UC Santa Cruz

UC Santa Cruz Previously Published Works

Title

Metatranscriptomics of N2-fixing cyanobacteria in the Amazon River plume

Permalink

https://escholarship.org/uc/item/2jq6j9q9

Journal

The ISME Journal: Multidisciplinary Journal of Microbial Ecology, 9(7)

ISSN

1751-7362

Authors

Hilton, Jason A
Satinsky, Brandon M
Doherty, Mary
et al.

Publication Date

2015-07-01

DOI

10.1038/ismej.2014.240

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Metatranscriptomics of N_2-fixing cyanobacteria in the Amazon River plume

Subject Category: Microbial ecology and functional diversity of natural habitats

Jason A. Hilton[1], Brandon M. Satinsky[2], Mary Doherty[3], Brian Zielinski[4], Jonathan P. Zehr[1*]

[1]University of California Department of Ocean Sciences, Santa Cruz, CA 95064, USA

[2]University of Georgia Department of Microbiology, Athens, GA 30602, USA

[3]Rhodes College Department of Biology, Memphis, TN 38112, USA

[4]University of South Florida College of Marine Science, St. Petersburg, FL 33701, USA

[*]Corresponding author
Ocean Sciences Department
Marine Microbiology Laboratory/Zehr Lab
1156 High Street, Santa Cruz, CA 95064
(831) 459-4009
zehrj@ucsc.edu

Abstract

Biological N_2 fixation is an important nitrogen source for surface ocean microbial communities. However, nearly all information on the diversity and gene expression of organisms responsible for oceanic N_2 fixation in the environment has come from targeted approaches that assay only a small number of genes and organisms. Using genomes of diazotrophic cyanobacteria to extract reads from extensive meta-genomic and -transcriptomic libraries, we examined diazotroph diversity and gene expression from the Amazon River plume, an area characterized by salinity and nutrient gradients. Diazotroph genome and transcript sequences were most abundant in the transitional waters compared to lower salinity or oceanic water masses. We were able to distinguish two genetically-divergent phylotypes 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 from several *Richelia* photosystem II genes were absent, indicating a prominent role for cyclic electron transport in *Richelia*. 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 us to identify 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.

Keywords

nitrogen fixation/marine/metagenome/metatranscriptome/
Richelia/*Trichodesmium*/Amazon plume

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 (Si, P, 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, 2014). Additionally, sequences implicating a novel bacterial group and a euryarchaeal population in deep sea nitrogen and carbon 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, we coupled metatranscriptomic and metagenomic approaches to analyze the N_2-fixing community that drives new production in the Amazon River plume.

Materials and methods

<u>Sample collection</u>

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 (20 L) were taken in duplicate for each of the sample types described below (DNA, RNA, and poly(A)-RNA) and pre-filtered (156 μm) to remove grazers 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^{o}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). In order to prevent the loss of any material that washed off of the filter, 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). The filter capsule was then triple-rinsed with PBS using a sterile syringe. Once the filters and the filtered suspension material were thoroughly rinsed, they were either broken or sliced into smaller pieces (see below) and recombined in 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 142mm, 0.22μm Supor filters 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 2.0 μm pore-size, 142 mm diameter polycarbonate membrane filters were sliced on a sterile cutting board with the filter folded in to prevent the cells from sliding off the surface of the filter. For Sterivex filters, the filter was removed from the casing by cracking the housing with pliers, sliced on a sterile cutting board, and added to the DNA extraction buffer with the original membrane filter. An internal genomic DNA standard (*Thermus thermophilus* HB8 genomic DNA) was also added as a means to normalize sequencing coverage across samples (Satinsky *et al.*, 2014). The standard genomic DNA was spiked into each individual sample in a known abundance (8.4 ng per liter filtered) prior to the initiation of cell lysis. The samples were then extracted as previously described (Crump *et al.*, 2003) with adjustments for the larger volumes associated with 142 mm filters.

Sample preparation for total community RNA

RNA extraction and DNA removal were carried out as previously described (Gifford *et al.*, 2010; Poretsky, Hewson, *et al.*, 2009; Poretsky, Gifford, *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 two synthesized mRNA standards, which were 916 nt and 970 nt in length were synthesized from the commercial vectors pTXB1 vector (New England Biolabs, Ipswich, MA) and pFN18A Halotag T7 Flexi Vector (Promega, Madison, WI) respectively, and were added individually to the prepared lysis tubes in known copy numbers (pTXB1 = 2.104 x 10^{10} copies; pFN18A = 1.172 x 10^{10} copies) prior to the initiation of cell lysis (Satinsky *et al.*, 2014). 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 and two aliquots of Turbo Dnase were added at different times to the samples in order to improve DNA removal. 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 synthesized 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). Low sequencing yield has previously been attributed to this kit (Yanmei Shi *et al.*, 2010), but multiple studies have reported high reproducibility (Francois *et al.*, 2007; Frias-Lopez *et al.*, 2008). Random primers were used with the Superscript III First Strand synthesis system (Invitrogen, Carlsbad, CA) to copy 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 (2.0 x 10^9 copies per tube) (Satinsky *et al.*, 2014). The poly(A) standard was created from an HAP-1 Protolomerase viral gene. An amplicon (544 bp) with a poly(A) tail and a T7 promoter was synthesized through PCR from the template DNA. The amplicon was then used as template for an in vitro transcription reaction to produce the standard sequence with a poly(A) tail. 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.

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 $\geq$50) (Gifford *et al.*, 2010). Those cDNA reads with BLAST hits were removed from the data set (**Table S1**). To remove internal standard sequences from the metagenome reads, DNA reads with a BLAST hit against the *Thermus thermophilus* HB8 genome (blastn, bit score $\geq$50) were queried against the RefSeq protein database. Reads with a BLAST hit matching a *T. thermophilus* protein (blastx, bit score $\geq$40) were designated as internal standard and removed.

More than 39 million DNA sequence reads were obtained, with more than 27 million reads remaining after sequence trimming and removal of standards (**Table S1**). 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 S1**). The DNA sequence reads, as well as the cDNA reads from the 0.2 µm size fraction, from the low salinity offshore station were unavailable at the time of the writing of this report, and thus are not included in this study. DNA reads were an average of 190 bp long, while cDNA averaged 173 bp each. An earlier version of these data than those deposited at NCBI (PRJNA237344) was used for this study.

Identification and analysis of diazotroph reads

A BLAST analysis of the DNA and cDNA reads against the genomes of six oceanic N_2-fixing cyanobacteria (**Table 1**) was conducted (blastn, bit score $\geq$50). The whole genome sequences were used in order to analyze the organisms in the context of all cellular processes rather than target specific pathways (e.g. N_2 fixation). Additionally, given that diversity varies depending on the open-reading frame (ORF) or intergenic spacer region (IGS), the inclusion of the whole genomes prevented a strong bias from any predetermined gene groups. 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 potential 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 ORFs and IGSs of that organism (blastn) in order to assign each read to a functional region. An ORF or IGS was considered to be detected in the dataset if at least one read was assigned to it. For each detected 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}. When transcript abundances are discussed throughout this study, they are presented in these units because the normalization provides absolute estimates, and, thus, tracks the relative number of reads that cover a given transcript just as sequence coverage depth, but can more appropriately be used to compare whole transcriptome expression of individual populations across several stations. 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, 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.

