH_13480	Npun_F1238
putative GPH family sugar transporter		*Aazo_3944	RintHH_3260	Npun_F1762
photosystem II reaction center L protein (psII 5 Kd protein)	*psbL*	*Aazo_1235	RintHH_21410	Npun_F5553
lysyl-tRNA synthetase	*lysS*	np	RintHH_10460	Npun_R5202
thymidylate synthase (FAD)	*thyX*	np	RintHH_16770	Npun_R2526
conserved hypothetical protein		*Aazo_2347	RintHH_8590	Npun_F0905
chromosomal replication initiator protein DnaA	*dnaA*	*Aazo_2080	*RintHH_4500	Npun_F0001
6-phosphofructokinase	*pfkA*	*Aazo_0224	np	Npun_R0482
putative L-cysteine/cystine lyase		*Aazo_2409	np	Npun_F1365
L-lactate dehydrogenase	*ldh*	np	np	Npun_F2517
dCTP deaminase	*dcd*	np	np	Npun_F2524
heterocyst-inhibiting signaling peptide	*patS*	np	np	Npun_R5353

np - gene is not present * - gene is present only as a pseudogene

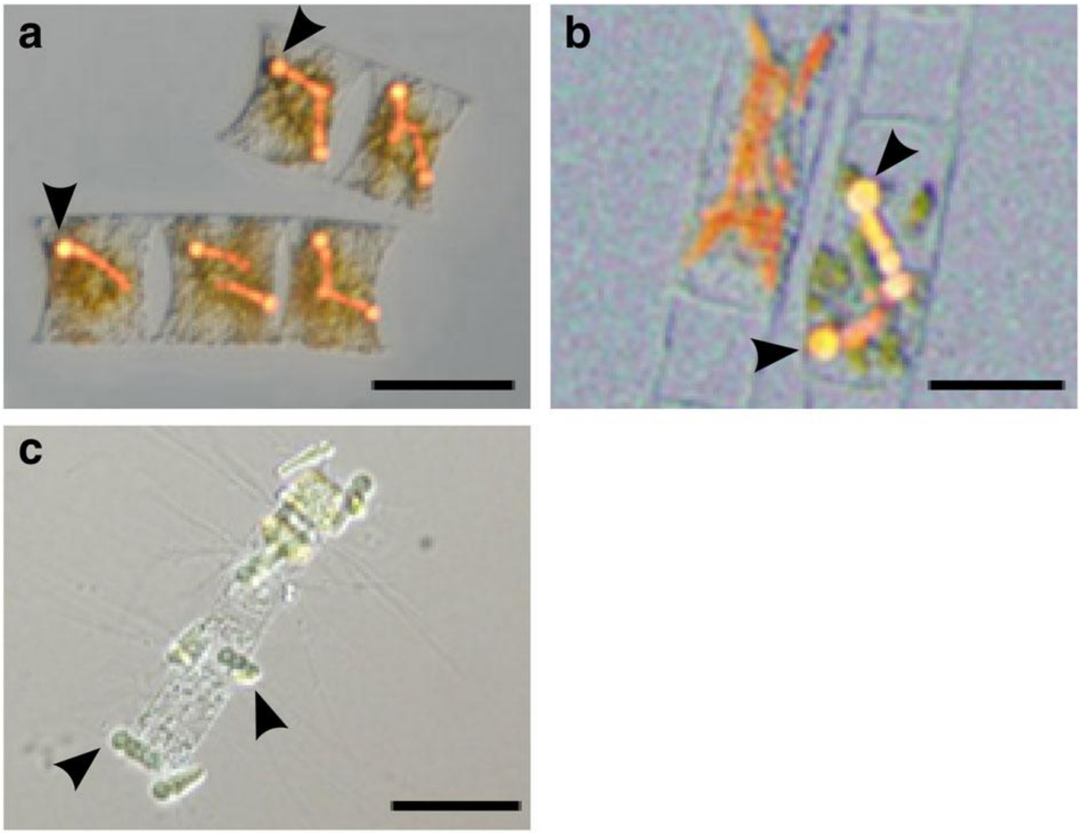


Figure 1. Diatom-diazotroph association images. Photomicrographs of cyanobacterial symbionts (denoted by arrows) representative of those sequenced in this study with host diatoms. Differential interference contrast bright field overlaid with blue light epifluorescence images of the diatoms *H. membranaceus* (a) and *H. hauckii* (b), with intracellular cyanobacterial symbionts. Bright-field microscopy image of epiphytic cyanobacterial symbiont *Ca. rhizosoleniae* SC01 attached to the host diatom *Chaetoceros* sp. (c). Scale bars, 50 µm.

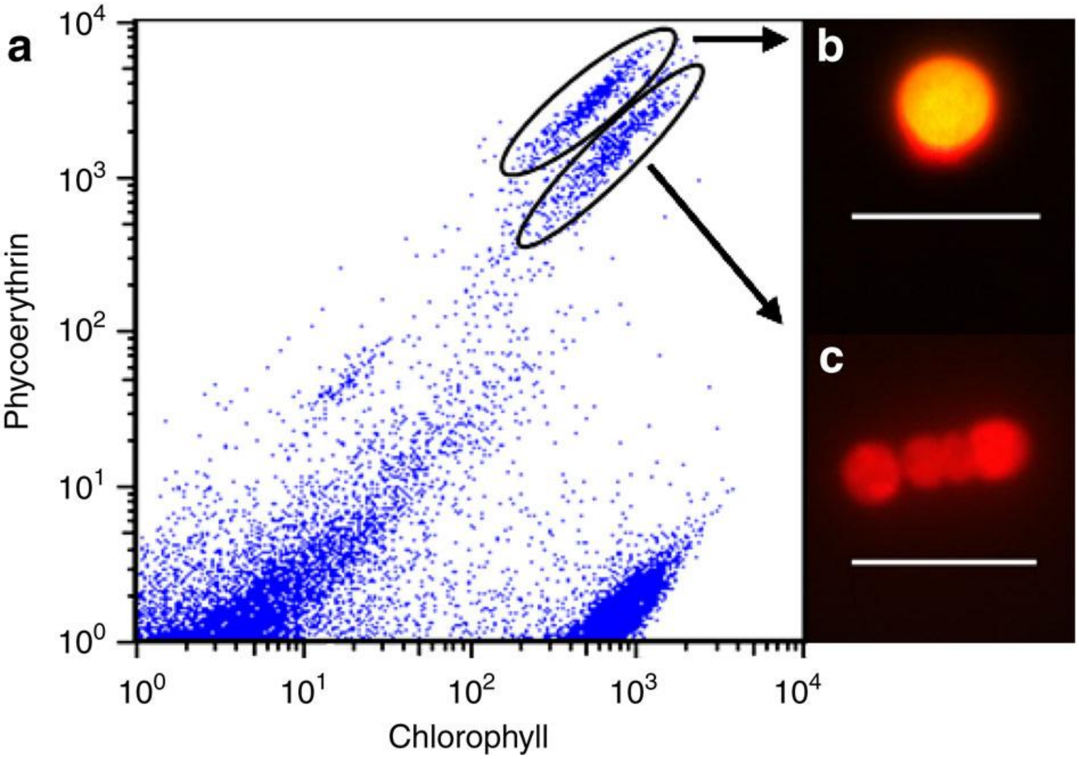


Figure 2. *Hemiaulus hauckii* symbionts on the flow cytometer. A cytogram displaying events gated based upon chlorophyll (692/40 nm) and phycoerythrin (572/27 nm) detection channels with the cyanobacteria symbiont populations easily separated (a). Insets are microscopy images under blue excitation of (b) a vegetative trichome and (c) a heterocyst cell representative of circled populations. Scale bars, 10 μm. The vegetative trichome population was sorted for genome sequencing samples.

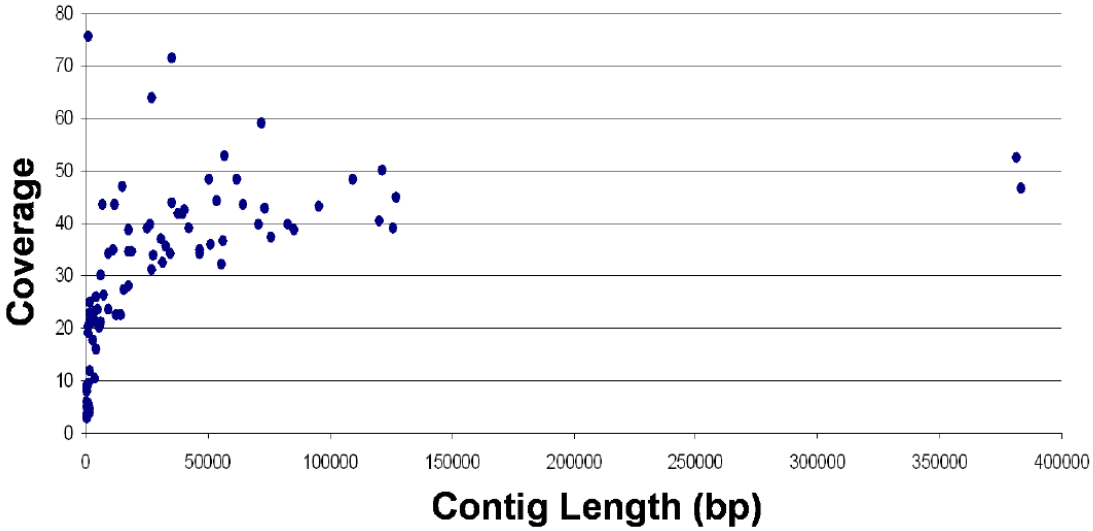


Figure 3. ***Richelia intracellularis*** **HH01 contig lengths.** The length and coverage depth of each of the 90 contigs that make up the *Richelia intracellularis* HH01 genome assembly.

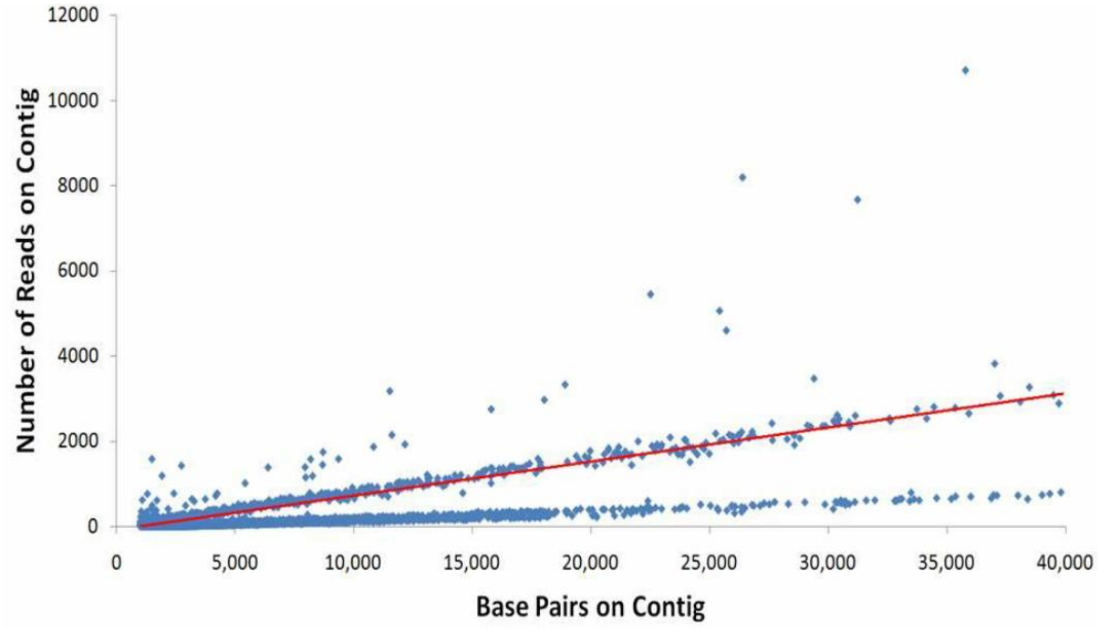


Figure 4. *Calothrix rhizosoleniae* SC01 contigs. Data from the initial assembly of 454 reads from *Ca. rhizosoleniae* SC01 samples (69,919 contigs, 81.4 Mb). For each contig, the number of reads assembled to each contig plotted against the length of the contig. The contigs lying along the red line (representing 30x coverage depth) are the ones presumed to be from *Ca. rhizosoleniae* SC01, based on BLAST (blastn) results.

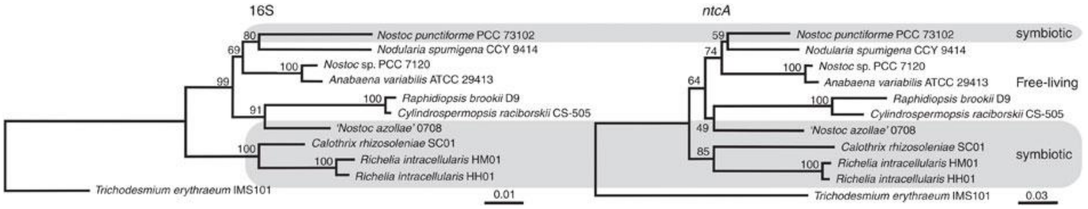

16S
80
69
99
100
91
100
100
100
Nostoc punctiforme PCC 73102
Nodularia spumigena CCY 9414
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS-505
'Nostoc azollae' 0708
Calothrix rhizosoleniae SC01
Richelia intracellularis HM01
Richelia intracellularis HH01
Trichodesmium erythraeum IMS101
0.01
ntcA
59
74
64
100
100
49
85
100
Nostoc punctiforme PCC 73102
Nodularia spumigena CCY 9414
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS-505
'Nostoc azollae' 0708
Calothrix rhizosoleniae SC01
Richelia intracellularis HM01
Richelia intracellularis HH01
Trichodesmium erythraeum IMS101
0.03
symbiotic
Free-living
symbiotic

Figure 5. Nostocales phylogeny. Neighbor-joining phylogenetic trees of 16S rRNA and *ntcA* sequences from seven previously sequenced cyanobacteria and three additional diatom symbionts from this study. The organisms observed in symbiotic relationships are shaded in grey. Both trees are rooted with *T. erythraeum* IMS101. Locus tags (16S, *ntcA*): '*N. azollae*' 0708 (Aazo_R0008, Aazo_1065), *Ra. brookii* (CRD_01297,CRD_00550), *Cy. raciborskii* CS-505 (CRC_01246, CRC_00858), *N.* sp. PCC 7120 (allrr01,alr4392), *N. spumigena* CCY9414 (N9414_r17988, N9414_19492), *N. punctiforme* PCC 73102 (Npun_r020, Npun_F5511), *A. variabilis* ATCC 29413 (Ava_R0006, Ava_3283), *T. erythraeum* IMS101 (Tery_R0014, Tery_2023), *Ca. rhizosoleniae* SC01 (CSC01_11477, CSC01_6586), *Ri. intracellularis* HH01 (RintHH_r10, RintHH_12150), *Ri. intracellularis* HM01 (RintHM_3660, RintHM_9700).

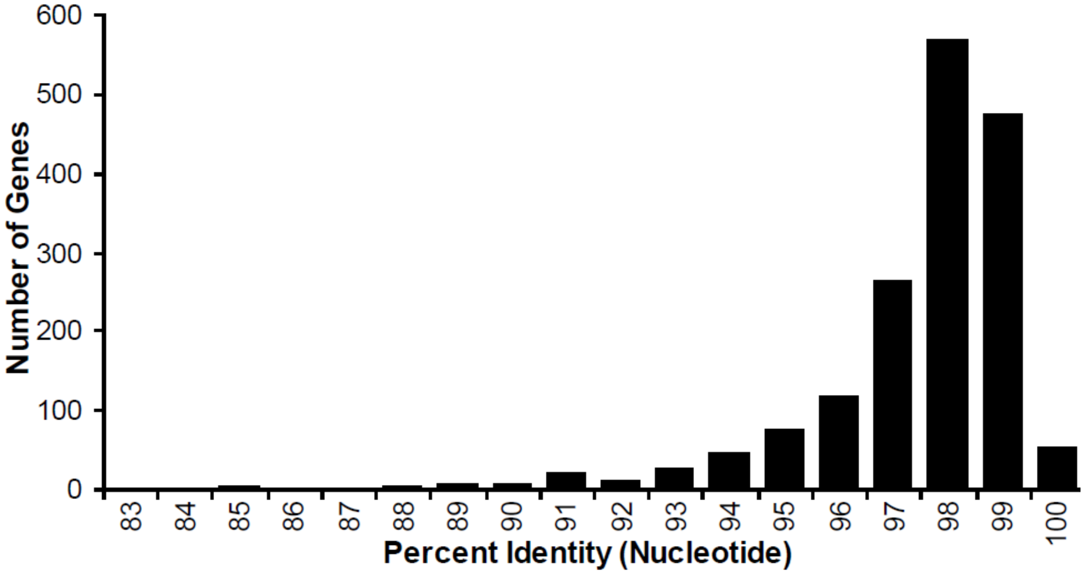


Figure 6. Gene similarity of *Hemiaulus* spp. symbionts. The percent identity at the nucleotide level of the 1,671 shared genes of *Ri. intracellularis* HM01 and HH01.

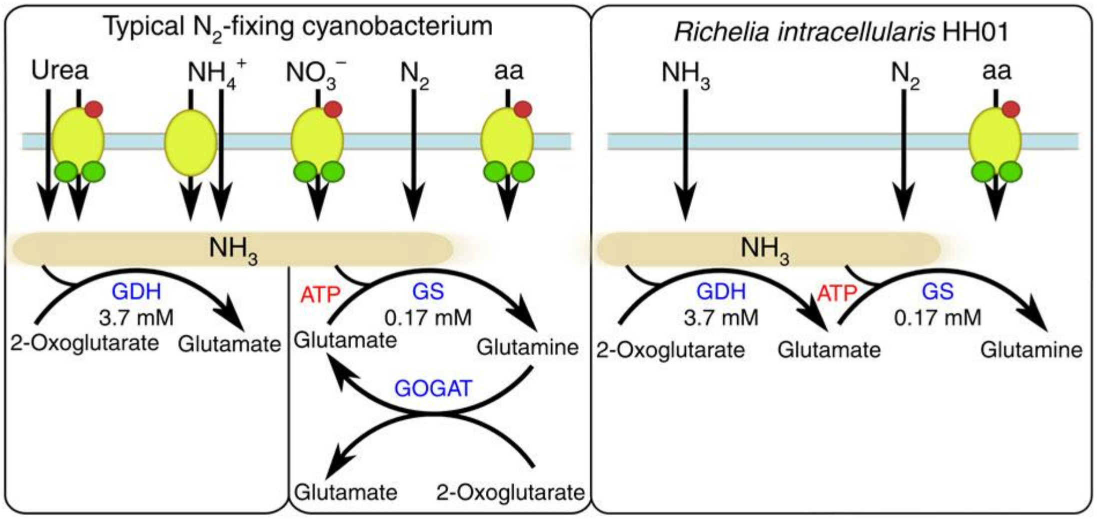


Figure 7. *Richelia intracellularis* HH01 nitrogen metabolism pathway. Nitrogen metabolism pathways common in N_2-fixing cyanobacteria compared with *Ri. intracellularis* HH01. Redrawn from **(Muro-Pastor & Florencio, 2003; Yan, 2007)**. K_m-values for GDH and GS are for ammonia in each reaction from the N_2-fixing cyanobacterium *Synechocystis* PCC 6803 **(Florencio *et al.*, 1987; Mérida *et al.*, 1990)**.

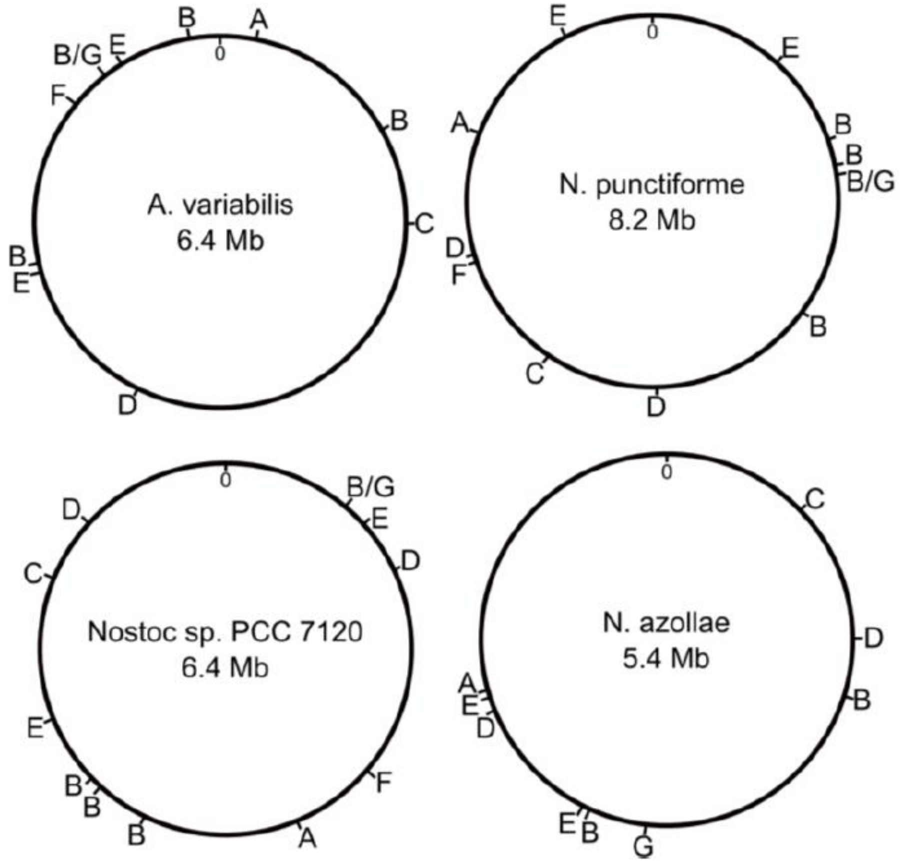


Figure 8. Nitrogen metabolism genes throughout closed genomes. The location of nitrogen metabolism-related genes across the four closed Nostocales genomes. A single letter may represent multiple genes (i.e. multiple adjacent nitrate transporters). A – GS inactivating factor 7, B – nitrate transporters, C – Fd-GOGAT, D – ammonium transporters, E – urease subunits, F – urea transporters, G – nitrate and nitrite reductase.

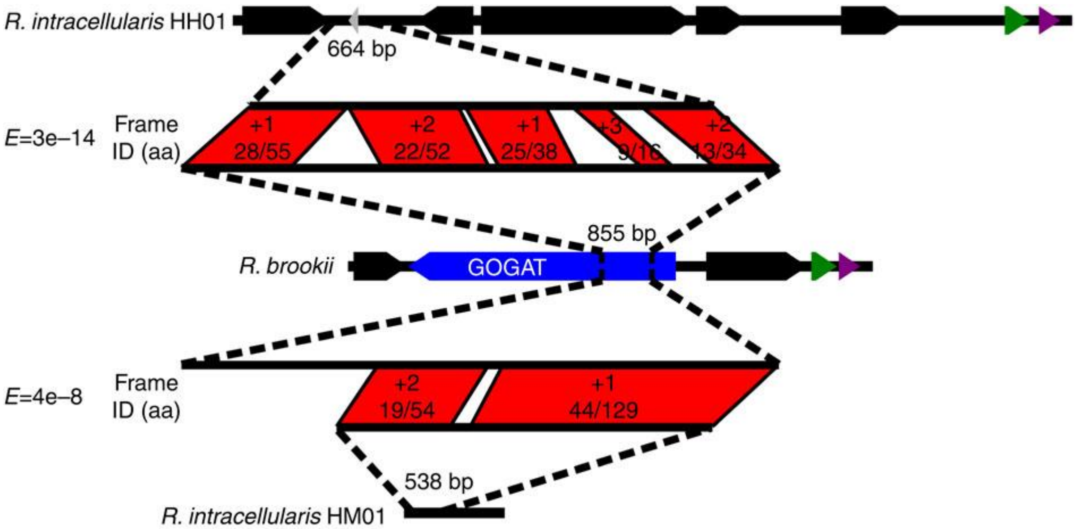


Figure 9. GOGAT remnants in *Richelia intracellularis* HH01. The intergenic space between two *Ri. intracellularis* HH01 ORFs, including an annotated 120 bp hypothetical protein (in grey, RintHH_15420) compared with the gene encoding GOGAT in *Raphidiopsis brookii* D9 (in blue, CRD_00957) and the genomic context of each. The middle alignment represents the top BLAST hit of the *Ri. intracellularis* HH01 intergenic space (1561, bp) in the nr database (blastx, e-value=3e^{-14}). The green and purple genes encode a mannose-6-phosphate isomerase and a probable iron binding protein, respectively.

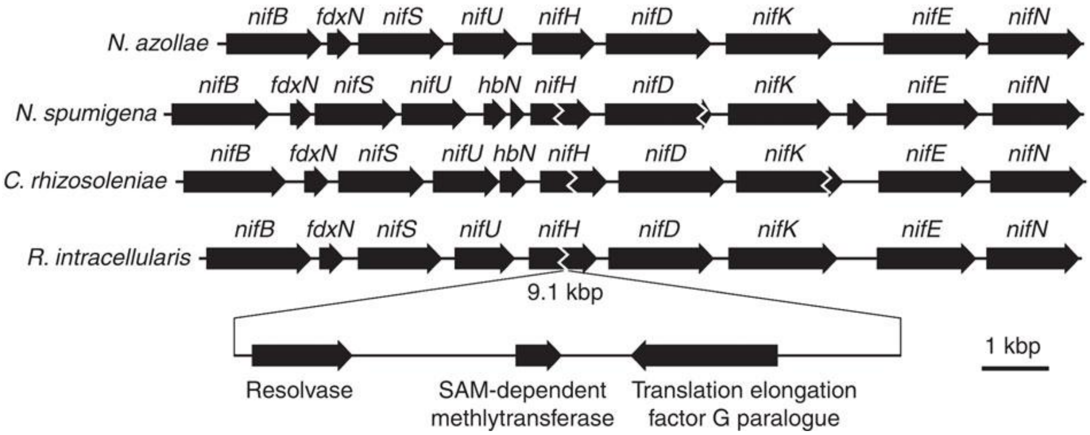


Figure 10. The *Richelia intracellularis* HH01 *nifH* element. The *nif* operon and surrounding genes of *Ri. intracellularis* HH01 and *Ca. rhizosoleniae* SC01 compared with other representative Nostocales cyanobacteria '*Nostoc azollae*' 0708 and *Nodularia spumigena* CCY9414, and the genes along the insertion element interrupting the *Ri. intracellularis* HH01 *nifH* gene.

Chapter 2

The genomic sequence of the *Rhizosolenia clevei* symbiont and comparison with other diatom symbionts

Abstract

Diatoms, a group of diverse unicellular algae, are commonly the dominant phytoplankton in high nutrient coastal waters, but some diatoms can thrive in oligotrophic waters by receiving fixed nitrogen from symbiotic N_2-fixing cyanobacteria. These symbioses are globally-significant for oceanic nutrient cycling and are commonly formed between diatoms of the genera *Hemiaulus, Rhizosolenia,* and *Chaetoceros* and cyanobacteria that form specialized N_2 fixation cells, called heterocysts. Previously, the genome of an internal *Hemiaulus* symbiont was shown to be reduced, reflective of the evolutionary pressure of an obligatory association, while an external *Chaetoceros* symbiont genome was more similar to free-living heterocyst-forming cyanobacteria. In the present report, the genome of an internal *Rhizosolenia* symbiont is shown to exhibit characteristics consistent with an obligate symbiont, although the obligatory nature of the association is likely more recent than that of *Hemiaulus*. Shared traits amongst all three diatom symbionts were also identified that reflect the influence of living in association with diatoms as well as inhabiting the open ocean environment. The genomic sequence of the *Rhizosolenia* symbiont represents an evolutionary intermediate between the two previously-sequenced diatom symbiont genomes and adds additional context for which to study the metabolic nature of bacterial symbioses, in general.

Introduction

Microbial productivity in much of the world's surface oceans is limited by the availability of fixed inorganic nitrogen (N) (Falkowski, 1997; Moore *et al.*, 2013). Some organisms, primarily cyanobacteria, are able to avoid N limitation by converting N_2 gas, the most abundant gas in the atmosphere, into ammonia (NH_3). These organisms, termed diazotrophs, support microbial growth by adding biologically-available N to the community, and fueling 'new' production (Eppley & Peterson, 1979; LaRoche & Breitbarth, 2005). Some diazotrophs even associate with other marine phytoplankton, including diatoms, and directly supply N to their partner (Carpenter & Foster, 2003; Foster *et al.*, 2011).

Diatoms, a diverse group of unicellular algae (Bowler *et al.*, 2008), normally comprise a large fraction of phytoplankton communities in nutrient-rich coastal waters (Goldman, 1993). Some diatoms, however, are able to thrive in oligotrophic waters by associating with, and receiving fixed N from, diazotrophic cyanobacteria (Heinbokel, 1986; Gomez *et al.*, 2005; Madhu *et al.*, 2013). Although a variety of cyanobacteria have been observed in association with diatoms (Carpenter & Janson, 2000; Rai *et al.*, 2000), the most common occurrences involve cyanobacteria that form specialized cells, termed heterocysts, for N_2 fixation. The enzyme that catalyzes N_2 fixation is sensitive to oxygen (O_2) (Wong & Burris, 1972), and heterocysts separate N_2 fixation from O_2 evolved from photosystem II (Wolk *et al.*, 2004).

These partnerships, termed diatom-diazotroph associations (DDAs), are found throughout global tropical and subtropical waters (Heinbokel, 1986; Villareal, 1994; Gomez *et al.*, 2005; Foster *et al.*, 2007, 2009; Madhu *et al.*, 2013), and have significant roles in supplying N to surface communities (Carpenter *et al.*, 1999; Wilson *et al.*, 2008; Padmakumar *et al.*, 2010) and exporting carbon (C) to the deep ocean (Scharek *et al.*, 1999; Subramaniam *et al.*, 2008; Karl *et al.*, 2012; Yeung *et al.*, 2012). Heterocyst-forming cyanobacteria have commonly been observed in association with three diatom genera: *Hemiaulus, Rhizosolenia*, and *Chaetoceros* (Heinbokel, 1986; Villareal, 1990; Janson *et al.*, 1999). The symbionts associated with each diatom genus are genetically distinct from the other diatom symbiont populations (Janson *et al.*, 1999; Foster & Zehr, 2006), and their cellular location on, or within, the diatom also varies depending on the genus of the diatom partner. The symbionts attach externally to *Chaetoceros* spp. (Janson *et al.*, 1999) and are found in the periplasmic space, within the frustule but external to the cell, of *Rhizosolenia* spp. (Villareal, 1990). The cyanobacteria associated with *Hemiaulus* spp. are also within the frustules (Heinbokel, 1986), but it is unclear if they are intra- or extra-cellular. Beyond that, little is known regarding how the three associations differ from each other.

The genomes of a *Hemiaulus hauckii*-associated symbiont (*Richelia intracellularis* HH01, referred to herein as RintHH) and a *Chaetoceros* sp.-associated symbiont (*Calothrix rhizosoleniae* SC01, CalSC) have been previously sequenced (Hilton et al. 2013). The CalSC genome characteristics resembled those of free-living heterocyst-forming cyanobacteria, while the RintHH genome exhibited significant genome reduction, implying it is an obligate symbiont. The streamlining was especially apparent in the N metabolism pathways, as RintHH was the first cyanobacterium found lacking a gene that encodes glutamate synthase (GOGAT). The differences between the genomes of external and internal diatom symbionts indicated the cellular location of the symbiont correlated with symbiont evolution and the host-dependency of the symbiont.

In the present study, the genomic sequence for the symbiont of *Rhizosolenia clevei*, a representative of the third common diatom-associated heterocyst-forming cyanobacterium is reported (Figure 1). The genome of *Richelia intracellularis* RC01 (RintRC) shares many features with both RintHH and CalSC, and adds resolution for which to study the nature of the diatom associations, and the evolution of symbionts.

