The ISME Journal (2009) 3, 1286–1300

www.nature.com/ismej

ORIGINAL ARTICLE

Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean

Ian Hewson[1], Rachel S Poretsky[2], Sonya T Dyhrman[3], Brian Zielinski[4], Angelicque E White[5], H James Tripp[1], Joseph P Montoya[6] and Jonathan P Zehr[1]
[1]*Department of Ocean Sciences, University of California Santa Cruz, Santa Cruz, CA, USA;* [2]*Department of Marine Science, University of Georgia, Athens, GA, USA;* [3]*Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA, USA;* [4]*Department of Marine Science, University of South Florida, St Petersburg, FL, USA;* [5]*College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA and* [6]*School of Biology, Georgia Institute of Technology, Atlanta, GA, USA*

Trichodesmium **are responsible for a large fraction of open ocean nitrogen fixation, and are often found in complex consortia of other microorganisms, including viruses, prokaryotes, microbial eukaryotes and metazoa. We applied a community gene expression (metatranscriptomic) approach to study the patterns of microbial gene utilization within colonies of** ***Trichodesmium*** **collected during a bloom in the Southwest Pacific Ocean in April 2007. The survey generated 5711-day and 5385-night putative mRNA reads. The majority of mRNAs were from the co-occurring microorganisms and not** ***Trichodesmium*****, including other cyanobacteria, heterotrophic bacteria, eukaryotes and phage. Most transcripts did not share homology with proteins from cultivated microorganisms, but were similar to shotgun sequences and unannotated proteins from open ocean metagenomic surveys.** ***Trichodesmium*** **transcripts were mostly expressed photosynthesis, N_2 fixation and S-metabolism genes, whereas those in the co-occurring microorganisms were mostly involved in genetic information storage and processing. Detection of** ***Trichodesmium*** **genes involved in P uptake and As detoxification suggest that local enrichment of N through N_2 fixation may lead to a *P*-stress response. Although containing similar dominant transcripts to open ocean metatranscriptomes, the overall pattern of gene expression in** ***Trichodesmium*** **colonies was distinct from free-living pelagic assemblages. The identifiable genes expressed by** ***Trichodesmium*** **and closely associated microorganisms reflect the constraints of life in well-lit and nutrient-poor waters, with biosynthetic investment in nutrient acquisition and cell maintenance, which is in contrast to gene transcription by soil and coastal seawater microbial assemblages. The results provide insight into aggregate microbial communities in contrast to planktonic free-living assemblages that are the focus of other studies.**
The ISME Journal (2009) **3,** 1286–1300; doi:10.1038/ismej.2009.75; published online 2 July 2009
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: Trichdoesmium; metatranscriptomics; aggregate; gene expression; community

Introduction

Over a decade of research has highlighted the critical importance of open ocean nitrogen fixation in supporting oligotrophic microbial food web N demands (Carpenter, 1983; Carpenter and Romans, 1991; Michaels *et al.*, 1996; Karl *et al.*, 1997; Capone, 2000; Codispoti *et al.*, 2001). Nitrogen fixation is carried out by a diverse suite of microorganisms, including unicellular cyanobacteria, other Bacteria, Archaea, and colonial cyanobacteria (Dugdale *et al.*, 1961; Carpenter, 1983; Villareal, 1990; Zehr *et al.*, 2003). Filamentous, nonheterocystous cyanobacteria of the genus *Trichodesmium* are biogeochemically important organisms in tropical waters, fixing at least 80 Gt of N per year, and are globally distributed in warm tropical and subtropical waters (Capone *et al.*, 1997; Capone and Carpenter, 1999; Subramaniam

Correspondence: I Hewson, Department of Microbiology, Cornell University, Wing Hall 403, Ithaca, NY 14853, USA.
E-mail: hewson@cornell.edu
Received 30 March 2009; revised 26 May 2009; accepted 30 May 2009; published online 2 July 2009

et al., 1999; Orcutt *et al.*, 2001; Westberry and Siegel, 2006). Filaments (trichomes) of the cyanobacterium form aggregates that are arranged as puffs loosely connected at the center of filaments, or as polarized rafts, both of which form dense aggregations at the surface during calm conditions (Bowman and Lancaster, 1967; Capone *et al.*, 1997). Blooms of *Trichodesmium* can be observed from space (Capone *et al.*, 1997; Subramaniam *et al.*, 1999; Westberry and Siegel, 2006), and represent a pseudobenthic surface on which other organisms may recruit (O'Neil and Roman, 1994; O'Neil *et al.*, 1996). Despite their global importance, studies of *Trichodesmium* gene expression have been restricted to a handful of genes involved in heterocyst differentiation, global nitrogen regulation, glutamine synthesis, nitrogen fixation, iron stress and phosphorus acquisition (Zehr and McReynolds, 1989; Kramer *et al.*, 1996; Webb *et al.*, 2001; Dyhrman *et al.*, 2002; El-Shehawy *et al.*, 2003; Dyhrman *et al.*, 2006).

Previous reports of the microflora associated with *Trichodesmium* colonies indicated that a wide spectrum of microorganisms are found closely associated with colonies, including viruses, bacteria, eukaryotic microorganisms and metazoa (Paerl *et al.*, 1989; Siddiqui *et al.*, 1992; Zehr, 1995; Ohki, 1999; Sheridan *et al.*, 2002). The abundances of closely associated organisms are elevated compared with surrounding seawater (Sheridan *et al.*, 2002), and the hydrolytic enzyme activities within the colonies are higher than outside colonies (Nausch, 1996). Therefore, colonies of *Trichodesmium* represent hotspots of biological activity in the low productivity and nutrient-poor waters where they inhabit. As a large amount of fixed N is released from the cells into the surrounding waters (Capone *et al.*, 1994; Mulholland *et al.*, 2006), co-occurring organisms can potentially benefit from localized N-enrichment surrounding and within the colonies (Capone *et al.*, 1994). The enhanced N conditions within colonies may cause P limitation within the colony infrastructure and selection for organisms capable of enhanced P uptake. However, knowledge of the function and physiology of organisms inhabiting *Trichodesmium* colonies, especially under bloom conditions, is unknown.

Recent application of whole-genome random transcript sequencing (metatranscriptomics) has shown utility for understanding coastal and open ocean microbial community ecophysiology (Poretsky *et al.*, 2005, 2009b; Frias-Lopez *et al.*, 2008; Gilbert *et al.*, 2008; Hewson *et al.*, 2009). Unlike metagenomic surveys which elucidate potential genetic capabilities (Tyson *et al.*, 2004; Venter *et al.*, 2004; Rusch *et al.*, 2007), and which are mostly focused on the discovery of novel metabolic pathways or extent of diversity, metatranscriptomics provide information on active processes in dominant microorganisms (Poretsky *et al.*, 2005, 2009b; Frias-Lopez *et al.*, 2008). Metatranscriptomic studies to date have elucidated that a large proportion of genes expressed in open ocean assemblages have no known function, or show less similarity to cultivated microorganisms than metagenomic surveys (Frias-Lopez *et al.*, 2008; Gilbert *et al.*, 2008). There are no published studies of community gene expression within particles, despite their importance as 'hotspots' of biological activity in the oligotrophic ocean.