A BLAST analysis of the non-duplicate reads that were not assigned to one of the six genomes was conducted against the nr database (NCBI, blastx, e-value ≤10, hit length ≥17 AA). The reads with a top BLAST hit in the nr database to a *nifH* gene sequence were pulled to assess the non-cyanobacterial diazotrophic populations in the dataset.

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 normalized transcript abundances of all the ORFs assigned to the given pathway in that organism.

Results

The four stations sampled are classified by the sea surface salinity at each, and referred to as oceanic (36.03), transitional (31.79), and low salinity (26.49 offshore and 22.55 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 reads that had a top BLAST hit to one of the diazotroph genomes aligned best with either the *Richelia intracellularis* HH01 genome (71.8%) or the *Trichodesmium erythraeum* IMS101 genome (19.2%). The reads that had a top BLAST hit to the *R. intracellularis* HH01 genome (163,293 DNA, 16,211 cDNA) were split into two populations, with 91.5% of those reads at least 98% identical (nucleotides) to the genome sequence and referred to as the *Hemiaulus-Richelia* (HR)-B population (**Figure 2**). An additional 7.6% of the *Richelia intracellularis* HH01 reads fell within the range of a secondary peak between 93-97% identity, which we termed the HR-A population (**Figure 2**). The diazotroph sequence reads that had a top BLAST hit to the *Trichodesmium erythraeum* IMS101 genome (33,038 DNA, 10,851 cDNA) exhibited 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 *R. 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.

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 7.1x10^{9} genome copies L^{-1} kbp^{-1}, with the majority at the transitional station (6.4x10^{9} genome copies L^{-1} kbp^{-1}) (**Figure 3**). Overall, the sequences from the HR-B population (98-100% identity to the genome) were the most abundant (6.0x10^{9} genome copies L^{-1} kbp^{-1}), and an order of magnitude greater than the sequences from the HR-A population (94-97% identity, 5.4x10^{8} genome copies L^{-1} kbp^{-1}) and the *Trichodesmium* population (5.1x10^{8} genome copies L^{-1} kbp^{-1}). RR population sequences were present at a lower abundance (7.9x10^{6} genome copies L^{-1} kbp^{-1}), and *Crocosphaera* population sequences were the least abundant in the diazotrophic cyanobacterium data set (7.9x10^{5} 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 3.01x10^{10} gene copies L^{-1} kbp^{-1}, and nearly all of that was from the transitional station (2.96x10^{10} gene copies L^{-1} kbp^{-1}). Similar to the normalized DNA abundance, normalized HR-B population cDNA from the four stations (2.6x10^{10} gene copies L^{-1} kbp^{-1}) was one order of magnitude greater than that of the HR-A population (1.1x10^{9} gene copies L^{-1} kbp^{-1}) or *Trichodesmium* (2.9x10^{9} gene copies L^{-1} kbp^{-1}). RR population cDNA (2.2x10^{7} gene copies L^{-1} kbp^{-1}) and *Crocosphaera* cDNA (3.7x10^{6} gene copies L^{-1} kbp^{-1}) were present at lower abundances.

The *R. intracellularis* HH01 genome contains 2,278 genes and 1,590 of them (69.8%) were detected in the HR-B population transcriptomes (15,311 reads) (**Figure S1**). By contrast, 2,233 of the *R. 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 *ndhD1* (RintHH_21740), which encodes the D1 subunit of NADH dehydrogenase I and *hisIE* (RintHH_14390), which encodes a fused phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase gene. 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 1,572 bp gene sequence (**Figure 4**). Similarly, all 171 *hisIE* reads from the transitional station covered only 232 bp of the 651 bp gene (**Figure 4**). 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 4**).

Only 265 *R. intracellularis* HH01 genes and 85 IGSs were found in the HR-A population transcriptomes (659 reads). Just 1,177 of the 2,278 *R. intracellularis* HH01 genes (51.7%) were detected in the metagenomes (1,177 reads). Of the HR-A transcripts detected, 85 genes 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. Just six RR cDNA reads were found in the transitional station sequences, while the rest were from the oceanic metatranscriptome. The most abundant transcripts found in the RR population at the oceanic station were a hypothetical protein (RintRC_2139) and the photosystem II *psbA* gene (RintRC_7737).

The *Trichodesmium* transcriptomes (9,892 reads) contained transcripts for 1,634 genes out of 5,076 in the *T. erythraeum* IMS101 genome (32.2%) and 247 IGSs (**Figure S1**). 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 4**). 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 4**). Genes involved in gas vesicles (Tery_2324, Tery_2325), photosystem II (Tery_4763), and other hypothetical proteins (Tery_0654, Tery_0835) were among the most abundant *Trichodesmium* transcripts at each station. Oceanic and low salinity offshore station reads from a photosystem II gene (Tery_4763) transcript were evenly distributed along the gene (**Figure 4**).

The transcriptomes of the unicellular *Crocosphaera* were comprised of 85 reads, 80 of which are from the oceanic station. Hypothetical proteins (CwatDRAFT_4329, CwatDRAFT_2191) and genes involved in photosynthesis (CwatDRAFT_0162, CwatDRAFT_1423) were the most abundant *Crocosphaera* transcripts at the oceanic station.

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. On account of 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. Photosynthesis was the most abundant KEGG pathway in the HR-B and *Trichodesmium* metatranscriptomes at each station. With the photosynthesis pathway, antenna proteins were 1.8-4.5% of HR-B transcription, and photosystem (PS) I proteins were 2.7% at the low salinity offshore station (**Figure 5**). PS-II genes were the most abundant photosynthesis group in *Trichodesmium* transcription at each station (2.9-10.3%), while antenna proteins were also abundant (0.8-3.2%) (**Figure 5**). All other gene groups for each population were no more than 2.0% of population transcription at any station (**Figure 5**).

nifH sequences

Three cDNA reads at the low salinity offshore station had top BLAST hits to gammaproteobacteria *nifH* genes, compared to 99 *nifH* transcript reads at that station that were assigned to a diazotrophic cyanobacteria genome. An additional three cDNA reads were found at the oceanic station with top hits to gammaproteobacteria *nifH* genes, while cyanobacteria *nifH* transcripts accounted for 43 reads at that station. None of the 214 *nifH* transcript reads at the transitional station, and no DNA reads, were attributed to heterotrophic *nifH* genes.

Discussion

At the time of sampling, the Amazon River plume had its maximum discharge rate for 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^{2-} and PO_4^{3-} 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 Si, and export efficiency (Yeung *et al.*, 2012).

Cyanobacteria comprised the majority of the diazotrophic community in the sequence dataset, and the distributions of the individual diazotroph populations in our study largely agree with previous observations from this region. However, 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 our data. 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 fewer diazotrophs are 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 P, Fe, and Si 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).

The two most prominent diatom symbionts in our data 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 we were not able to confirm that this population is within the het2A clade.