Methods

Sample collection and DNA preparation

Rhizosolenia-Richelia cultures were isolated from the western Gulf of Mexico and grown in modified MET-44 with no added N resulting in ambient levels of DIN (~0.2 µM) (Villareal, 1990). Cultures were filtered on to a 3-µm pore size, 25-mm diameter polyester filter (Sterlitech) and frozen for storage. SO media (1 mL) was added and once the filter thawed, the cells were re-suspended by vortexing for 1 min. In order to ensure the symbionts had been re-suspended and were no longer within the diatoms (Figure 2), 10 µL were placed on a slide and observed under the microscope. The sample was then analyzed on the Influx flow cytometer and cell sorter (BD Biosciences) and cyanobacterial cells were distinguished from other cells by their phycoerythrin pigmentation. One hundred forty-four events were sorted into 60 µL

SO media, and 10 µL were removed for microscopy observation. One filament was seen on this slide along with twelve single heterocysts. Genomic DNA in the sorted sample was then amplified by multiple displacement amplification using the Repli-g Midi kit (Qiagen). The manufacturer's protocol for 0.5 µL of cell material was followed with one exception: after Buffer D2 was added, the samples were incubated for 5 min at 65ºC and then put on ice for 1 min.

To ensure an uncontaminated sample, the amplified DNA sample was rRNA genes were amplified using universal 16S rRNA primers 27F (5'-AGAGTTTGATCMTGGCTCAG-3') and 1492R (5'-GGTTACCTTGTTACGACTT-3') (Lane, 1991). The PCR was carried out in 50 µL reactions consisting of 1x PCR Buffer, 2 mM $MgCl_2$, 200 µM dNTPs, 0.2 µM of each primer, and 1.5 U of Platinum Taq DNA polymerase (Invitrogen). A touchdown PCR was performed as follows: an initial denaturing step at 94°C for 5 min, followed by 30 cycles of three 1-min steps (denaturation at 94°C, annealing at 53-41°C, and elongation at 72°C), and a final elongation step at 72°C for 10 min. The first cycle annealing took place at 53°C, and was lowered by 0.4°C each cycle to reach 41°C for the final cycle. The resulting product was run on a 1.2% agarose gel, the distinct band of ~1500 bp was excised, and recovered using the Zymoclean DNA Recovery Kit (Zymo Research). The recovered DNA was then ligated and plated for blue/white screening using the pGem-T and pGem-T Easy Vectors Systems (Promega). Fifty colonies were picked and grown over-night at 37°C in 2xLB media with carbenicillin (200 µg/mL). The Montáge Plasmid Miniprep$_{HTS}$ Kit (Millipore) was used following manufacturer’s instructions for the full lysate protocol for plasmid DNA miniprep. Samples were sequenced at UC-Berkeley DNA Sequencing Facility and each sequence was subjected to BLAST analysis against the nr/nt database (blastn). All sequences were identical to each other and had top hits to 16S rRNA sequences of heterocyst-forming cyanobacteria.

DNA concentration and quality were checked (Agilent 2100 Bioanalyzer, Agilent Technologies) prior to submission for 454 Titanium sequencing (Roche) at GenoSeq, the UCLA Genotyping and Sequencing Core.

Genome assembly

A total of 611,821 reads were sequenced (379 Mb) and assembled with the GS *De Novo* Assembler (Roche) to 8.7 Mb. BLAST analysis of the 2,382 contigs against the nr/nt database (blastn, NCBI) revealed a large number of contigs with top hits to humans or other primates (1,305 contigs, 3.0 Mb), indicating significant contamination in the sequence data. Thus, RintRC contigs were defined as those which had a better BLAST hit to a cyanobacterium than to any other organism in the nr/nt database (blastn). The resulting draft genome consisted of 857 contigs totaling 5,488,377 bp (Table 1) with a sequencing coverage depth of 42x (CBZS010000001-CBZS010000857).

Genomic analysis

After assembly and contamination screening, the genomes were submitted to RAST (Rapid Annotation using Subsystem Technology) (Aziz *et al.*, 2008) for annotation.

Annotations were searched for the N metabolism genes that were lacking in the RintHH genome (Table 2) (Hilton *et al.*, 2013). For the genes not located in the RintRC annotations, the orthologs were compiled from each Nostocales genome, and each gene was subjected to BLAST analysis against a database of all 611,821 reads (tblastx, e-value<10, >25% identity). No read had a BLAST hit that was at least 50% the length of the read or gene, whichever was shorter.

Predicted open reading frames (ORFs) in the genome that had a significant BLAST hit (blastp, e-value $<10^{-19}$) in the Transporter Classification database (Saier *et al.*, 2009) were counted as transporter genes.

Results and Discussion

Genome characteristics

The RintRC genome has a reduced genome relative to most heterocyst-forming cyanobacteria, but to a lesser extent than RintHH. The coding percentage of the RintRC genome (65%), as well as the transporter count (240), are among the lowest among the genomes of heterocyst-forming cyanobacteria (Table 1). The similarity of the RintRC percent coding to that of RintHH and *Nostoc azollae* '0708', the obligate symbiont of the water fern *Azolla filiculoides*, implies the association with *Rhizosolenia clevei* is obligate. The low RintRC transporter count is similar to that of RintHH. This indicates that neither symbiont population is exposed to the oceanic environment at any stage of their life cycle. However, the RintRC genome size (5.5 Mb) and GC content (39%) are not reduced to the extent as in RintHH, but are similar to those in CalSC and the majority of genomes from heterocyst-forming cyanobacteria (Table 1). Furthermore, *N. azollae* '0708' is the only heterocyst-forming cyanobacterium that contains more transposases (532) than RintRC (297) (Table 1). Increased symbiont transposase counts are often seen in the initial stages of symbiosis, possibly providing a mechanism for gene inactivation and deletion, and facilitating genome reduction (Moran & Plague, 2004). Thus, the general symbiont genome characteristics indicate that the *Rhizosolenia-Richelia* association is likely obligate for the symbiont, but is at an earlier evolutionary stage than the *Hemiaulus-Richelia* symbiosis.

Nitrogen metabolism

The genome reduction previously seen in RintHH was especially noticeable in the N metabolism pathways, and RintRC shares some of the same N acquisition deletions (Figure 3). Similar to the streamlined RintHH genome, the RintRC genome lacks transporters for ammonium and nitrate, and nitrate and nitrite reductase. These deletions prevent the symbionts from utilizing fixed N sources from the environment, thus ensuring continuous N_2 fixation and persistence of the symbiotic association. On the other hand, the exposure of CalSC directly to the open ocean environment likely provides the evolutionary pressure to preserve the N acquisition genes within the genome. The differences in N acquisition capabilities reflect the environments in which each of the symbionts inhabit and distinguish the external symbionts from those located within the diatom frustule.

The RintRC genome reduction does not extend to the streamlining of the N assimilation pathways that was observed in RintHH. Similar to CalSC, the RintRC genome has N assimilation genes encoding glutamate synthase (GOGAT) and glutaminase that are found in all heterocyst-forming cyanobacteria except RintHH. GOGAT uses glutamine, produced by glutamine synthetase (GS), and 2-oxoglutarate to produce two glutamate molecules, one of which is recycled to be utilized in ammonium assimilation by GS. This pathway is known as the GS-GOGAT cycle and has been identified as the main N assimilation pathway in cyanobacteria (Muro-Pastor *et al.*, 2005). Glutaminase also produces glutamate from glutamine, but does not add the freed amide group onto a C skeleton as GOGAT does, and very little is known about the role of glutaminase in N metabolism in cyanobacteria (Zhou *et al.*, 2008). The difference in N metabolism capabilities may indicate that different mechanisms are present for N exchanges between the *Hemiaulus* association and the other two common diatom symbioses. This variation from RintHH is yet another example that the RintRC symbiont has had less evolutionary pressure to streamline its genome, possibly due to a more recent development of the *Rhizosolenia* association, relative to *Hemiaulus* associations.

Common deletions within N assimilation pathways distinguish all of the diatom symbionts from other heterocyst-forming cyanobacteria. None of the diatom symbionts have urea transporters or urease, the enzyme responsible for the acquisition and assimilation of urea. Genes encoding these proteins are present in all other heterocyst-forming cyanobacteria except for *Calothrix* sp. PCC 7507. The internal diatom symbionts, RintHH and RintRC, are likely not exposed to any urea present in the environment, but it is unclear why the external symbiont CalSC would not retain these genes. Urea can be an important N source for oceanic cyanobacteria (Moore *et al.*, 2002; Glibert *et al.*, 2004; Heil *et al.*, 2007), but the terrestrial environment is a main source for urea in the water column (Glibert *et al.*, 2006). Thus, urea may not an important N source in the subtropical Pacific Ocean where the CalSC associations are commonly found (Marumo & Asaoka, 1974; Gomez *et al.*, 2005; Kitajima *et al.*, 2009), or this may be an initial deletion towards streamlining N metabolism in the external symbiont. Additionally, RintHH, RintRC, and CalSC each lack *gifA*, the gene that encodes a protein-to-protein GS inactivating factor (GSIF) and is present in all other heterocyst-forming cyanobacteria except *Calothrix* sp. PCC 6303. *Synechocystis* sp. PCC 6803 mutants incapable of forming GSIFs could not inactivate GS, leading to an increase in glutamine through GS activity (Muro-Pastor *et al.*, 2001). The lack of GSIFs in the diatom symbionts, and subsequent elevated glutamine pools, implies that glutamine could be the form of N transferred to the diatom in each symbiosis. The absence of key N metabolism genes shows parallels between the symbionts and the associations they form with diatoms.

Additional unique DDA gene presence/absence

The N metabolism streamlining likely affects other metabolic pathways in the diatom symbiont genomes. RintHH, RintRC, and CalSC, along with *Mastigocoleus testarum* BC008, are the only sequenced heterocyst-forming cyanobacteria that possess genes encoding agmatine deiminase (AgD) and N-carbamoylputrescine amidase (NCPA). However, the presence of three genetically-distinct 16S rRNA sequences in the *Mastigocoleus testarum* BC008 assembly, and similarity of its AgD and NCPA genes to non-cyanobacteria, suggests these sequences are likely from a contaminant organism. AgD and NCPA form a two-step synthesis of putrescine from agmatine, while producing one ammonia molecule per step. Agmatinase carries out the same reaction in one step, but produces urea instead of ammonia, and is found all heterocyst-forming cyanobacteria except the diatom symbionts. In the diatom symbionts, agmatinase activity would lead to a build-up of urea, which could not be assimilated without urease. Thus, the metabolic substitution in the diatom symbionts is likely due to the streamlining of the N metabolism pathways.

The absence of some genes in all three diatom symbiont genomes indicates that similarities amongst all DDAs has led to the deletion of genes made dispensable through the nature of the associations. RintHH, RintRC, and CalSC were the only heterocyst-forming cyanobacteria that lack genes encoding gas vesicle proteins. The absence of gas vesicles in *Rh. clevei* symbionts has been noted (Carpenter & Janson, 2000), and this supports the hypothesis that buoyancy is a key advantage for the cyanobacteria that results from forming associations with diatoms, thus rendering gas vesicles expendable (Padmakumar *et al.*, 2010). Additionally, RintHH, RintRC, and CalSC were each lacking *hglT* (formerly referred to as *all5341*), a glycosyltransferase gene that is found in all other heterocyst-forming cyanobacteria except for *Cylindrospermopsis raciborskii* CS-506. Without a functional *hglT*, *Nostoc* sp. PCC 7120 mutants were unable to form a heterocyst glycolipid layer and could not fix N_2 in the presence of oxygen (Awai & Wolk, 2007). The lack of this gene in diatom symbionts is consistent with results from a study that found *Hemiaulus* spp.-associated cyanobacteria synthesized different glycolipids than free-living cyanobacteria, possibly as a result of elevated oxygen levels within the diatom (Schouten *et al.*, 2013). A BLAST analysis of glycosyltransferase genes against all heterocyst-forming cyanobacteria reveals that the diatom symbiont genomes are the only ones without *all0143*, and they are three of the five that are missing *alr0776*. No studies have been done on the specific functions of *all0143* or *alr0776*, but each has been identified as a family 2 glycosyltransferase (Yang *et al.*, 2007). However, each is also found in non-heterocyst-forming cyanobacteria, and thus, is likely not tied directly to heterocyst glycolipid layers like *all5341*. The lack of gas vesicle proteins and key glycosyltransferase genes are examples of the results of the evolutionary selection in diatom symbionts, relative to other heterocyst-forming cyanobacteria.

Some shared characteristics of the diatom symbionts reflect influences of living in the oceanic environment, rather than being in association with diatoms. The RintHH, RintHM, and CalSC genomes each contain a gene annotated as a phycocyanobilin lyase alpha subunit (*pecE*), but do not have the gene for the corresponding beta subunit (*pecF*), as seen in all other heterocyst-forming cyanobacterial genomes. However, the protein sequences of the annotated alpha subunits are actually more similar to a PecE/F fused protein first discovered in *Synechococcus* sp. Strain WH8102 (Six *et al.*, 2005). In addition to the diatom symbionts, the fused lyase gene, *rpcG*, is present in eight cyanobacterial genomes (four *Synechococcus* genomes, three *Crocosphaera* strains, and *Rubidibacter lacunae*). All eight of these organisms are oceanic cyanobacteria, thus, it has been hypothesized that the fused gene provides an advantage in oceanic waters (Blot *et al.*, 2009). A phycocyanin-associated rod-capping polypeptide linker gene is found adjacent to *pecE* and *pecF* in nearly every heterocyst-forming cyanobacteria, but this gene is not located anywhere in the RintHH, RintRC, or CalSC genomes. The four marine *Synechococcus* genomes that possess *rpcG* also lack the rod-capping linker gene, supporting the hypothesis that these genetic traits are additional adaptations to the blue-green light that is prevalent in the open ocean. The alteration of light harvesting proteins in the diatom symbionts sets them apart from other heterocyst-forming cyanobacteria, and these differences are likely due to environmental factors, rather than symbiotic associations.

A metabolic substitution in the diatom symbionts may be a result of nutrient limitation caused by a combination of the oceanic environment and their associations. The RintHH, RintRC, and CalSC genomes each contain a Ni-dependent superoxide dismutase (NiSOD) that is only found in two other heterocyst-forming cyanobacteria: *Mastigocoleus testarum* BC008 and *Rivularia* sp. PCC 7116. The diatom symbionts additionally contain a Mn-dependent SOD (MnSOD) like most other heterocyst-forming cyanobacterial genomes, but lack an Fe-dependent SOD (FeSOD) that is found in all other heterocyst-forming cyanobacterial genomes. Cyanobacteria create the superoxide radical (O_2^-) through several cellular processes, and SODs reduce O_2^- in order to prevent its toxicity from damaging the cell (Fay, 1992). FeSOD gene transcript activity in a filamentous cyanobacterium has been shown to be dependent on Fe availability (Campbell & Laudenbach, 1995). The diatom symbionts have been shown to fix much more N than is required for their own growth (Foster *et al.*, 2011). It is likely the high Fe demand of this N_2 fixation (Rueter, 1988), coupled with scarce Fe availability in the surface oceans (Fung *et al.*, 2000), has led to the preservation of a NiSOD gene over FeSOD in the diatom-associated cyanobacteria. The unique associations formed by diatom symbionts and the low nutrient environments they inhabit each represent an evolutionary pressure not experienced by the majority of other heterocyst-forming cyanobacteria, and these differences are apparent in the metabolism of diatom symbionts.

As with any genome assembly that is not closed, it is possible some of the genes not found in the RintRC draft genome are present in the genome, but were not sequenced. The draft genome of RintRC does not contain a tRNA for glutamic acid, global nitrogen regulatory gene *ntcA*, or several nitrogenase genes that are conserved in diazotrophs. A BLAST analysis revealed one unassembled read that was 90% identical (nucleotide) to the RintHH *ntcA* gene, confirming its presence in RintRC.

Thus, the current RintRC assembly may be somewhat incomplete due to low coverage in some regions.

Conclusions

The sequencing of the *Rhizosolenia clevei* symbiont provides a representative genome from a phytoplankton group that is significant to oceanic nutrient cycling. Comparison with the two previously sequenced diatom symbiont genomes revealed the *Rhizosolenia clevei* symbiont is an evolutionary intermediate, sharing general genome characteristics and N metabolism capabilities with each of the *Hemiaulus* and *Chaetoceros* symbionts. The *Rhizosolenia* symbiont is likely obligate, but has entered a symbiotic state more recently than the *Hemiaulus* symbiont. However, all three diatom symbionts have traits in common that separate them from all other heterocyst-forming cyanobacteria. This comparative analysis showed that these metabolic differences can be a result of the diatom associations or from inhabiting the low nutrient open ocean. The *Rhizosolenia* symbiont genome provides a snapshot of symbiont evolution and provides more resolution for which to study how partnerships between bacteria and eukaryotic partners are formed and evolve over time.

Acknowledgements

This work was sponsored in part by the NSF and the NSF Center for Microbial Oceanography: Research and Education. We are grateful to B. Carter, R. Foster and J. Tripp for technical help and discussions, and J. Meeks for comments on the manuscript.

Tables and Figures

Table 1. Nostocales genome statistics. The genome characteristics of the 28 genomes from nostocalean cyanobacteria available at the time of this study. TC - transporter classification.

Cyanobacterium	Genome Size	GC %	Coding %	TC Count	Transposase Count
Scytonema hofmanni PCC 7110	12.0	41.5	82.3	773	78
Calothrix sp. PCC 7103	11.6	38.5	82.5	720	79
Calothrix desertica PCC 7102	11.4	38.5	80.7	730	81
Nostoc punctiforme PCC 73102	9.1	41.4	77.4	558	67
Rivularia sp. PCC 7116	8.7	37.5	78.5	566	32
Scytonema hofmanni UTEX 2349	8.1	41.1	80.7	512	62
Cylindrospermum stagnale PCC 7417	7.6	42.2	80.2	492	73
Nostoc sp. PCC 7120	7.2	41.3	82.5	589	47
Anabaena variabilis ATCC 29413	7.1	41.4	82.3	553	108
Anabaena cylindrica PCC 7122	7.1	38.8	80.3	481	30
Calothrix sp. PCC 7507	7.0	42.2	79.6	532	21
Calothrix sp. PCC 6303	7.0	39.8	79.3	457	32
Nostoc sp. PCC 7524	6.7	41.5	82.2	523	28
Nostoc sp. PCC 7107	6.3	40.4	81.3	479	26
Calothrix rhizosoleniae SC01	6.0	39.5	76.5	406	45
Anabaena sp. PCC 7108	5.9	38.8	80.8	427	26
Microchaete sp. PCC 7126	5.7	42.2	81.5	454	22
Richelia intracellularis RC01	5.5	39.2	64.9	240	297
Nostoc azollae 0708	5.5	38.4	52.1	337	532
Nodularia spumigena CCY9414	5.3	41.3	81.8	431	74
Anabaena sp. 90	5.1	38.1	83.3	328	27
Anabaena circinalis AWQC131C	4.4	37.0	82.1	339	8
Anabaena circinalis AWQC310F	4.4	37.3	81.9	342	6
Cylindrospermopsis raciborskii CS-506	4.2	41.7	86.1	448	2
Cylindrospermopsis raciborskii CS-509	4.0	41.3	85.2	395	4
Cylindrospermopsis raciborskii CS-505	3.9	40.2	84.9	353	59
Richelia intracellularis HH01	3.2	33.7	55.8	190	2

Table 2. Nitrogen metabolism genes in three diatom symbiont genomes. The presence and absence of key nitrogen metabolism genes in *Richelia intracellularis* compared to the two previously-sequenced diatom genomes. x - not present

	***Richelia* HH01**	***Richelia* RC01**	***Calothrix* SC01**
NH_4^+ transporter	x	x	present
NO_3 transporter & reductase	x	x	present
NO_2^- reductase	x	x	present
Urea transporter & urease	x	x	x
GS inactivating factor	x	x	x
GOGAT	x	present	present
Glutaminase	x	present	present

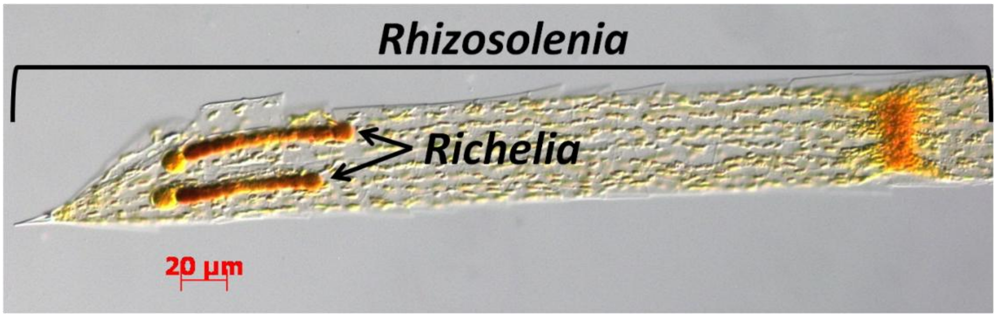


Figure 1. *Richelia* in association with *Rhizosolenia*. Bright-field microscopy image of cyanobacterial symbiont *Ri. intracellularis* RC01 within the frustule of *Rh. clevei*.

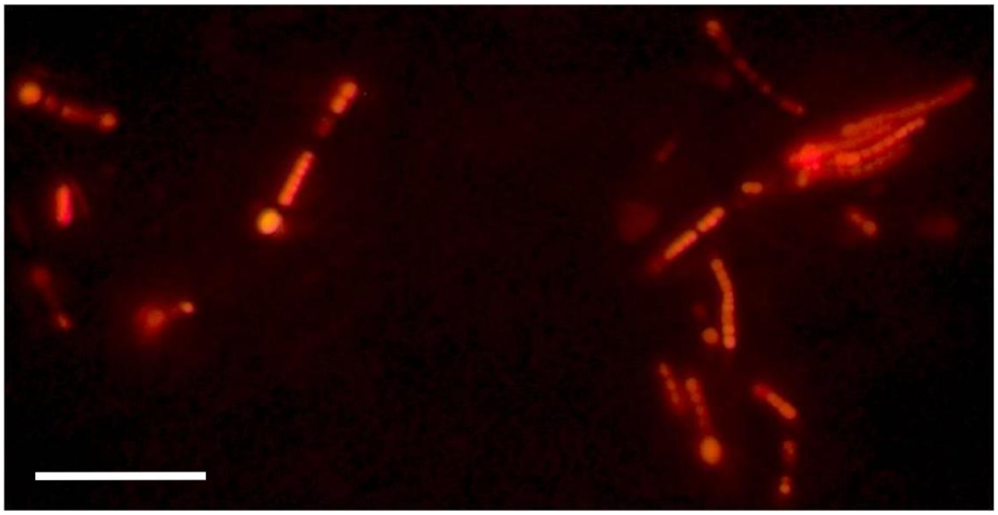


Figure 2. ***Richelia intracellularis* RC01.** Microscopy image of *Ri. intracellularis* RC01 filaments under blue excitation after removal from within *Rh. clevei* frustules. Scale bar is 100 µm.

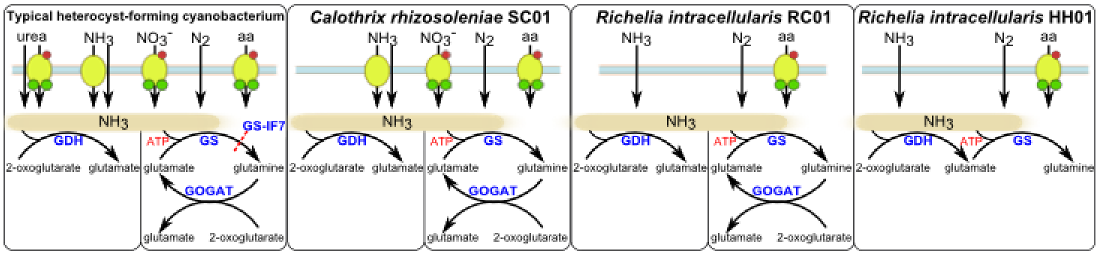


Figure 3. Nitrogen metabolism pathways of diatom symbionts. Nitrogen metabolism pathways of *Ca. rhizosoleniae* SC01, *Ri. intracellularis* RC01, and *Ri. intracellularis* HH01 compared with other heterocyst-forming cyanobacteria. Redrawn from **(Muro-Pastor & Florencio, 2003; Yan, 2007)**.

Chapter 3

Metatranscriptomics of N_2-fixing cyanobacteria in the Amazon River plume

Abstract

Biological N_2 fixation is an important nitrogen (N) input for surface ocean microbial communities. However, nearly all information on the diversity of organisms responsible for oceanic N_2 fixation and their gene expression in the environment has come from targeted approaches that assay only a small number of genes and organisms. Using sequenced genomes of diazotrophic cyanobacteria to extract reads from extensive meta-genomic and -transcriptomic libraries, diazotroph diversity and gene expression were examined from the Amazon River plume, an area characterized by large salinity and nutrient gradients. Diazotroph genome and transcript sequences (raw and normalized data) were most abundant in the transitional waters compared to lower salinity or oceanic water masses. Through variations in sequence identities, two genetically-divergent phylotypes were distinguished within the *Hemiaulus*-associated *Richelia* sequences, which were the most abundant diazotroph sequences in the data set. Photosystem II transcripts in *Richelia* populations were much less abundant than those in *Trichodesmium*, and transcripts belonging to several *Richelia* photosystem II genes were absent. Additionally, there were several abundant regulatory transcripts, including one that targets a gene involved in photosystem I cyclic electron transport in *Richelia*. High sequence coverage of the *Richelia* transcripts, as well as those from *Trichodesmium* populations, allowed for the identification of expressed regions of the genomes that had been overlooked by genome annotations. High-coverage genomic and transcription analysis enabled the characterization of distinct phylotypes within

diazotrophic populations, revealed a distinction in a core process between dominant populations, and provided evidence for a prominent role for non-coding RNAs in microbial communities.

Introduction

The productivity of a large fraction of the ocean's surface waters is limited by the availability of fixed inorganic nitrogen (N) (Zehr & Kudela 2011). Some organisms, termed diazotrophs, have the ability to assimilate, or fix, N_2 gas, thus avoiding N limitation. N_2 fixation is an important source of ‘new’ N to maintain primary production in oligotrophic oceans (Dugdale & Goering 1967).

Diazotrophic cyanobacteria have been shown to comprise a large fraction of microbial communities in the Amazon River plume and surrounding waters (Foster et al. 2007; Goebel et al. 2010). As the high-nutrient riverine water mixes with oligotrophic oceanic waters, NO_3 and NO_2 are rapidly taken up by microbial communities dominated by coastal diatoms (Shipe et al. 2007; Subramaniam et al. 2008; Goes et al. 2014). Further along the mixing gradient, some nutrients (SiO_3, PO_4, Fe) persist in relatively high concentrations, but N is depleted, providing an advantage to the diazotrophs (Foster et al. 2007; Shipe et al. 2007; Subramaniam et al. 2008; Goes et al. 2014). The cyanobacterium *Richelia,* located within the cell wall of the diatom *Hemiaulus*, is the most abundant N_2-fixer in transitional waters (30-35 psu), while the colony-forming, filamentous *Trichodesmium* is the dominant diazotroph in more oceanic waters (>35 psu) (Carpenter et al. 1999; Subramaniam et al. 2008). The free-living unicellular cyanobacterium *Crocosphaera*, the picoeukaryotic alga-associated UCYN-A, and *Richelia* associated with the diatom *Rhizosolenia* have also been detected in and around the Amazon River plume (Foster et al. 2007; Goebel et al. 2010).

The abundance of diazotrophic cyanobacteria strongly influences surface communities and nutrient cycling in this area. A bloom of *Richelia*-harboring *Hemiaulus* in transitional waters, accompanied by *Trichodesmium*, accounted for an estimated input of nearly 0.5 Tg N to the surface community over just a 10 day period (Carpenter et al. 1999). Another study found that the particulate export at transitional stations was dominated by *Richelia-Hemiaulus* associations which were estimated to be responsible for the sequestration of 20 Tg Carbon (C) to the deep ocean annually (Subramaniam et al. 2008). These studies show the significance of the Amazon River plume diazotroph community, as a whole, but provide little information about the organisms that comprise the populations within that community.

Prior studies of oceanic diazotroph diversity, abundance, and activity have mostly been based on microscopic observations or molecular biology methods targeting a specific gene (e.g. *nifH*, *hetR*). In contrast, metatranscriptomics avoid potential bias stemming from targeting predetermined organisms or processes while providing a full transcription snapshot of microorganisms comprising the entire microbial community. Studying metatranscriptomes of marine microbial communities, in general, have revealed the abundance of novel transcripts and small RNAs (sRNAs) (Gilbert et al. 2008; Shi et al. 2009), the intricacies of diatom population response to iron limitation (Marchetti et al. 2012), and the synchronicity of diel transcription amongst bacterial and archaeal populations (Ottesen et al. 2013). Additionally, sequences implicating a novel bacterial group and a euryarchaeal population in deep sea nitrogen and carbon (C) cycling were found to be abundant in a Gulf of California metatranscriptome (Baker et al. 2013).

Although more community-based research is enabled through the use of metatranscriptomes, only a few studies have utilized this tool to elucidate the physiological state of cells within diazotrophic populations. Important information such as the expression of key nutrient limitation response genes, as well as highly-expressed genes of unknown function, were obtained from metatranscriptomic analyses of *Crocosphaera* (Hewson, Poretsky, Beinart, et al. 2009) and *Trichodesmium* populations (Hewson, Poretsky, Dyhrman, et al. 2009). In the current study, metatranscriptomic and metagenomic approaches were coupled to analyze the N_2-fixing community that drives new production in the Amazon River plume.

Materials and methods

Sample collection

Samples were collected in May-June, 2010 as part of the Amazon Influence on the Atlantic: Carbon Export from Nitrogen Fixation by Diatom Symbioses (ANACONDAS) project. Surface waters were sampled aboard the R/V *Knorr* from four stations (Figure 1). Samples were taken in duplicate for each of the sample types described below (DNA, RNA, and poly(A)-RNA) and pre-filtered (156 μm) before filtration through a 2.0 μm pore-size, 142 mm diameter polycarbonate membrane filter (Sterlitech Corporation, Kent, CWA). For all samples but the poly(A)-RNA, the

2.0 μm filter was in-line with a 0.22 μm pore-size, 142 mm diameter Supor membrane filter (Pall, Port Washington, NY). Immediately after filtration, and within 30 min of water collection, filters were stored in RNAlater (Applied Biosystems, Austin, TX). They were incubated overnight at room temperature, and stored at -80°C.

Sample preparation for DNA sequencing

DNA extraction and purification was conducted as previously described (Zhou et al. 1996; Crump et al. 1999, 2003) with some modification. Briefly, once each filter thawed, it was removed from RNAlater. In order to clean any residual RNAlater, the filter was rinsed three times in autoclaved, filter-sterilized, 0.1% phosphate-buffered saline (PBS). The liquid from the rinses was pooled with the RNAlater used for storage and pushed through a 0.2 μm Sterivex-GP filter capsule (Millipore). To shatter the sample filters, they were placed in Whirl-Pak® bags (Nasco, Fort Atkinson, WI), flash-frozen in liquid nitrogen, and broken into small pieces using a rubber mallet. The filter pieces were then placed in a tube containing DNA extraction buffer [DEB: 0.1 M Tris-HCl (pH 8), 0.1 M Na-EDTA (pH 8), 0.1 M $Na_2H_2PO_4$ (pH 8), 1.5 M NaCl, 5% CTAB]. The filter used for the RNAlater and rinse liquid was sliced into pieces and added to the DNA extraction buffer with the original membrane filter. An internal genomic DNA standard was also added as a means to normalize sequencing coverage across samples (Gifford *et al.*, 2010), as discussed further below. The samples were then treated with proteinase-K, lysozyme, and sodium dodecyl sulfate. The DNA was purified via phenol:chloroform extraction and isopropanol precipitation.