The aim of this study was to examine the *in situ* gene expression of microbial assemblages associated with *Trichodesmium* during a bloom to determine the composition of active components of the associated microflora and to identify major metabolic characteristics of tightly associated microorganisms. We applied a metatranscriptomic approach to independent samples of *Trichodesmium* aggregates collected in the day and night. These results provide new information on the diversity and ecophysiology of microorganisms closely associated with *Trichodesmium* blooms.

Materials and methods

Sampling of Trichodesmium *bloom sample*

Samples for metatranscriptomic analysis were collected on board the *R/V Kilo Moana* during an intense bloom of the cyanobacterium at station KM070324 (15°S, 178°45′E) north of the Fiji Islands on 12 April 2007. The sea state at the time of sampling was calm (glassy surface) and *Trichodesmium* colonies were observed at the surface in a thin layer. *Trichodesmium* was collected using a 64-μm mesh plankton net. Immediately after the net was retrieved, the tow material was placed in a seawater-rinsed bucket. Subsamples of *Trichodesmium* were then collected within 5 min by skimming the floating material (*Trichodesmium* and associated organisms) into 50-ml centrifuge tubes which were immediately frozen in liquid nitrogen. The night sample was collected at 0100 h, whereas the day sample was collected at 0900 h. Samples for enumeration of viruses and bacteria were collected at 5 m using a Niskin bottle mounted on a CTD rosette. Seawater from the Niskin was retrieved into 50-ml centrifuge tubes to which 2 ml of 0.02-μm filtered formaldehyde was added, and samples processed immediately following an established protocol (Patel *et al.*, 2007). Samples were transported on liquid nitrogen to the University of California Santa Cruz for analysis.

RNA extraction

The *Trichodesmium* samples were thawed in the laboratory by centrifuging tubes at $5000 \times g$ for 20 min, which pelleted the *Trichodesmium* and associated microorganisms. The supernatant was decanted and cells (approximate volume was 5 ml, containing $\sim 5 \times 10^6$ *Trichodesmium* cells) were immediately placed on ice. Subsamples (~ 0.1 ml)

of cell pellets were removed using a pipette tip, placed into RNase-free 2-ml cryovials, then subjected to the RNeasy Plant Mini kit (Qiagen, Valencia, CA, USA) with the following modification. Glass beads (100 μl) and 450 μl of buffer RLT containing 1% β-mercaptoethanol were added, and the tubes placed in a bead beater for 2 min. After mechanical lysis, the homogenized material was processed according to the Qiagen protocol. The resulting RNA was eluted in deionized water and subsequently treated to remove DNA using the RNase-free DNase kit (Zymo Research, Orange, CA, USA).

mRNA enrichment

The total RNA sample containing rRNA and mRNA was subjected to two protocols to enrich the fraction of mRNA relative to rRNA following an approach applied previously to open ocean microbial communities (Poretsky *et al.*, 2009a, b). RNA was first subjected to terminator exonuclease treatment (which removes 5′-monophosphate-capped RNA) using the mRNA-ONLY protocol (Epicentre, Madison, WI, USA). rRNA was further reduced by subtractive hybridization using the MicrobExpress kit (Ambion, Austin, TX, USA) following the manufacturer's protocols. The resulting mRNA enrichment was purified (according to the MicrobExpress protocol) by precipitating, washing and resuspending the RNA in 20 μl of nuclease-free deionized water.

RNA in vitro *amplification*

The mRNA-enriched samples were amplified using *in vitro* transcription after mRNAs were polyadenylated, as part of the MessageAmp II—Bacteria aRNA kit (Ambion). The polyadenylation step was performed on 180 ng of the mRNA-enriched samples. The protocol does not select for prokaryotic mRNAs as all mRNAs are polyadenylated (even those which are already polyadenylated, that is, eukaryotic mRNA); hence the resulting aRNA contains eukaryotic, prokaryotic and phage transcripts. The aRNA was prepared according to the manufacturer's protocols, and eluted in a final volume of 150 μl of deionized water. Samples were subsequently concentrated in a speed evaporator (Thermo Savant, Waltham, MA, USA) to a volume of 50 μl and quantified using a spectrophotometer (Nanodrop Technologies, Wilmington, DE, USA).

Double-stranded cDNA synthesis

The aRNA (15 μg) was converted to double-stranded cDNA (ds cDNA) by reverse transcription and second strand synthesis reactions. First, the aRNA was diluted to 500 ng μl^{-1} (10 μl total). Triplicate samples were then treated with 400 U reverse transcriptase (Superscript III; Invitrogen, Carlsbad, CA, USA), 10 nmol dNTPs (Invitrogen), 1 × First Strand Buffer (Invitrogen), 0.1 μmol dithiothreitol (DTT) and 500 ng random primers (Promega, Madison, WI, USA). The reactions were first heated to 70 °C for 10 min in the presence of the random primers and dNTPs, then cooled on ice before addition of the remaining reagents. The reactions were then incubated at 50 °C for 50 min. After first strand synthesis, 1 × Second Strand Buffer (Invitrogen), 6 μmol dNTPs, 10 U *E. coli* DNA ligase, 40 U *E. coli* DNA polymerase and 2 U RNAse H were added to the reactions, which were then incubated at 16 °C for 2 h. At the conclusion of the reaction, each tube was amended with 10 U of T4 DNA ligase and incubated at 14 °C for 10 min. Following the second strand synthesis, the samples were treated with 20 μg of RNAse A at 37 °C for 30 min. RNAse A treatment was terminated by mixing the samples with 50% phenol:chloroform:isoamyl alcohol (24:1:0.1) by inversion, after which the aqueous layer was removed. Nucleic acids in the aqueous layer were precipitated with 16 μl of 7.5 M NH_4COOH, 35 μg Glycogen (Ambion) and 326 μl of 100% EtOH, and precipitated overnight at −20 °C. The samples were then centrifuged at 15,000 × *g* for 1 h. Supernatant was decanted and pellets were rinsed with 70% EtOH at −20 °C. The pellets were dried in a speed evaporator (Savant) for 10 min, resuspended in 20 μl of distilled H_2O, and replicate reactions were combined. Small ds cDNA fragments and other materials were removed by passing the ds cDNA through a cleanup kit (Zymo Clean & Concentrator −5). The cleaned ds cDNA was quantified by spectrophotometer at 260 nm, and size range and concentration of fragments were determined by running 1 μl on an Agilent (Santa Clara, CA, USA) Bioanalyzer DNA 1000 chip. The samples were then pyrosequenced as described elsewhere in picoliter reactors on a GS FLX platform (454 Life Sciences, Branford, CT, USA) (Margulies *et al.*, 2005). Sequence reads from the two metatranscriptomes are available at the Community Cyberinfrastructure of Advanced Marine Microbial Ecology Research and Analysis under accession CAM_P0000051.