The high coverage of the HR-B and *Trichodesmium* metagenomes across their respective genomes shows that these populations were well-represented in the sampled data. The relatively lower similarity between the *Trichodesmium* populations and the representative genome is similar to previous studies that investigated the diversity of *Trichodesmium hetR* gene fragments (Janson, Bergman, *et al.*, 1999; Lundgren *et al.*, 2005; Hynes *et al.*, 2012). Additionally, if the gene content of the *Trichodesmium* populations varies from the *T. erythraeum* IMS 101 reference genome just as the percent identity does, some of the *Trichodesmium* genes may be absent from the metagenome because they are not present in the genomes of the natural populations. Thus, the *Trichodesmium* population coverage may actually be higher than the metagenomic coverage indicates. The diversity of the *Trichodesmium* populations relative to other reference sequences is explored in Supplemental Materials. The metatranscriptomics analysis was focused on the two populations that were well-represented in the datasets. It should be noted that while the presence of *Crocosphaera* was anticipated, the unicellular cyanobacterium fixes N_2 at night (Mohr *et al.*, 2010; Tuo Shi *et al.*, 2010), and thus, N_2 fixation gene transcripts from this population were not expected to be found in the morning samples.

The HR-B and *Trichodesmium* populations exhibited very different abundances of photosystem (PS) II gene transcripts relative to the total normalized transcription abundance for the given population in three different environments, making it more likely that this is a trend with biological implications rather than a chance sampling occurrence. Two *Trichodesmium psbA* copies, coding the PS-II D1 subunit, were among the 11 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 has multiple gene copies in 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 18 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 transcripts within HR-B photosynthesis gene groups other than PS-II, however, was comparable, and often greater than that of *Trichodesmium*, relative to the total normalized transcription abundance for the given population. 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 not only to meet their own N needs, but also those of their host diatom (Foster *et al.*, 2011).

Intriguingly, the second most abundant transcript in HR-B transitional transcriptome may regulate cyclic electron transport. We hypothesize that this transcript is 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 could provide for 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 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 our study.

The HR-B population transcriptomes were also characterized by an abundance of transcripts involved in N_2 fixation. Both the *Hemiaulus* symbiont and *Rhizosolenia* symbiont genomes lack ammonium transporters and the genes that encode the enzymes required to assimilate nitrate, nitrite, and urease, limiting the N sources available to the symbionts (Hilton *et al.*, 2013; Hilton, 2014). 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.

Trichodesmium nitrogenase gene transcripts were detected in the transcriptome, but not in high abundance. However, there was little indication of *Trichodesmium* utilizing other nitrogen sources as nitrate and nitrite reductase genes were not detected in the transcript libraries. Furthermore, only one cDNA read was assigned to an ammonium transporter transcript and one other cDNA read to a urease accessory protein, each at the oceanic station. Transcripts involved in 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 from gas vesicle protein genes 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). Gas vesicles are 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 *R. 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.

The 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, we found that the majority of diazotroph populations in this environment were 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* 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. We also identified a need for studies on non-coding transcripts and their function in regulating a variety of metabolic processes of N_2-fixing cyanobacteria, and of microbial communities, in general. Additionally, our 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, we utilized extensive community DNA and RNA sequencing to study 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 (J.P.Z.), ROCA award (P. Yager), 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 (EF-0424599) (J.P.Z.). 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.

Supplementary information

Supplementary information is available at ISMEJ's website.

Conflict of interest statement

The authors declare that there is no conflict of interest regarding this manuscript.

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.

Aro E-M, Virgin I, Andersson B. (1993). Photoinhibition of photosystem II. Inactivation, protein damage and turnover. *Biochim Biophys Acta BBA-Bioenerg* **1143**:113–134.

Baker BJ, Sheik CS, Taylor CA, Jain S, Bhasi A, Cavalcoli JD, *et al.* (2013). Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *ISME J* **7**:1962–1973.

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.

Crump BC, Armbrust EV, Baross JA. (1999). Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia River, its estuary, and the adjacent coastal ocean. *Appl Environ Microbiol* **65**:3192–3204.

Crump BC, Kling GW, Bahr M, Hobbie JE. (2003). Bacterioplankton community shifts in an arctic lake correlate with seasonal changes in organic matter source. *Appl Environ Microbiol* **69**:2253–2268.

Dugdale RC, Goering JJ. (1967). Uptake of new and regenerated forms of nitrogen in primary productivity. *Limnol Oceanogr* **12**:196–206.

Falgueras J, Lara AJ, Fernández-Pozo N, Cantón FR, Pérez-Trabado G, Claros MG. (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* **11**:38.

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.

Francois P, Garzoni C, Bento M, Schrenzel J. (2007). Comparison of amplification methods for transcriptomic analyses of low abundance prokaryotic RNA sources. *J Microbiol Methods* **68**:385–391.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, *et al.* (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci* **105**:3805.

Georg J, Hess WR. (2011). *cis*-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev MMBR* **75**:286–300.

Georg J, Honsel A, Voss B, Rennenberg H, Hess WR. (2010). A long antisense RNA in plant chloroplasts. *New Phytol* **186**:615–622.

Georg J, Voß B, Scholz I, Mitschke J, Wilde A, Hess WR. (2009). Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol Syst Biol* **5**:305.

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.

Goebel NL, Turk KA, Achilles KM, Paerl R, Hewson I, Morrison AE, *et al.* (2010). Abundance and distribution of major groups of diazotrophic cyanobacteria and their potential contribution to N_2 fixation in the tropical Atlantic Ocean. *Environ Microbiol* **12**:3272–3289.

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.

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. (2014). Ecology and evolution of diatom-associated cyanobacteria through genetic analyses. Ph.D., University of California: Santa Cruz, CA.

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.

Hynes AM, Webb EA, Doney SC, Waterbury JB. (2012). Comparison of cultured *Trichodesmium* (Cyanophyceae) with species characterized from the field. *J Phycol* **48**:196–210.

Ikeuchi M, Eggers B, Shen G, Webber A, Yu J, Hirano A, *et al.* (1991). Cloning of the *psbK* gene from *Synechocystis* sp. PCC 6803 and characterization of photosystem II in mutants lacking PSII-K. *J Biol Chem* **266**:11111–11115.

Janson S, Bergman B, Carpenter EJ, Giovannoni SJ, Vergin K. (1999). Genetic analysis of natural populations of the marine diazotrophic cyanobacterium *Trichodesmium*. *FEMS Microbiol Ecol* **30**:57–65.

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.

Levitan O, Sudhaus S, LaRoche J, Berman-Frank I. (2010). The influence of pCO_2 and temperature on gene expression of carbon and nitrogen pathways in *Trichodesmium* IMS101. *PloS One* **5**:e15104.

Ljones T. (1979). Nitrogen fixation and bioenergetics: the role of ATP in nitrogenase catalysis. *FEBS Lett* **98**:1–8.

Lundgren P, Janson S, Jonasson S, Singer A, Bergman B. (2005). Unveiling of novel radiations within *Trichodesmium* cluster by *hetR* gene sequence analysis. *Appl Environ Microbiol* **71**:190–196.

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.

Mayes SR, Dubbs JM, Vass I, Hideg E, Nagy L, Barber J. (1993). Further characterization of the *psbH* locus of *Synechocystis* sp. PCC 6803: inactivation of *psbH* impairs Q_A to Q_B electron transport in photosystem 2. *Biochemistry (Mosc)* **32**:1454–1465.