Sample preparation for total community RNA

RNA extraction and DNA removal were carried out as previously described (Gifford et al. 2010; Poretsky, Gifford, et al. 2009; Poretsky, Hewson, et al. 2009). In brief, after the filters were broken, as described above for DNA sample filters, they were transferred to a lysis solution consisting of 8 mL of RLT Lysis Solution (Qiagen, Valencia, CA), 3 grams of RNA PowerSoil beads (Mo-Bio, Carlsbad, CA), and internal standards. Tubes containing the filter pieces and lysis solution were vortexed for 10 min, and RNA was purified from cell lysate using the RNeasy Kit (Qiagen, Valencia, CA). To remove residual DNA, the Turbo DNA-free kit (Invitrogen, Carlsbad, CA) was used in two successive treatments. Ribosomal RNA (rRNA) was removed using community-specific probes prepared with DNA from a simultaneously-collected sample (Stewart et al. 2010). Biotinylated-rRNA probes were created for bacterial and archaeal 16S and 23S rRNA and eukaryotic 18S and 28S rRNA, and probe-bound rRNA was removed via hybridization to streptavidin-coated magnetic beads (New England Biolabs, Ipswich, MA). Successful removal of rRNA from the samples was confirmed using either an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA) or a Bioanalyzer (Agilent Technologies, Santa Clara, CA). Samples were then linearly amplified using the MessageAmp II-Bacteria Kit (Applied Biosystems, Austin, TX). Random primers were used with the Superscript III First Strand synthesis system (Invitrogen, Carlsbad, CA) to convert the amplified mRNA to cDNA, followed by the NEBnext mRNA second strand synthesis module (New England Biolabs, Ipswich, MA). The QIAquick PCR purification kit (Qiagen, Valencia, CA) was used to purify the double-stranded cDNA, followed by ethanol precipitation. The nucleic acids were resuspended in 100 μL of TE buffer and stored at -80^{o} C.

Sample preparation for poly(A)-tail-selected RNA

An additional metatranscriptome protocol that selectively sequenced RNA sequences with poly(A)-tails was conducted on the 2.0 μm pore-size filter samples only. The samples were prepared as described above for the total community RNA samples with the following exceptions. The lysis solution for poly(A)-tail-selected RNA contained 9 mL of RLT Lysis Solution, 250 μL of zirconium beads (OPS Diagnostics, Lebanon, NJ, USA), and an internal poly(A)-tailed mRNA standard. The Oligotex mRNA kit (Qiagen, Valencia, CA) was used to isolate poly(A)-tailed mRNA from total RNA. The poly(A)-tailed mRNA was then linearly amplified with the MessageAmp II-aRNA Amplification Kit (Applied Biosystems, Austin, TX). Double-stranded cDNA was prepared as described above for total community RNA with the exception that no ethanol precipitation was done.

Internal Standards

Internal standards were used in order to normalize across samples and enable quantitative analysis of the metagenomic and metatranscriptomic sequences (Gifford *et al.*, 2010). A known number of standard sequences are added to each sample. All of the sequences in the library were then normalized to the standard sequences in the library relative to the number added during sample preparation (Table 1). For the DNA samples in this chapter, a known number of *Thermus thermophilus* HB27 genome copies were added to each sample as a standard.

Sequencing and post-sequencing screening

Nucleic acids from all samples were ultrasonically sheared to fragments (~200-250 bp) and TruSeq libraries (Illumina Inc., San Diego, CA) were constructed for paired-end sequencing (2 x 150 bp) using the Illumina Genome Analyzer IIx sequencing platform (Illumina Inc., San Diego, CA). SHE-RA (Rodrigue et al. 2010) was used to join paired-end reads with a quality metric score of 0.5, and paired reads were then trimmed using SeqTrim (Falgueras et al. 2010). A BLAST analysis of metatranscriptome reads was conducted against a database containing representative rRNA sequences along with the internal standard sequences (blastn, bit score ≥50) (Gifford et al. 2010). Those cDNA reads with BLAST hits were removed from the data set. To remove internal standard sequences from the metagenome reads, DNA reads with a BLAST hit against the *Thermus thermophilus* HB8 genome (blastn, bit score ≥50) were queried against the RefSeq protein database. Reads with a BLAST hit matching a *T. thermophilus* protein (blastx, bit score ≥40) were designated as internal standard and removed.

More than 39 million DNA sequence reads were obtained from samples from three stations (the offshore low salinity station was not sampled for DNA), with more than 27 million reads remaining after sequence trimming and removal of standards (Table 1). A total of 162 million cDNA reads were sequenced from the four stations, and over 53 million reads remained after trimming, and removal of standards, rRNA, and tRNA reads (Table 1). DNA reads were an average of 190 bp long, while cDNA averaged 173 bp each.

Identification and analysis of diazotroph reads

A BLAST analysis of the DNA and cDNA reads against six oceanic N_2-fixing cyanobacteria (Table 2) was conducted (blastn, bit score $\geq$50). Replicate reads, defined as those that matched another read from the same sample across the first 100 bp, were removed. A BLAST analysis of non-duplicate diazotrophic reads was then conducted against the nr/nt database (NCBI, blastn, e-value ≤10, hit length ≥50 bp). The percent identities of each read with a top BLAST hit to one of the diazotrophic cyanobacterial genomes was plotted in order to determine a cut-off percent identity value for each organism (Figure 2). DNA reads with hits above these cut-off values for each organism at each station were summed and normalized to the internal standard recovery percentage for that sample and the genome length of the organism, resulting in genome copies L^{-1} kbp^{-1}. A BLAST analysis of the cDNA reads above the percent identity cut-off for a given organism was conducted against a database of open-reading frames (ORFs) and intergenic spacer regions (IGSs) of that organism (blastn) in order to assign each read to a functional region. For each ORF, the number of reads assigned was normalized for the gene length and the sample internal standard, as described above, to arrive at transcript copies L^{-1} kbp^{-1}. For IGSs with fewer than ten reads assigned, the entire IGS length was used for normalization. For those IGSs with at least ten reads assigned, the reads were mapped to the IGS in order to get a more accurate transcript length. The mapping was done using the GS Reference Mapper (Roche) with default settings. Mapping of cDNA reads to the gene sequence was done in the same manner for abundant diazotroph transcripts.

KEGG orthology K numbers were assigned to *Richelia intracellularis* HH01 ORFs by submitting the protein sequences to the KEGG Automatic Annotation Server (KAAS) (Moriya et al. 2007) using the best bi-directional hit (BBH) method. The *Trichodesmium* K numbers were obtained through the DOE Joint Genome Institute (JGI) Integrated Microbial Genomes (img) annotation table for *T. erythraeum* IMS 101. The transcript abundance for each KEGG pathways was then calculated by summing the transcript abundances of all the ORFs assigned to the given pathway in that organism.

The BLAST analysis (blastn, e-value $\leq 10^{-4}$) of *Trichodesmium* cDNA and DNA reads was conducted against a database containing the *T. erythraeum* IMS 101 genome and two publicly available *Trichodesmium* metagenomes from BATS (Bermuda Atlantic Time Series) and the North Pacific Subtropical Gyre (IMG Genome IDs: 2156126005 and 2264265224, respectively).

Results

The four stations sampled are classified by the sea surface salinity at each, and referred to as oceanic (36.03 psu), transitional (31.79 psu), and low salinity (26.49 psu offshore and 22.55 psu coastal) (Figure 1). The sea surface temperatures ranged between 28.4°C (oceanic) and 29.36°C (coastal) and all samples were taken in the morning between 07:00-09:30 within a one-month span (Figure 1).

Environmental sequence similarity to references

Most of the diazotroph sequence reads had a top BLAST hit to the *Richelia intracellularis* HH01 genome (163,293 DNA, 16,211 cDNA), with 91.5% of those reads at least 98% identical (nucleotides) to the genome sequence, referred to as the *Hemiaulus-Richelia* (HR)-B population (Figure 2). An additional 7.6% of the reads fell within the range of a secondary peak between 93-97% identity, termed the HR-A population (Figure 2). Many reads also had a top BLAST hit to the *Trichodesmium erythraeum* IMS101 genome (33,038 DNA, 10,851 cDNA) with a peak at 92% identity. All but 26 reads were above the determined cut-off of 80% identity to the genome sequence (Figure 2). Fewer reads had a top BLAST hit to the *Crocosphaera watsonii* WH8501 genome (998 DNA, 532 cDNA) or the *Rhizosolenia*-associated *Richelia intracellularis* RC01 genome (907 DNA, 440 cDNA), but both sets of reads had a peak at 100% identity to genome sequences (Figure 2). The *Crocosphaera* population consisted of reads that were at least 98% identical to the *C. watsonii* WH8501 genome. Reads at least 97% identical to the *Ri. intracellularis* RC01 genome were analyzed for the *Rhizosolenia-Richelia* (RR) population. A fraction of reads had a top BLAST hit to the unicellular haptophyte-associated UCYN-A cyanobacteria genome (664 DNA, 488 cDNA) and the heterocyst-forming external diatom symbiont *Calothrix rhizosoleniae* SC01 genome (591 DNA, 215 cDNA), but neither had more than 50 reads at least 95% identical to the genome sequence (data not shown). These reads were not analyzed further.

Comparing the *Trichodesmium* cDNA and DNA reads with *Trichodesmium* metagenomes from other areas, 56% of reads had a top BLAST hit to the Bermuda Atlantic Time-series Study (BATS) metagenome, and 34% of reads had a top hit to the North Pacific metagenome. The remaining 10% of reads were more similar to the *T. erythraeum* IMS 101 genome than to either metagenome. The reads with a top hit to either metagenome were, on average, 7.0% more identical to the metagenome than to the *T. erythraeum* IMS 101 genome. At the transitional and oceanic stations, a majority of reads had a better BLAST hit to the *Trichodesmium* BATS metagenome, while most low salinity offshore reads had a top hit of *T. erythraeum* IMS 101 genome (Figure 3).

Eight *Trichodesmium* cDNA reads and one DNA read aligned with the PCR amplified *hetR* gene fragments in NCBI. Each of the nine reads was most similar to one of five sequences amplified from four different *Trichodesmium* species (Table 3). *T. thiebautii* was the top hit to five reads, the most of any *Trichodesmium* species, and all five of those reads were from the oceanic station sequences. *T. erythraeum* (low salinity offshore), *T. hildebrandtii* (transitional and oceanic), and *T.aurem* (transitional) were also the most similar to one or more reads (Table 3).

Diazotroph metagenomes

The oceanic metagenome consisted of 0.95% diazotroph reads (89,683 reads), and 1.17% of the transitional metagenome was comprised of diazotrophic reads (105,153 reads). The low salinity coastal metagenome was 0.01% diazotrophic reads (514 reads). Total normalized diazotrophic cyanobacterium DNA from three stations was 1.5×10^{13} genome copies L^{-1} kbp^{-1}, with the majority at the transitional station (9.9×10^{12} genome copies L^{-1} kbp^{-1}) (Figure 4). Overall, the sequences from the HR-B population (98-100% identity to the genome) were the most abundant (1.3×10^{13} genome copies L^{-1} kbp^{-1}), and an order of magnitude greater than the sequences from the HR-A population (94-97% identity, 1.1×10^{12} genome copies L^{-1} kbp^{-1}) and *Trichodesmium* population (1.2×10^{12} genome copies L^{-1} kbp^{-1}). RR population sequences were present at a lower abundance (2.2×10^{10} genome copies L^{-1} kbp^{-1}), and *Crocosphaera* population sequences were the least abundant in the diazotrophic cyanobacterium data set (2.3×10^{9} genome copies L^{-1} kbp^{-1}).

Diazotroph transcriptomes

Diazotroph reads (14,557 reads) were 0.10% of the transitional metatranscriptome, while 0.05% of each of the low salinity offshore and oceanic metatranscriptomes were diazotroph reads (5,132 reads and 6,230 reads, respectively). Less than 0.01% of the reads in the low salinity coastal metatranscriptome was diazotrophic (281 reads). The total normalized diazotrophic cDNA from four stations was 3.01×10^{10} gene copies L^{-1} kbp^{-1}, and nearly all of that was from the transitional station (2.96×10^{10} gene copies L^{-1} kbp^{-1}). Similar to the DNA abundance, HR-B population cDNA from the four stations (2.6×10^{10} gene copies L^{-1} kbp^{-1}) was one order of magnitude greater than that of the HR-A population (1.1×10^{9} gene copies L^{-1} kbp^{-1}) or *Trichodesmium* (2.9×10^{9} gene copies L^{-1} kbp^{-1}). RR population cDNA (2.2×10^{7} gene copies L^{-1} kbp^{-1}) and *Crocosphaera* cDNA (3.7×10^{6} gene copies L^{-1} kbp^{-1}) were present at lower abundances.

The *Ri. intracellularis* HH01 genome contains 2278 genes and 1590 of them (69.8%) were detected in the HR-B population transcriptomes (15,311 reads). By contrast, 2233 of the *Ri. intracellularis* HH01 genes (98.0%) were detected in the metagenomes (148,968 reads). Most of the genes not found in the transcriptomes were hypothetical proteins (401 out of 688). There were also 689 IGSs with at least one cDNA read, including several that were among the most abundant transcripts. The two most abundant ORFs at the transitional station were the gene that encodes the D1 subunit of NADH dehydrogenase I (*ndhD1*, RintHH_21740) and a fused phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase gene (*hisIE*, RintHH_14390). A total of 1081 reads from the transitional station were assigned to *ndhD1* and they mapped mostly to a 397 bp region in the middle of the 1572 bp gene sequence (Figure 5). Similarly, all 171 *hisIE* reads from the transitional station covered only 232 bp of the 651 bp gene (Figure 5).

Nitrogenase gene transcripts were most abundant at the transitional station (Figure 6A). However, nitrogenase transcripts comprised a smaller fraction of normalized transcript abundance at this station than the low salinity offshore (Figure 6B). Only 21 oceanic station cDNA reads were attributed to nitrogenase genes, and thus were not included in this comparison. HR-B population *nifH* cDNA reads from the transitional and low salinity offshore stations displayed even distribution along the gene, relative to *ndhD1* and *hisIE* (Figure 5).

Only 265 *Ri. intracellularis* HH01 genes and 85 IGSs were found in the HR-A population transcriptomes (659 reads). Just 1177 of the 2278 *Ri. intracellularis* HH01 genes (51.7%) were detected in the metagenomes (1,177 reads). Of the HR-A transcripts detected, 85 ORFs and 39 IGSs did not appear among the HR-B population transcript sequences. The most abundant transcript was at the transitional station and coded a cyanobacteria-specific hypothetical protein (RintHH_13740).

The RR population transcriptomes consisted of 253 reads, which were assigned to 129 ORFs and 46 IGSs. The most abundant gene found in the RR population transcripts at the oceanic station was a hypothetical protein (RintRC_2139).

The *Trichodesmium* transcriptomes (9,892 reads) contained transcripts for 1634 genes out of 5076 in the *T. erythraeum* IMS101 genome (32.2%) and 247 IGSs. The *Trichodesmium* metagenomes (33,017 reads) contained 3,772 of the genes in the genome (74.3%). The most abundant *Trichodesmium* transcript at each of the transitional and oceanic stations was a hypothetical protein (Tery_2611). Reads from each of those stations only mapped to a small region of the gene (Figure 5). Reads assigned to a gene that encodes an S-adenosylmethionine--tRNA-ribosyltransferase isomerase (*queA*, Tery_0731) were found mostly at the transitional station, and also mapped to just a small portion of the gene (Figure 5).

Genes involved in gas vesicles, photosystem II, and other hypothetical proteins were among the most abundant *Trichodesmium* transcripts at each station. Oceanic and low salinity offshore station reads from an abundant gas vesicle protein transcript were evenly distributed along the gene (Figure 5).

The transcriptomes of the unicellular *Crocosphaera* were comprised of 85 reads, 80 of which are from the oceanic station. Hypothetical proteins (33 reads) and genes involved in photosynthesis (14 reads) were the most abundant *Crocosphaera* transcripts.

Given the low coverage of the transcripts from the HR-A, RR, and *Crocosphaera* populations, the transcription profiles of only the HR-B and *Trichodesmium* populations were compared more closely. Due to the lack of diazotrophic abundance in the low salinity coastal data sets, the populations were compared only amongst the other three stations. KEGG pathways were identified for the ORFs of 10,560 HR-B reads and 5,001 *Trichodesmium* reads within the three metatranscriptomes. The energy metabolism fraction of the normalized transcript abundance from the HR-B reads at a station ranged from 29.6-50.5% for the *Trichodesmium* population, and 30.8-42.3% for the HR-B populations (Figure 7). Within energy metabolism, 17.5-36.1% of *Trichodesmium* KEGG pathway transcripts at a station belonged to photosynthesis pathways and 4.0-4.9% were involved in nitrogen metabolism pathways (Figure 8). For the HR-B population, photosynthesis comprised 12.9-23.1% of total transcript abundance, and nitrogen metabolism was 5.5-9.5% (Figure 8). A large fraction of transcripts for both populations at each station were also attributed to genetic information processing (17.8-29.2%) (Figure 7).

The gene groups within photosynthesis pathways were examined individually. Antenna proteins were 3.6-8.7% of HR-B transcription, and photosystem (PS) I proteins were 5.0% at the low salinity offshore station (Figure 9). PS-II genes were the most abundant photosynthesis group in *Trichodesmium* transcription at each station (12.4-23.5%), while antenna proteins were also abundant (2.7-7.2%) (Figure 9). All other gene groups for each population were no more than 3.2% of population transcription at any station (Figure 9).

Discussion

At the time of sampling, the Amazon River plume had its maximum discharge rate of 2010 (Yeung et al. 2012). The plume flowed NW and was defined by reduced sea surface salinity and elevated chlorophyll-*a* relative to surrounding water (Yeung et al. 2012; Goes et al. 2014). The riverine discharge had low concentrations of NO_3 and NO_2, but SiO_3 and PO_4 within the plume were higher than surrounding waters (Goes et al. 2014). Additionally, there was a coupling between the diatom-associated diazotrophs, drawdown of C and SiO_3, and export efficiency (Yeung et al. 2012).

The distributions of diazotrophs in this study largely agree with previous observations from this region. The riverine fixed N concentration is high enough in low salinity waters to negate the advantage of N_2 fixation (Subramaniam et al. 2008), and thus the least diazotrophic activity is found in these waters. Furthest from the Amazon River influence, *Trichodesmium* is the dominant diazotroph in the more oceanic environment, as has been observed previously (Foster et al. 2007; Turk-Kubo et al. 2012). In transitional waters between the river input and open ocean, enough fixed N has been assimilated by the community, but riverine PO_4, Fe, and SiO_3 are still in sufficiently high concentrations to create ideal conditions for diazotrophs, especially those in association with diatoms (Yeung et al. 2012; Goes et al. 2014). It is possible that the 156 μm pre-filtration may have removed some long-chain diatoms harboring diazotrophs and large *Trichodesmium* colonies from the sequenced samples, altering the representation of these populations in the data.

The two most prominent diatom symbionts were each associated with diatoms of the genus *Hemiaulus*. These two distinct symbiont populations were separated by a slight difference in sequence similarity, and likely represent symbionts of different *Hemiaulus* species. The use of the *H. hauckii* symbiont as the reference genome, and the high similarity between it and the *H. membranaceus* symbiont genome (Hilton et al. 2013), place the symbionts of these two diatoms within the high percent identity range of the Amazon River plume HR-B population. The less similar HR-A population was likely made up of the symbionts of *H. indicus* and/or *H. sinensis,* each of which have also been observed harboring heterocyst-forming symbionts (Sundström 1984; Villareal 1991). Previous phylogenetic analysis has reported two distinct clades within the *Hemiaulus* symbionts, het2A and het2B, that exhibit a similar genetic distance as HR-A and HR-B (Janson, Wouters, et al. 1999; Foster & Zehr 2006). All of the HR-B reads that aligned with the *hetR* region used in these previous studies (49 DNA, 6 cDNA reads) exhibited more similarity to het2B sequences than het2A sequences. However, no HR-A population DNA or cDNA reads mapped to the *hetR* region amplified in these studies, so this population was not confirmed to be within the het2A clade.

In contrast to the diatom-associated cyanobacteria, the sequences most related to free-living *Trichodesmium* had a much wider range of divergence from the representative genome with no distinct separation of obvious populations. Like the diatom symbionts, *hetR* has been a common genetic marker used to study *Trichodesmium* diversity (Janson, Bergman, et al. 1999; Schiefer et al. 2002; Lundgren et al. 2005; Hynes et al. 2012). Two DNA reads and seven cDNA reads from the *Trichodesmium* population aligned within the 448 bp *hetR* region amplified in most of these studies. Five of the six oceanic station reads were identical to *T. thiebautii hetR* sequences, and this species has previously been the dominant *Trichodesmium* species in the tropical North Atlantic (Carpenter et al. 2004; Sohm et al. 2008). *T. erythraeum* has also been observed in the area (Webb et al. 2007), and its presence is supported by a *hetR* read from the offshore low salinity station that was identical to the sequenced *T. erythraeum* IMS 101 genome. *T. aureum* and *T. hildebrandtii hetR* fragments were also the most similar to at least one read each. Therefore, there are likely at least four different phylotypes that comprised the *Trichodesmium* metagenomes and transcriptomes.

The Amazon *Trichodesmium* populations showed considerably more similarity to populations from the North Pacific and North Atlantic Oceans than to the *T. erythraeum* IMS 101 genome. The offshore low salinity station was the only station where more reads were more similar to the *T. erythraeum* IMS 101 genome than either of the *Trichodesmium* metagenomes. This is also the station where the single *hetR* read with 100% identity to *T. erythraeum* was found. Therefore, the *T.*

erythraeum IMS 101 genome appears to be representative of the natural population at one station, but is not representative of the majority of *Trichodesmium* populations in the Amazon River plume, or from the North Atlantic and North Pacific subtropical gyres. Additional sequencing of a variety of *Trichodesmium* isolates is necessary to determine how the metabolic capabilities and ecological roles differ amongst the various *Trichodesmium* populations.

The high coverage of the HR-B and *Trichodesmium* metagenomes across the respective genome signifies that these populations were adequately sequenced in the data. Thus, it is these two populations that the majority of the conclusions are focused on. Additionally, the *Trichodesmium* coverage may actually be higher than the metagenomic coverage indicates, as the total gene content of the natural populations may vary from the *T. erythraeum* IMS 101 reference genome.

Energy metabolism encompassed a relatively large fraction of transcripts in the HR-B and *Trichodesmium* populations at each of the three stations analyzed. Within energy metabolism, the higher fraction of transcription of photosynthesis genes in the *Trichodesmium* populations relative to HR-B populations is due to the overall abundance of *Trichodesmium* photosystem (PS) II gene transcripts. Two *Trichodesmium psbA* copies, coding the PS-II D1 subunit, were each among the eleven most abundant transcripts in the *Trichodesmium* low salinity offshore and oceanic transcriptomes. Additionally, one of the *psbA* copies was the 14th most abundant gene in the *Trichodesmium* transitional transcriptome. High expression of PS-II genes, relative to other photosynthesis genes, has been commonly observed

(Levitan et al. 2010; Mohr et al. 2010) due to a high rate of PS-II protein turnover as a result of photodamage (Aro et al. 1993). Only one *psbA* gene copy is present in the *Richelia intracellularis* HH01 genome assembly, but it is alone on a contig. This is indicative that it could not be assembled among other sequences because it represents multiple gene copies within the genome. The transcripts of *psbA* were among the 15 most abundant transcripts in the HR-B low salinity offshore and oceanic transcriptomes and detected in the transitional transcriptome, albeit at low abundance. However, PS-II genes *psbH* and *psbK* were not detected in any HR-B transcriptome, despite *psbH* transcripts among the 20 most abundant *Trichodesmium* transcripts in each of the low salinity offshore and transitional transcriptomes. Additionally, *psbH* and *psbK* were each detected in the *Trichodesmium* oceanic transcriptome. In the diazotrophic cyanobacterium *Synechocystis*, neither *psbH* nor *psbK* were essential to photoautotrophic growth, but the loss of either resulted in reduced growth rates (Ikeuchi et al. 1991; Mayes et al. 1993). The PS-II transcript differences may reflect the morphological difference between *Richelia* and *Trichodesmium*, or indicate the *Hemiaulus* symbiont has reduced growth rates, as seen with heterocyst-forming cyanobacteria in other associations (Peters & Meeks 1989; Adams et al. 2006). It is also possible that *Richelia* is better protected from photodamage within the diatom, resulting in a lower PS-II protein turnover rate, and thus reduced PS-II gene expression relative to free-living oceanic cyanobacteria. However, *psbH* and *psbK* were each detected in one HR-A transcriptome, indicating that photosynthetic activity may differ between the two closely-related *Hemiaulus* symbiont populations.

The HR-B fraction of transcripts within photosynthesis gene groups other than PS-II, however, was comparable, and often greater than that of *Trichodesmium*. Thus, the HR-B populations may have been investing more energy towards cyclic electron transport around PS-I, rather than linear electron transport which requires PS-II activity. Cyclic electron transport can generate ATP by recycling electrons through the reduction of NADPH by NADH dehydrogenase (Mi et al. 1995). Even though elevated transcription does not necessarily equate to increased activity, it is reasonable to assume that diatom symbionts may require additional ATP from cyclic electron transport. N_2 fixation is an energetically expensive process (Ljones 1979), and the symbionts increase N_2 fixation rates in order to meet not only their own N needs, but also those of their diatom partners (Foster et al. 2011).

Intriguingly, the second most abundant transcript in HR-B transitional transcriptome may regulate cyclic electron transport. This transcript is likely an antisense RNA (asRNA), since it had only partial coverage of the NADH dehydrogenase D1 subunit gene. asRNAs are transcribed in the opposite direction to an mRNA target, can up- or down-regulate that gene, and require rho-independent termination mechanisms (Georg & Hess 2011). A T-tail following a stem-loop secondary structure that would qualify as such a termination mechanism was located by mfold (Zuker 2003) near the predicted end of the HR-B *ndhD1* asRNA. It is unclear if this abundant transcript up-regulates or down-regulates the expression of *ndhD1*. Additionally, NADH dehydrogenases have other functions in cyanobacteria (Ogawa & Mi, 2007), and thus, it is unclear what affect the asRNA has on the symbiont or the association, as a whole. However, asRNAs have been identified for genes encoding other NADH dehydrogenase subunits in *Synechocystis* (Georg et al. 2009) and chloroplasts (Georg et al. 2010), indicating this level of regulation is not restricted to diatom symbionts.

Similar to HR-B *ndhD1*, other abundant transcripts in each of the *Trichodesmium* and HR-B transcriptomes showed only partial coverage on coding sequences. These reads may also belong to non-coding RNA (ncRNA) transcripts, such as asRNAs. No stem-loop structure could be found near the end of the other transcripts in question, but other rho-independent termination mechanisms are possible (Georg & Hess 2011). Significant expression has been observed for more than 400 asRNAs in *Synechocystis* (Mitschke et al. 2011), thus, it would not be surprising to detect additional regulatory transcripts in the cyanobacterial populations in this study.

The HR-B population transcriptomes were also characterized by an abundance of transcripts involved in N_2 fixation. Two of the most abundant HR-B transcripts were *nifH* and *nifD,* which encode the iron protein and alpha chain, respectively, of the MoFe protein of nitrogenase, the enzyme that catalyzes N_2 fixation. Similarly, *nifH* was the 9th most abundant transcript in the RR transcriptome, highlighting the metabolic importance of N_2 fixation in each diatom-diazotroph association. The abundance of HR-B nitrogenase transcripts from the transitional station indicates *Hemiaulus* symbionts in these waters have the greatest impact on new production in the surface community. However, N_2 fixation genes comprised a greater fraction of transcripts at the low salinity offshore station, possibly reflecting higher per cell N_2 fixation rates at this station.

Trichodesmium nitrogenase gene transcripts were detected in the transcriptome, but not in high abundance. Rather, transcripts highlighting important processes such as gas vesicle formation were more highly expressed in the *Trichodesmium* transcriptomes. Two of the most abundant transcripts in the low salinity offshore, transitional, and oceanic *Trichodesmium* transcriptomes were gas vesicle proteins found adjacent to each other in the genome. Gas vesicles provide buoyancy to return to surface waters after *Trichodesmium* sinks to depth, possibly to acquire phosphorus (Villareal & Carpenter 2003). Given that all samples were taken during the day, gas vesicles were likely important for remaining in the photic zone.

Unexpectedly, several of the highly abundant transcripts in the diazotroph metatranscriptomes corresponded to regions of the genome that have not been annotated as coding regions. Some of the IGS regions were between genes known to constitute an operon, and thus included in the transcript (e.g. *nifHDK*). However, three of the top five most abundant transcripts in the HR-B transcriptome did not correspond to known operons. A BLAST analysis of these three IGS regions resulted in high similarity to a transfer messenger RNA (NZ_CAIY01000044_209707_211231), an RNA subunit of RNase P (NZ_CAIY01000027_241244_243250), and a leucine transfer RNA intron sequence (NZ_CAIY01000027_330123_331418). These functional regions have been poorly annotated in previously sequenced genomes, and thus were initially unidentified in the *Ri. intracellularis* HH01 genome. Similarly, an abundant *Trichodesmium* IGS region (NC_008312__1642616_1643889) showed similarity to transposases, which can be difficult to annotate, further demonstrating the value of transcription sequences in genome annotations.

Conclusions

The deep sequencing of metagenomes and metatranscriptomes in this study has made it possible to analyze diazotrophic populations that cannot be achieved through targeted assays such as PCR. With the ability to compare genetic markers from across the genome, the majority of diazotroph populations in this environment were found to be similar to the genomes currently available. However, the *Trichodesmium* population was an exception to this, and was not representative of *T. erythraeum* IMS 101, the only currently sequenced *Trichodesmium* sp. genome. This suggests that genomic sequencing of a variety of *Trichodesmium* species is needed to more accurately depict natural populations, their metabolic capabilities, and their roles in surface communities. A need was also identified for studies on non-coding transcripts and their function in regulating a variety of metabolic processes within N_2-fixing cyanobacteria, and throughout microbial communities, in general. Additionally, analysis revealed a stark contrast within the distribution of transcripts amongst vital cellular processes, such as photosynthesis and N_2 fixation, between the free-living *Trichodesmium* and the diatom-associated *Richelia*. In this study, extensive community DNA and RNA sequencing was utilized to highlight individual diazotroph populations, and the metabolic pathways within those populations, to elucidate the community composition and cellular state of the diazotrophs in the Amazon River plume.

Acknowledgements

This work was sponsored in part by the Gordon and Betty Moore Foundation (GBMF) Marine Investigator award, ROCA award, the UCSC Microbial Environmental, Genomics, Application, Modeling, Experimental, Remote sensing (MEGAMER) facility (supported by the GBMF) and the NSF Center for Microbial Oceanography: Research and Education. We are very grateful to M.A. Moran, B. Crump, and J. Paul for their assistance throughout this project and the writing of this report. We also thank D. Bombar for comments on the manuscript, and I. Shilova and J. Robidart for technical help and discussions.

Tables and Figures

Table 1. Sample info and sequencing statistics for the Amazon River plume samples.