Bioinformatic analysis

The pyrosequence reads were initially analyzed to remove replicate sequences by comparing the first 100 bp of sequence in Microsoft Excel. As random primers were used in the creation of cDNA libraries, and furthermore the cDNA was sheared before sequencing, it is unlikely that replicate sequences represent true replicate transcripts, but rather are an artifact of the sequencing protocol. Furthermore, poor-quality sequence towards the end of the sequence read length precludes removal of replicate sequence over the entire read length. After replicate sequences were removed (leaving one sequence to represent each replicate), the dereplicated libraries were compared against the Ribosomal Database Project (RDP II; (Cole *et al.*, 2007)), a boutique database of 23S rRNAs and a boutique database containing the 5.8S, 18S and 28S of common

marine eukaryotic microorganisms, using BLASTn (Altschul *et al*., 1997). Sequences matching at E-values of $<10^{-3}$ were discarded from the dereplicated sequence library as they represented microbial rRNAs. In addition, sequences <75 bp or containing $>60\%$ of any single base were discarded.

The mRNA libraries were compared by BLASTx against the All Prokaryotic Proteins, All Eukaryotic Microbial Proteins and All Viral Proteins databases in the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) server (http://camera.calit2.net/) (Rusch *et al*., 2007). The putative taxonomic affiliation and function were assigned as the top match for each read which had an E-value of $<10^{-3}$. Sequence reads with no matches in the three protein databases were compared by BLASTn analysis against the non-redundant (nr) database at NCBI and all assembled metagenomic reads of the Global Ocean Survey (GOS) to further resolve hits. These sequence reads were compared by BLASTx against all assembled proteins at CAMERA, and the functional category of protein matches was determined by comparing the sequence reads against the Kyoto Encyclopedia for Genes and Genomes (KEGG) database. Furthermore, remaining unclassified sequences were compared with the nr protein sequence database at NCBI using BLASTx. Prokaryotic and microbial eukaryotic reads with E-values $<10^{-3}$ were compiled with the previous hits. Data from additional remaining reads with bit score $\geqslant 40$ (E-value ~ 0.06) were considered separately, as this has been the lower cutoff value used in previous studies (Frias-Lopez *et al*., 2008; John *et al*., 2009).

Results and Discussion

Microscopic observations of Trichodesmium *bloom samples*

Trichodesmium abundance in surface waters was $\sim 1.13 \times 10^4$ trichomes per liter and was primarily of raft morphology with a large number of free filaments. The abundance of heterotrophic bacteria immediately below the bloom at 5 m, as measured by SYBR Green I staining and epifluorescence microscopy (Noble and Fuhrman, 1998; Patel *et al*., 2007), was 2.8×10^6 cells per ml and virus abundance of 2.4×10^7 viruses per ml (I Hewson, unpublished data). The abundance of *Synechococcus* and *Prochlorococcus* at 15 m, as measured by phycoerythrin autofluorescence flow cytometry, was 6×10^3 cells per ml and 3×10^1 cells per ml, respectively (B Carter, unpublished data).

Owing to the extremely heterogeneous nature of *Trichodesmium* colonies, it was not possible to quantify the associated organisms accurately, but we qualitatively describe the biomass of co-occurring microorganisms observed at night in a SYBR Green I-stained sample prepared from the bloom depth and from light microscopic observations of bloom samples. In SYBR Green I-stained samples, we observed bacteria attached to single *Trichodesmium* trichomes and high concentrations of long filaments between trichomes in colonies of *Trichodesmium* (Figure 1). We also observed cells resembling *Synechococcus* and non-phycoerythrin-containing cells within the colonies. Under light microscopic examination, we observed multiple morphologies of *Trichodesmium* and other cyanobacteria, including *Phormidium*-like filaments. Morphologically diverse eukaryotic flora were also seen among the *Trichodesmium* colonies, including dinoflagellates, diatoms, ciliates, silicoflagellates, a foram, and radiolarians (Figure 2). In addition, at least one unidentified metazoan was observed within raft colonies of *Trichodesmium*.

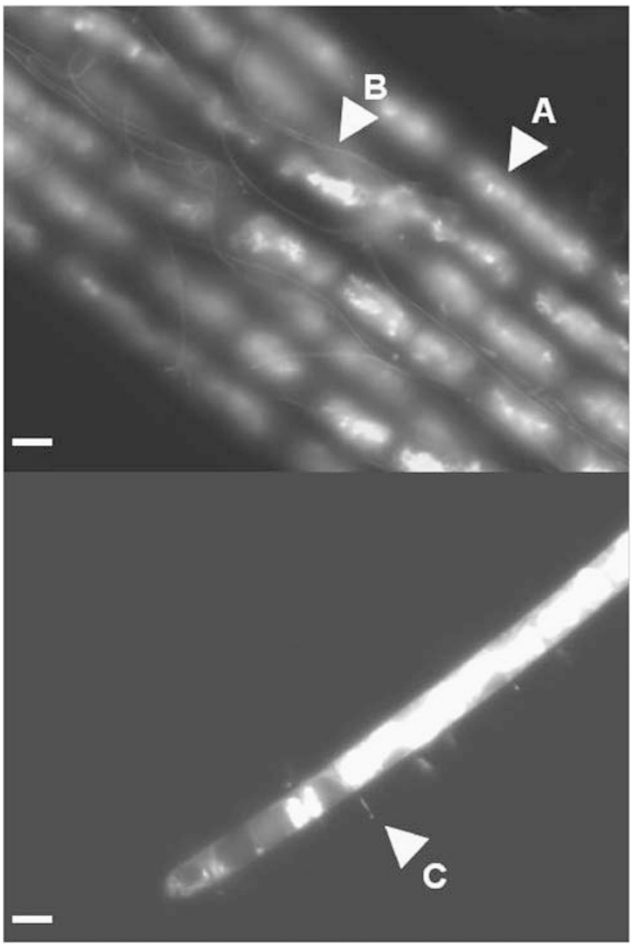

Figure 1 Photomicrographs of SYBR Green I-stained samples of *Trichodesmium* colonies at night. The filaments (A) forming rafts (likely *Trichodesmium erythraeum*) had a large number of filamentous bacteria (B) intertwined with filaments, whereas other filaments had individual bacterial cells (C) attached to the outside of the filaments. Scale bar = 10 μm.

Metatranscriptome library characteristics

We obtained a total of 5711 sequences in the day metatranscriptome and 5385 sequences in the night