Mi H, Endo T, Ogawa T, Asada K. (1995). Thylakoid membrane-bound, NADPH-specific pyridine nucleotide dehydrogenase complex mediates cyclic electron transport in the cyanobacterium *Synechocystis* sp. PCC 6803. *Plant Cell Physiol* **36**:661–668.

Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, *et al.* (2011). An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc Natl Acad Sci* **108**:2124–2129.

Mohr W, Intermaggio MP, LaRoche J. (2010). Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium *Crocosphaera watsonii* WH8501. *Environ Microbiol* **12**:412–421.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**:W182–W185.

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

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.

Ottesen EA, Young CR, Gifford SM, Eppley JM, Marin R, Schuster SC, *et al.* (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* **345**:207–212.

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

Poretsky RS, Gifford S, Rinta-Kanto J, Vila-Costa M, Moran MA. (2009). Analyzing gene expression from marine microbial communities using environmental transcriptomics. *J Vis Exp* **24**:1086.

Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11**:1358–1375.

Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, *et al.* (2010). Unlocking short read sequencing for metagenomics. *PLoS One* **5**:e11840.

Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski B, Doherty M, *et al.* (2014). Microspatial gene expression patterns in the Amazon River plume. *Proc Natl Acad Sci* **111**:11085–11090.

Shi T, Ilikchyan I, Rabouille S, Zehr JP. (2010). Genome-wide analysis of diel gene expression in the unicellular N_2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *ISME J* **4**:621–632.

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

Shi Y, Tyson GW, Eppley JM, DeLong EF. (2010). Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J* **5**:999–1013.

Shipe RF, Carpenter EJ, Govil SR, Capone DG. (2007). Limitation of phytoplankton production by Si and N in the western Atlantic Ocean. *Mar Ecol Prog Ser* **338**:33–45.

Stewart FJ, Ottesen EA, DeLong EF. (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J* **4**:896–907.

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.

Sundström BG. (1984). Observations on *Rhizosolenia clevei* Ostenfeld (Bacillariophyceae) and *Richelia intracellularis* Schmidt (Cyanophyceae). *Bot Mar* **27**:345–356.

Turk-Kubo KA, Achilles KM, Serros TRC, Ochiai M, Montoya JP, Zehr JP. (2012). Nitrogenase (*nifH*) gene expression in diazotrophic cyanobacteria in the Tropical North Atlantic in response to nutrient amendments. *Front Microbiol* **3**:386.

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

Villareal TA, Carpenter EJ. (2003). Buoyancy regulation and the potential for vertical migration in the oceanic cyanobacterium *Trichodesmium*. *Microb Ecol* **45**:1–10.

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.

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

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**:316–322.

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

Figure Legends

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

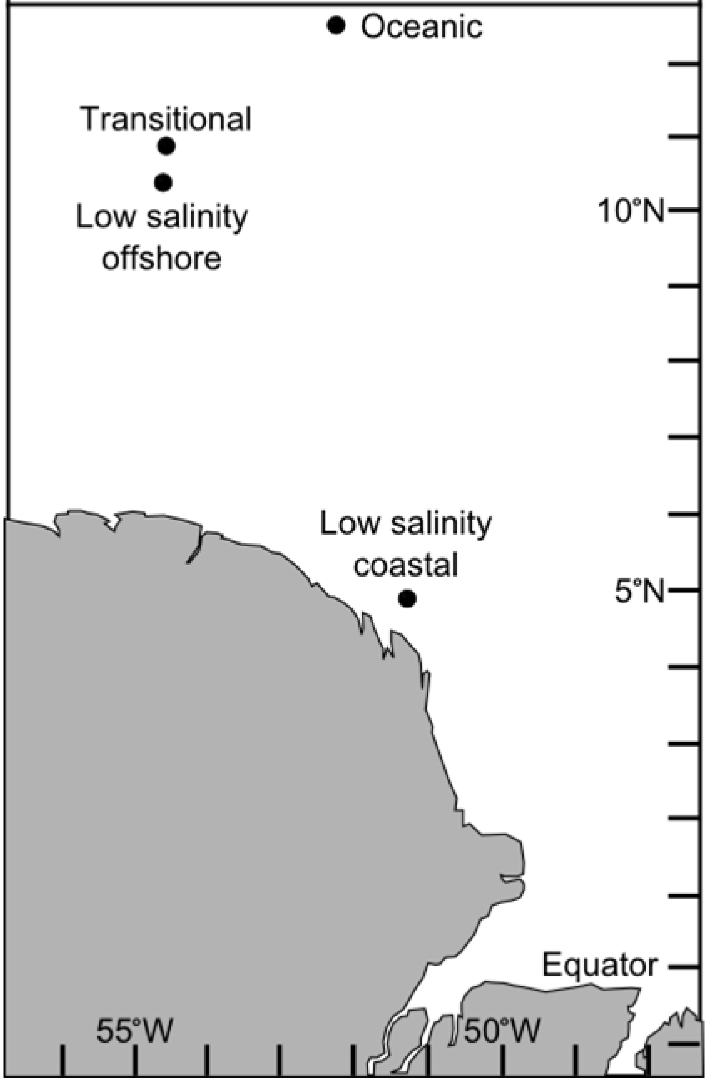

Figure 1. Amazon River plume stations.