Sample ID	Station	Filter Size (µm)	Type	Reads	Reads after Trim	Std Reads	rRNA Reads	Remaining Reads
ACM1	transitional	2	DNA	5428695	3650345	26592	n/a	3623753
ACM2	transitional	0.2	DNA	7488350	5355957	10543	n/a	5345414
ACM3	low salinity coastal	2	DNA	6095193	4211422	3701	n/a	4207721
ACM4	low salinity coastal	0.2	DNA	6587526	4656473	3196	n/a	4653277
ACM5	oceanic	2	DNA	5138361	3212733	57600	n/a	3155133
ACM6	oceanic	0.2	DNA	8854420	6343198	16054	n/a	6327144
ACM7	transitional	2	Euk cDNA	15944969	10261507	113810	9916	10137781
ACM8	low salinity coastal	2	Euk cDNA	18327975	11572252	17455	25672	11529125
ACM9	low salinity off-shore	2	Euk cDNA	15427448	9050683	273423	11481	8765779
ACM10	oceanic	2	Euk cDNA	15930794	9679397	203133	16502	9459762
ACM11	low salinity coastal	0.2	Prok cDNA	15058449	5694376	37903	3222353	5656473
ACM12	low salinity coastal	2	Prok cDNA	10001411	2435885	41397	1360889	2394488
ACM13	low salinity off-shore	2	Prok cDNA	17035424	4908011	354928	2594703	4553083
ACM14	transitional	0.2	Prok cDNA	16379485	4696328	54005	2478789	4642323
ACM15	transitional	2	Prok cDNA	11055118	2575008	2772	628451	2572236
ACM16	oceanic	0.2	Prok cDNA	11436878	3167818	179159	1416551	2988659
ACM17	oceanic	2	Prok cDNA	15529442	4626581	274916	2150303	4351665

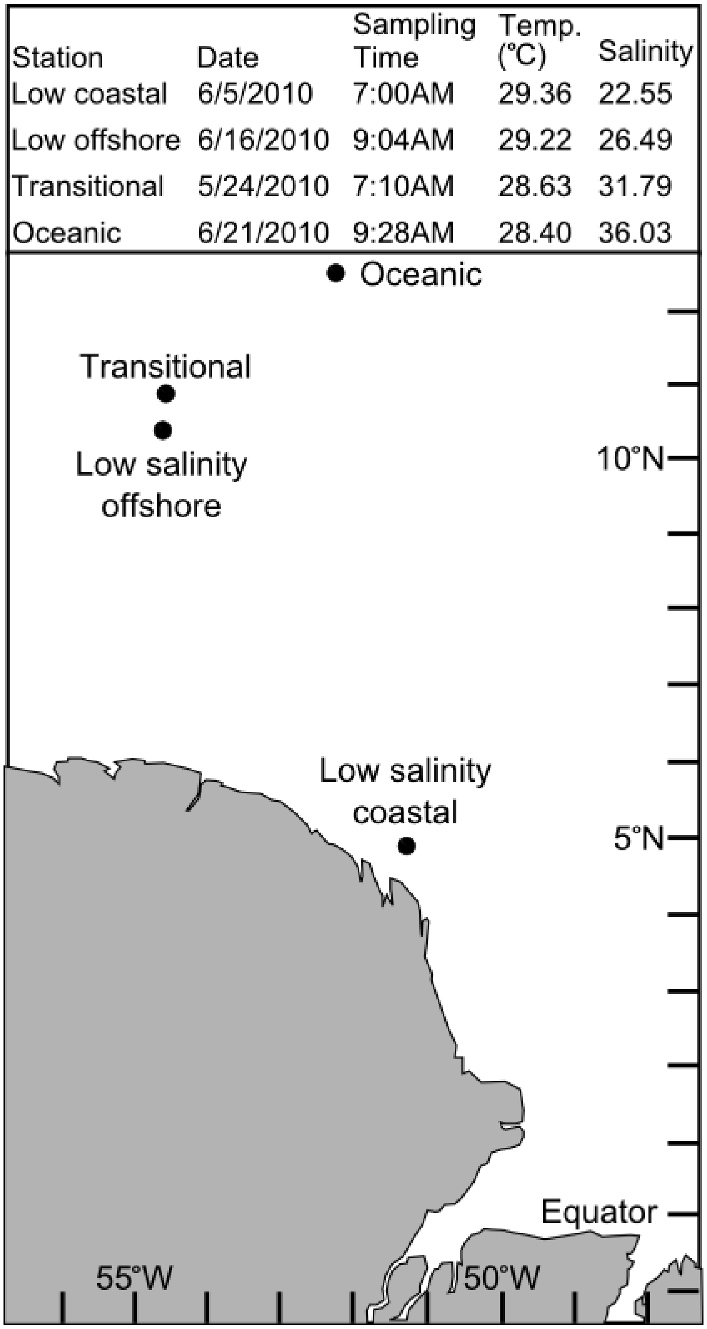
Normalization Factor
326737
824111
2347636
2718586
200129
606156
1690
20833
697
938
168983
188525
12896
85944
2132430
20944
16386

Table 2. Oceanic cyanobacterial diazotroph genomes. The four diazotrophic cyanobacterial genomes used as references for the Amazon River plume populations, and two additional genomes (*) that were not found in these data.

Diazotrophic Cyanobacterium	**Genome Size (Mb)**	**Morphology**	**Lifestyle**
Richelia intracellularis HH01	3.2	filamentous, heterocyst-forming	*Hemiaulus*-associated
Richelia intracellularis RC01	5.5	filamentous, heterocyst-forming	*Rhizosolenia*-associated
Trichodesmium erythraeum IMS 101	7.8	filamentous	free-living
Crocosphaera watsonii WH 8501	6.2	unicellular	free-living
**Calothrix rhizosoleniae* SC01	6.0	filamentous, heterocyst-forming	*Chaetoceros*-associated
*UCYN-A	1.4	unicellular	prymnesiophyte-associated

Table 3. *Trichodesmium hetR* sequences in the Amazon River plume. The percent identity (nucleotide) of nine *Trichodesmium* Amazon River plume sequences and each of the *hetR* fragments that are the best hit to at least one of the reads. The highest percent hit is highlighted in bold.

	Read1	**Read2**	**Read3**	**Read4**	**Read5**	**Read6**	**Read7**	**Read8**	**Read9**	
	low off	**oceanic**	**oceanic**	**oceanic**	**oceanic**	**oceanic**	**oceanic**	**trans**	**trans**	
	cDNA	**cDNA**	**cDNA**	**cDNA**	**cDNA**	**cDNA**	**cDNA**	**cDNA**	**DNA**	
Organism	**229 bp**	**222 bp**	**222 bp**	**177 bp**	**174 bp**	**151 bp**	**140 bp**	**163 bp**	**152 bp**	**Accession**
T. erythraeum IMS101	**100**	87.3	87.3	88.1	93.1	93.3	85.7	87.5	87.4	AF410432
T. thiebautii II-3	88.2	**100**	**100**	**100**	98.8	98.7	95.7	97.4	94.7	HM486692
T. thiebautii	89.5	99.5	99.5	**100**	**100**	**100**	95.7	97.4	95.4	AF490684
T. hildebrandtii	89	98.1	98.1	96.6	98.3	98	**98.6**	**100**	93.4	AF490679
T. aureum strain B49	88.2	95.3	95.3	95.5	99.4	99.3	91.4	93.4	**98.7**	AF490680



Station	Date	Sampling Time	Temp. (°C)	Salinity
Low coastal	6/5/2010	7:00AM	29.36	22.55
Low offshore	6/16/2010	9:04AM	29.22	26.49
Transitional	5/24/2010	7:10AM	28.63	31.79
Oceanic	6/21/2010	9:28AM	28.40	36.03

Figure 1. Stations sampled in the Amazon River plume.

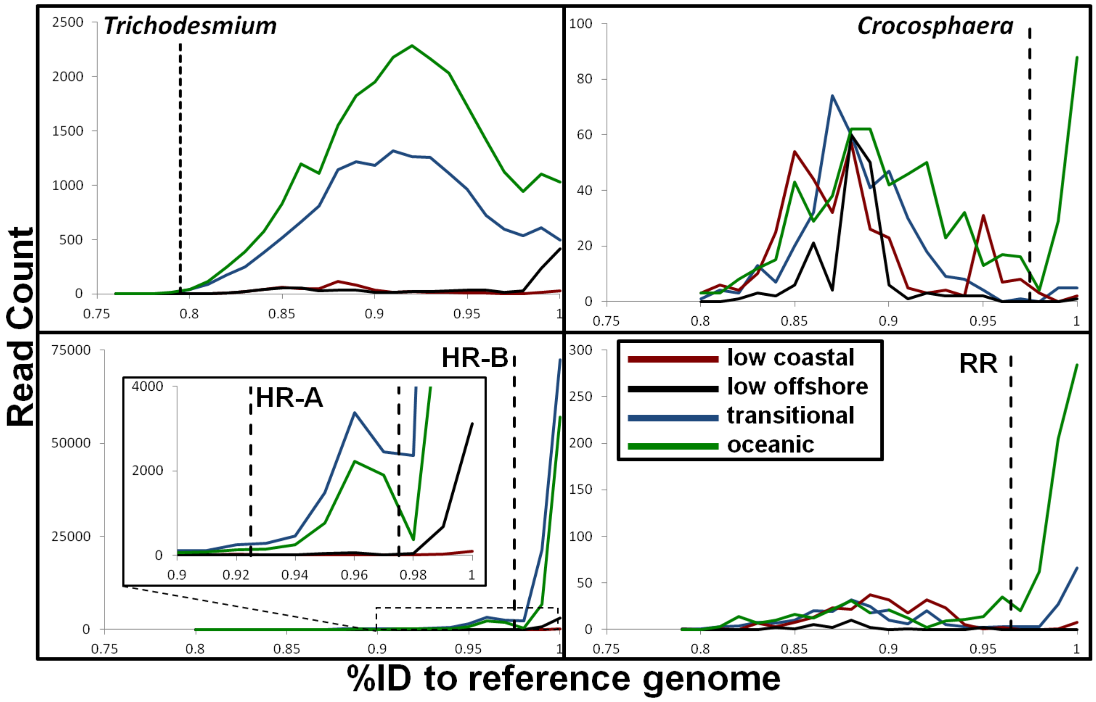


Figure 2. Similarity of Amazon River plume diazotroph populations to reference genomes. Histograms of the percent identity of reads with a top hit to each of four diazotrophic cyanobacteria genomes. The dotted lines mark the cut-off used in this study for each population.

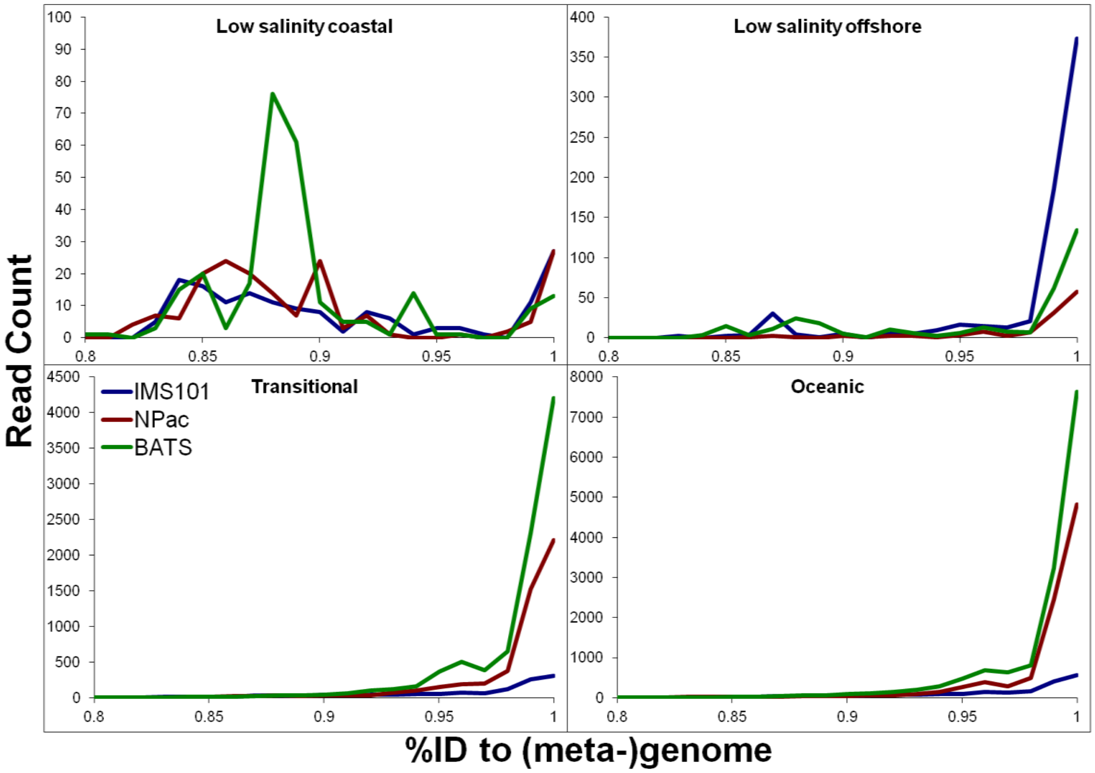


Figure 3. Similarity of Amazon River plume *Trichodesmium* populations to other *Trichodesmium* sequences. Percent identity histograms between all *Trichodesmium* DNA and cDNA reads and the most similar *Trichodesmium* reference sequence.

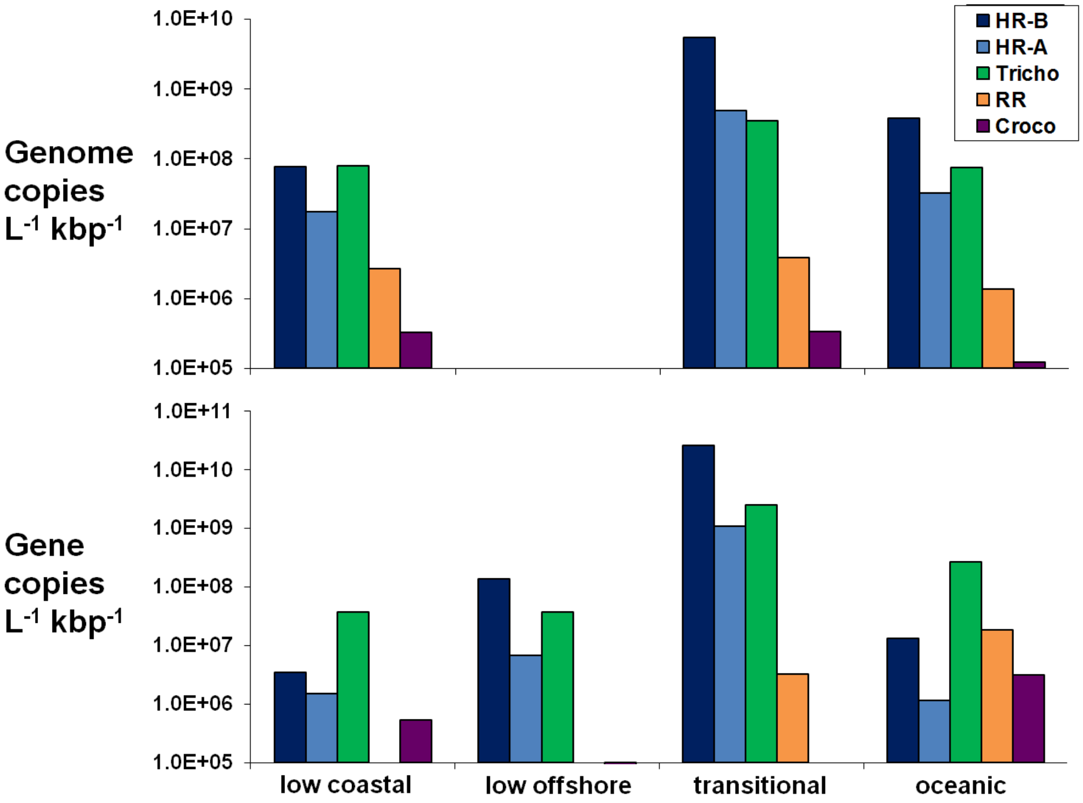


Figure 4. Diazotroph DNA and cDNA abundances. Normalized DNA and cDNA data for the five diazotrophic cyanobacteria populations at each of the four stations, with the exception of the low offshore station (DNA not sampled).

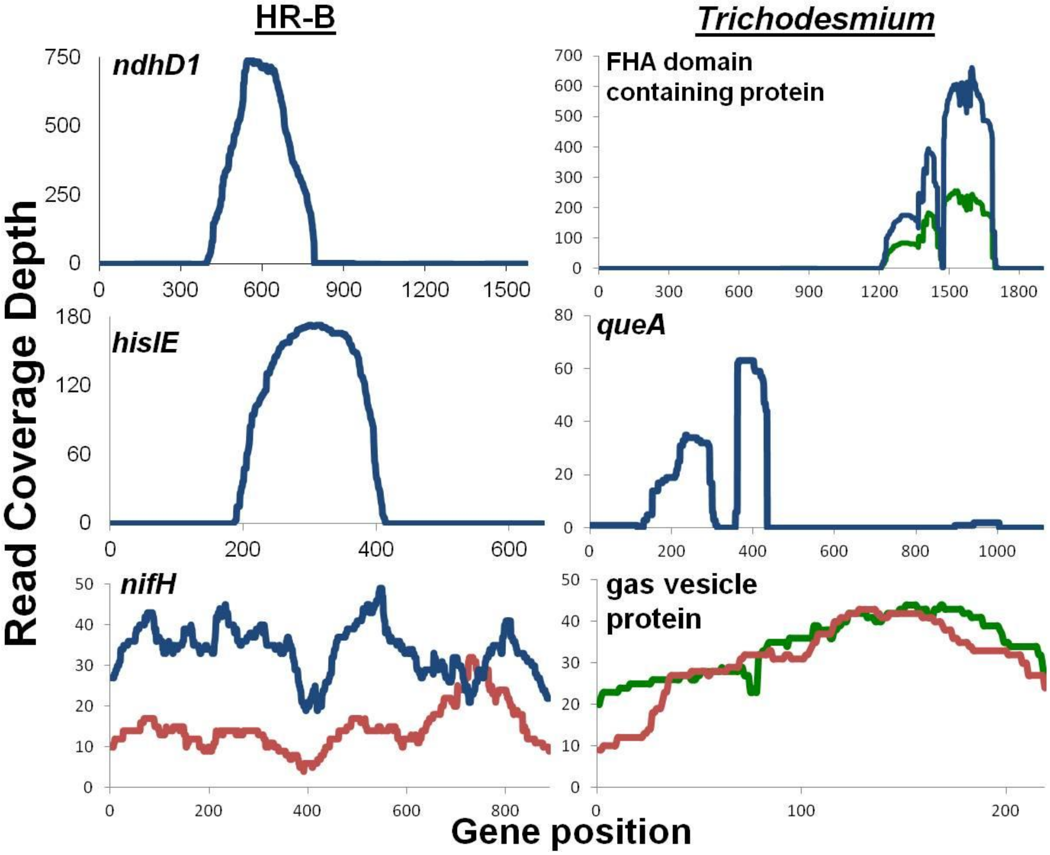


Figure 5. cDNA read coverage on abundant diazotroph transcripts. cDNA reads from the transitional (blue), oceanic (green), and low salinity offshore (red), stations mapped to abundant genes from the HR-B (left column) and Trichodesmium (right column) metatranscriptomes.

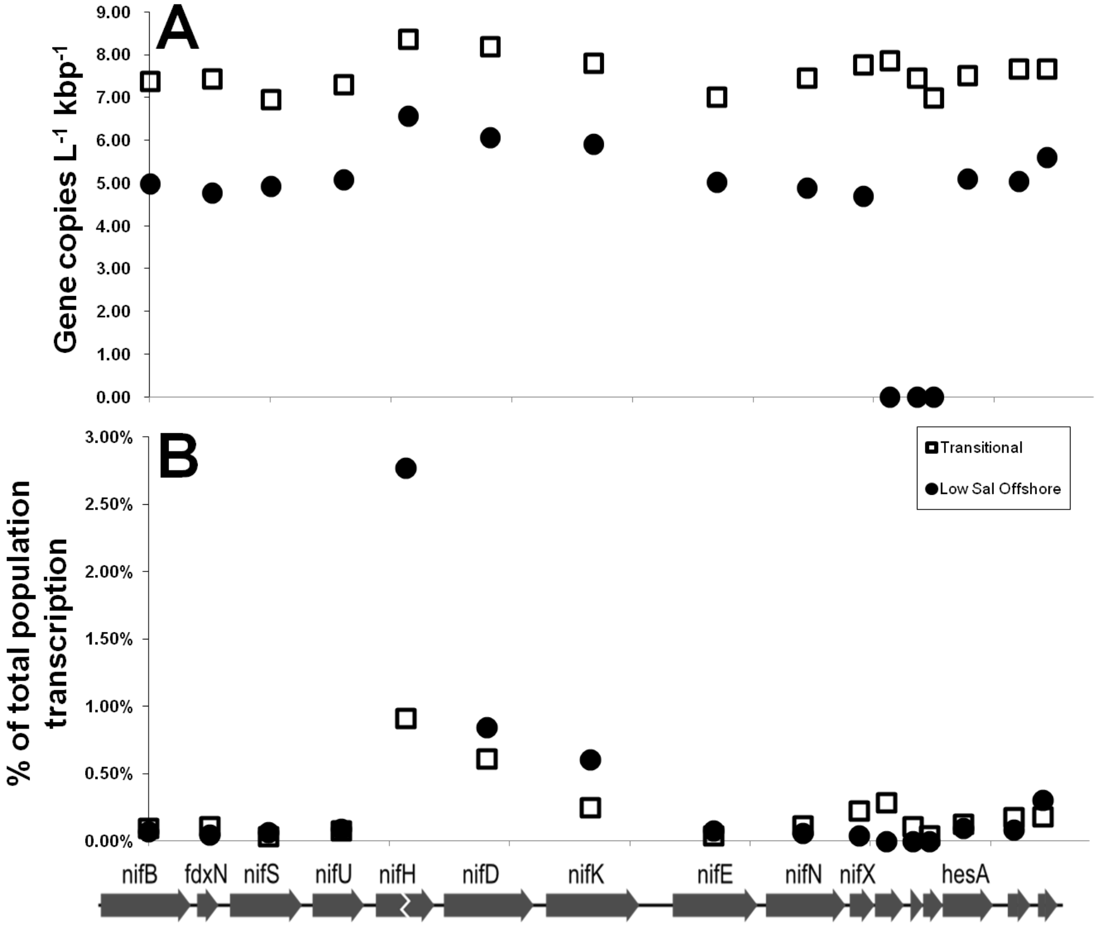


Figure 6. HR-B population nitrogenase transcripts. The normalized abundances of transcripts belonging to nitrogenase genes in the HR-B population at two stations (A) and the percent of total HR-B transcripts of those same genes (B).

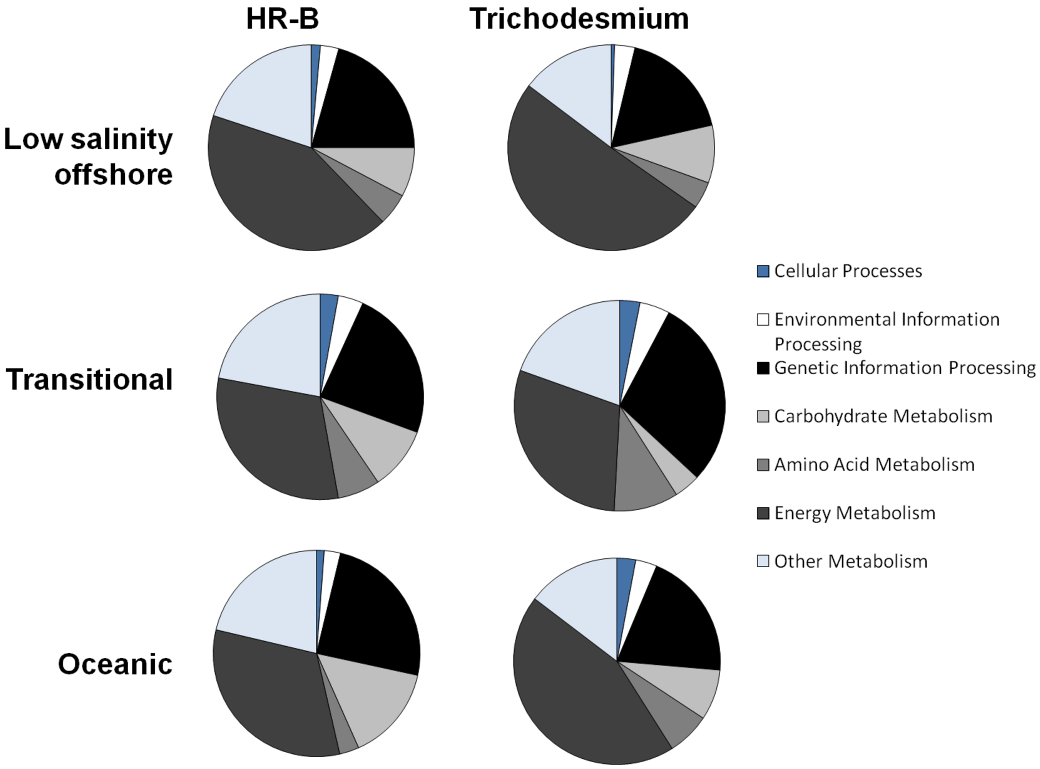


Figure 7. Diazotroph categorized transcription. The distribution of HR-B and *Trichodesmium* normalized transcript abundances within KEGG categories at each station.

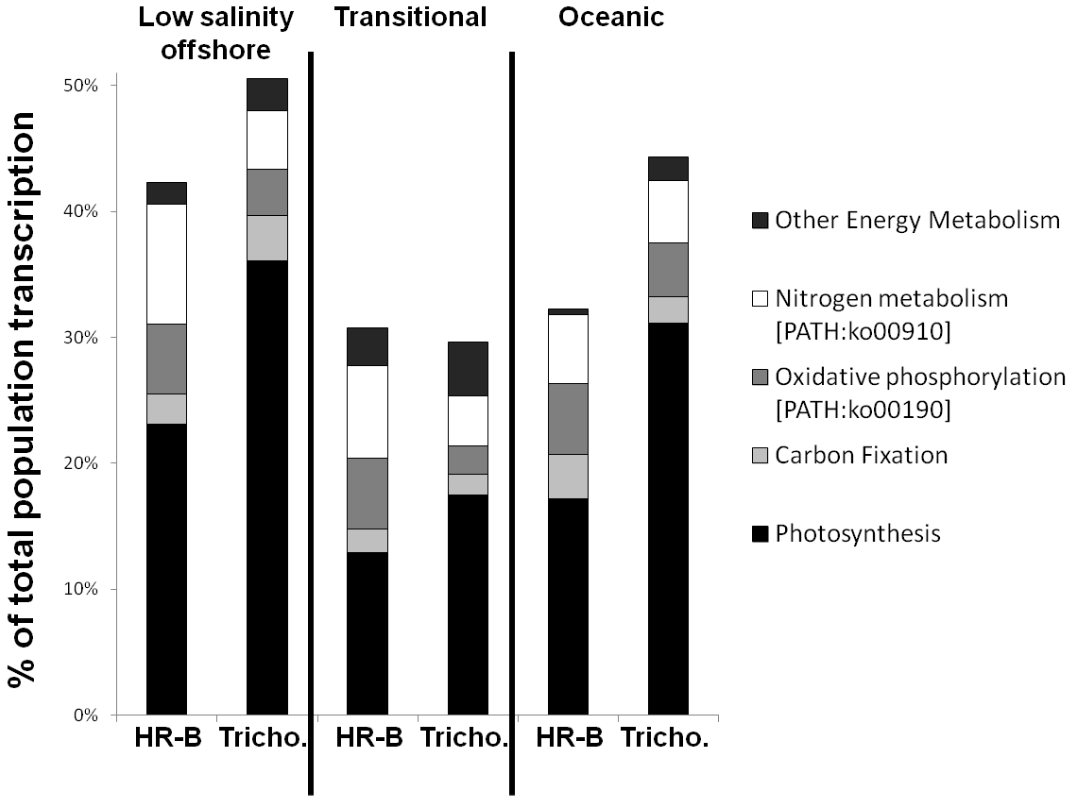


Figure 8. Diazotroph energy metabolism transcription. The distribution of HR-B and *Trichodesmium* normalized transcript abundances within KEGG Energy Metabolism pathways at each station.

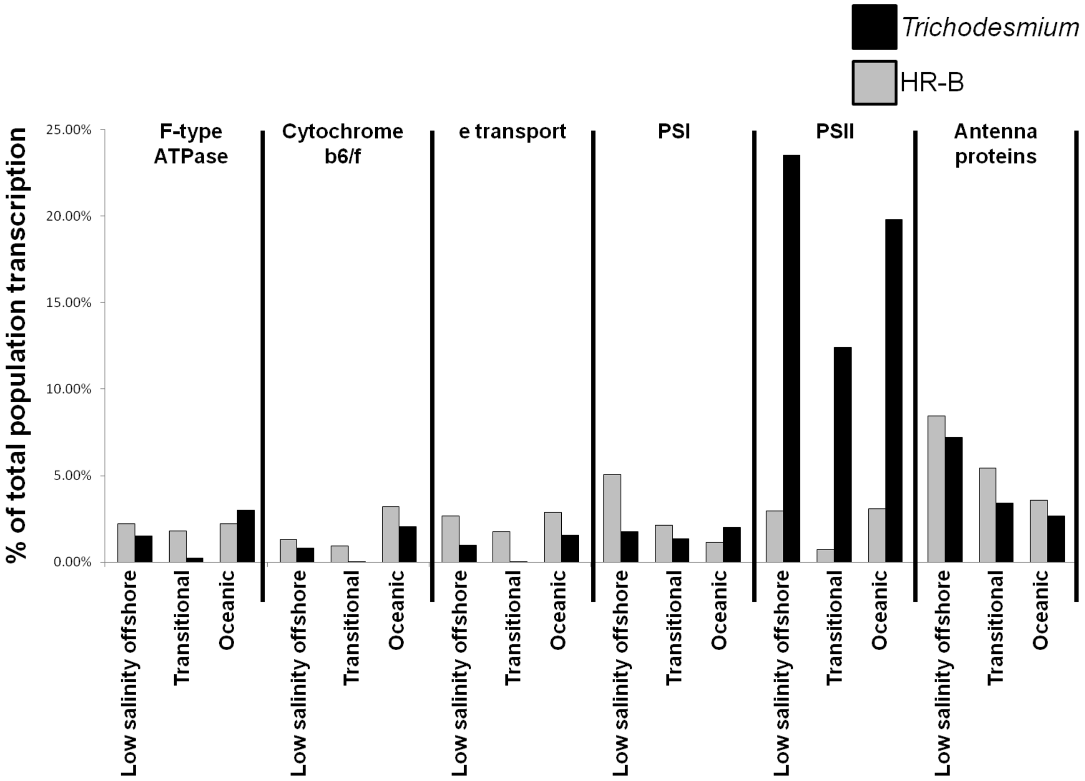


Figure 9. Diazotroph photosynthesis transcription. The distribution of HR-B and *Trichodesmium* normalized transcript abundances within KEGG Photosynthesis pathways at each station.

Chapter 4

Distribution and diversity of DNA elements within functional genes of heterocyst-forming cyanobacteria

Abstract

Some heterocyst-forming cyanobacteria excise DNA sequences from within key N_2 fixation genes during heterocyst differentiation, but the origin of these sequences is largely unknown. To examine the evolutionary relationships and possible function of DNA sequences that interrupt genes of heterocyst-forming cyanobacteria, 101 sequences, termed interruption elements, were identified and compared within genes from 28 heterocyst-forming cyanobacterial genomes. The interruption element lengths ranged from those long enough to contain only the recombinase gene responsible for element excision up to nearly 1 Mb. The genes contained on the interruption elements were variable between elements. The recombinase sequences were the sole genetic markers that were common across the interruption elements and could be used for analysis of the evolutionary history of recombinases. Elements were found that interrupted 22 different orthologs, most of which encoded proteins previously shown to have heterocyst-specific activity. The presence of interruption elements within genes with no known role in N_2-fixation, as well as in three non-heterocyst-forming cyanobacteria, suggests the excision of elements can be triggered by processes in addition to heterocyst development. This comprehensive analysis provides the framework to study the history and behavior of these unique sequences.

Introduction

Cyanobacteria have been shown to alter their cell shape and size, cell wall thickness, and filament orientation in response to environmental conditions varying from nutrient limitation to predation (Marcus *et al.*, 1982; Mühling *et al.*, 2003; Pattanaik & Montgomery, 2010). Some filamentous cyanobacteria can also differentiate distinct cell types. In addition to vegetative cells, they can form akinetes, hormogonia, and heterocysts. Akinetes are spore-like cells formed during unfavorable growth conditions (Kaplan-Levy *et al.*, 2010), while hormogonia are small-celled, motile filaments that are especially important in symbiosis initiation (Herdman & Rippka, 1988). Heterocysts are N_2 fixation cells that are formed under nitrogen (N) limitation (Wolk *et al.*, 2004).

N_2 fixation, the process of converting N_2 to NH_3, can be a means of avoiding a common nutrient limitation for microbial growth in a variety of environments (LeBauer & Treseder, 2008; Zehr & Kudela, 2011). Nitrogenase, the enzyme that catalyzes N_2 fixation is sensitive to inactivation by oxygen (O_2) (Wong & Burris, 1972), and most cyanobacteria evolve O_2 through photosynthesis, thus, establishing a paradox. Heterocysts create and maintain a microoxic microenvironment by shutting down the activity of O_2-evolving photosystem II, increasing respiratory O_2 uptake and creating a thick envelope around the cell wall to restrict gas diffusion (Wolk *et al.*, 2004).