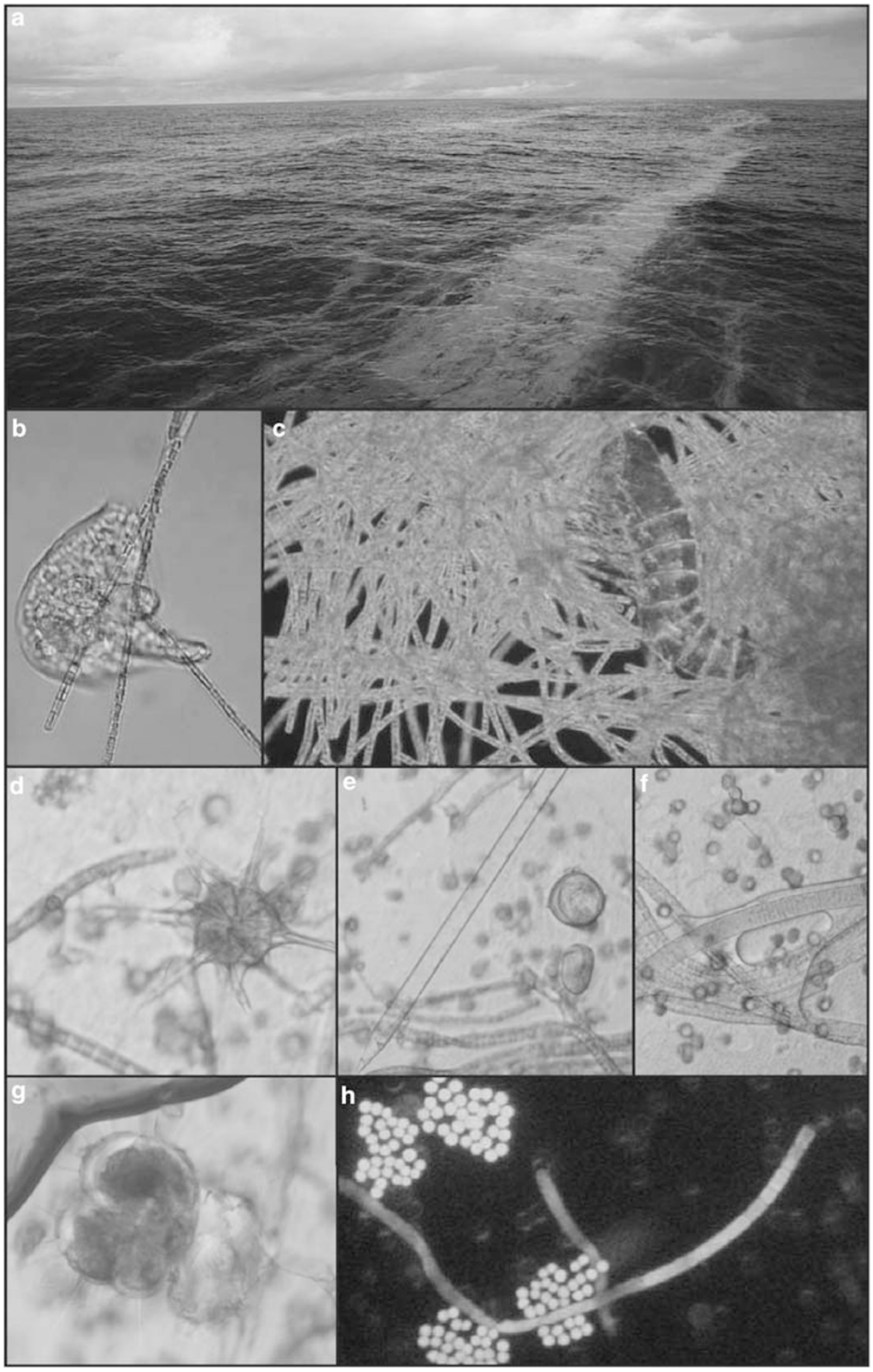

Figure 2 Photograph of the *Trichodesmium* bloom at KM070324 (**a**) and photomicrographs of organisms associated with *Trichodesmium* colonies at night: (**b**) copepod nauplii larva , (**c**) harpactacoid copepod, (**d**) radiolarian, (**e**) diatoms and dinoflagellates, (**f**) *Lyngbya*-like cyanobacterium among *Trichodesmium* filaments, (**g**) Foraminifera and (**h**) unidentified coccoid microorganisms. Scale bar = 50 μm.

metatranscriptome that were putatively mRNAs >75 bp and with individual base contents <60% of total metatranscript sequence (Table 1). The mean sequence length was 210 bp in the day and 213 bp in the night. The mean G+C content of metatranscripts was 50.9 ± 0.1% in the day and 51.6 ± 0.1% in the night. The number of sequence reads matching prokaryotic, eukaryotic and viral databases at

Table 1 Summary of number of sequence reads matching proteins and metagenomes retrieved from *Trichodesmium* metatranscriptomes in the day and night

	Day	*Night*
Putative mRNAs	5711	5385
Possible protein-encoding genes		
Prokaryotic proteins	667	536
Trichodesmium	198	19
Eukaryotic microbial proteins	231	142
Phage proteins	33	9
Eukaryotic metazoan proteins	165	203
Sequences not matching annotated proteins or rRNAs at E <0.001	4417	4476
Sequences matching nr at E <0.06 and bit score ⩾40		
Prokaryotic proteins	92	61
Eukaryotic proteins	510	658
GOS comparison		
BLASTn to reads	2216	2836
BLASTn to assembled sequences	2083	2763
BLASTx to unannotated proteins	929	1144
Non-rRNA sequences not matching annotated proteins or GOS	2580	1770

Abbreviations: GOS, Global Ocean Survey; nr, non-redundant database.

CAMERA at E-values <0.001 and Genbank nr at bit score $\geqslant 40$ (which for sequence alignments in our study equates to an E-value of ~ 0.06) varied in the day and night (Table 1).

The majority of matches to proteins were at E-values $>10^{-20}$ (63 and 57% of all hits in the day and night, respectively). Previous metagenomic and metatranscriptomic studies have used various E-value cutoffs to annotate libraries of sequences with approximately the same length as those in this study, including $\leqslant 10^{-4}$ (Thurber *et al.*, 2008), $\leqslant 10^{-3}$ (Culley *et al.*, 2006; Dinsdale *et al.*, 2008; Gilbert *et al.*, 2008), $\leqslant 10^{-2}$ (Poretsky *et al.*, 2009a, b), and bit score cutoff of $\geqslant 40$ (Frias-Lopez *et al.*, 2008). As the E-value of read is related to the length of alignment (shorter read matches have higher E-values; Wommack *et al.*, 2008), we chose to use an E-value cutoff of $E<0.001$, and separately analyze reads matching at bit score $\geqslant 40$ in line with previous metatranscriptomic efforts. Furthermore, as we focused our study on the frequency of individual genes among the randomly sampled transcript pool, we chose not to assemble the metatranscriptomic sequence data.

The taxonomic affiliation of bacterial mRNAs was different between the two metatranscriptomes (Figure 3a). *Trichodesmium* mRNAs made up the dominant fraction of all annotated mRNAs in the day sample (30%), but made up only a small fraction of all night prokaryotic reads (4%). Eukaryotic microbial mRNAs (Figure 3b) were primarily most closely related to proteins in Fungi (30 and 27% of all recognized eukaryotic mRNAs in day and night, respectively), followed by Alveolates (25% day and 22% night), *Viridiplantae* (11% day and 7% night) and Stramenopiles (10% day and 7% night). Viral mRNAs in the day and night metatranscriptomes were primarily cyanophage, with fewer eukaryotic virus and bacteriophage mRNAs. Lower stringency (bit score $\geqslant 40$) results from comparisons against the nr database were heavily affiliated with metazoans (80% day and 90% night; Table 2).