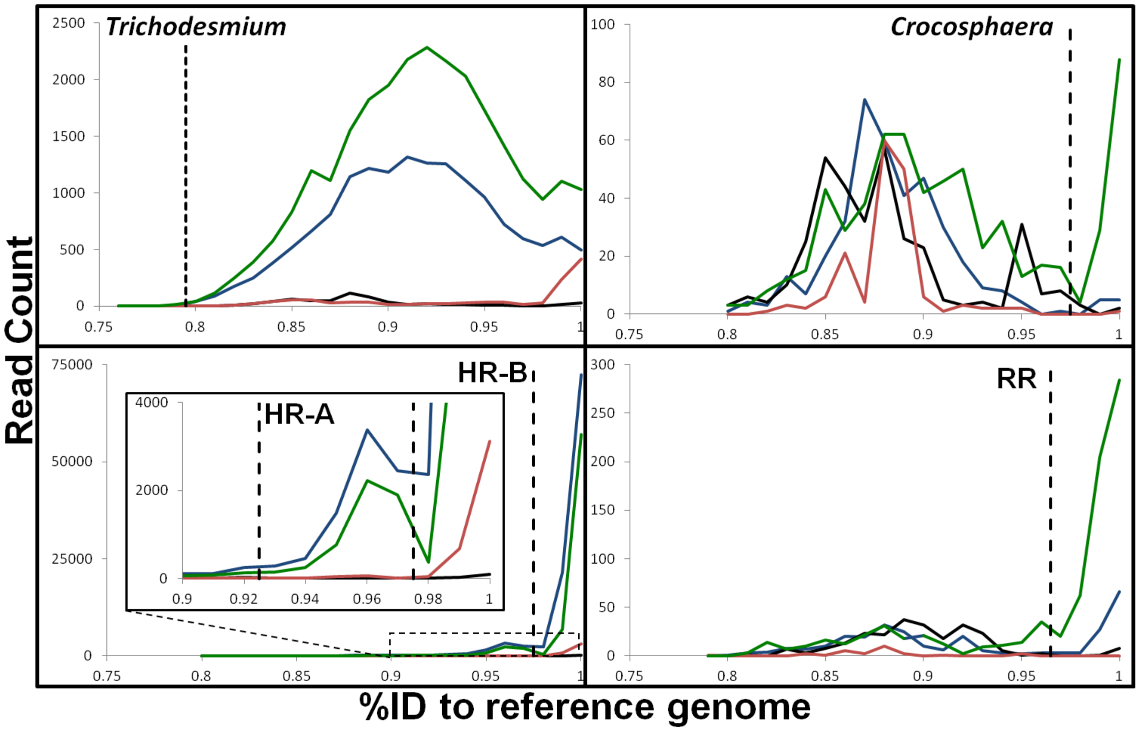

Figure 2. Natural populations similarity to genomes. Histograms of the percent identity of reads with a top hit to each of four diazotrophic cyanobacteria genomes from the transitional (blue), oceanic (green), low salinity offshore (red), and low salinity coastal (black) stations. The dotted lines mark the cut-off used in this study for each population. HR - *Hemiaulus*-associated *Richelia* (split in "A" and "B" populations as discussed in the text). RR - *Rhizosolenia*-associated *Richelia*.

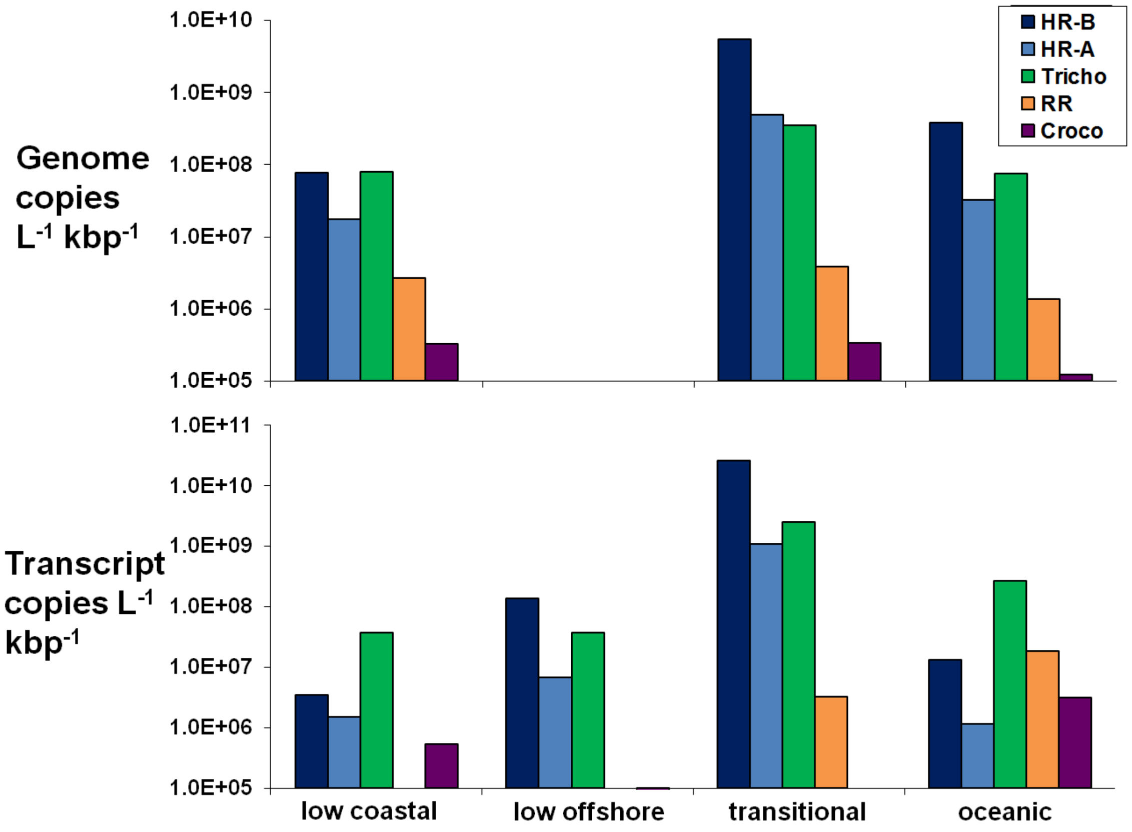

Figure 3. Diazotrophic cyanobacterial DNA and cDNA abundance. Normalized DNA (above) and cDNA (below) data for the five diazotrophic cyanobacterial populations at each of the four stations, with the exception of the low offshore station (DNA not sampled).