Heterocyst differentiation requires a multitude of signaling pathways (Zhang *et al.*, 2006) and the regulation of the expression of 500 to 1,000 genes (Christman *et al.*, 2011; Ehira, 2013). Genome rearrangement can also be necessary in order to form fully functional heterocysts. Key N_2 fixation genes are often interrupted in the genome of vegetative cells by DNA sequences, referred to herein as interruption elements, previously shown to be up to 80 kbp in length (Wang *et al.*, 2012). Excision of the interruption element during heterocyst formation results in a contiguous, functional gene in heterocyst genomes (Golden & Wiest, 1988) (Figure 1). These elements are commonly found within *hupL*, which encodes one subunit of an uptake hydrogenase responsible for recycling H_2 produced during N_2 fixation, and *nifD*, which encodes the nitrogenase alpha subunit (Golden *et al.*, 1985; Golden & Wiest, 1988; Carrasco *et al.*, 1995; Carrasco & Golden, 1995; Henson *et al.*, 2011; Wang *et al.*, 2012). Additionally, elements have been reported within the *fdxN* gene, which encodes a ferredoxin and is located in an operon with nitrogenase genes *nifU, nifS,* and *nifB* (Golden *et al.*, 1985; Golden & Wiest, 1988; Mulligan & Haselkorn, 1989; Carrasco & Golden, 1995). More recently, elements have been found in the nitrogenase iron subunit gene *nifH*, and in *nifK*, the nitrogenase beta subunit gene (Vintila *et al.*, 2011; Wang *et al.*, 2012; Hilton *et al.*, 2013).

The excised DNA is bracketed by excision sites that contain direct repeats, and the elements are removed during the later stages of heterocyst development by a site-specific recombinase, which is encoded by a gene located within the element (Golden *et al.*, 1985; Lammers *et al.*, 1986). Genes for these recombinases have been identified for *nifD*, *hupL*, and *fdxN* elements as *xisA*, *xisC*, and *xisF*, respectively (Golden & Wiest, 1988; Carrasco *et al.*, 1994, 2005). The *nifH* element recombinase is most similar to *xisA* (Vintila *et al.*, 2011), and the recombinase on the *nifK* element is most closely related to *xisF* (Hilton *et al.*, 2013). *xisA* expression is likely regulated by the transcriptional regulators for nitrogen, NtcA, and ferric uptake, FurA, however; little is known of the details of the regulatory mechanisms and other potential regulating factors of interruption element recombinases (Chastain *et al.*, 1990; Ramasubramanian *et al.*, 1994; López-Gomollón *et al.*, 2007). Two additional genes, *xisH* and *xisI*, located in the *Nostoc* sp. PCC 7120 *fdxN* element are also required for excision of that element, although the specific role of these genes is unknown (Ramaswamy *et al.*, 1997). It is also unclear when or how these interruption elements originated, but it has been hypothesized that they are the remnants of viruses or phages (Haselkorn, 1992; Henson *et al.*, 2011). Additionally, it is unknown if they provide organisms with any advantages or disadvantages.

As of the writing of this report, there were 38 heterocyst-forming cyanobacteria with sequenced genomes; 27 from the order Nostocales and 11 from the Stigonematales (Figure 2). We defined 101 interruption elements in these genomes, including many previously-unknown elements, and compared them in an attempt to determine their evolutionary paths. We also discovered three interruption elements within genes of cyanobacteria that are not capable of forming heterocysts.

Materials and Methods

Annotations of the genes previously shown to contain interruption elements (*nifH, nifD, nifK, hupL,* and *fdxN*) were searched in all 38 currently sequenced genomes of heterocyst-forming cyanobacteria (Figure 2). The gene synteny surrounding each of these genes is highly conserved in heterocyst-forming cyanobacteria. Any genes found between *nifH, nifD, nifK, hupL,* or *fdxN* and the expected adjacent gene were examined for the presence of a gene with similarity to one of the known *xis* genes. A BLAST analysis (blastp) was conducted with these *xis* genes against all protein sequences of the 38 genomes. Each gene with a hit to any of the *xis* genes (with an *in silico*-determined threshold e-value $\leq 10^{-20}$) was examined as follows. A BLAST analysis (blastx, e-value ≤ 10, maximum matches in a query range=15) was conducted for each *xis* candidate and surrounding nucleotides (3 kbp before and after) against the nr database (NCBI). Genes that were homologous to the flanking region, yet aligned only partially, were used as references for the possible interrupted gene. Each reference gene was submitted to a BLAST analysis (tblastn, e-value ≤ 10) against the entire genome in order to locate sequences homologous to different sections of the reference gene. Those genome regions were then aligned with the reference gene to confirm they formed an orthologous gene. Once aligned, the direct repeats were identified in the overlapping bases of the two genome regions. As more interruption elements were confirmed, the *xis* sequences from each were added to the set, and the BLAST analysis was repeated to find more potential *xis* genes.

A BLAST analysis was then conducted with all of the confirmed *xis* genes from the heterocyst-forming cyanobacterial genomes against all non-heterocyst-forming cyanobacterial genomes (blastp, e-value $<10^{-50}$). All BLAST hits were checked for interrupted genes as described above.

For the recombinase gene and 16S rRNA phylogenetic trees, sequences were aligned with ClustalW (Thompson *et al.*, 1994). The recombinase sequence alignments were done with respect to codons, and thus, no gaps were inserted within a codon. Maximum likelihood phylogenetic trees were constructed with MEGA5 using the Tamura-Nei model and only those alignment sites with at least 90% coverage were used (Tamura *et al.*, 2011). Statistical support for nodes was based on 1,000 bootstrap replicates (Felsenstein, 1985).

Results

101 interruption elements were identified within genes of heterocyst-forming cyanobacteria, ranging between 1.3 to 984 kbp, and averaging 43 kbp in length (Figure 3,4). The elements were found in genes of 28 genomes of heterocyst-forming cyanobacteria, while an additional ten genomes did not have an element confirmed (Figure 2). Five out of eleven cyanobacteria within Stigonematales each had an element within *nifD*, and *Fischerella* sp. PCC 9339 had an additional *nifK* element. None of the interruption elements were found in the other six Stigonematales cyanobacteria. This result is consistent with a report in 1987 on the lack of elements interrupting the *nifHDK* genes in *Fischerella* sp., a Stigonematales isolate (Saville *et al.*, 1987). On the other hand, 23 out of 27 Nostocales organisms possessed at least one of the interruption elements, with as many as eleven elements in one genome (*Calothrix* sp. PCC 7103). The Nostocales that lack elements include three different *Cylindospermopsis raciborski* isolates and the obligate symbiont of the water fern *Azolla* spp. Sequence analysis previously demonstrated lack of *nifHDK* interruption elements in the symbiont in *A. filicoloides* (Ran *et al.*, 2010), while restriction fragment length polymorphisms indicated lack of a *nifHDK* element in that associated

with *A. Caroliniana* (Meeks *et al.*, 1988). Thus, lack of interruption elements in Nostocales is rare.

There were 86 individual genes interrupted by interruption elements, each belonging to one of 22 different ortholog groups (Table 1). The *Rivularia* sp. PCC 7116 *nifD* was interrupted by four different elements, the most within a single gene. The *Anabaena cylindrica* PCC 7122 primase P4 interruption element was on a *nifD* element, the only instance of an element located within another.

A total of 100 *xis* genes were confirmed in interruption elements, and one additional *nifH* element in *Richelia intracellularis* RC01 was identified, but due to a contig break, no *xis* could be identified. Seven of the *xis* genes were located in elements that interrupted a gene with the two gene sections oriented in opposite directions. These *xis* sequences were included in phylogenetic analysis but no further characterization of the interruption elements was done. There were 44 additional genes with a BLAST hit (e-value <10^{-20}) to a *xis* gene, but an interrupted gene could not be located near them, and, thus, could not be verified as element excision genes. 52 of the verified *xis* genes belonged to the tyrosine recombinase superfamily, while the remaining 48 were of the serine superfamily. *xis* genes of the serine superfamily were mostly located within 100 bp of the beginning or end of the element (72%), while most of the tyrosine *xis* genes were between 100-200 bp of the beginning or end (74%). The 69 *xis* genes aligned in the same orientation as the interrupted genes were located towards the beginning of the element. All but one of the 24 *xis* genes that were oriented in the opposite direction to the interrupted gene were found closer to the end of the element. The *xis* gene on the *Calothrix* sp. PCC 6303 arabinose efflux permease interruption element is the exception to this, and although oriented in the opposite direction of the interrupted gene, it is located at the beginning of the element.

Not all elements within a given ortholog interrupted the gene at the same location. All elements that interrupt the same ortholog at roughly the same position along that ortholog were considered to be an element variant. A slight fluctuation in the start positions of direct repeats within a variant is due to the varying lengths of the direct repeat (Table 2). The 34 *nifD* elements belonged to five different element variants, while six element variants were found amongst the 19 *nifH* elements. Two variants each were found for *nifK, nifB, hupL, hupS,* and *coxA3* elements. The interruption element lengths, as well as GC content, were analyzed by element variant, but few patterns were seen (Figures 3-6).

The majority of *xis* sequences formed well-supported phylogenetic clusters with other *xis* genes found in interruption elements of the same element variant (Figures 7,8). There were a few exceptions. The three *xis* sequences of elements that interrupted *nifH* at position 150 (± 7 bp) do not cluster together (Figure 7). The *nifD* elements with insertion locations at position 895 (± 1 bp) contain *xis* sequences that cluster together, however, only four of the genes are within a subcluster with strong bootstrap support, while the *Rivularia* sp. PCC 7116 and *Calothrix* sp. PCC 7103 genes are phylogenetically distant (Figure 8). Lastly, *Rivularia* sp. PCC 7116 and *Calothrix* sp. PCC 7103 are the only two genomes surveyed that possess *nifD* elements with insertion locations at position 227. The *xis* sequences from these two *nifD* elements cluster together, but exhibit long branch lengths and low bootstrap support (Figure 8). The three element variants just described as containing relatively diverse *xis* genes are also three of the four variants that have variable *xis* locations on the element and *xis* directions relative to the interrupted gene. The *flv3B* elements of *Scytonema hofmanni* PCC 7110 and *Calothrix rhizosoleniae* SC01 also have *xis* genes of differing location and orientation (Figure 7). The *Scytonema hofmanni* PCC 7110 *xis* gene is located at the end of the element and is oriented in the opposite direction of the *flv3B* that it interrupts while the *Calothrix rhizosoleniae* SC01 *xis* gene is at the beginning of the element and is read in the same direction as *flv3B*.

In some cases, *xis* sequences group phylogenetically by organism. For example, seven genes in the *Anabaena cylindrical* PCC 7122 were identical to each other, and while no interruption element could be identified for four of the genes, the three other genes were identified as encoding element excision proteins (*Anabaena cylindrical* PCC 7122 #1-3, Figure 7). Similarly, the *xis* gene of the interruption element within a *Scytonema hofmanni* UTEX 2349 transposase is identical to the *xis* gene in an oxidoreductase gene interruption element in the same genome (Table 2). However, the oxidoreductase gene interruption element *xis* sequence is shortened by a gap in the genome sequence, and, thus, was left out of the analysis. Although more divergent than the *A. cylindrical* PCC 7122 and *S. hofmanni* UTEX 2349 *xis* gene sets, the *Rivularia* sp. PCC 7116 *xis* sequences responsible for the *nifB* and *hupS* elements cluster together (Figure 8).

A hypothetical protein, *alr1449*, has been previously identified as commonly-occurring in elements that interrupt *nifD* at position 1355 (± 2 bp) (Lammers *et al.*, 1990; Henson *et al.*, 2005, 2011). This hypothetical protein was present in 18 out of the 24 *nifD* elements of that variant. *alr1449* was also present in 11 other genomes, but was not located on a interruption element.

Interruption elements were also found within genes of three unicellular non-heterocyst-forming cyanobacteria, including a hypothetical membrane protein of *Chroococcidiopsis thermalis* PCC 7203. A *Synechococcus* sp. PCC 7502 DNA helicase gene and a *Cyanothece* sp. PCC 7822 hypothetical protein were also interrupted by interruption elements (Table 2). The *xis* gene found on each of these elements belonged to the serine recombinase superfamily (Figure 7).

16S rRNA sequences from the stigonematalean *Mastigocladopsis repens* PCC 10914 and *Mastigocoleus testarum* BC008 were interspersed among sequences from Nostocales (Figure 2). However, the rest of the sequences from Stigonematales clustered together with strong bootstrap support, and *Fischerella* strains formed two subclusters within this group (Figure 2). The tree structure largely agrees with a 16S rRNA phylogenetic tree that included a majority of the genomes used in the present study (Shih *et al.*, 2013).

The presence of element variants with more than three occurrences in the heterocyst-forming cyanobacterial genomes was plotted along the 16S rRNA phylogeny (Figure 9). The organisms that possessed a variant that was found in more than six genomes were spread out throughout the 16S rRNA phylogenetic tree.

However, the organisms with less frequently observed variants (six or less occurrences) generally clustered together. The organisms that possess the element variant that interrupts *nifD* at position 895 (± 1 bp) are the exception to this trend. The 16S rRNA sequences of *Calothrix* sp. PCC 7103 and *Rivularia* sp. PCC 7116 are relatively distant from other four organisms that contain this element variant.

Discussion

Although interruption elements have been documented within genes of heterocyst-forming cyanobacteria for many years (Rice *et al.*, 1982; Golden *et al.*, 1985; Lammers *et al.*, 1986), questions still remain about these genomic features. The origin of the elements is unknown, as well as how the elements have evolved since their integration into cyanobacterial genomes. It is also unclear if the elements provide an advantage or disadvantage to heterocyst-forming cyanobacteria or if they are selfish DNA, and have little or no effect on the organism (Orgel *et al.*, 1980). A *nifD* element has been deleted in the laboratory from two different *Nostoc* strains. Expression of *xisA* in trans lead to deletion of the *nifD* element in all cells of *Nostoc* 7120, yet the strain grew identical to the wild type in the presence of combined N or with N_2 as the sole N source (Brusca *et al.*, 1990). Cultures of *Nostoc* sp. strain Mac form defective heterocysts and fix N_2 only under microoxic incubation conditions (Meeks *et al.*, 1994). Spontaneous revertants of *Nostoc* Mac can be isolated that grow and fix N_2 under oxic conditions. One out of 18 revertant strains (and out of 205 additional non N_2-fixing clones in oxic conditions) had deleted the *nifD* element; repression of heterocyst differentiation and nitrogenase expression by combined N and induction of differentiation and N_2 fixation following the immediate removal of combined N were identical in the revertants possessing or lacking a *nifD* element. Thus, the presence or absence of this element has no apparent selective advantage under laboratory growth conditions. In the present study, all of the sequenced heterocyst-forming cyanobacterial genomes for interruption elements in order to investigate these sequences further.

<u>Interrupted genes</u>

New variants of interruption elements were found in the analysis. Many interruption elements were identified within orthologs that were not previously reported to contain elements. Among these are several genes known to be important to N_2 fixation or maintenance of a microoxic microenvironment in heterocysts. The products of *nifE* and *nifB* are involved in the synthesis of the nitrogenase Fe/Mo cofactor (Roberts *et al.*, 1978). *hupS* encodes an uptake hydrogenase small subunit and is in an operon with *hupL*. *nifJ* encodes an oxidoreductase that is responsible for electron transfer to nitrogenase, and is required for N_2 fixation when iron is limiting (Shah *et al.*, 1983; Bauer *et al.*, 1993). A previous study of *Nostoc* 7120 indicated that expression of genes orthologous to the two element-containing cytochrome c oxidase subunit I copies, *coxA2* and *coxA3*, are specific to heterocysts, while a third copy, *coxA1*, is expressed only in vegetative cells (Valladares *et al.*, 2003). The role of cytochrome c oxidase in heterocysts is in respiratory O_2 consumption (Valladares *et al.*, 2003). Similarly, the *flv3B* gene product of *Nostoc* 7120 is a heterocyst-specific flavin reductase that reduces O_2 (Ermakova *et al.*, 2013). The fatty acid synthase encoded by *hglE* is involved in the formation of heterocyst glycolipids, which form a layer in the envelope that acts to prevent gas diffusion into the heterocyst (Campbell *et al.*, 1997). Given the heterocyst-specific roles of these genes, the excision of their interruption elements is likely dependent on transcription of the respective *xis* during heterocyst formation, as is assumed to be the case with all of the previously-reported elements (Golden *et al.*, 1985).

The presence of interruption elements within genes with no known role in N_2 fixation indicates the excision of some elements could be brought on by factors other than heterocyst development. These interrupted genes may indeed have heterocyst-specific roles, or the element excision could be triggered by differentiation of other cell types. Many heterocyst-forming cyanobacteria also form akinetes (Kaplan-Levy *et al.*, 2010) and hormogonia (Herdman & Rippka, 1988). Additionally, the unicellular *Chroococcidiopsis thermalis* PCC 7203 forms survival cells upon nitrogen limitation (Billi & Grilli-Caiola, 1996); the excision of the interruption element from a membrane protein could be triggered by the formation of these cells. This idea is supported by a phylogenetic relationship between *Chroococcidiopsis* and heterocyst-forming cyanobacteria that provides a possible evolutionary link between these survival cells and heterocysts (Fewer *et al.*, 2002). Thus, interruption element excision triggered by cell differentiation may be shared between these organisms.

Excision of an interruption element during differentiation of non-heterocyst cell types may require additional mechanisms for the element to persist. Excision of an element during development of a non-terminal cell type, such as hormogonia, akinetes, or the *Chroococcidiopsis* survival cells, would require reintegration into the chromosome upon differentiation back into a vegetative cell, and Tn3 family transposases are capable of integrating a donor sequence into a target sequence (Grindley *et al.*, 2006). Tn3 family transposases are found adjacent to the recombinase genes of the interruption elements within the *Calothrix* sp. PCC 6303 arabinose efflux permease and the *Chroococcidiopsis thermalis* PCC 7203 membrane protein, making each a likely candidate to be excised from the chromosome in a non-terminal manner. Reintegration of these elements could also explain why the *xis* on these two elements are the only two recombinases identified in this study that are oriented in different ways than the others. While element excision may be triggered by differentiation of various cell types, the mechanisms could differ depending on the cell type.

Factors other than cell differentiation should also be considered for triggers of interruption element excision in heterocyst-forming cyanobacteria. However, similar to the case of a non-terminal cell type, the excision of interruption elements throughout a population during a single event would result in the element being present in future generations exclusively as extrachromosomal DNA. Reintegration would again be required, but the interruption elements found in unicellular cyanobacteria do not contain transposases. The presence of interruption elements in *Synechococcus* sp. PCC 7502 and *Cyanothece* sp. PCC 7822, neither of which has been shown to differentiate cells of any kind, supports the possibility of other factors prompting genome rearrangements.

Evolution of elements

There was considerable difference in frequency of elements between the two orders of heterocyst-forming cyanobacteria. Based on the genomic sequences available, interruption elements are more frequent in the order Nostocales than those within the Stigonematales. Heterocyst-forming cyanobacteria are assigned to either order by their phenotypic ability (Stigonematales), or inability (Nostocales), to form filament branches by undergoing cellular division in more than one plane (Gugger & Hoffmann, 2004; Henson *et al.*, 2004; Singh *et al.*, 2013).

Assuming that formation of a branched filament with cells of smaller size than the primary trunk is more evolutionarily advanced than an unbranched filament, many, or most, of the interruption elements may have originated within the heterocyst-forming cyanobacterial lineage after a common ancestor of the majority of branching organisms had separated from the unbranched Nostocales; the difference in interruption element frequency could correspond to an evolutionary distinction between the two orders.

The two recombinase superfamilies responsible for interruption element excision from heterocyst-forming cyanobacterial genes have different evolutionary histories. The tyrosine and serine recombinase superfamilies are named for the amino acid residue of each that covalently binds to the DNA, and each superfamily uses a different mechanism for recombination (Grindley *et al.*, 2006). The lack of an evolutionary relationship between the recombinase superfamilies indicates there have been at least two events in which the heterocyst-forming cyanobacterial lineage has acquired interruption elements, one involving each type of recombinase.

The relatedness of the recombinase sequences within each superfamily provides clues as to how the interruption elements have evolved in heterocyst-forming cyanobacteria. In general, the recombinase sequences clustered with those from the same element variant in other organisms. The phylogeny of the sequences within those clusters largely agrees with 16S rRNA phylogeny, indicating element acquisition through past generations as previously hypothesized (Henson *et al.*, 2011). However, the *Anabaena cylindrica* PCC 7122 and *Scytonema hofmanni* UTEX 2349 sets of recombinase sequences clustered with others from the same organism, and are likely the result of duplication events within the genome. The *Rivularia* sp. PCC 7116 *xis* sequences responsible for the *nifB* and *hupS* elements are the only example of tyrosine recombinase sequences clustering together, and they are much more divergent than those serine recombinase clusters. Additionally, serine superfamily recombinases are found in interruption elements interrupting a wider variety of orthologs than tyrosine superfamily recombinases, including all of the genes with no known N_2 fixation role. The relationships of the available recombinase sequences provide evidence that serine superfamily recombinases have a greater tendency to integrate into a genome or replicon, or replicate within a genome than tyrosine superfamily recombinases. This replication is a characteristic of selfish DNA (Orgel *et al.*, 1980), and is likely the mechanism for the origin of an element variant.

A relative timeline for the origin of each element variant is provided by the frequency of the variant, as well as the phylogeny of the organisms that possess it. The variants that were found in many genomes were concurrently found in organisms widely spread throughout the 16S rRNA phylogeny. In contrast, the variants that were found in fewer organisms generally exhibited a clear pattern of presence only in closely-related organisms. Thus, the frequently occurring variants likely originated earlier in heterocyst-forming cyanobacterial ancestry, while others have more recently formed. The element variant that interrupts *nifD* at position 1355 (± 2 bp) was, by far, the most common variant observed in the genomes studied here, and likely the first interruption element to integrate and persist in heterocyst-forming cyanobacteria. Therefore, it is this element variant that warrants special attention when examining the origin of interruption elements within this lineage.

Element gene content

Based on the relatedness of their recombinase genes, the interruption elements within each variant appear to have a strong evolutionary connection, yet there are very few physical similarities within variant groups. There seem to be no consistencies in length or GC content among all interruption elements of a single variant, indicating interruption elements are dynamic sequences that have undergone changes within the individual organisms. The highly variable gene content encoded on each interruption element is also indicative of a dynamic sequence with gene deletions and insertions taking place over time. Although the transcription of genes in elements before and after excision has not been well studied, they are present, and presumably functional, in the heterocyst as circular extrachromosomal elements (Golden *et al.*, 1985), so there may be little or no evolutionary pressure to preserve a gene specifically on an element. The transformations the interruption elements have undergone over time have erased any possible genetic signature that would have linked them to their origin.

However, there are some genes located on elements of multiple organisms that indicate a function specific to the element. The *xis* genes are an obvious example of this. A gene encoding a translation elongation factor G (TEFG) paralog might also have an element-specific function. TEFG paralogs were found on the *Richelia intracellularis* HH01 and *Ri. intracellularis* RC01 *nifH* elements, the *Calothrix* sp. PCC 7103 and *Ca. desertica* PCC 7102 *nifK* elements, and the *Rivularia* sp. PCC 7116 *nifD* element. The *Ca. rhizosoleniae* SC01 TEFG paralog is in an 11.9 kbp contig, and may be in one of several interruption elements that are interrupted by a contig break (*hupL, nifH, nifK,* or *flv3B*) or not in any interruption element. Similarly, the *Mastigocoleus testarum* BC008 TEFG paralog is located in a 40.5 kbp contig. No interruption element has been confirmed in that genome, but there are two possible *xis* genes that are on contig edges, and thus, are possible element recombinases. Therefore, the TEFG paralog may be exclusively located on interruption elements in heterocyst-forming cyanobacteria. However, the interruption elements containing the TEFG paralogs are not closely related, based on recombinase sequences, indicating the TEFG paralog was not simply inherited with that interruption element through speciation. In bacteria, TEFGs are involved in elongation during protein synthesis and ribosome recycling (Rodnina *et al.*, 1997; Hirokawa *et al.*, 2005; Zavialov *et al.*, 2005). In addition to a TEFG of cyanobacterial ancestry, a TEFG paralog is found in some cyanobacteria (Atkinson & Baldauf, 2011), including seven closely-related heterocyst-forming strains. Thus, the TEFG paralog may be preserved on interruption elements due to an element-specific function, such as a role in the synthesis of proteins encoded on the excised element in the heterocyst.

A gene present in multiple interruption elements may also be a result of conservation of the gene in the element as it is likely passed through the heterocyst-forming cyanobacterial lineage. *alr1449* is found only in one *nifD* element variant, indicative that it has been conserved on this specific element. The presence of this hypothetical protein outside of an interruption element in an additional eleven genomes, including four organisms that possessed no interruption elements in the entire genome, indicates the function of *alr1449* in heterocyst-forming cyanobacteria is not element-specific. Although the interruption elements in heterocyst-forming cyanobacteria display very little conservation, there are some genes that are exceptions to this.

Conclusions

The extensive analysis reported here provides structure for which to study interruption elements in cyanobacteria. Given the general lack of conserved gene content on the interruption elements, the recombinase sequences provide the sole evolutionary link amongst the elements. A lack of homology between the two recombinase superfamilies implies at least two distinct origins and evolutionary paths of interruption elements in heterocyst-forming cyanobacterial ancestry. However, the relatedness of recombinase sequences within each superfamily sheds light on the origin of element variants through recombinase replication within a genome, and shows a greater tendency for serine superfamily recombinases to undergo this replication. While the previously-reported *nifD* and *hupL* interruption elements are among the most frequently-occurring elements in heterocyst-forming cyanobacteria, analysis uncovered elements interrupting other gene orthologs, as well as variants of the interruption elements in previous reports. My findings indicate the possibility of factors beyond heterocyst differentiation as controlling excision of interruption elements. Through expansion of the interruption element data set, we have begun to trace the evolutionary paths of these unique genetic features and identified that their impact is broader than previously thought.

Acknowledgements

This work was sponsored in part by a Marine Investigator grant to J. Zehr and the MEGAMER facility (supported by the Gordon and Betty Moore Foundation, GBMF), NSF and the NSF Center for Microbial Oceanography: Research and Education. We thank I. Shilova and J. Robidart for comments on the manuscript.

Tables and Figures

Table 1. Genes interrupted by elements. The 22 orthologs that were interrupted by interruption elements in heterocyst-forming cyanobacteria, and three orthologs that were interrupted in non-heterocyst-forming cyanobacteria.

Gene Product	Gene Symbol	Reference Locus Tag	Elements	Variations
nitrogenase molybdenum-iron protein alpha subunit	*nifD*	Aazo_1353	34	5
nitrogenase iron protein	*nifH*	Aazo_1354	19	6
uptake hydrogenase large subunit	*hupL*	Aazo_3865	11	2
nitrogenase molybdenum-iron protein beta subunit	*nifK*	Aazo_1352	8	2
uptake hydrogenase small subunit	*hupS*	Aazo_3866	4	2
ferredoxin	*fdxN*	Aazo_1357	3	1
nitrogenase cofactor biosynthesis protein	*nifB*	Aazo_1358	2	2
nitrogenase molybdenum-iron cofactor biosynthesis protein	*nifE*	Aazo_1350	2	1
pyruvate ferredoxin/flavodoxin oxidoreductase	*nifJ*	Cal7507_5433	2	1
flavin reductase domain-containing FMN-binding protein	*flv3B*	Aazo_4140	2	1
cytochrome c oxidase subunit I	*coxA*	Aazo_2640	3	3
polyketide-type polyunsaturated fatty acid synthase	*hglE*	Aazo_3917	1	1
NADPH-dependent FMN reductase		Aazo_5221	1	1
FAD-dependent oxidoreductase		Cal7507_5656	1	1
phospholipase D/transphosphatidylase		Cal7507_0570	1	1
primase P4		Ana7108_2845	1	1
arabinose efflux permease		Mic7126DRAFT_5075	1	1
integrase family protein		Aazo_2682	1	1
transposase		Ava_B0242	1	1
transposase (ISSoc8)		Mas10914DRAFT_5058	1	1
caspase domain-containing protein		Cal7103DRAFT_00047390	1	1
hypothetical protein		CylstDRAFT_1988	1	1
Genes interrupted in non-heterocyst-forming cyanobacteria				
predicted integral membrane protein		Pse6802_3453	1	1
hypothetical protein		Cya7822_6696	1	1
ATP-dependent DNA helicase	*recQ*	Pse6802_0098	1	1

Table 2. Interruption elements. The 101 elements found in heterocyst-forming cyanobacteria and additional three elements in non-heterocyst-forming cyanobacteria. Direct repeat start positions are relative to the reference gene. The xis direction and location are reported as if the interrupted gene is in the positive direction. The length and GC content of elements that span multiple contigs (blue) may vary if additional contigs are included between the two identified here. Genome sequences not found in NCBI are identified instead by their JGI Scaffold ID (purple). The *Calothrix rhizosoleniae* SC01 genome used for this study has not been submitted to a public database (yellow). The Tol9009DRAFT_00004090 sequence is shortened due to a gap in sequencing and thus is not included on any phylogenetic tree (orange).