Our results highlight the richness of microorganisms inhabiting *Trichodesmium* colonies under bloom conditions. Previous studies have observed the presence and microbial activities of bacteria on the surface of *Trichodesmium* colonies (Paerl *et al.*, 1989; Zehr, 1995; Nausch, 1996; Dyhrman *et al.*, 2002). However, there have been few studies documenting the composition of bacteria inhabiting the *Trichodesmium* colonies. The large number of γ-proteobacterial transcripts, especially from *Alteromonas, Marinobacter* and *Pseudoalteromonas,* as well as from members of the Bacteroidetes suggests that *Trichodesmium* is colonized by fast-growing and opportunistic microorganisms, and composition of assemblages is different from free-living bacterioplankton communities (Fuhrman and Ouverney, 1998; Giovannoni and Rappe, 2000; Venter *et al.*, 2004; Rusch *et al.*, 2007). In a study of coral mucus additions to seawater (Allers *et al.*, 2008), a rapid increase in the abundance of *Alteromonas* bacteria concomitant with significant decreases in the C:N ratio of particulate organic matter was found, suggesting that they rapidly assimilated dissolved organic carbon. *Marinobacter* is the most common heterotrophic organism associated with libraries of the assimilatory nitrate uptake gene *nasB* in nitrate-rich waters (Allen *et al.*, 2001). *Trichodesmium* is known to exude fixed N and C into surrounding waters (Capone *et al.*, 1994; Mulholland *et al.*, 2006). The Bacteroidetes group includes those which are surface-associated in the marine environment and those which are capable of degrading macromolecules by hydrolytic enzyme production (Giovannoni and Rappe, 2000). Within the trichosphere, it is hypothesized that organic C and N concentrations are several fold higher than the surrounding seawater, which is reflected in higher heterotrophic activity and increased hydrolysis of several compounds within colonies relative to surrounding waters (Nausch, 1996). This environment may select for fast growing opportunistic taxa. Our results are consistent with a previous study using fluorescent *in situ* hybridization on *Trichodesmium*-associated bacteria, which found large numbers of Flavobacteria and enteric bacteria (presumably γ-proteobacteria) (Zehr, 1995). We also observed α-proteobacterial mRNAs, which is in contrast to observations of *Trichodesmium* colonies reported by Zehr, 1995.

Transcripts from cyanobacteria other than *Trichodesmium* included those matching *Synechococcus*, *Prochlorococcus*, *Crocosphaera*, *Nostoc* and *Lyngbya* proteins. *Prochlorococcus*, *Crocosphaera* and

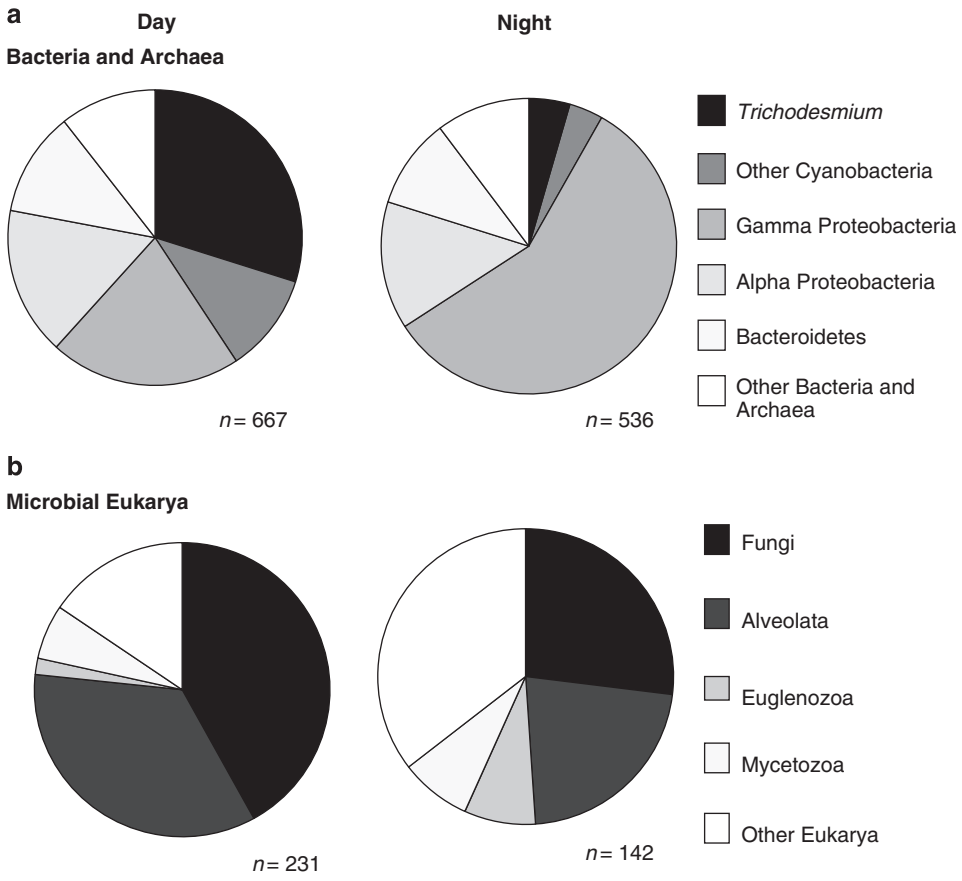

Figure 3 Taxonomic affiliation of (**a**) prokaryotic and (**b**) eukaryotic *Trichodesmium* colony-associated putative mRNAs as a percentage of each group. Taxonomic affiliation was assigned as the best BLASTx hit of mRNAs to the All Prokaryotic Proteins and All Eukaryotic Proteins at CAMERA, with E-values $<10^{-3}$. Where mRNAs matched both prokaryotic and eukaryotic proteins, the lowest E-value match was assigned.

Synechococcus transcripts may have been the contaminants from the surrounding seawater, which contained abundant picocyanobacteria (unpublished data) or cyanobacteria attached to the *Trichodesmium* colonies. Cyanobacteria resembling *Synechococcus* and *Phormidium* were observed microscopically within bloom samples (Figure 2), consistent with previous observations of these genera within *Trichodesmium* colonies (Siddiqui *et al.*, 1992; Sheridan *et al.*, 2002).

Our observations of microbial eukaryotic mRNA taxonomic affiliation are consistent with microscopic observations, which showed the presence of morphologically diverse eukaryotic flora among the *Trichodesmium* colonies (Figure 2), as well as previous studies demonstrating a diverse eukaryotic flora and fauna associated with *Trichodesmium* (Sheridan *et al.*, 2002). The large number of putative fungal transcripts in the mRNA libraries is consistent with previous observations of fungi in *Trichodesmium* colonies (Sheridan *et al.*, 2002). However, it may also reflect greater representation of fungal genomes in the eukaryotic microbial proteins database. Diatoms were also observed colonizing *Trichodesmium*, where it was speculated that the benthic-like environment allowed heavily silicified diatom taxa to exist (Sheridan *et al.*, 2002).

The phage and viral sequences recovered may have originated in either *Trichodesmium* or the associated organisms. The large number of cyanophage-like sequences in the day, when *Trichodesmium* comprised the greatest proportion of total mRNAs, suggests that some of the viral sequences may be associated with *Trichodesmium*. Previous reports of virus-like particles in dying and mitomycin C-treated *Trichodesmium* cells (Ohki, 1999), and the production of virus-like particles from filaments (Hewson *et al.*, 2004) suggest that the cyanobacterium may be affected by viruses in nature. However, no genomic information on potential *Trichodesmium* cyanophages is available. The viral genes detected in this study may be attractive targets for future study of virus–*Trichodesmium* interactions.