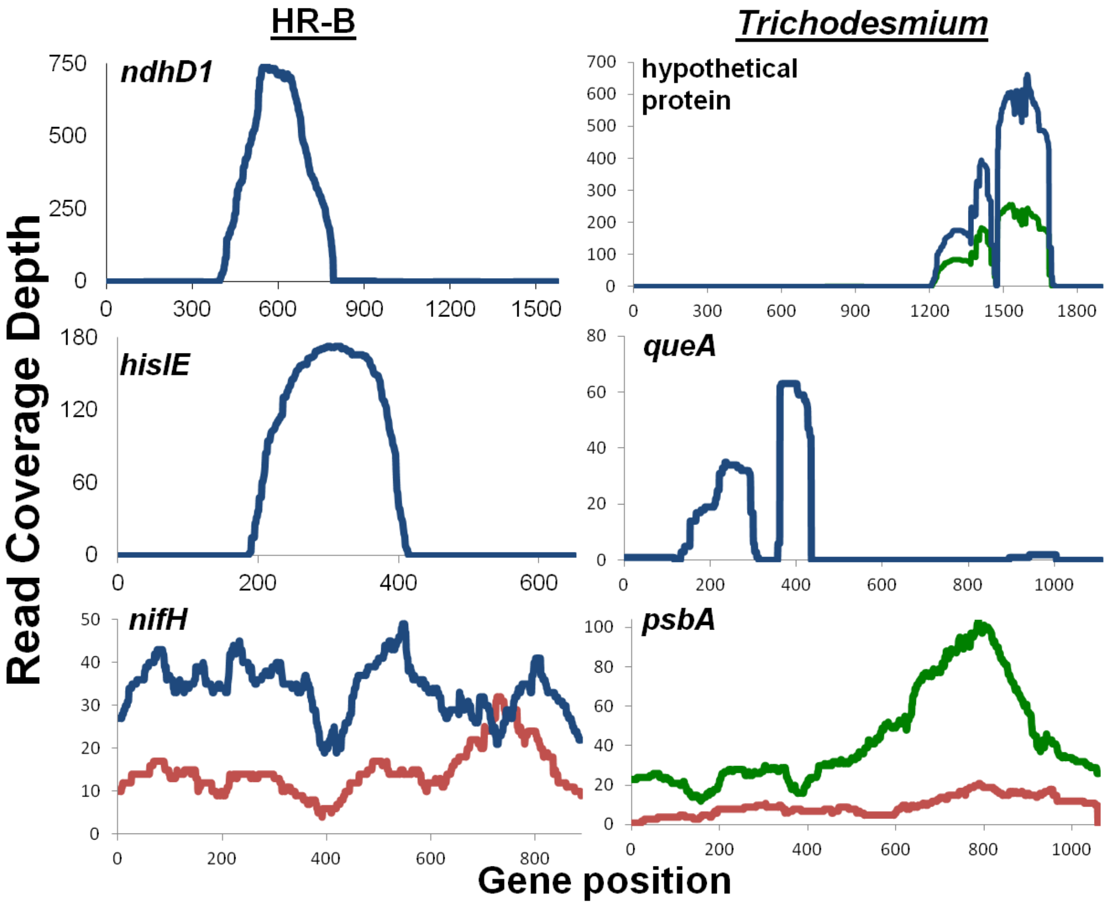

Figure 4. Transcript coverage of abundant genes. cDNA reads from the transitional (blue), oceanic (green), and low salinity offshore (red) stations mapped to abundant genes from the HR-B (left column) and *Trichodesmium* (right column) metatranscriptomes. *ndhD1* (RintHH_21740, NADH dehydrogenase I subunit D1), *hisIE* (RintHH_14390, fused phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase), *nifH* (RintHH_3070, nitrogenase iron protein), hypothetical protein (Tery_2611, FHA domain containing protein,), *queA* (Tery_0731, S-adenosylmethionine--tRNA-ribosyltransferase isomerase), *psbA* (Tery_4763, photosystem II protein D1)

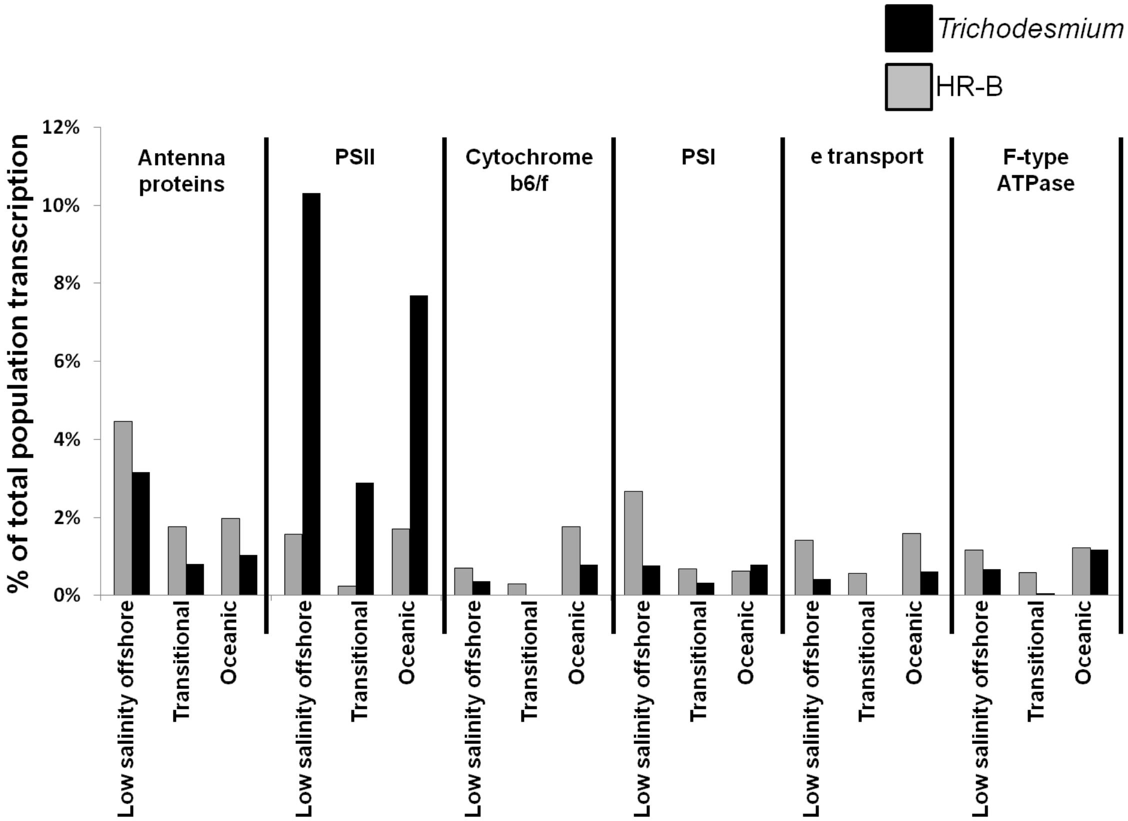

Figure 5. Photosynthesis component transcription. The normalized abundance of transcripts within six KEGG-defined photosynthesis components, relative to the total normalized transcript abundance for a population at a given station.

Table 1. 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 S1. Sample information and sequencing statistics of the Amazon River plume samples.

Sample ID	Station	Filter Size (µm)	Type	Reads	Reads after Trim	Std Reads	rRNA Reads	Remaining Reads	Normalization Factor
ACM1	transitional	2	DNA	5428695	3650345	26592	n/a	3623753	326737
ACM2	transitional	0.2	DNA	7488350	5355957	10543	n/a	5345414	824111
ACM3	low salinity coastal	2	DNA	6095193	4211422	3701	n/a	4207721	2347636
ACM4	low salinity coastal	0.2	DNA	6587526	4656473	3196	n/a	4653277	2718586
ACM5	oceanic	2	DNA	5138361	3212733	57600	n/a	3155133	200129
ACM6	oceanic	0.2	DNA	8854420	6343198	16054	n/a	6327144	606156
ACM7	transitional	2	Euk cDNA	15944969	10261507	113810	9916	10137781	1690
ACM8	low salinity coastal	2	Euk cDNA	18327975	11572252	17455	25672	11529125	20833
ACM9	low salinity off-shore	2	Euk cDNA	15427448	9050683	273423	11481	8765779	697
ACM10	oceanic	2	Euk cDNA	15930794	9679397	203133	16502	9459762	938
ACM11	low salinity coastal	0.2	Prok cDNA	15058449	5694376	37903	3222353	5656473	168983
ACM12	low salinity coastal	2	Prok cDNA	10001411	2435885	41397	1360889	2394488	188525
ACM13	low salinity off-shore	2	Prok cDNA	17035424	4908011	354928	2594703	4553083	12896
ACM14	transitional	0.2	Prok cDNA	16379485	4696328	54005	2478789	4642323	85944
ACM15	transitional	2	Prok cDNA	11055118	2575008	2772	628451	2572236	2132430
ACM16	oceanic	0.2	Prok cDNA	11436878	3167818	179159	1416551	2988659	20944
ACM17	oceanic	2	Prok cDNA	15529442	4626581	274916	2150303	4351665	16386