Heterocyst-forming Cyanobacterium	Element #	*xis* locus tag	Direct Repeat	Sequence Accession	Element Start	Element End	2nd Sequence Accession	Element Start	Element End
Anabaena circinalis AWQC131C	#2	131C_1121	AGCAGTTATATGG	KE384601	203587	205154	KE384600	1	6424
Anabaena circinalis AWQC131C	#1	131C_1156	TACTCCG	KE384600	47031	28589			
Anabaena circinalis AWQC310F	#4	310F_2868		KE384676	odd assembly		KE384689	odd assembly	
Anabaena circinalis AWQC310F	#2	310F_2836		KE384676	odd assembly		KE384689	odd assembly	
Anabaena circinalis AWQC310F	#3	310F_1894	CCGTGAAG	KE384676	61142	68895			
Anabaena circinalis AWQC310F	#1	310F_2885		KE384689	odd assembly		KE384676	odd assembly	
Anabaena cylindrica PCC 7122	#7	Anacy_1762	GCAGTTATATGG	NC_019771	2051506	2030665			
Anabaena cylindrica PCC 7122	#5	Anacy_2118	TATACCCTG	NC_019771	2414668	2489653			
Anabaena cylindrica PCC 7122	#4	Anacy_2214	TACTCCG	NC_019771	2490117	2505306			
Anabaena cylindrica PCC 7122	#6	Anacy_2116	CCGTGAAG	NC_019771	2407440	2413175			
Anabaena cylindrica PCC 7122	#3	Anacy_6026	AGTAGT	NC_019773	72642	112639			
Anabaena cylindrica PCC 7122	#2	Anacy_2144		NC_019771	odd assembly				
Anabaena cylindrica PCC 7122	#1	Anacy_2120	AGTATATG	NC_019771	2477490	2418266			
Anabaena sp. 90	#4	ANA_C11423	AGCAGTTATATGG	NC_019427	1576390	1565144			
Anabaena sp. 90	#2	ANA_C13532	TACTCCG	NC_019427	3981255	4001945			
Anabaena sp. 90	#1	ANA_C13501	ACTCTACTCGTTT	NC_019427	3893223	3973114			
Anabaena sp. 90	#3	ANA_C13510	CCGTGAAG	NC_019427	3973401	3979285			
Anabaena sp. PCC 7108	#4	Ana7108_2755	GCAGTTATATGG	NZ_KB235896	3144035	3147824			
Anabaena sp. PCC 7108	#2	Ana7108_2847	TATTCCCTG	NZ_KB235896	3252808	3232998			
Anabaena sp. PCC 7108	#1	Ana7108_2785	TACTCCG	NZ_KB235896	3232534	3174501			
Anabaena sp. PCC 7108	#3	Ana7108_2850	TCCGTGAAG	NZ_KB235896	3263265	3254311			
Anabaena variabilis ATCC 29413	#1	Ava_3928	GGATTACTCCG	NC_007413	4872417	4883490			
Calothrix desertica PCC 7102	#6	Cal7102DRAFT_02973	TCACT	2509891660	1997435	1985286			
Calothrix desertica PCC 7102	#2	Cal7102DRAFT_03503	GAATTTAC	2509891660	2568150	2570434			
Calothrix desertica PCC 7102	#7	Cal7102DRAFT_00422	ATTACTCCG	2509891660	2665820	2989987	2509891655	1	4513
Calothrix desertica PCC 7102	#8	Cal7102DRAFT_03508	TGGCTGA	2509891660	2574041	2575946			
Calothrix desertica PCC 7102	#1	Cal7102DRAFT_03510	GGCGGTTTCGCTA	2509891660	2576264	2605417			
Calothrix desertica PCC 7102	#5	Cal7102DRAFT_03532	ACTCC	2509891660	2605790	2664218			
Calothrix desertica PCC 7102	#4	Cal7102DRAFT_07832	GTCTTCAAAAGA	2509891662	2142499	2136292			
Calothrix desertica PCC 7102	#3	Cal7102DRAFT_00467	GTTC	2509891655	6150	70961			
Calothrix rhizosoleniae SC01	#2	15818.peg.4808	TAGTTCATG	contig69776	3940	1	contig69372	1	4208
Calothrix rhizosoleniae SC01	#4	15818.peg.2817	CACAGCAGTTATATGG	contig69757	5582	8787	contig68560	1	3384
Calothrix rhizosoleniae SC01	#1	15818.peg.3760*	GTGGCGGTTTCGCTA	contig69345	21875	1	contig00132	1	62
Calothrix rhizosoleniae SC01	#3	15818.peg.1389	TAG	contig00132	3692	21713	contig01068	16608	14554
Calothrix sp. PCC 6303	#3	Cal6303_2021	TTT	NC_019751	2407697	3392169			
Calothrix sp. PCC 6303	#4	Cal6303_1693	AGGTAA	NC_019751	2039332	2017055			
Calothrix sp. PCC 6303	#8	Cal6303_0402	CCCACAGCAGTTATATGGTG	NC_019751	484448	475144			
Calothrix sp. PCC 6303	#7	Cal6303_0224	GGATTACTCCG	NC_019751	275147	260910			
Calothrix sp. PCC 6303	#6	Cal6303_0220		NC_019751	odd assembly				
Calothrix sp. PCC 6303	#5	Cal6303_0394	GTGGCGGTTTC	NC_019751	465954	277108			
Calothrix sp. PCC 6303	#2	Cal6303_1564	TTA	NC_019751	1839406	1833940			
Calothrix sp. PCC 6303	#1	Cal6303_5631	GGTACT	NC_019727	4602	743			
Calothrix sp. PCC 7103	#5	Cal7103DRAFT_00055580	TAAAGGTGGAAACC	NZ_KB217478	3797020	3776214			
Calothrix sp. PCC 7103	#8	Cal7103DRAFT_00041270	ATATTAG	NZ_KB217478	2155864	2170547			
Calothrix sp. PCC 7103	#4	Cal7103DRAFT_00041060	GTC	NZ_KB217478	2124376	2154808			
Calothrix sp. PCC 7103	#11	Cal7103DRAFT_00042040	CAATCAAAGA	NZ_KB217478	2256871	2255122			
Calothrix sp. PCC 7103	#10	Cal7103DRAFT_00042020	TGCCTTGGATG	NZ_KB217478	2254452	2252784			
Calothrix sp. PCC 7103	#9	Cal7103DRAFT_00041710	TACTCCG	NZ_KB217478	2252322	2217103			
Calothrix sp. PCC 7103	#2	Cal7103DRAFT_00044010	GTGGT	NZ_KB217478	2473187	2378226			
Calothrix sp. PCC 7103	#3	Cal7103DRAFT_00042750	GGCGGTTTCGCTA	NZ_KB217478	2378073	2257718			
Calothrix sp. PCC 7103	#7	Cal7103DRAFT_00006510	TTCAAAAGA	NZ_KB217483	736187	769815			
Calothrix sp. PCC 7103	#6	Cal7103DRAFT_00041480	GTTC	NZ_KB217478	2215522	2186759			
Calothrix sp. PCC 7103	#1	Cal7103DRAFT_00003030	ATACCT	NZ_KB217483	548754	345459			
Calothrix sp. PCC 7507	#1	Cal7507_0111	ATTACTCCG	NC_019682	83026	106377			
Chlorogloeopsis fritschii PCC 6912	#1	UYCDRAFT_06247	GGATTACTCCGG	NZ_AJLN01000143	8072	15970			
Chlorogloeopsis sp. PCC 9212	#1	UYEDRAFT_00081	GGATTACTCCGG	NZ_AJLM01000008	76091	68193			
Cylindrospermum stagnale PCC 7417	#3	CylstDRAFT_5455	AGCAGTTATATGG	NC_019757	6139921	6133099			
Cylindrospermum stagnale PCC 7417	#2	CylstDRAFT_5410	TACCCTG	NC_019757	6098326	6106942			
Cylindrospermum stagnale PCC 7417	#1	CylstDRAFT_5431	TACTCCG	NC_019757	6107404	6116724			
Fischerella sp. PCC 9339	#2	PCC9339DRAFT_03888		JH992898	odd assembly				
Fischerella sp. PCC 9339	#1	PCC9339DRAFT_03891		JH992898	odd assembly				
Fischerella sp. PCC 9431	#1	Fis9431DRAFT_4159	GGATTACTCCG	KE650771	4811296	4802026			
Mastigocladopsis repens PCC 10914	#1	Mas10914DRAFT_0641	GGATTACTCCGG	NZ_JH992901	700103	687410			
Microchaete sp. PCC 7126	#2	Mic7126DRAFT_0221	TATTCCCTG	NZ_KB235931	231670	237545			
Microchaete sp. PCC 7126	#1	Mic7126DRAFT_0254	GTTCT	NZ_KB235931	239631	275825			
Nodularia spumigena CCY9414	#3	N9414_14930	AGCAGTTATATGGT	NZ_AAVW01000003	73105	80653			
Nodularia spumigena CCY9414	#1	N9414_15040	TACTCCG	NZ_AAVW01000003	102063	97724			
Nodularia spumigena CCY9414	#2	N9414_15065	CCGTGAAG	NZ_AAVW01000003	109184	104012			
Nostoc punctiforme PCC 73102	#2	Npun_F0392	GGATTACTCCG	NC_010628	511393	487671			
Nostoc punctiforme PCC 73102	#1	Npun_R2637	CTACA	NC_010628	3198696	3271351			
Nostoc sp. PCC 7107	#1	Nos7107_3361	TATTC	NC_019676	3842092	3883301			
Nostoc sp. PCC 7107	#2	Nos7107_3373	GGATTACTCCG	NC_019676	3888954	3897528			
Nostoc sp. PCC 7120	#1	alr1459	TATTC	NC_003272	1776227	1716800			
Nostoc sp. PCC 7120	#3	alr0677	CACAGCAGTTATATGG	NC_003272	794972	785538			
Nostoc sp. PCC 7120	#2	alr1442	GGATTACTCCG	NC_003272	1711911	1700623			
Nostoc sp. PCC 7524	#1	Nos7524_1241	CTATTCACAGAAATATT	NC_019684	1422173	1397486			
Nostoc sp. PCC 7524	#2	Nos7524_1212	GGATTACTCCG	NC_019684	1386811	1370962			
Nostoc sp. PCC 7524	#4	Nos7524_1231	TCCGTGAAG	NC_019684	1394473	1388762			
Nostoc sp. PCC 7524	#3	Nos7524_1209	ATCGGTGA	NC_019684	1370150	1364533			

Table 2 (cont.)

Heterocyst-forming Cyanobacterium	Element #	xis locus tag	Direct Repeat	Sequence Accession	Element Start	Element End	2nd Sequence Accession	Element Start	Element End
Richelia intracellularis HH01	#1	RINTHH_3080	ATGCGGTGGTTTTGCTA	NZ_CAIY01000015	10856	20002			
Richelia intracellularis RC01	#1	RintRC_2216	TTACTCCG	CBZS010000430	14931	17603			
Richelia intracellularis RC01	not on tree	xis gene not located	ATGCGGTGGTTT	CBZS010000660	772	847	CBZS010000430	1	12918
Rivularia sp. PCC 7116	#1	Riv7116_6196	ATGTCAT	NC_019678	7756636	7847669			
Rivularia sp. PCC 7116	#7	Riv7116_6198	ACAAGAA	NC_019678	7848231	7850058			
Rivularia sp. PCC 7116	#8	Riv7116_6362	ACTATGCAAGAA	NC_019678	8063805	8061938			
Rivularia sp. PCC 7116	#9	Riv7116_6312	CGATTA	NC_019678	7994320	7992586			
Rivularia sp. PCC 7116	#4	Riv7116_6308	GTGGTCTGGTGA	NC_019678	7992035	7983223			
Rivularia sp. PCC 7116	#6	Riv7116_6305	ATTCCCTGGAT	NC_019678	7983104	7979758			
Rivularia sp. PCC 7116	#5	Riv7116_6281	GGATTACTCCG	NC_019678	7979299	7955109			
Rivularia sp. PCC 7116	#3	Riv7116_6354	ACACC	NC_019678	8056757	7994807			
Rivularia sp. PCC 7116	#2	Riv7116_6219	GTTCT	NC_019678	7953571	7868842			
Scytonema hofmanni PCC 7110	#2	WA1DRAFT_02505	AACCTAG	NZ_ANNX01000031	60296	12813			
Scytonema hofmanni PCC 7110	#5	WA1DRAFT_03870	TTACTCCG	NZ_ANNX01000057	50146	52257	NZ_ANNX01000058	1	17029
Scytonema hofmanni PCC 7110	#4	WA1DRAFT_03874	TGAACC	NZ_ANNX01000058	20259	29807			
Scytonema hofmanni PCC 7110	#3	WA1DRAFT_03849	GTTT	NZ_ANNX01000057	41897	47918			
Scytonema hofmanni PCC 7110	#1	WA1DRAFT_10544	ACTTATAA	NZ_ANNX01000202	11217	1	NZ_ANNX01000081	64817	21184
Scytonema hofmanni UTEX 2349	#2	Tol9009DRAFT_00061200	ATGTCATTTCTC	2507267910	680077	422069			
Scytonema hofmanni UTEX 2349	#4	Tol9009DRAFT_00060810	GGATTACTCCG	2507267910	365062	388701			
Scytonema hofmanni UTEX 2349	#3	Tol9009DRAFT_00060970	TCTTT	2507267910	390307	406345			
Scytonema hofmanni UTEX 2349	not on tree	Tol9009DRAFT_00004090	GCTAAG	2507267906	87754	89054			
Scytonema hofmanni UTEX 2349	#1	Tol9009DRAFT_00005880	GGTAGG	2507267907	33324	29243			
Non-heterocyst-forming Cyanobacterium									
Chroococcidiopsis thermalis PCC 7203	#1	Chro_3279	GAATTG	NC_019695	3668489	3664613			
Cyanothece sp. PCC 7822	#1	Cyan7822_6640	TCTATG	NC_014504	442	4175			
Synechococcus sp. PCC 7502	#1	Syn7502_02805	GGGAA	NC_019702	2770351	2818033			

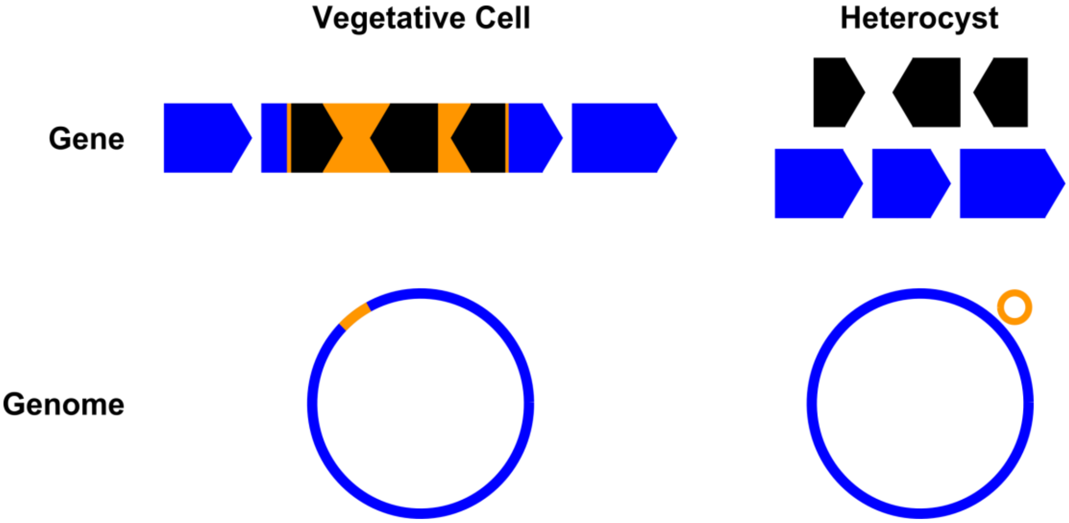
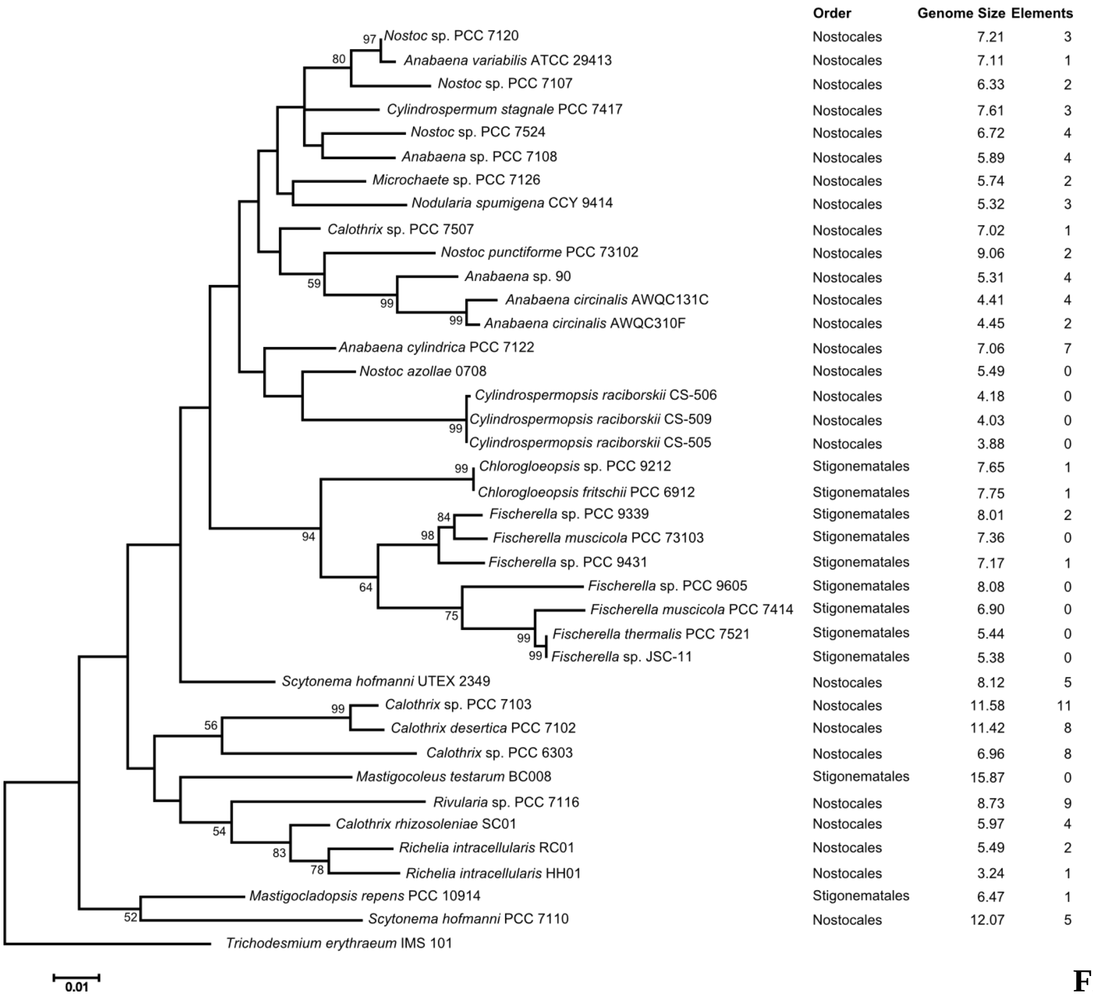


Figure 1. Element excision cartoon. Cartoon depiction of an element (orange, genes on element are black) within the chromosome (blue) in vegetative cells and its presence in heterocyst cells as a separate DNA element, as viewed at the gene and genome levels.



	Order	Genome Size	Elements
Nostoc sp. PCC 7120	Nostocales	7.21	3
Anabaena variabilis ATCC 29413	Nostocales	7.11	1
Nostoc sp. PCC 7107	Nostocales	6.33	2
Cylindrospermum stagnale PCC 7417	Nostocales	7.61	3
Nostoc sp. PCC 7524	Nostocales	6.72	4
Anabaena sp. PCC 7108	Nostocales	5.89	4
Microchaete sp. PCC 7126	Nostocales	5.74	2
Nodularia spumigena CCY 9414	Nostocales	5.32	3
Calothrix sp. PCC 7507	Nostocales	7.02	1
Nostoc punctiforme PCC 73102	Nostocales	9.06	2
Anabaena sp. 90	Nostocales	5.31	4
Anabaena circinalis AWQC131C	Nostocales	4.41	4
Anabaena circinalis AWQC310F	Nostocales	4.45	2
Anabaena cylindrica PCC 7122	Nostocales	7.06	7
Nostoc azollae 0708	Nostocales	5.49	0
Cylindrospermopsis raciborskii CS-506	Nostocales	4.18	0
Cylindrospermopsis raciborskii CS-509	Nostocales	4.03	0
Cylindrospermopsis raciborskii CS-505	Nostocales	3.88	0
Chlorogloeopsis sp. PCC 9212	Stigonematales	7.65	1
Chlorogloeopsis fritschii PCC 6912	Stigonematales	7.75	1
Fischerella sp. PCC 9339	Stigonematales	8.01	2
Fischerella muscicola PCC 73103	Stigonematales	7.36	0
Fischerella sp. PCC 9431	Stigonematales	7.17	1
Fischerella sp. PCC 9605	Stigonematales	8.08	0
Fischerella muscicola PCC 7414	Stigonematales	6.90	0
Fischerella thermalis PCC 7521	Stigonematales	5.44	0
Fischerella sp. JSC-11	Stigonematales	5.38	0
Scytonema hofmanni UTEX 2349	Nostocales	8.12	5
Calothrix sp. PCC 7103	Nostocales	11.58	11
Calothrix desertica PCC 7102	Nostocales	11.42	8
Calothrix sp. PCC 6303	Nostocales	6.96	8
Mastigocoleus testarum BC008	Stigonematales	15.87	0
Rivularia sp. PCC 7116	Nostocales	8.73	9
Calothrix rhizosoleniae SC01	Nostocales	5.97	4
Richelia intracellularis RC01	Nostocales	5.49	2
Richelia intracellularis HH01	Nostocales	3.24	1
Mastigocladopsis repens PCC 10914	Stigonematales	6.47	1
Scytonema hofmanni PCC 7110	Nostocales	12.07	5
Trichodesmium erythraeum IMS 101			

Figure 2. **Heterocyst-forming cyanobacteria 16S phylogeny**. Maximum likelihood phylogenetic tree of 16S rRNA from 38 heterocyst-forming cyanobacterial genomes, rooted with *Trichodesmium erythraeum* IMS 101.

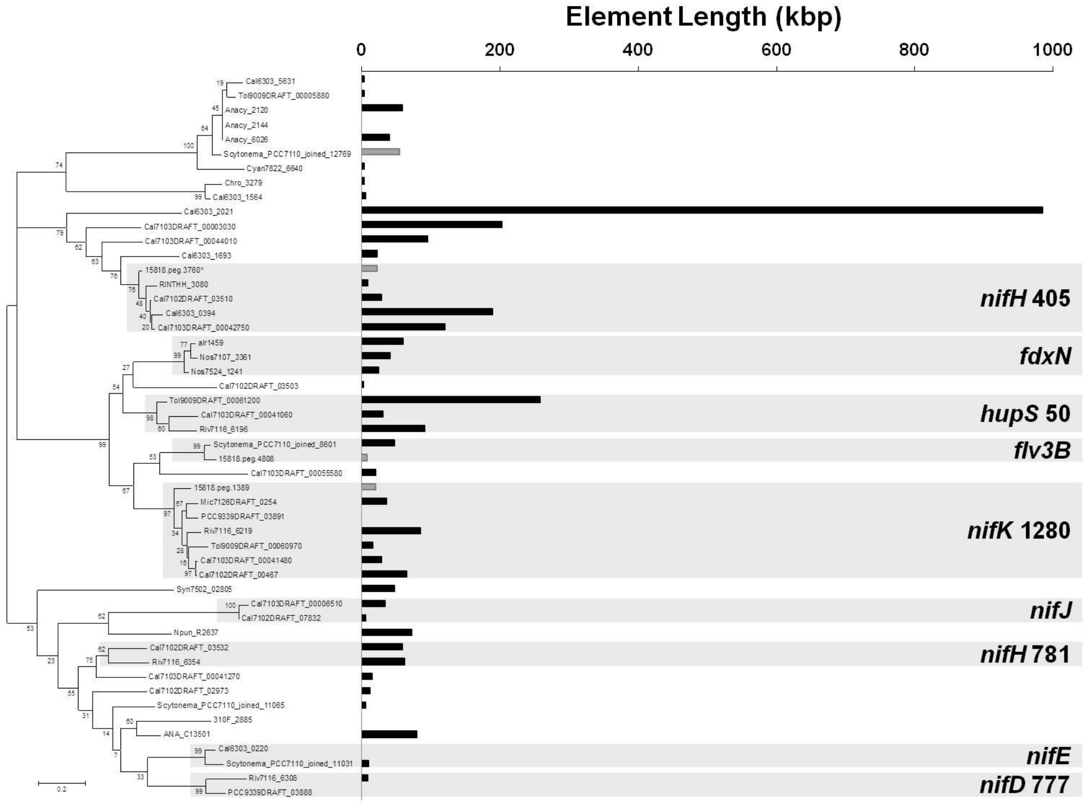


Figure 3. Serine recombinase phylogeny with element lengths. Maximum likelihood phylogenetic tree of serine recombinase protein sequences and the length of the element each *xis* gene is found on. Elements in gray bars are not on a single contig, and thus, may be longer.

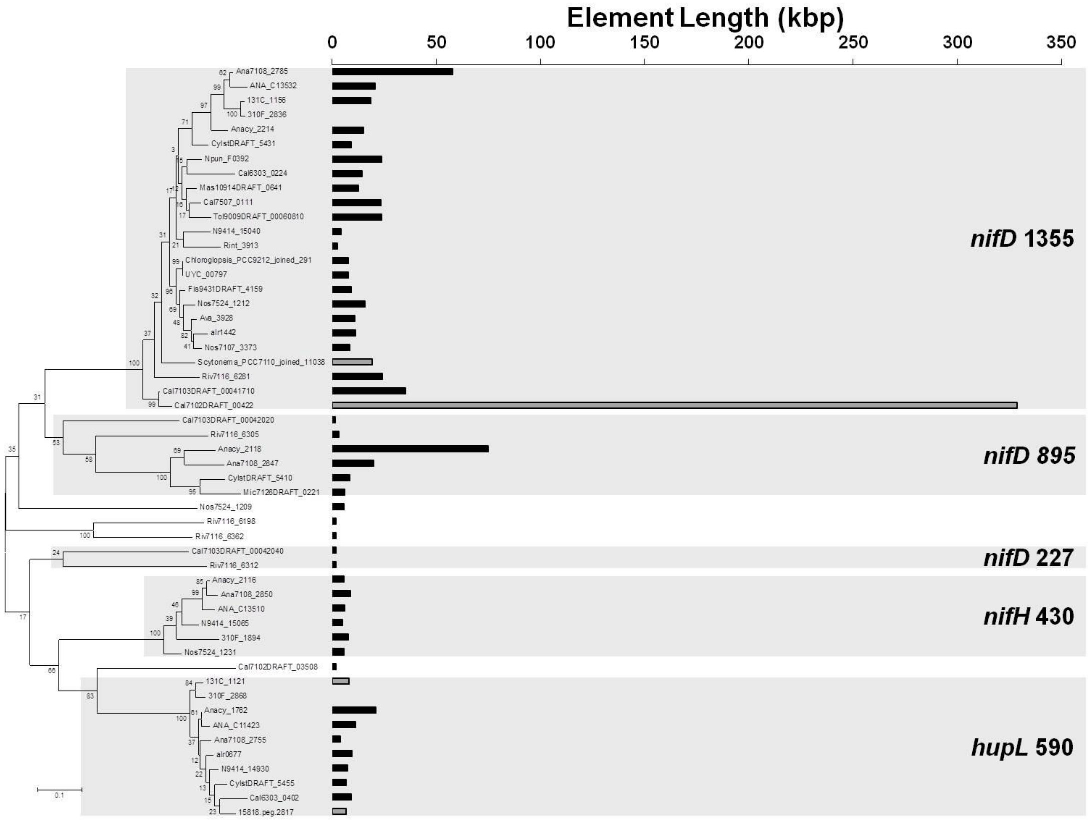


Figure 4. Tyrosine recombinase phylogeny with element lengths. Maximum likelihood phylogenetic tree of tyrosine recombinase protein sequences and the length of the element each *xis* gene is found on. Elements in gray bars are not on a single contig, and thus, may be longer.

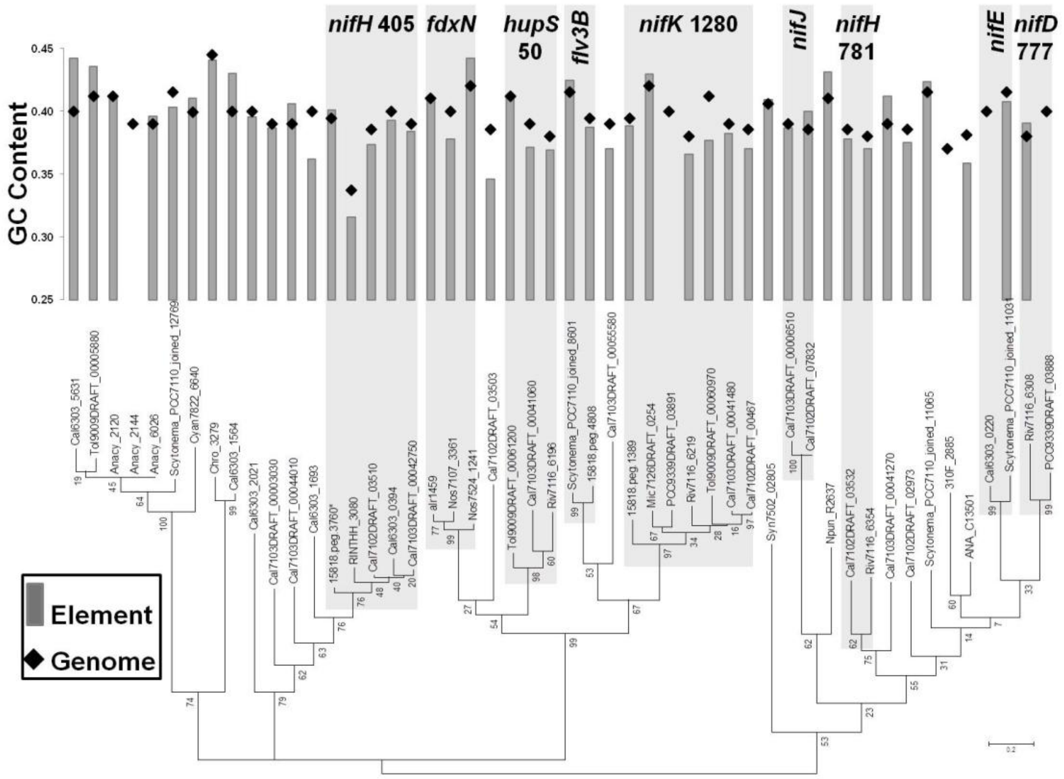


Figure 5. Serine recombinase phylogeny with element GC content. Maximum likelihood phylogenetic tree of serine recombinase protein sequences and the GC content of the element each *xis* gene is found on (gray bar) and the genome it is in (black diamond).