Transcript library gene orthologs

The detected transcripts of *Trichodesmium* show the key importance of nitrogen fixation and photosynthesis in the ecology of the microorganism under

Table 2 Top 10 most hit proteins to the non-redundant protein database in both the day and night samples, using the sequences not matching annotated proteins or rRNAs at E <0.001

Time	*Genbank reference*	*Organism*	*Protein*	*# Hits*	*Bit score range*	*E-value range*
Day	ref|XP_002210989.1|	*Branchiostoma floridae*	Hypothetical protein BRAFLDRAFT_73440	84	46.2–40	0.001–0.066
Day	ref|XP_002206263.1|	*Branchiostoma floridae*	Hypothetical protein BRAFLDRAFT_118873	71	46.2–40	0.001–0.065
Day	gb|ABN15002.1|	*Taenia asiatica*	Cytoplasmic antigen 1	49	45.4–40	0.002–0.066
Day	ref|XP_002137676.1|	*Drosophila pseudoobscura*	GA26398	34	61.6–40	2E-08–0.066
Day	ref|XP_002221622.1|	*Branchiostoma floridae*	Hypothetical protein BRAFLDRAFT_123626	18	75–40.4	1E-12–0.05
Day	gb|ACH48223.1|	*Ornithoctonus huwena*	Tumor differentially expressed protein	13	49.7–40.4	8E-05–0.05
Day	gb|AAL18703.1|	*Spirometra erinaceieuropaei*	Cytoplasmic antigen 1	9	48.5–40	2E-04–0.065
Day	ref|NP_032871.3|	*Mus musculus*	Phosphatidylinositol-4-phosphate 5-kinase type II-α	8	41.6–40.4	0.023–0.049
Day	ref|XP_001618287.1|	*Nematostella vectensis*	Hypothetical protein NEMVEDRAFT_v1g155059	7	56.2–40.8	9E-07–0.038
Day	ref|XP_001490517.2|	*Equus caballus*	PREDICTED: similar to E2F-associated phosphoprotein	6	56.6–42.7	7E-07–0.01
Night	ref|XP_002206263.1|	*Branchiostoma floridae*	Hypothetical protein BRAFLDRAFT_118873	159	49.3–40	1E-04–0.066
Night	ref|XP_002137676.1|	*Drosophila pseudoobscura pseudoobscura*	GA26398	82	68.2–40	2E-10–0.066
Night	ref|XP_002210989.1|	*Branchiostoma floridae*	Hypothetical protein BRAFLDRAFT_73440	57	44.7–40	0.003–0.066
Night	gb|ABN15002.1|	*Taenia asiatica*	Cytoplasmic antigen 1	54	45.4–40.4	0.002–0.049
Night	ref|NP_032871.3|	*Mus musculus*	Phosphatidylinositol-4-phosphate 5-kinase type II alpha	36	41.6–40	0.023–0.066
Night	gb|AAL18703.1|	*Spirometra erinaceieuropaei*	Cytoplasmic antigen 1	20	54.7–40	3E-06–0.065
Night	ref|XP_002221622.1|	*Branchiostoma floridae*	Hypothetical protein BRAFLDRAFT_123626	19	82.8–40.4	9E-15–0.05
Night	ref|XP_001618287.1|	*Nematostella vectensis*	Hypothetical protein NEMVEDRAFT_v1g155059	15	55.1–43.1	2E-06–0.029
Night	gb|ACH48223.1|	*Ornithoctonus huwena*	Tumor differentially expressed protein	11	55.5–40.4	2E-06–0.05
Night	gb|EEB19804.1|	*Pediculus humanus corporis*	Hypothetical protein Phum_PHUM590900	6	54.7–42	3E-06–0.017

Abbreviation: CAMERA, Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis.
A hit was considered any bit score ⩾40.

bloom conditions, whereas heterotrophic bacteria were primarily involved in growth and energy metabolism. Of the 198-day and 19-night *Trichodesmium* mRNAs (Figure 4), the most common orthologs were hypothetical proteins or proteins of unknown function (31 in the day and eight at night), transcripts related to the photosynthetic apparatus (28 in the day and two at night) or involved in nitrogenase enzyme activity (18 in the day and one at night; Figure 4). Prokaryotic genes, which did not match *Trichodesmium* proteins (Figure 4), were primarily hypothetical proteins or proteins of unknown function. Among the 449-day and 508-night transcripts, 101 during the day and 96 at night were conserved hypothetical proteins or proteins of unknown function, whereas 53 in the day and 83 at night were either ribosome components or genes involved in DNA replication and repair. The remaining mRNAs were mostly present as singletons in the libraries and involved in photosystem apparatus biosynthesis, oxidative phosphorylation, and transport, P acquisition, signal processing, and As detoxification.

Our results show that the dominant gene expression of *Trichodesmium* colonies is similar to free-living pelagic marine microorganisms (Frias-Lopez *et al.*, 2008; Hewson *et al.*, 2009; Poretsky *et al.*, 2009b), with heavy genetic machinery investment in energy metabolism and growth. *Trichodesmium* colony metatranscripts were mostly associated with ribosome synthesis, RNA polymerase and other replication and repair enzymes (14% of all metatranscripts). Transcripts of genes involved in energy metabolism, including photosynthesis, oxygenic phosphorylation and the citrate cycle also comprised a large component of total *Trichodesmium* colony metatranscripts (15%). These results show that energy acquisition and growth were critical to *Trichodesmium* and other closely associated organisms within the bloom.

Trichodesmium transcripts included those implicated in sulfur metabolism, including methionine

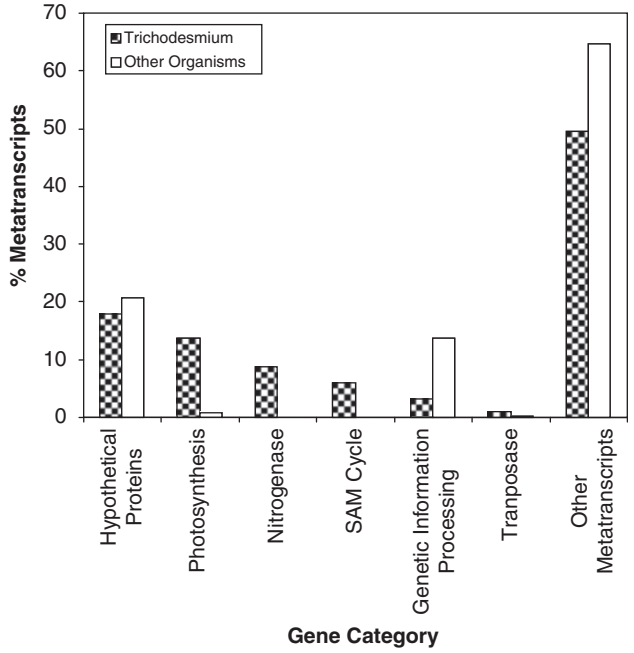

Figure 4 Comparison of the orthology of total number of metatranscripts between day and night assigned to *Trichodesmium* and prokaryotic organisms other than *Trichodesmium*.

synthase and adenosylhomocysteinase, both enzymes involved in the S-adenosyl methionine (SAM) cycle, used in transmethylation reactions in proteins, nucleic acids and lipids (Koshiishi *et al.*, 2001). *Trichodesmium* has extensively methylated adenine in its genomic DNA (15 mol%) (Zehr *et al.*, 1991). *Trichodesmium* transcripts also included the *arsA* component of the *arsA/B*-encoded arsenite efflux pump, and thus a potentially active pathway for arsenate detoxification in this bloom population (see below).