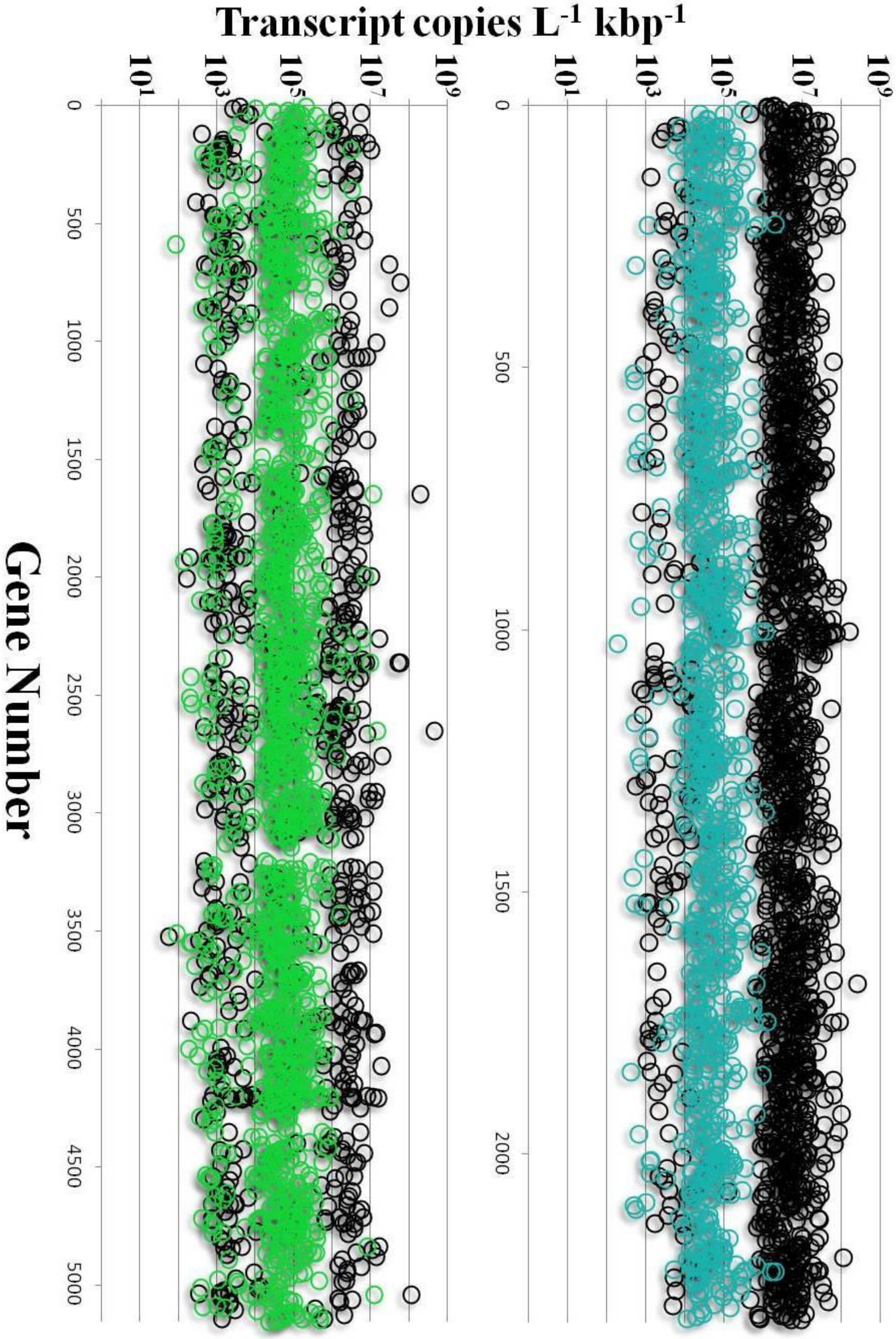

Figure S1. The transcript abundances of the genes across the genomes of *Richelia intracellularis* HH01 (HR-B population, above) and *Trichodesmium erythraeum* IMS101 (below) at the two stations where each population was most abundant; low salinity

offshore (blue) for HR-B, and oceanic (green) for *Trichodesmium*, and transitional (black) for both populations.

***Trichodesmium* population diversity**

In contrast to the diatom-associated cyanobacteria, the sequences of free-living *Trichodesmium* populations had a much wider range of nucleotide sequence divergence from the representative genome, and there was no distinct separation among the *Trichodesmium* populations. As for the diatom symbionts, *hetR* has been a common genetic marker used to study *Trichodesmium* diversity (Janson *et al.*, 1999; Schiefer *et al.*, 2002; Lundgren *et al.*, 2005; Hynes *et al.*, 2012). Eight *Trichodesmium* cDNA reads and one DNA read aligned within the 448 bp *hetR* region amplified in most of these studies, and these nine reads were used to more fully characterize the *Trichodesmium* populations in the Amazon River plume. A BLAST analysis of the nine reads was conducted against the nr/nt database (blastn, NCBI), and each of the nine reads was most similar to one of five sequences amplified from four different *Trichodesmium* species (**Table S2**). 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, we believe there could be up to four different phylotypes that comprised the *Trichodesmium* metagenomes and transcriptomes.

Table S2. The percent identity (nucleotide) of nine *Trichodesmium* reads and each of the *hetR* partial sequences that are the best hit to at least one of the reads. The highest percent hit is highlighted in bold. All reads have the ID prefix "HWI-EAS165_0077_FC70822AAXX:" omitted for conciseness.

		Accession	AF410432	AF490680	AF490679	HM486692	AF490684
		Reference	Schiefer et al. 2002	Lundgren et al. 2005	Lundgren et al. 2005	Hynes et al. 2012	Lundgren et al. 2005
Read ID	**Sample Type**	**Strain**	***Trichodesmium erythraeum* IMS101**	***Trichodesmium aureum* strain B49**	***Trichodesmium hildebrandtii***	***Trichodesmium thiebautii* II-3**	***Trichodesmium thiebautii***
5:116:4695:6318	cDNA	Low Sal Off	**100.0**	88.2	89.0	88.2	89.5
1:7:17644:14996	DNA	Transitional	87.4	**98.7**	93.4	94.7	95.4
6:61:1375:11580	cDNA		87.5	93.4	**100.0**	97.4	97.4
7:27:1567:4520	cDNA	Oceanic	85.7	91.4	**98.6**	95.7	95.7
7:32:12859:11505	cDNA		87.3	95.3	98.1	**100.0**	99.5
7:32:12853:11523	cDNA		87.3	95.3	98.1	**100.0**	99.5
7:19:18183:14579	cDNA		88.1	95.5	96.6	**100.0**	**100.0**
7:76:10525:12982	cDNA		93.1	99.4	98.3	98.8	**100.0**
7:56:14891:3783	cDNA		93.3	99.3	98.0	98.7	**100.0**