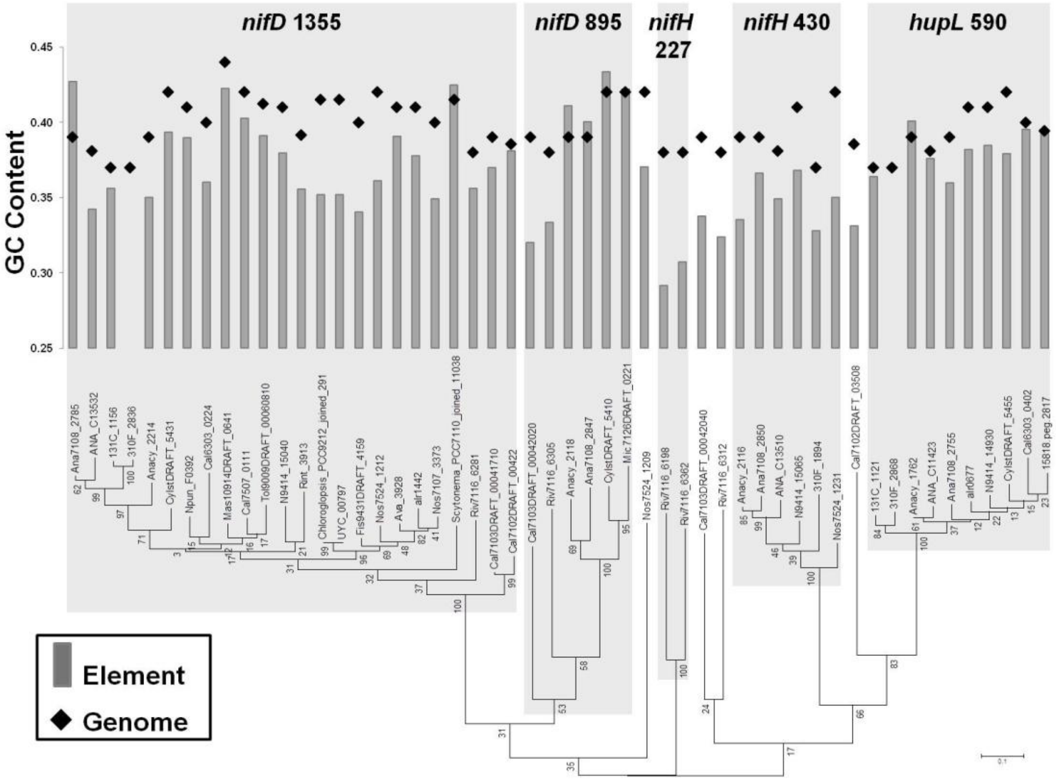
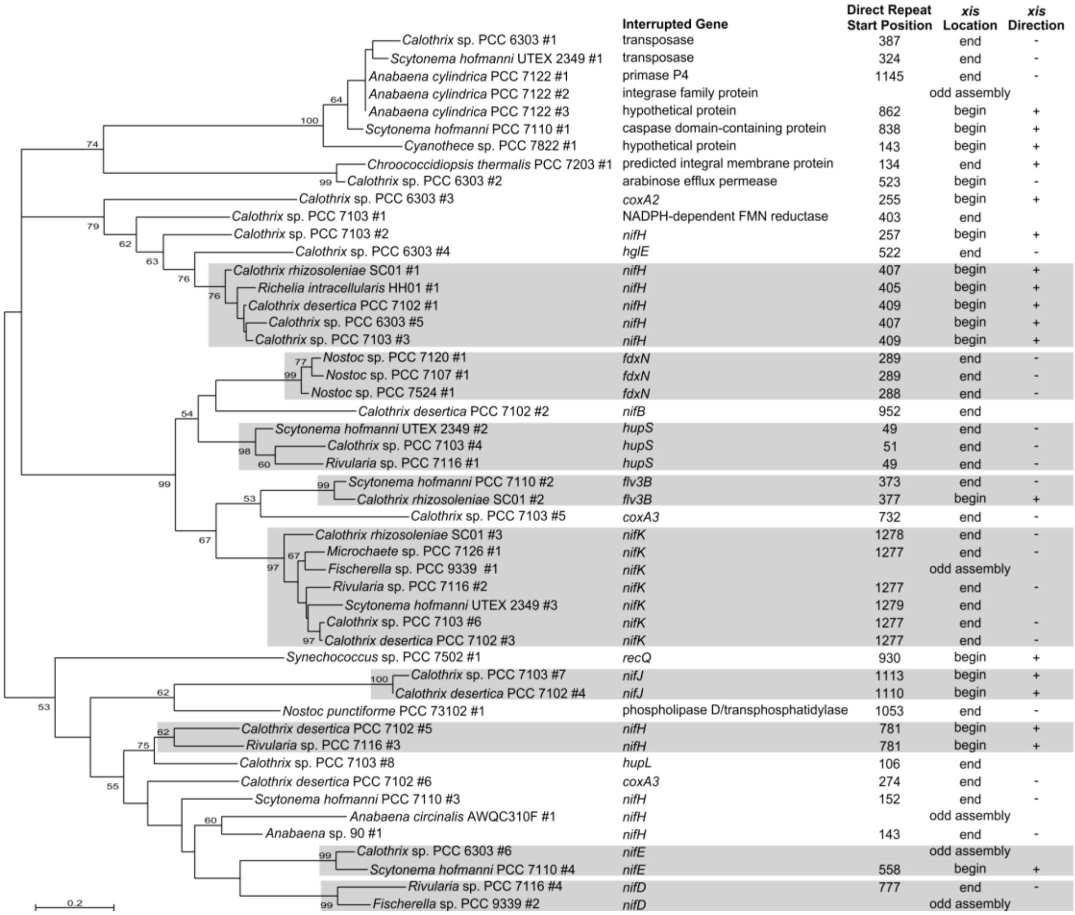
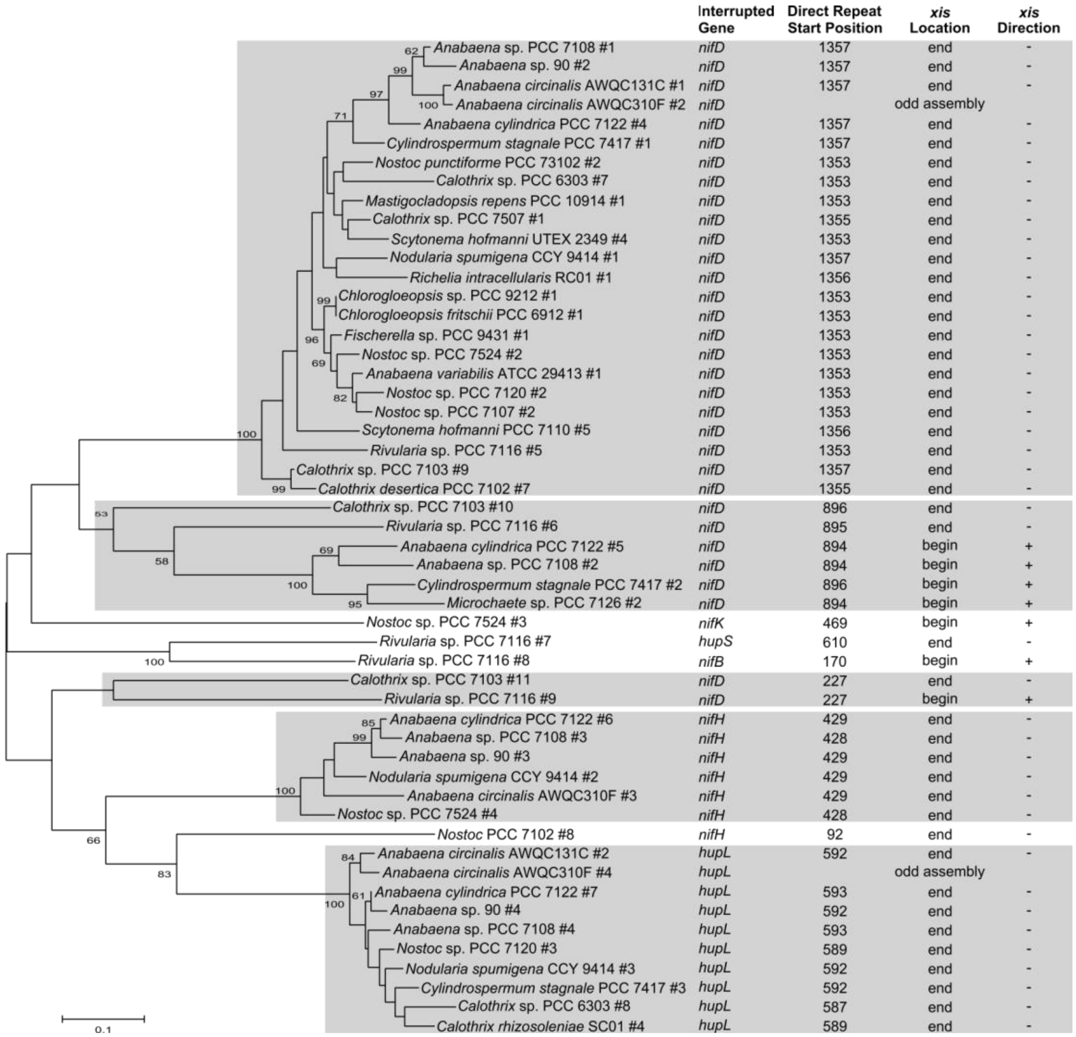


Figure 6. Tyrosine recombinase phylogeny with element GC content. Maximum likelihood phylogenetic tree of tyrosine recombinase protein sequences and the GC content of the element each *xis* gene is found on (gray bar) and the genome it is in (black diamond).



	Interrupted Gene	Direct Repeat Start Position	*xis* Location	*xis* Direction
Calothrix sp. PCC 6303 #1	transposase	387	end	-
Scytonema hofmanni UTEX 2349 #1	transposase	324	end	-
Anabaena cylindrica PCC 7122 #1	primase P4	1145	end	-
Anabaena cylindrica PCC 7122 #2	integrase family protein		odd assembly	
Anabaena cylindrica PCC 7122 #3	hypothetical protein	862	begin	+
Scytonema hofmanni PCC 7110 #1	caspase domain-containing protein	838	begin	+
Cyanothece sp. PCC 7822 #1	hypothetical protein	143	begin	+
Chroococcidiopsis thermalis PCC 7203 #1	predicted integral membrane protein	134	end	+
Calothrix sp. PCC 6303 #2	arabinose efflux permease	523	begin	-
Calothrix sp. PCC 6303 #3	*coxA2*	255	begin	+
Calothrix sp. PCC 7103 #1	NADPH-dependent FMN reductase	403	end	
Calothrix sp. PCC 7103 #2	*nifH*	257	begin	+
Calothrix sp. PCC 6303 #4	*hglE*	522	end	-
Calothrix rhizosoleniae SC01 #1	*nifH*	407	begin	+
Richelia intracellularis HH01 #1	*nifH*	405	begin	+
Calothrix desertica PCC 7102 #1	*nifH*	409	begin	+
Calothrix sp. PCC 6303 #5	*nifH*	407	begin	+
Calothrix sp. PCC 7103 #3	*nifH*	409	begin	+
Nostoc sp. PCC 7120 #1	*fdxN*	289	end	-
Nostoc sp. PCC 7107 #1	*fdxN*	289	end	-
Nostoc sp. PCC 7524 #1	*fdxN*	288	end	-
Calothrix desertica PCC 7102 #2	*nifB*	952	end	
Scytonema hofmanni UTEX 2349 #2	*hupS*	49	end	-
Calothrix sp. PCC 7103 #4	*hupS*	51	end	-
Rivularia sp. PCC 7116 #1	*hupS*	49	end	-
Scytonema hofmanni PCC 7110 #2	*flv3B*	373	end	-
Calothrix rhizosoleniae SC01 #2	*flv3B*	377	begin	+
Calothrix sp. PCC 7103 #5	*coxA3*	732	end	-
Calothrix rhizosoleniae SC01 #3	*nifK*	1278	end	-
Microchaete sp. PCC 7126 #1	*nifK*	1277	end	-
Fischerella sp. PCC 9339 #1	*nifK*		odd assembly	
Rivularia sp. PCC 7116 #2	*nifK*	1277	end	-
Scytonema hofmanni UTEX 2349 #3	*nifK*	1279	end	
Calothrix sp. PCC 7103 #6	*nifK*	1277	end	-
Calothrix desertica PCC 7102 #3	*nifK*	1277	end	-
Synechococcus sp. PCC 7502 #1	*recQ*	930	begin	+
Calothrix sp. PCC 7103 #7	*nifJ*	1113	begin	+
Calothrix desertica PCC 7102 #4	*nifJ*	1110	begin	+
Nostoc punctiforme PCC 73102 #1	phospholipase D/transphosphatidylase	1053	end	-
Calothrix desertica PCC 7102 #5	*nifH*	781	begin	+
Rivularia sp. PCC 7116 #3	*nifH*	781	begin	+
Calothrix sp. PCC 7103 #8	*hupL*	106	end	
Calothrix desertica PCC 7102 #6	*coxA3*	274	end	-
Scytonema hofmanni PCC 7110 #3	*nifH*	152	end	-
Anabaena circinalis AWQC310F #1	*nifH*		odd assembly	
Anabaena sp. 90 #1	*nifH*	143	end	-
Calothrix sp. PCC 6303 #6	*nifE*		odd assembly	
Scytonema hofmanni PCC 7110 #4	*nifE*	558	begin	+
Rivularia sp. PCC 7116 #4	*nifD*	777	end	-
Fischerella sp. PCC 9339 #2	*nifD*		odd assembly	

Figure 7. Serine recombinase phylogeny. Maximum likelihood phylogenetic tree of serine recombinase protein sequences and data characterizing the element each *xis* gene is found on.



	Interrupted Gene	Direct Repeat Start Position	*xis* Location	*xis* Direction
Anabaena sp. PCC 7108 #1	*nifD*	1357	end	-
Anabaena sp. 90 #2	*nifD*	1357	end	-
Anabaena circinalis AWQC131C #1	*nifD*	1357	end	-
Anabaena circinalis AWQC310F #2	*nifD*		odd assembly	
Anabaena cylindrica PCC 7122 #4	*nifD*	1357	end	-
Cylindrospermum stagnale PCC 7417 #1	*nifD*	1357	end	-
Nostoc punctiforme PCC 73102 #2	*nifD*	1353	end	-
Calothrix sp. PCC 6303 #7	*nifD*	1353	end	-
Mastigocladopsis repens PCC 10914 #1	*nifD*	1353	end	-
Calothrix sp. PCC 7507 #1	*nifD*	1355	end	-
Scytonema hofmanni UTEX 2349 #4	*nifD*	1353	end	-
Nodularia spumigena CCY 9414 #1	*nifD*	1357	end	-
Richelia intracellularis RC01 #1	*nifD*	1356	end	-
Chlorogloeopsis sp. PCC 9212 #1	*nifD*	1353	end	-
Chlorogloeopsis fritschii PCC 6912 #1	*nifD*	1353	end	-
Fischerella sp. PCC 9431 #1	*nifD*	1353	end	-
Nostoc sp. PCC 7524 #2	*nifD*	1353	end	-
Anabaena variabilis ATCC 29413 #1	*nifD*	1353	end	-
Nostoc sp. PCC 7120 #2	*nifD*	1353	end	-
Nostoc sp. PCC 7107 #2	*nifD*	1353	end	-
Scytonema hofmanni PCC 7110 #5	*nifD*	1356	end	-
Rivularia sp. PCC 7116 #5	*nifD*	1353	end	-
Calothrix sp. PCC 7103 #9	*nifD*	1357	end	-
Calothrix desertica PCC 7102 #7	*nifD*	1355	end	-
Calothrix sp. PCC 7103 #10	*nifD*	896	end	-
Rivularia sp. PCC 7116 #6	*nifD*	895	end	-
Anabaena cylindrica PCC 7122 #5	*nifD*	894	begin	+
Anabaena sp. PCC 7108 #2	*nifD*	894	begin	+
Cylindrospermum stagnale PCC 7417 #2	*nifD*	896	begin	+
Microchaete sp. PCC 7126 #2	*nifD*	894	begin	+
Nostoc sp. PCC 7524 #3	*nifK*	469	begin	+
Rivularia sp. PCC 7116 #7	*hupS*	610	end	-
Rivularia sp. PCC 7116 #8	*nifB*	170	begin	+
Calothrix sp. PCC 7103 #11	*nifD*	227	end	-
Rivularia sp. PCC 7116 #9	*nifD*	227	begin	+
Anabaena cylindrica PCC 7122 #6	*nifH*	429	end	-
Anabaena sp. PCC 7108 #3	*nifH*	428	end	-
Anabaena sp. 90 #3	*nifH*	429	end	-
Nodularia spumigena CCY 9414 #2	*nifH*	429	end	-
Anabaena circinalis AWQC310F #3	*nifH*	429	end	-
Nostoc sp. PCC 7524 #4	*nifH*	428	end	-
Nostoc PCC 7102 #8	*nifH*	92	end	-
Anabaena circinalis AWQC131C #2	*hupL*	592	end	-
Anabaena circinalis AWQC310F #4	*hupL*		odd assembly	
Anabaena cylindrica PCC 7122 #7	*hupL*	593	end	-
Anabaena sp. 90 #4	*hupL*	592	end	-
Anabaena sp. PCC 7108 #4	*hupL*	593	end	-
Nostoc sp. PCC 7120 #3	*hupL*	589	end	-
Nodularia spumigena CCY 9414 #3	*hupL*	592	end	-
Cylindrospermum stagnale PCC 7417 #3	*hupL*	592	end	-
Calothrix sp. PCC 6303 #8	*hupL*	587	end	-
Calothrix rhizosoleniae SC01 #4	*hupL*	589	end	-

Figure 8. Tyrosine recombinase phylogeny. Maximum likelihood phylogenetic tree of tyrosine recombinase protein sequences and data characterizing the element each *xis* gene is found on.

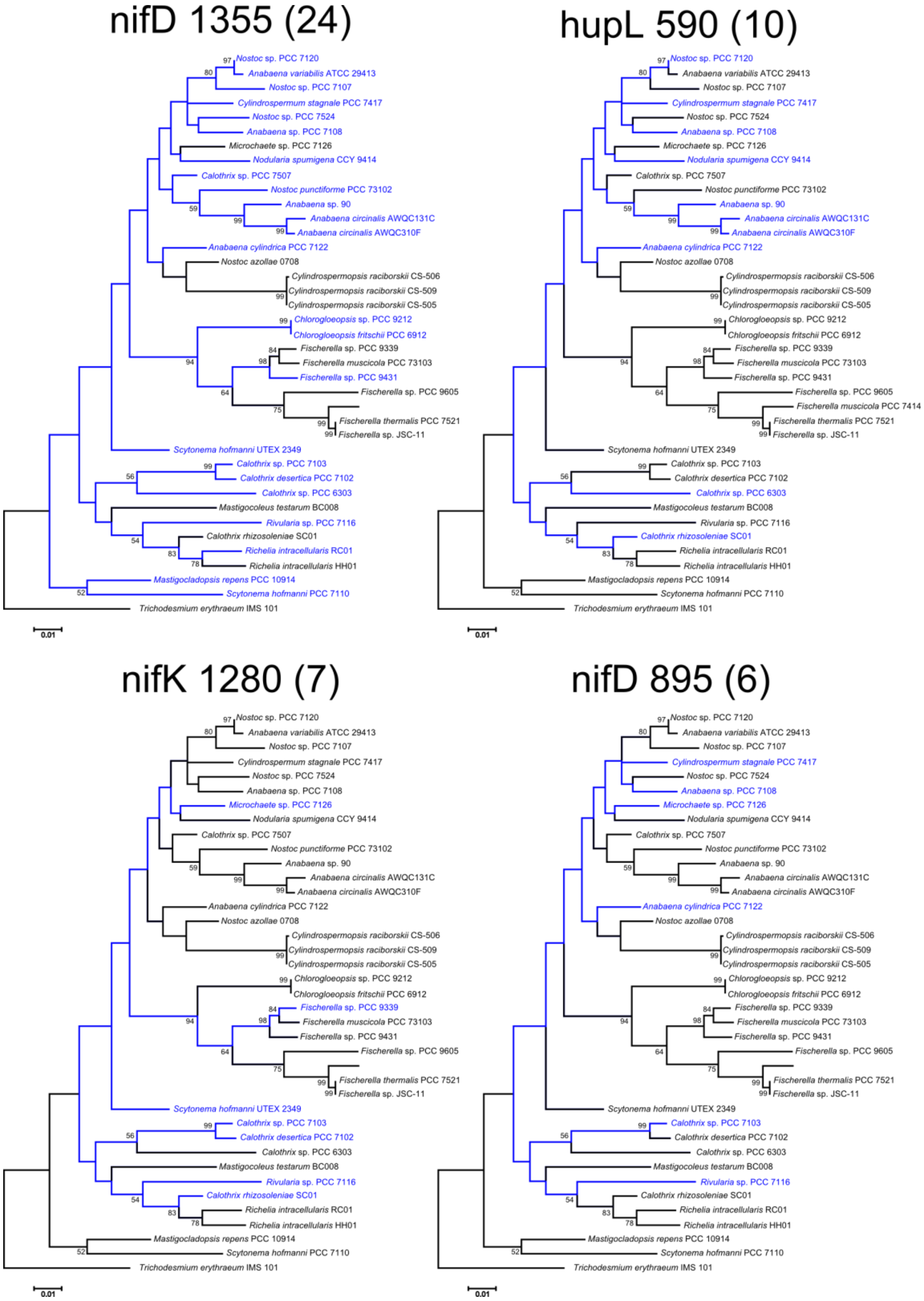

nifD 1355 (24)
hupL 590 (10)
nifK 1280 (7)
nifD 895 (6)
97
Nostoc sp. PCC 7120
80
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Cylindrospermum stagnale PCC 7417
Nostoc sp. PCC 7524
Anabaena sp. PCC 7108
Microchaete sp. PCC 7126
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Anabaena sp. 90
59
Anabaena circinalis AWQC131C
99
99
Anabaena circinalis AWQC310F
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Cylindrospermopsis raciborskii CS-506
Cylindrospermopsis raciborskii CS-509
99
Cylindrospermopsis raciborskii CS-505
99
Chlorogloeopsis sp. PCC 9212
Chlorogloeopsis fritschii PCC 6912
84
Fischerella sp. PCC 9339
94
98
Fischerella muscicola PCC 73103
Fischerella sp. PCC 9431
64
Fischerella sp. PCC 9605
Fischerella muscicola PCC 7414
75
99
Fischerella thermalis PCC 7521
99
Fischerella sp. JSC-11
Scytonema hofmanni UTEX 2349
99
Calothrix sp. PCC 7103
56
Calothrix desertica PCC 7102
Calothrix sp. PCC 6303
Mastigocoleus testarum BC008
Rivularia sp. PCC 7116
54
Calothrix rhizosoleniae SC01
83
Richelia intracellularis RC01
78
Richelia intracellularis HH01
Mastigocladopsis repens PCC 10914
52
Scytonema hofmanni PCC 7110
Trichodesmium erythraeum IMS 101
0.01

Figure 9. 16S phylogeny of heterocyst-forming cyanobacteria. The presence of interruption element variants (with at least three occurrences) in organisms, in relation to16S rRNA phylogeny. The headings are labeled as the interrupted gene and approximate position of interruption, followed by the number of occurrences of that variant. Blue lines and text highlight the organisms that possess each variant.

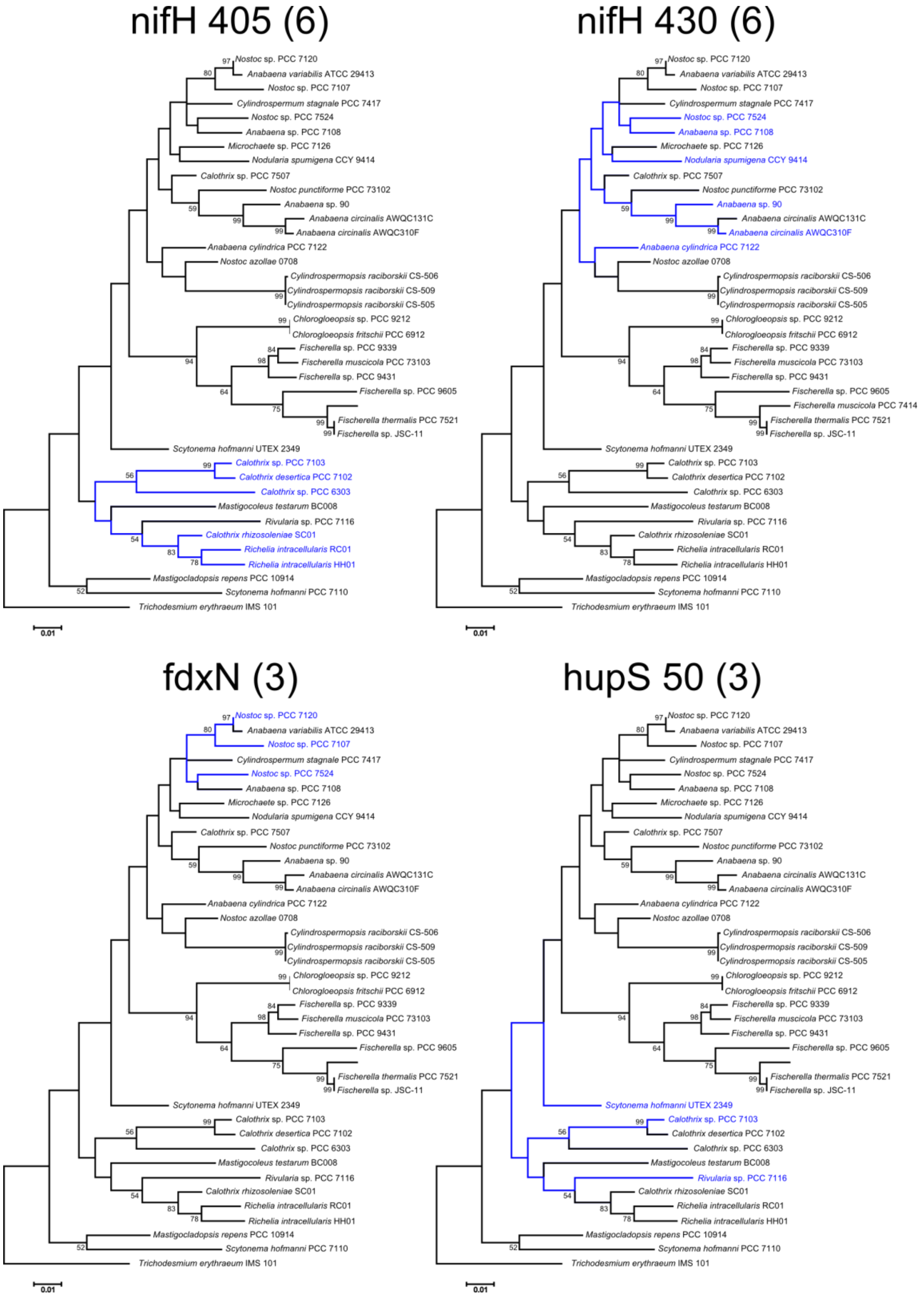

nifH 405 (6)
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Cylindrospermum stagnale PCC 7417
Nostoc sp. PCC 7524
Anabaena sp. PCC 7108
Microchaete sp. PCC 7126
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Anabaena sp. 90
Anabaena circinalis AWQC131C
Anabaena circinalis AWQC310F
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Cylindrospermopsis raciborskii CS-506
Cylindrospermopsis raciborskii CS-509
Cylindrospermopsis raciborskii CS-505
Chlorogloeopsis sp. PCC 9212
Chlorogloeopsis fritschii PCC 6912
Fischerella sp. PCC 9339
Fischerella muscicola PCC 73103
Fischerella sp. PCC 9431
Fischerella sp. PCC 9605
Fischerella thermalis PCC 7521
Fischerella sp. JSC-11
Scytonema hofmanni UTEX 2349
Calothrix sp. PCC 7103
Calothrix desertica PCC 7102
Calothrix sp. PCC 6303
Mastigocoleus testarum BC008
Rivularia sp. PCC 7116
Calothrix rhizosoleniae SC01
Richelia intracellularis RC01
Richelia intracellularis HH01
Mastigocladopsis repens PCC 10914
Scytonema hofmanni PCC 7110
Trichodesmium erythraeum IMS 101
0.01
nifH 430 (6)
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Cylindrospermum stagnale PCC 7417
Nostoc sp. PCC 7524
Anabaena sp. PCC 7108
Microchaete sp. PCC 7126
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Anabaena sp. 90
Anabaena circinalis AWQC131C
Anabaena circinalis AWQC310F
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Cylindrospermopsis raciborskii CS-506
Cylindrospermopsis raciborskii CS-509
Cylindrospermopsis raciborskii CS-505
Chlorogloeopsis sp. PCC 9212
Chlorogloeopsis fritschii PCC 6912
Fischerella sp. PCC 9339
Fischerella muscicola PCC 73103
Fischerella sp. PCC 9431
Fischerella sp. PCC 9605
Fischerella muscicola PCC 7414
Fischerella thermalis PCC 7521
Fischerella sp. JSC-11
Scytonema hofmanni UTEX 2349
Calothrix sp. PCC 7103
Calothrix desertica PCC 7102
Calothrix sp. PCC 6303
Mastigocoleus testarum BC008
Rivularia sp. PCC 7116
Calothrix rhizosoleniae SC01
Richelia intracellularis RC01
Richelia intracellularis HH01
Mastigocladopsis repens PCC 10914
Scytonema hofmanni PCC 7110
Trichodesmium erythraeum IMS 101
0.01
fdxN (3)
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Cylindrospermum stagnale PCC 7417
Nostoc sp. PCC 7524
Anabaena sp. PCC 7108
Microchaete sp. PCC 7126
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Anabaena sp. 90
Anabaena circinalis AWQC131C
Anabaena circinalis AWQC310F
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Cylindrospermopsis raciborskii CS-506
Cylindrospermopsis raciborskii CS-509
Cylindrospermopsis raciborskii CS-505
Chlorogloeopsis sp. PCC 9212
Chlorogloeopsis fritschii PCC 6912
Fischerella sp. PCC 9339
Fischerella muscicola PCC 73103
Fischerella sp. PCC 9431
Fischerella sp. PCC 9605
Fischerella thermalis PCC 7521
Fischerella sp. JSC-11
Scytonema hofmanni UTEX 2349
Calothrix sp. PCC 7103
Calothrix desertica PCC 7102
Calothrix sp. PCC 6303
Mastigocoleus testarum BC008
Rivularia sp. PCC 7116
Calothrix rhizosoleniae SC01
Richelia intracellularis RC01
Richelia intracellularis HH01
Mastigocladopsis repens PCC 10914
Scytonema hofmanni PCC 7110
Trichodesmium erythraeum IMS 101
0.01
hupS 50 (3)
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Cylindrospermum stagnale PCC 7417
Nostoc sp. PCC 7524
Anabaena sp. PCC 7108
Microchaete sp. PCC 7126
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Nostoc punctiforme PCC 73102
Anabaena sp. 90
Anabaena circinalis AWQC131C
Anabaena circinalis AWQC310F
Anabaena cylindrica PCC 7122
Nostoc azollae 0708
Cylindrospermopsis raciborskii CS-506
Cylindrospermopsis raciborskii CS-509
Cylindrospermopsis raciborskii CS-505
Chlorogloeopsis sp. PCC 9212
Chlorogloeopsis fritschii PCC 6912
Fischerella sp. PCC 9339
Fischerella muscicola PCC 73103
Fischerella sp. PCC 9431
Fischerella sp. PCC 9605
Fischerella thermalis PCC 7521
Fischerella sp. JSC-11
Scytonema hofmanni UTEX 2349
Calothrix sp. PCC 7103
Calothrix desertica PCC 7102
Calothrix sp. PCC 6303
Mastigocoleus testarum BC008
Rivularia sp. PCC 7116
Calothrix rhizosoleniae SC01
Richelia intracellularis RC01
Richelia intracellularis HH01
Mastigocladopsis repens PCC 10914
Scytonema hofmanni PCC 7110
Trichodesmium erythraeum IMS 101
0.01

Figure 9 (cont.)

Conclusions

The genomic sequences of the diatom symbionts display an evolutionary gradient, ranging from the genome reduction seen in the *Hemiaulus hauckii* symbiont to the external symbiont of *Chaetoceros* sp. that possesses a genome very similar to those of free-living cyanobacteria. The small genome size indicates the *H. hauckii* symbiont has been obligate for a very long time, and the streamlined nitrogen metabolism pathway highlights the importance of nitrogen transfer to maintain the association and provides the diatom a mechanism for which to regulate the symbiont. The *Rhizosolenia clevei* symbiont, on the other hand, exhibits some genome streamlining, but not to the extent of the *H. hauckii* symbiont. The genome of the *Rh. clevei* symbiont provides strong evidence that it has become an obligate symbiont much more recently. The differences amongst the diatom symbiont genomes provide an evolutionary gradient that shows the alterations a symbiont genome undergoes through the evolution of an association. The similarities of the three symbiont genomes reveal the pressures for heterocyst-forming cyanobacteria living in an oceanic environment, and in association with an unicellular eukaryotic partner. These genomic characteristics carry ecological implications for globally significant symbioses.

Metagenome and metatranscriptome sequencing allowed for an unbiased examination of natural populations of these unique associations relative to other oceanic N_2-fixers. The high similarity of the environmental sequences to the symbiont genomes confirmed that the genomes were good representatives of the Amazon River plume populations, unlike the *Trichodesmium* populations. The diatom symbiont populations also exhibited a higher fraction of transcripts involved in N_2 fixation than *Trichodesmium*, and decreased transcription in photosystem II, but not in other photosynthesis processes. Surprisingly, regulatory transcripts were abundant throughout the diazotrophic communities, and potentially play a role in the symbiont photosynthesis activity. The DNA from this diazotroph hot spot allowed an in-depth examination of genome-wide diversity of the diazotrophic populations, while the RNA sequencing provided a snapshot of the metabolic processes within those populations.

The collection of the interruption elements throughout heterocyst-forming cyanobacteria provided higher resolution to study these unique genetic features than has previously been available. The analysis done on this expanded collection has implicated the interruption elements in processes in addition to the previously-noted heterocyst differentiation. Variants to the interruption elements previously found within N_2 fixation genes were also identified. The dynamic nature of these sequences leaves little trace of an evolutionary origin, and the recombinase sequences, coding the excision enzyme, are the only markers for which to link all of the interruption elements. The two recombinase superfamilies responsible for removal of the elements share no evolutionary link, indicating two distinct evolutionary histories of interruption elements within heterocyst-forming cyanobacteria. However, the similarity of some recombinases within each superfamily supports the origin of elements through the replication of a recombinase gene within a genome. The examination of interruption elements in heterocyst-forming cyanobacteria has identified several evolutionary clues, broadened their impact beyond N_2 fixation, and give much-needed structure for which to continue the study of these mysterious genetic sequences.

The studies completed in the efforts of this dissertation have greatly expanded the knowledge on globally significant oceanic microbes. They also bear great significance for the study of plant-microbe interactions and the evolution of symbioses in all environments.

References

Adams DG, Bergman B, Nierzwicki-Bauer SA, Rai AN, Schüssler A. (2006). Cyanobacterial–plant symbioses. In:*The Prokaryotes. A Handbook on the Biology of Bacteria*, Vol. 1, Springer Science: New York, NY, pp. 331–363.

Atkinson GC, Baldauf SL. (2011). Evolution of elongation factor G and the origins of mitochondrial and chloroplast forms. *Mol Biol Evol* **28**:1281–1292.

Awai K, Wolk CP. (2007). Identification of the glycosyl transferase required for synthesis of the principal glycolipid characteristic of heterocysts of *Anabaena* sp. strain PCC 7120. *FEMS Microbiol Lett* **266**:98–102.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**:75.

Bauer CC, Scappino L, Haselkorn R. (1993). Growth of the cyanobacterium *Anabaena* on molecular nitrogen: NifJ is required when iron is limited. *Proc Natl Acad Sci* **90**:8812–8816.