The large number of *Trichodesmium* transcripts associated with nitrogenase activity (8% of all transcrtipts), including *nifH*, *nifD* and *nifE* confirms previous studies demonstrating the high expression of nitrogenase relative to other genes of diazotrophs (Stoeckel *et al.*, 2008), and indicates that it is an active biosynthetic pathway. It is interesting to note that no *nif* genes were recovered from organisms other than *Trichodesmium*. Two transposases were detected among the 198 *Trichodesmium* metatranscripts. The function of transposases in open ocean cyanobacteria is unknown, however in other related diazotrophic cyanobacteria, tranposases are present in high genome copy number and recently have been observed as dominant transcripts *in situ (*Hewson *et al.*, 2009*)*. The large number of conserved hypothetical proteins or proteins of unknown function (31 in the day and eight at night) is consistent with observations in other metatranscriptomic surveys (Poretsky *et al.*, 2005, 2009b; Frias-Lopez *et al.*, 2008; Gilbert *et al.*, 2008; McGrath *et al.*, 2008), and shows a need for prioritized studies of gene function.

In contrast to *Trichodesmium* metatranscripts, undersampling of the mRNA pool resulted in primarily singleton or doubleton sequences in the library, with no dominant transcript pathway observed. Most gene transcripts were of unknown function or conserved hypothetical proteins. A dominant set of recognizable non-*Trichodesmium* transcripts were those for DNA repair and replication, suggesting that the populations of bacteria surrounding *Trichodesmium* colonies were actively growing. There were also a number of transcripts related to transport and signal processing (for example, *tonB*-dependent receptors), suggesting the dominance of these processes to the physiology of cells in this consortium or microniche.

Given the dearth of functional sequences that can be identified, it is striking that there are also a number of genes with known responsiveness to phosphorus deficiency. This sample was taken from an oligotrophic region of relatively low phosphorus and increased *Trichodesmium* derived nitrogen inputs, which may have increased phosphorus demand in the bloom sample. Consistent with this, there is a transcript encoding an alkaline phosphatase enzyme, which hydrolyzes phosphate from dissolved organic phosphorus, and there are several transcripts encoding exopolyphosphatases (*ppX*), an enzyme involved in the breakdown of polyphosphate stores. A recent proteome study in the low-phosphorus Sargasso Sea identified phosphate binding and transport, a dominant feature of the SAR11, *Prochlorococcus* and *Synechococcus* metaproteome (Sowell *et al.*, 2009). In this study, there are similarly a number of transcripts associated with the high-affinity uptake of phosphate, including *pstsS/sphX* and a possible *pstB* transcript in *Crocosphaera*. These genes are typically encoded in a co-transcribed cluster (*pstSCAB*) that is upregulated to cope with low phosphorus availability. The *Crocosphaera pstB* gene is in a cluster with *pstS* in the WH8501 genome, the later of which has been shown to be upregulated by P deficiency in culture studies (Dyhrman and Haley 2006). Taken together, the data suggest that this consortium express phosphorus-regulated transcripts to meet phosphorus demand through the breakdown of phosphorus stores, the hydrolysis of dissolved organic phosphorus, and the high-affinity uptake of phosphate.

In oligotrophic systems, arsenate concentrations in the ocean can be similar to that of phosphate, and it has widely been hypothesized that under these conditions, arsenate would be transported into marine microbial cells through high-affinity phosphate transport systems, thus necessitating the induction of arsenate detoxification strategies (Figure 5). The speciation of arsenate in these systems suggests microbial transformation and variability in these transformations coincident with *Trichodesmium* blooms (Cutter and Cutter, 2006), but the genes for these pathways have not been

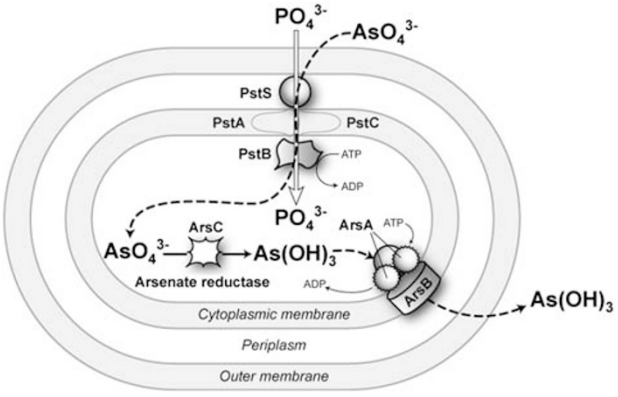

Figure 5 A schematic of a putative phosphate/arsenate uptake and detoxification pathway. Transcripts for the highlighted proteins, including the high-affinity binding (PstS) and uptake (PstB) of phosphate, the reduction of arsenate (ArsC) and an ArsAB arsenite efflux pump are present in the prokaryotic sequences from the *Trichodesmium* bloom samples.

examined in any detail in field populations. In addition to the transcripts for high-affinity phosphate transport described above, the metatranscriptome contains transcripts for arsenate reductase, and the *arsA* and *arsB* genes, which encode an arsenite efflux pump. These arsenate-related genes are typically induced by phosphorus deficiency to cope with the transport of arsenate through the *pstSCAB*-encoded high-affinity phosphate transport system. Although all of the transcripts in this pathway (Figure 5) are not encoded in the same taxa, their presence highlights the potential coupling of phosphorus stress and arsenate detoxification in the upper water column of oligotrophic systems and in association with *Trichodesmium* communities. The presence of *ars* and *pst* genes, like all genes in this study, among transcripts may also be influenced by the relative taxonomic composition of *Trichodesmium* colonies.

In contrast to prokaryotic genes, reads matching microbial eukaryotic genomes were well annotated, with only 26% of day and 16% of night sequences matching genes of unknown function. Photosynthesis-related genes among eukaryotic microbial genes represented a much larger portion of the day sample (20 of 199) than the night (one of 128). Other informative genes like cytochrome c oxidase 1 (*cox1*) and cytochrome b (*cob*) were also abundant in the day (18 of 199) and night (15 of 128) samples. *cox1* and *cob* are the mitochondrial genes that have been shown useful in DNA barcoding of fungi (Seifert *et al.*, 2007), microalgae (Robba *et al.*, 2006) and potentially unicellular algae (Evans *et al.*, 2007; Lin *et al.*, 2009).

Most mRNAs matching sequences in nr at E-values >0.001 but bit scores $\geqslant 40$ were proteins of unknown function (398 of 602 day and 513 of 719 night). Several genes had a large number of matches, comprising up to 22 % of all mRNAs in this category (Table 2). Cytoplasmic antigen 1 was the most expressed annotated protein in the nr database libraries, representing about 10% of the category in both day and night. The large transcript abundance of some highly represented organisms in this comparison, such as *Branchiostoma floridae* (Table 2), likely does not represent their actual abundance in the sample. It is possible that as the *Trichodesmium* bloom was raft-like (Figure 2a), it could provide shelter to *B. floridae* eggs or larvae. However, it is more likely that sequences most closely related to *B. floridae* are an artifact of disproportionately distributed eukaryotic genes in the non-redundant database (John *et al.*, 2009). Model organisms studied in molecular genetics (for example, *Xenopus laevis*, *Mus musculus* and *Drosophila sp.*) and organism with their whole genomes sequenced (for example, *B. floridae*; Putnam *et al.*, 2008) represent a large amount of Genbank's sequence coverage. Therefore, many of the sequences we annotated could represent different organisms where genes have not yet been characterized.