T. erythraeum IMS 101 is currently the only fully sequenced *Trichodesmium* genome, but there are two environmental sequence data sets available to compare to the Amazon sequences. A BLAST analysis (blastn, e-value $\leq 10^{-4}$) of Amazon *Trichodesmium* cDNA and DNA reads was conducted against a database containing the *T. erythraeum* IMS 101 genome and metagenomes from the North Atlantic (BATS, Bermuda Atlantic Time Series) and the North Pacific Subtropical Gyres (IMG Genome IDs: 2156126005 and 2264265224, respectively). Overall, more than half of the reads (56%) had a top BLAST hit to the Bermuda Atlantic Time-series Study (BATS) metagenome, and about one third of the reads (34%) had a top hit to the North Pacific metagenome. The remaining one tenth of reads were more similar to the *T. erythraeum* IMS 101 genome than to either metagenome. At the transitional and oceanic stations, a majority of reads (59% and 56%, respectively) had a best BLAST hit to the *Trichodesmium* BATS metagenome, with many of the rest most similar to the North Pacific metagenome (33% and 36%) (**Figure S2**). Few reads at either of these stations had a top BLAST hit to the *T. erythraeum* IMS101 genome (8% at each station). This indicates that the taxonomic composition of the *Trichodesmium* populations at the transitional and oceanic stations may be very similar to each other. The offshore low salinity station was the only station where more reads were more similar to the *T. erythraeum* IMS 101 genome (61%) than the BATS (29%) or North Pacific (11%) metagenomes (**Figure S2**). Additionally, the only offshore low salinity station *hetR* read was 100% identical to *T. erythraeum* IMS 101 (**Table S2**). Therefore, the *T. erythraeum* IMS 101 genome appears to be representative of the natural population at the offshore low salinity 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. On average, the reads with a top hit to either metagenome were 7.0% more identical to the metagenome than to the *T. erythraeum* IMS 101 genome. Sequencing of a variety of *Trichodesmium* isolates is necessary to determine if the metabolic capabilities and ecological roles differ as significantly as the genetic diversity of the various *Trichodesmium* populations.

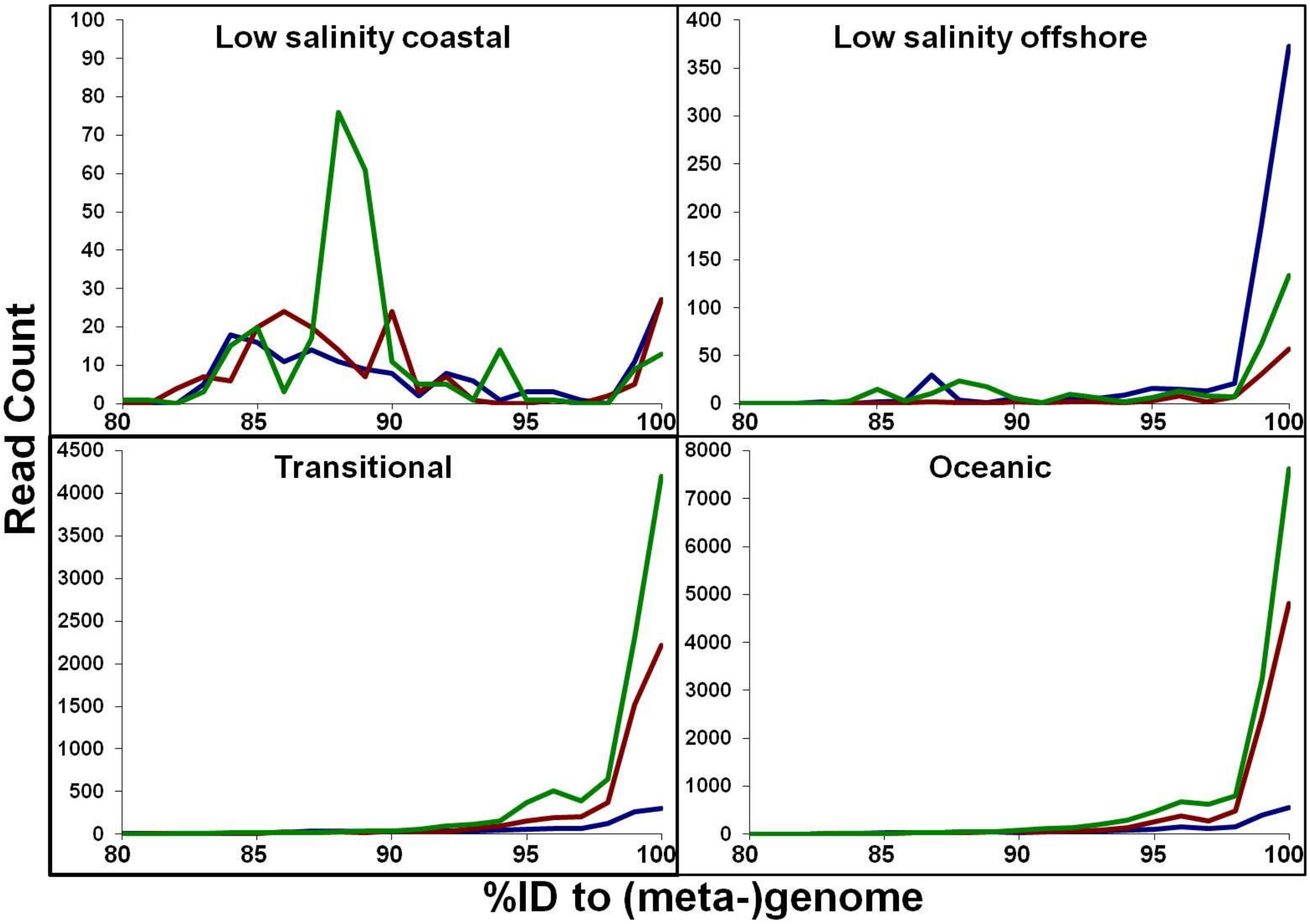

Figure S2. Percent identity histograms of *Trichodesmium* DNA and cDNA reads at each station to the top BLAST hit within a database comprised of the *T. erythraeum* IMS 101 genome (blue), and *Trichodesmium* metagenomes from BATS (green) and the North Pacific (red).

Supplementary Information References

Carpenter EJ, Subramaniam A, Capone DG. (2004). Biomass and primary productivity of the cyanobacterium *Trichodesmium* spp. in the tropical N Atlantic ocean. *Deep Sea Res Pt I* **51**:173–203.

Hynes AM, Webb EA, Doney SC, Waterbury JB. (2012). Comparison of cultured *Trichodesmium* (Cyanophyceae) with species characterized from the field. *J Phycol* **48**:196–210.

Janson S, Bergman B, Carpenter EJ, Giovannoni SJ, Vergin K. (1999). Genetic analysis of natural populations of the marine diazotrophic cyanobacterium *Trichodesmium*. *FEMS Microbiol Ecol* **30**:57–65.

Lundgren P, Janson S, Jonasson S, Singer A, Bergman B. (2005). Unveiling of novel radiations within *Trichodesmium* cluster by *hetR* gene sequence analysis. *Appl Environ Microbiol* **71**:190–196.

Schiefer W, Schütz K, Hachtel W, Happe T. (2002). Molecular cloning and characterization of *hetR* genes from filamentous cyanobacteria. *Biochim Biophys Acta BBA-Gene Struct Expr* **1577**:139–143.

Sohm JA, Mahaffey C, Capone DG. (2008). Assessment of relative phosphorus limitation of *Trichodesmium* spp. in the North Pacific, North Atlantic, and the north coast of Australia. *Limnol Oceanogr* **53**:2495–2502.

Webb EA, Wisniewski Jakuba R, Moffett JW, Dyhrman ST. (2007). Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol Oceanogr* **52**:2221–2232.