Berman-Frank I, Quigg A, Finkel ZV, Irwin AJ, Haramaty L. (2007). Nitrogen-fixation strategies and Fe requirements in cyanobacteria. *Limnol Oceanogr* **52**:2260–2269.

Billi D, Grilli-Caiola M. (1996). Effects of nitrogen limitation and starvation on *Chroococcidiopsis* sp. (Chroococcales). *New Phytol* **133**:563–571.

Blot N, Wu X-J, Thomas J-C, Zhang J, Garczarek L, Böhm S, *et al.* (2009). Phycourobilin in trichromatic phycocyanin from oceanic cyanobacteria is formed post-translationally by a phycoerythrobilin lyase-isomerase. *J Biol Chem* **284**:9290–9298.

Bombar D, Moisander PH, Dippner JW, Foster RA, Voss M, Karfeld B, *et al.* (2011). Distribution of diazotrophic microorganisms and *nifH* gene expression in the Mekong River plume during intermonsoon. *Mar Ecol Prog Ser* **424**:39–52.

Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, *et al.* (2008). The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**:239–244.

Brusca JS, Chastain CJ, Golden JW. (1990). Expression of the *Anabaena* sp. strain PCC 7120 *xisA* gene from a heterologous promoter results in excision of the *nifD* element. *J Bacteriol* **172**:3925–3931.

Brzezinski MA, Villareal TA, Lipschultz F. (1998). Silica production and the contribution of diatoms to new and primary production in the central North Pacific. *Mar Ecol Prog Ser* **167**:89–104.

Burford MA, Rothlisberg PC, Wang YG. (1995). Spatial and temporal distribution of tropical phytoplankton species and biomass in the Gulf of Carpentaria, Australia. *Mar Ecol Prog Ser Oldendorf* **118**:255–266.

Campbell EL, Cohen MF, Meeks JC. (1997). A polyketide-synthase-like gene is involved in the synthesis of heterocyst glycolipids in *Nostoc punctiforme* strain ATCC 29133. *Arch Microbiol* **167**:251–258.

Campbell WS, Laudenbach DE. (1995). Characterization of four superoxide dismutase genes from a filamentous cyanobacterium. *J Bacteriol* **177**:964–972.

Carpenter E, Foster R. (2003). Marine cyanobacterial symbioses. In:*Cyanobacteria in Symbiosis*, Kluwer Publishers: Dordrecht, The Netherlands, pp. 11–17.

Carpenter EJ, Janson S. (2000). Intracellular cyanobacterial symbionts in the marine diatom *Climacodium frauenfeldianum* (Bacillariophyceae). *J Phycol* **36**:540–544.

Carpenter EJ, Montoya JP, Burns J, Mulholland M, Subramaniam A, Capone DG. (1999). Extensive bloom of a N_2 fixing symbiotic association in the tropical Atlantic Ocean. *Mar Ecol Prog Ser* **185**:273–283.

Carrasco CD, Buettner JA, Golden JW. (1995). Programmed DNA rearrangement of a cyanobacterial *hupL* gene in heterocysts. *Proc Natl Acad Sci* **92**:791–795.

Carrasco CD, Golden JW. (1995). Two heterocyst-specific DNA rearrangements of *nif* operons in *Anabaena* cylindrica and *Nostoc* sp. strain Mac. *Microbiology* **141**:2479–2487.

Carrasco CD, Holliday SD, Hansel A, Lindblad P, Golden JW. (2005). Heterocyst-specific excision of the *Anabaena* sp. strain PCC 7120 *hupL* element requires *xisC*. *J Bacteriol* **187**:6031–6038.

Carrasco CD, Ramaswamy K, Ramasubramanian T, Golden JW. (1994). *Anabaena xisF* gene encodes a developmentally regulated site-specific recombinase. *Genes Dev* **8**:74–83.

Chastain CJ, Brusca JS, Ramasubramanian TS, Wei T-F, Golden JW. (1990). A sequence-specific DNA-binding factor (VF1) from *Anabaena* sp. strain PCC 7120 vegetative cells binds to three adjacent sites in the *xisA* upstream region. *J Bacteriol* **172**:5044–5051.

Christman HD, Campbell EL, Meeks JC. (2011). Global transcription profiles of the nitrogen stress response resulting in heterocyst or hormogonium development in *Nostoc punctiforme*. *J Bacteriol* **193**:6874–6886.

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messié M, Tringe SG, *et al.* (2010). Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc Natl Acad Sci* **107**:14679–14684.

Ehira S. (2013). Transcriptional regulation of heterocyst differentiation in *Anabaena* sp. strain PCC 7120. *Russ J Plant Physiol* **60**:443–452.

Enderlin CS, Meeks JC. (1983). Pure culture and reconstitution of the *Anthoceros-Nostoc* symbiotic association. *Planta* **158**:157–165.

Eppley RW, Peterson BJ. (1979). Particulate organic matter flux and planktonic new production in the deep ocean. *Nature* **282**:677–680.

Ermakova M, Battchikova N, Allahverdiyeva Y, Aro E-M. (2013). Novel heterocyst-specific flavodiiron proteins in *Anabaena* sp. PCC 7120. *FEBS Lett* **587**:82–87.

Falkowski PG. (1997). Evolution of the nitrogen cycle and its influence on the biological sequestration of CO_2 in the ocean. *Nature* **387**:272–275.

Fay P. (1992). Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol Rev* **56**:340–373.

Felsenstein J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783–791.

Fewer D, Friedl T, Büdel B. (2002). *Chroococcidiopsis* and heterocyst-differentiating cyanobacteria are each other's closest living relatives. *Mol Phylogenet Evol* **23**:82–90.

Florencio F, Marques S, Candau P. (1987). Identification and characterization of a glutamate dehydrogenase in the unicellular cyanobacterium *Synechocystis* PCC 6803. *FEBS Lett* **223**:37–41.

Foster R, Subramaniam A, Zehr J. (2009). Distribution and activity of diazotrophs in the Eastern Equatorial Atlantic. *Environ Microbiol* **11**:741–750.

Foster RA, Goebel NL, Zehr JP. (2010). Isolation of *Calothrix rhizosoleniae* (cyanobacteria) strain SC01 from *Chaetoceros* (Bacillariophyta) spp. diatoms of the Subtropical North Pacific Ocean. *J Phycol* **46**:1028–1037.

Foster RA, Kuypers MMM, Vagner T, Paerl RW, Musat N, Zehr JP. (2011). Nitrogen fixation and transfer in open ocean diatom–cyanobacterial symbioses. *ISME J* **5**:1484–1493.

Foster RA, Subramaniam A, Mahaffey C, Carpenter EJ, Capone DG, Zehr JP. (2007). Influence of the Amazon River plume on distributions of free-living and symbiotic cyanobacteria in the western tropical north Atlantic Ocean. *Limnol Oceanogr* **52**:517–532.

Foster RA, Zehr JP. (2006). Characterization of diatom-cyanobacteria symbioses on the basis of *nifH, hetR* and 16S rRNA sequences. *Environ Microbiol* **8**:1913–1925.

Fung IY, Meyn SK, Tegen I, Doney SC, John JG, Bishop JK. (2000). Iron supply and demand in the upper ocean. *Glob Biogeochem Cycles* **14**:281–295.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA. (2010). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* **5**:461–472.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, others. (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* **3**:e3042.

Glibert PM, Harrison J, Heil C, Seitzinger S. (2006). Escalating worldwide use of urea–a global change contributing to coastal eutrophication. *Biogeochemistry* **77**:441–463.

Glibert PM, Heil CA, Hollander DJ, Revilla M, Hoare A, Alexander J, *et al.* (2004). Evidence for dissolved organic nitrogen and phosphorus uptake during a cyanobacterial bloom in Florida Bay. *Mar Ecol Prog Ser* **280**:73–83.

Goes JI, Gomes H do R, Chekalyuk AM, Carpenter EJ, Montoya JP, Coles VJ, *et al.* (2014). Influence of the Amazon River discharge on the biogeography of phytoplankton communities in the western tropical north Atlantic. *Prog Oceanogr* **120**:29–40.

Golden JW, Robinson SJ, Haselkorn R. (1985). Rearrangement of nitrogen fixation genes during heterocyst differentiation in the cyanobacterium *Anabaena*. *Nature* **314**:419–423.

Golden JW, Wiest DR. (1988). Genome rearrangement and nitrogen fixation in *Anabaena* blocked by inactivation of *xisA* gene. *Science* **242**:1421–1423.

Goldman JC. (1993). Potential role of large oceanic diatoms in new primary production. *Deep Sea Res Part Oceanogr Res Pap* **40**:159–168.

Gomez F, Furuya K, Takeda S. (2005). Distribution of the cyanobacterium *Richelia intracellularis* as an epiphyte of the diatom *Chaetoceros compressus* in the western Pacific Ocean. *J Plankton Res* **27**:323–330.

Grindley ND, Whiteson KL, Rice PA. (2006). Mechanisms of site-specific recombination. *Annu Rev Biochem* **75**:567–605.

Gugger MF, Hoffmann L. (2004). Polyphyly of true branching cyanobacteria (Stigonematales). *Int J Syst Evol Microbiol* **54**:349–357.

Hakoyama T, Niimi K, Watanabe H, Tabata R, Matsubara J, Sato S, *et al.* (2009). Host plant genome overcomes the lack of a bacterial gene for symbiotic nitrogen fixation. *Nature* **462**:514–517.

Haselkorn R. (1992). Developmentally regulated gene rearrangements in prokaryotes. *Annu Rev Genet* **26**:113–130.

Heil CA, Revilla M, Glibert PM, Murasko S. (2007). Nutrient quality drives differential phytoplankton community composition on the southwest Florida shelf. *Limnol Oceanogr* **52**:1067–1078.

Heinbokel JF. (1986). Occurence of *Richelia Intracellularis* (Cyanophyta) within the diatoms *Hemiaulus hauckii* and *H. membranaceus* off Hawaii. *J Phycol* **22**:399–403.

Henson BJ, Hartman L, Watson LE, Barnum SR. (2011). Evolution and variation of the *nifD* and *hupL* elements in the heterocystous cyanobacteria. *Int J Syst Evol Microbiol* **61**:2938–2949.

Henson BJ, Hesselbrock SM, Watson LE, Barnum SR. (2004). Molecular phylogeny of the heterocystous cyanobacteria (subsections IV and V) based on *nifD*. *Int J Syst Evol Microbiol* **54**:493–497.

Henson BJ, Watson LE, Barnum SR. (2005). Characterization of a 4 kb variant of the *nifD* element in *Anabaena* sp. strain ATCC 33047. *Curr Microbiol* **50**:129–132.

Herdman M, Rippka R. (1988). Cellular differentiation: Hormogonia and baeocytes. *Methods Enzymol* **167**:232–242.

Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR, *et al.* (2009). *In situ* transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*. *ISME J* **3**:618–631.

Hewson I, Poretsky RS, Dyhrman ST, Zielinski B, White AE, Tripp HJ, *et al.* (2009). Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J* **3**:1286–1300.

Hilton JA, Foster RA, Tripp HJ, Carter BJ, Zehr JP, Villareal TA. (2013). Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nat Commun* **4**:1767.

Hirokawa G, Nijman RM, Raj VS, Kaji H, Igarashi K, Kaji A. (2005). The role of ribosome recycling factor in dissociation of 70S ribosomes into subunits. *RNA* **11**:1317–1328.

Jabir T, Dhanya V, Jesmi Y, Prabhakaran MP, Saravanane N, Gupta GVM, *et al.* (2013). Occurrence and distribution of a diatom-diazotrophic cyanobacteria association during a *Trichodesmium* bloom in the southeastern Arabian Sea. *Int J Oceanogr* **2013**:350594.

Janson S, Rai AN, Bergman B. (1995). Intracellular cyanobiont *Richelia intracellularis*: ultrastructure and immuno-localisation of phycoerythrin, nitrogenase, Rubisco and glutamine synthetase. *Mar Biol* **124**:1–8.

Janson S, Wouters J, Bergman B, Carpenter EJ. (1999). Host specificity in the *Richelia*-diatom symbiosis revealed by *hetR* gene sequence analysis. *Environ Microbiol* **1**:431–438.

Johansson C, Bergman B. (1994). Reconstitution of the *Gunnera manicata* Linde symbioses: cyanobacterial specificity. *New Phytol* **126**:643–652.

Kaplan-Levy RN, Hadas O, Summers ML, Rücker J, Sukenik A. (2010). Akinetes: Dormant cells of cyanobacteria. In:*Dormancy and Resistance in Harsh Environments*, Springer: New York, NY, pp. 5–27.

Karl DM, Church MJ, Dore JE, Letelier RM, Mahaffey C. (2012). Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc Natl Acad Sci* **109**:1842–1849.

Kimor B, Gordon N, Neori A. (1992). Symbiotic associations among the microplankton in oligotrophic marine environments, with special reference to the Gulf of Aqaba, Red Sea. *J Plankton Res* **14**:1217–1231.

Kitajima S, Furuya K, Hashihama F, Takeda S, Kanda J. (2009). Latitudinal distribution of diazotrophs and their nitrogen fixation in the tropical and subtropical western North Pacific. *Limnol Oceanogr* **54**:537–547.

Knapke EM. (2012). Influence of the Mississippi River plume on diazotroph distributions in the northern Gulf of Mexico during summer 2011. M.S., The University of Texas at Austin: Austin, Texas.

Kneip C, Lockhart P, Voß C, Maier U-G. (2007). Nitrogen fixation in eukaryotes–new models for symbiosis. *BMC Evol Biol* **7**:55.

Kulkarni VV, Chitari RR, Narale DD, Patil JS, Anil AC. (2010). Occurrence of cyanobacteria–diatom symbiosis in the Bay of Bengal: implications in biogeochemistry. *Curr Sci* **99**:736–737.

Lammers PJ, Golden JW, Haselkorn R. (1986). Identification and sequence of a gene required for a developmentally regulated DNA excision in *Anabaena*. *Cell* **44**:905–911.

Lammers PJ, McLaughlin S, Papin S, Trujillo-Provencio C, Ryncarz AJ. (1990). Developmental rearrangement of cyanobacterial *nif* genes: nucleotide sequence, open reading frames, and cytochrome P-450 homology of the *Anabaena* sp. strain PCC 7120 *nifD* element. *J Bacteriol* **172**:6981–6990.

Lane DJ. (1991). 16S/23S rRNA sequencing. In:*Nucleic Acid Techniques in Bacterial Systematics*, John Wiley & Sons: Chichester, United Kingdom, pp. 115–175.

LaRoche J, Breitbarth E. (2005). Importance of the diazotrophs as a source of new nitrogen in the ocean. *J Sea Res* **53**:67–91.

Larsson J, Nylander J, Bergman B. (2011). Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evol Biol* **11**:187.

LeBauer DS, Treseder KK. (2008). Nitrogen limitation of net primary productivity in terrestrial ecosystems is globally distributed. *Ecology* **89**:371–379.

Lindblad P, Bergman B, Hofsten AV, Hallbom L, Nylund JE. (1985). The cyanobacterium-*Zamia* symbiosis: an ultrastructural study. *New Phytol* **101**:707–716.

Lindblad P, Rai AN, Bergman B. (1987). The *Cycas revoluta-Nostoc* symbiosis: enzyme activities of nitrogen and carbon metabolism in the cyanobiont. *Microbiology* **133**:1695–1699.

Lohrenz SE, Fahnenstiel GL, Redalje DG, Lang GA, Dagg MJ, Whitledge TE, *et al.* (1999). Nutrients, irradiance, and mixing as factors regulating primary production in coastal waters impacted by the Mississippi River plume. *Cont Shelf Res* **19**:1113–1141.

López-Gomollón S, Hernández JA, Pellicer S, Angarica VE, Peleato ML, Fillat MF. (2007). Cross-talk between iron and nitrogen regulatory networks in *Anabaena* (*Nostoc*) sp. PCC 7120: identification of overlapping genes in FurA and NtcA regulons. *J Mol Biol* **374**:267–281.

Lyimo TJ. (2011). Distribution and abundance of the cyanobacterium *Richelia intracellularis* in the coastal waters of Tanzania. *J Ecol Nat Environ* **3**:85–94.

Madhu NV, Paul M, Ullas N, Ashwini R, Rehitha TV. (2013). Occurrence of cyanobacteria (*Richelia intracellularis*)-diatom (*Rhizosolenia hebetata*) consortium in the Palk Bay, southeast coast of India. **42**:453–457.

Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, *et al.* (2012). Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proc Natl Acad Sci* **109**:E317–E325.

Marcus Y, Zenvirth D, Harel E, Kaplan A. (1982). Induction of HCO_3^- transporting capability and high photosynthetic affinity to inorganic carbon by low concentration of CO_2 in *Anabaena variabilis*. *Plant Physiol* **69**:1008–1012.

Marumo R, Asaoka O. (1974). Distribution of pelagic blue-green algae in the North Pacific Ocean. *J Oceanogr* **30**:77–85.

Meeks JC, Campbell EL, Bisen PS. (1994). Elements interrupting nitrogen fixation genes in cyanobacteria: presence and absence of a *nifD* element in clones of *Nostoc* sp. strain Mac. *Microbiology* **140**:3225–3232.

Meeks JC, Enderlin CS, Joseph CM, Chapman JS, Lollar MWL. (1985). Fixation of [^{13}N] N_2 and transfer of fixed nitrogen in the *Anthoceros-Nostoc* symbiotic association. *Planta* **164**:406–414.

Meeks JC, Joseph CM, Haselkorn R. (1988). Organization of the *nif* genes in cyanobacteria in symbiotic association with *Azolla* and *Anthoceros*. *Arch Microbiol* **150**:61–71.

Meeks JC, Steinberg N, Joseph CM, Enderlin CS, Jorgensen PA, Peters GA. (1985). Assimilation of exogenous and dinitrogen-derived $^{13}NH_4^+$ by *Anabaena azollae* separated from *Azolla caroliniana* Willd. *Arch Microbiol* **142**:229–233.

Mérida A, Leurentop L, Candau P, Florencio FJ. (1990). Purification and properties of glutamine synthetases from the cyanobacteria *Synechocystis* sp. strain PCC 6803 and *Calothrix* sp. strain PCC 7601. *J Bacteriol* **172**:4732–4735.

Moore CM, Mills MM, Arrigo KR, Berman-Frank I, Bopp L, Boyd PW, *et al.* (2013). Processes and patterns of oceanic nutrient limitation. *Nat Geosci* **6**:701–710.

Moore LR, Post AF, Rocap G, Chisholm SW. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol Oceanogr* **47**:989–996.

Moran NA. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr Opin Microbiol* **6**:512–518.

Moran NA, Plague GR. (2004). Genomic changes following host restriction in bacteria. *Curr Opin Genet Dev* **14**:627–633.

Mühling M, Harris N, Belay A, Whitton BA. (2003). Reversal of helix orientation in the cyanobacterium *Arthrospira*. *J Phycol* **39**:360–367.

Mulligan ME, Haselkorn R. (1989). Nitrogen fixation (*nif*) genes of the cyanobacterium *Anabaena* species strain PCC 7120. The *nifB-fdxN-nifS-nifU* operon. *J Biol Chem* **264**:19200–19207.

Muro-Pastor MI, Florencio FJ. (2003). Regulation of ammonium assimilation in cyanobacteria. *Plant Physiol Biochem* **41**:595–603.

Muro-Pastor MI, Reyes JC, Florencio FJ. (2005). Ammonium assimilation in cyanobacteria. *Photosynth Res* **83**:135–150.

Muro-Pastor MI, Reyes JC, Florencio FJ. (2001). Cyanobacteria perceive nitrogen status by sensing intracellular 2-oxoglutarate levels. *J Biol Chem* **276**:38320–38328.

Ogawa T, Mi H. (2007). Cyanobacterial NADPH dehydrogenase complexes. *Photosynth Res* **93**:69–77.

Orgel LE, Crick FHC, Sapienza C. (1980). Selfish DNA. **288**:645–646.

Ottesen EA, Young CR, Eppley JM, Ryan JP, Chavez FP, Scholin CA, *et al.* (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc Natl Acad Sci* **110**:E488–E497.

Padmakumar KB, Menon NR, Sanjeevan VN. (2010). Occurrence of endosymbiont *Richelia intracellularis* (Cyanophyta) within the diatom *Rhizosolenia hebetata* in the Northern Arabian Sea. *Int J Biodivers Conserv* **2**:70–74.

Pattanaik B, Montgomery BL. (2010). FdTonB is involved in the photoregulation of cellular morphology during complementary chromatic adaptation in *Fremyella diplosiphon*. *Microbiology* **156**:731–741.

Peters G, Meeks J. (1989). The *Azolla-Anabaena* symbiosis - basic biology. *Annu Rev Plant Physiol Plant Mol Biol* **40**:193–210.

Pyle A. (2011). Light dependant growth and nitrogen fixation rates in the *Hemiaulus hauckii* and *Hemiaulus membranaceus* diatom-diazotroph associations. M.S., The University of Texas at Austin: Austin, Texas.

Rai AN, Rowell P, Stewart WD. (1983). Interactions between cyanobacterium and fungus during $^{15}N_2$-incorporation and metabolism in the lichen *Peltigera canina*. *Arch Microbiol* **134**:136–142.

Rai AN, Söderbäck E, Bergman B. (2000). Cyanobacterium-plant symbioses. *New Phytol* **147**:449–481.

Ramasubramanian TS, Wei TF, Golden JW. (1994). Two *Anabaena* sp. strain PCC 7120 DNA-binding factors interact with vegetative cell-and heterocyst-specific genes. *J Bacteriol* **176**:1214–1223.

Ramaswamy K, Carrasco CD, Fatma T, Golden JW. (1997). Cell-type specificity of the *Anabaena fdxN*-element rearrangement requires *xisH* and *xisI*. *Mol Microbiol* **23**:1241–1249.

Ran L, Larsson J, Vigil-Stenman T, Nylander JAA, Ininbergs K, Zheng W-W, *et al.* (2010). Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE* **5**:e11486.

Rice D, Mazur BJ, Haselkorn R. (1982). Isolation and physical mapping of nitrogen fixation genes from the cyanobacterium *Anabaena* 7120. *J Biol Chem* **257**:13157–13163.

Roberts GP, MacNeil T, MacNeil D, Brill WJ. (1978). Regulation and characterization of protein products coded by the *nif* (nitrogen fixation) genes of *Klebsiella pneumoniae*. *J Bacteriol* **136**:267–279.

Rodnina MV, Savelsbergh A, Katunin VI, Wintermeyer W. (1997). Hydrolysis of GTP by elongation factor G drives tRNA movement on the ribosome. *Nature* **385**:37–41.

Rueter J. (1988). Iron stimulation of photosynthesis and nitrogen fixation in *Anabaena* 7120 and *Trichodesmium* (Cyanophyceae). *J Phycol* **24**:249–254.

Saier MHJ, Yen MR, Noto K, Tamang DG, Elkan C. (2009). The transporter classification database: recent advances. *Nucleic Acids Res* **37**:D274–D278.

Saitou N, Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**:406–425.

Saville B, Straus N, Coleman JR. (1987). Contiguous organization of nitrogenase genes in a heterocystous cyanobacterium. *Plant Physiol* **85**:26–29.

Scharek R, Tupas LM, Karl DM. (1999). Diatom fluxes to the deep sea in the oligotrophic North Pacific gyre at Station ALOHA. *Mar Ecol Prog Ser* **182**:55–67.

Schouten S, Villareal TA, Hopmans EC, Mets A, Swanson KM, Sinninghe Damsté JS. (2013). Endosymbiotic heterocystous cyanobacteria synthesize different heterocyst glycolipids than free-living heterocystous cyanobacteria. *Phytochemistry* **85**:115–121.

Schüssler A, Bonfante P, Schnepf E, Mollenhauer D, Kluge M. (1996). Characterization of the *Geosiphon pyriforme* symbiosome by affinity techniques: confocal laser scanning microscopy (CLSM) and electron microscopy. *Protoplasma* **190**:53–67.

Shah VK, Stacey G, Brill WJ. (1983). Electron transport to nitrogenase. Purification and characterization of pyruvate: flavodoxin oxidoreductase. The *nifJ* gene product. *J Biol Chem* **258**:12064–12068.

Shi Y, Tyson GW, DeLong EF. (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**:266–269.

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, *et al.* (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci* **110**:1053–1058.

Silvester W, Parsons R, Watt P. (1996). Direct measurement of release and assimilation of ammonia in the *Gunnera-Nostoc* symbiosis. *New Phytol* **132**:617–625.

Singh P, Singh SS, Elster J, Mishra AK. (2013). Molecular phylogeny, population genetics, and evolution of heterocystous cyanobacteria using *nifH* gene sequences. *Protoplasma* **250**:751–764.

Six C, Thomas J-C, Thion L, Lemoine Y, Zal F, Partensky F. (2005). Two novel phycoerythrin-associated linker proteins in the marine cyanobacterium *Synechococcus* sp. strain WH8102. *J Bacteriol* **187**:1685–1694.

Smith Jr WO, Demaster DJ. (1996). Phytoplankton biomass and productivity in the Amazon River plume: correlation with seasonal river discharge. *Cont Shelf Res* **16**:291–319.

Söderbäck E, Bergman B. (1993). The *Nostoc-Gunnera* symbiosis: Carbon fixation and translocation. *Physiol Plant* **89**:125–132.

Staal M, Meysman FJR, Stal LJ, others. (2003). Temperature excludes N_2-fixing heterocystous cyanobacteria in the tropical oceans. *Nature* **425**:504–507.

Stewart WDP, Haystead A, Pearson HW. (1969). Nitrogenase activity in heterocysts of blue–green algae. *Nature* **224**:226–228.

Stucken K, John U, Cembella A, Murillo AA, Soto-Liebe K, Fuentes-Valdés JJ, *et al.* (2010). The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS ONE* **5**:e9235.

Subramaniam A, Yager P, Carpenter E, Mahaffey C, Björkman K, Cooley S, *et al.* (2008). Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci* **105**:10460–10465.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**:2731–2739.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, *et al.* (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**:1546–1550.

Thompson JD, Higgins DG, Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**:4673–4680.

Tripp H, Bench S, Turk K, Foster R, Desany B, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**:90–94.

Usher KM, Bergman B, Raven JA. (2007). Exploring cyanobacterial mutualisms. *Annu Rev Ecol Evol Syst* **38**:255–273.

Vaillancourt RD, Marra J, Seki MP, Parsons ML, Bidigare RR. (2003). Impact of a cyclonic eddy on phytoplankton community structure and photosynthetic competency in the subtropical North Pacific Ocean. *Deep Sea Res Part Oceanogr Res Pap* **50**:829–847.

Valladares A, Herrero A, Pils D, Schmetterer G, Flores E. (2003). Cytochrome c oxidase genes required for nitrogenase activity and diazotrophic growth in *Anabaena* sp. PCC 7120. *Mol Microbiol* **47**:1239–1249.

Venrick EL. (1974). The distribution and significance of *Richelia intracellularis* Schmidt in the North Pacific Central Gyre. *Limnol Oceanogr* **19**:437–445.

Villareal T. (1989). Division cycles in the nitrogen-fixing *Rhizosolenia* (Bacillariophyceae) *Richelia* (Nostocaceae) symbiosis. *Br Phycol J* **24**:357–365.

Villareal T. (1990). Laboratory culture and preliminary characterization of the nitrogen-fixing *Rhizosolenia-Richelia* symbiosis. *Mar Ecol Prog Ser* **11**:117–132.

Villareal TA. (1992). Marine nitrogen-fixing diatom-cyanobacteria symbioses. In:*Marine Pelagic Cyanobacteria:* Trichodesmium *and Other Diazotrophs*, Kluwer: Netherleands, pp. 163–175.

Villareal TA. (1991). Nitrogen-fixation by the cyanobacterial symbiont of the diatom genus *Hemiaulus*. *Mar Ecol Prog Ser* **76**:201–204.

Villareal TA. (1994). Widespread occurrence of the *Hemiaulus*-cyanobacterial symbiosis in the southwest North Atlantic Ocean. *Bull Mar Sci* **54**:1–7.

Vintila S, Selao T, Norén A, Bergman B, El-Shehawy R. (2011). Characterization of *nifH* gene expression, modification and rearrangement in *Nodularia spumigena* strain AV1. *FEMS Microbiol Ecol* **77**:449–459.

Vitousek PM, Cassman K, Cleveland C, Crews T, Field CB, Grimm NB, *et al.* (2002). Towards an ecological understanding of biological nitrogen fixation. *Biogeochemistry* **57**:1–45.

Wang H, Sivonen K, Rouhiainen L, Fewer DP, Lyra C, Rantala-Ylinen A, *et al.* (2012). Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium *Anabaena* sp. strain 90. *BMC Genomics* **13**:613.

Weare NM, Azam F, Mague TH, Holm-Hansen O. (1974). Microautoradiographic studies of the marine phycobionts *Rhizosolenia* and *Richelia*. *J Phycol* **10**:369–371.

White AE, Prahl FG, Letelier RM, Popp BN. (2007). Summer surface waters in the Gulf of California: Prime habitat for biological N_2 fixation. *Glob Biogeochem Cycles* **21**:GB2017.

White AE, Spitz YH, Letelier RM. (2007). What factors are driving summer phytoplankton blooms in the North Pacific Subtropical Gyre? *J Geophys Res* **112**:C12006.

Wilson C, Villareal TA, Maximenko N, Bograd SJ, Montoya JP, Schoenbaechler CA. (2008). Biological and physical forcings of late summer chlorophyll blooms at 30 N in the oligotrophic Pacific. *J Mar Syst* **69**:164–176.

Wolk CP, Ernst A, Elhai J. (2004). Heterocyst metabolism and development. In:*The Molecular Biology of Cyanobacteria*, Springer: New York, NY, pp. 769–823.

Wong PP, Burris RH. (1972). Nature of oxygen inhibition of nitrogenase from *Azotobacter vinelandii*. *Proc Natl Acad Sci* **69**:672–675.

Yan D. (2007). Protection of the glutamate pool concentration in enteric bacteria. *Proc Natl Acad Sci* **104**:9475–9480.

Yang Y, Qin S, Zhao F, Chi X, Zhang X. (2007). Comparison of envelope-related genes in unicellular and filamentous cyanobacteria. *Comp Funct Genomics* **2007**:25751.

Yeung LY, Berelson WM, Young ED, Prokopenko MG, Rollins N, Coles VJ, *et al.* (2012). Impact of diatom-diazotroph associations on carbon export in the Amazon River plume. *Geophys Res Lett* **39**:L18609.

Zavialov AV, Hauryliuk VV, Ehrenberg M. (2005). Splitting of the posttermination ribosome into subunits by the concerted action of RRF and EF-G. *Mol Cell* **18**:675–686.

Zeev EB, Yogev T, Man-Aharonovich D, Kress N, Herut B, Béjà O, *et al.* (2008). Seasonal dynamics of the endosymbiotic, nitrogen-fixing cyanobacterium *Richelia intracellularis* in the eastern Mediterranean Sea. *ISME J* **2**:911–923.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, *et al.* (2008). Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**:1110–1112.

Zehr JP, Kudela RM. (2011). Nitrogen cycle of the open ocean: from genes to ecosystems. *Annu Rev Mar Sci* **3**:197–225.

Zhang C-C, Laurent S, Sakr S, Peng L, Bédu S. (2006). Heterocyst differentiation and pattern formation in cyanobacteria: a chorus of signals. *Mol Microbiol* **59**:367–375.

Zhang Y, Zhao Z, Sun J, Jiao N. (2011). Diversity and distribution of diazotrophic communities in the South China Sea deep basin with mesoscale cyclonic eddy perturbations. *FEMS Microbiol Ecol* **78**:417–427.

Zhao L, Denis M, Barani A, Beker B, Mante C, Xiao T, *et al.* (2012). Possible bloom of free trichomes in the Bay of Marseille, NW Mediterranean Sea: an anomaly evidenced by flow cytometry. *J Plankton Res* **34**:711–718.

Zhou JX, Zhou J, Yang HM, Chen M, Huang F. (2008). Characterization of two glutaminases from the filamentous cyanobacterium *Anabaena* sp. PCC 7120. *FEMS Microbiol Lett* **289**:241–249.