Unannotated sequence reads

The large number of putative mRNA sequence reads with no strong matches to microbial proteins suggests that the majority of putative mRNA in *Trichodesmium* colonies is made up from microorganisms for which genome data are not available, are transcripts from non-encoding intergenic regions (including small RNAs; Shi *et al.*, 2009), or that the length of the sequence reads (175–177 bp average, shorter than the average for all sequence reads) was too short to assign reads to proteins at our E-value cutoff. Our results are in line with recent reports of putative small RNAs that indicated that they can comprise of a substantial proportion of transcripts from marine plankton (Gilbert *et al.*, 2008; Shi *et al.*, 2009). We compared sequences not matching the cultivated prokaryotic, eukaryotic or viral proteins with the GOS assembled protein and nucleotide databases (Venter *et al.*, 2004; Rusch *et al.*, 2007) at CAMERA to determine whether they were present in other open ocean microbial populations. Of 4417-day and 4476-night reads not matching microbial, metazoan or viral proteins, 2083-day and 2762-night sequences matched assembled nucleotide sequences at $E < 10^{-3}$ in the GOS survey. Of these, 929-day and 1144-night reads matched unannotated proteins in the assembled proteins database from GOS. In addition 2216-day and 2838-night sequences matched GOS reads, of which most were from the Sargasso Sea, followed by those from the Indian Ocean and the Galapagos Islands. However, the mean similarity of metatranscripts was not the same among different regions (Figure 6). The least similarity between metatranscripts and GOS reads was for matches to the Galapagos Island

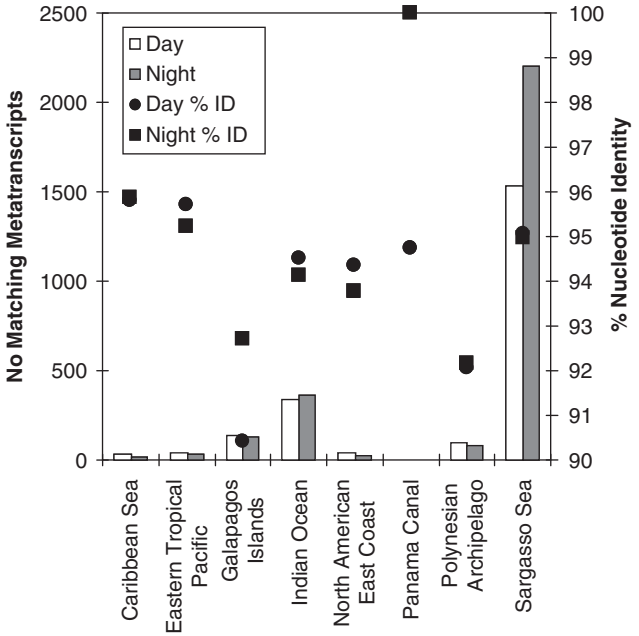

Figure 6 Number of matches to the Global Ocean Survey sequence reads at different locations (left axis) and mean similarity to reads (right axis).

and Polynesian Archipelago sequences. However at E-values <0.001, all matches were $>91\%$ similar at the nucleotide level to the GOS sequences. This observation is surprising as these regions are more likely similar to the habitat from which the *Trichodesmium* samples were retrieved. These results confirm that the majority of sequences from the *Trichodesmium* colonies match uncultivated microorganisms (the GOS samples were from the 0.1–0.8-μm size fraction), which use proteins that share little homology to cultivated microorganisms.

The large proportion of unrecognized mRNAs (that is, those not matching either eukaryotic microbial or prokaryotic proteins or the GOS dataset) may in part be explained by the presence of genes from metazoa inadvertently collected along with *Trichodesmium* colonies. Microscopic examination of *Trichodesmium* colonies from this station showed the presence of several metazoa (Figure 2), which is not surprising as the sample was collected at night when there are more zooplankton in surface waters. *Trichodesmium* is also colonized by several genera of copepods, which utilize the colonies as pseudobenthic substrates for laying eggs and for food (O'Neil and Roman, 1994; O'Neil *et al.*, 1996). No complete marine zooplankton genomes are publicly available against which the metatranscriptomes could be compared. However, we compared the unannotated metatranscripts with the non-redundant protein database at NCBI, which showed 123-day and 178-night mRNAs having homology to metazoan proteins, including those from arthropods and echinoderms. It is possible that the remaining matches with no homology to sequenced metazoa may be proteins from organisms for which no close relatives are available, or to eukaryotic organisms with no fully sequenced genomes (for example, radiolarians or dinoflagellates).

Comparison of metatranscripts to metatranscriptomes in other studies

Despite having similar dominant transcript orthologs, the overall profile of gene expression of the *Trichodesmium* colonies was different from free-living pelagic communities (Frias-Lopez *et al.*, 2008; Hewson *et al.*, 2009; Poretsky *et al.*, 2009b; M Vila-Costa and MA Moran, unpublished data) and coastal seawater (Gilbert *et al.*, 2008; R Poretsky and MA Moran, unpublished data) investigated elsewhere (Figure 7). The difference between free-living and *Trichodesmium* transcript inventories was mostly driven by higher frequency of ribosome synthesis and oxidative phosphorylation genes, but fewer nitrogen metabolism, propanoate metabolism, porphyrin and chlorophyll metabolism, and pyruvate metabolism genes. As previous coastal seawater (Poretsky *et al.*, 2005), eukaryotic (John *et al.*, 2009) and soil (McGrath *et al.*, 2008) metatranscriptomes had limited sequence depth (400, 232 and 48 sequences, respectively), direct comparison with our metatranscriptomes on the basis of gene orthology is not possible. However, the metatranscriptomes in these studies generally had dominant fractions of genes involved in central metabolism, protein synthesis, and transport and binding proteins, and were therefore distinct from the *Trichodesmium* colony gene expression profiles in this study.

Conclusions

Our data are the first on the composition of a transcriptionally active community of a pelagic microniche and suggest that *Trichodesmium* colonies represent microhabitats for diverse co-occurring heterotrophic bacteria, eukaryotes and phage. Although overall types of genes that are expressed as a dominant fraction of total mRNA inventories within colonies reflects those of free-living microorganisms, the distinct pattern of gene expression suggests that the habitat within colonies leads to different metabolic processes. For example, *Trichodesmium* transcripts related to P stress (and consequently As detoxification) response indicate that local enrichment of N leads to local P limitation. Therefore, the pervasive observation of variable and diverse populations residing on the *Trichodesmium* colonies with different metabolism from organisms in pelagic habitats suggests that the relevant scale of processes such as elemental cycling is intimately associated with colonies in the phycosphere or 'trichosphere' of the cyanobacterium. Finally, our

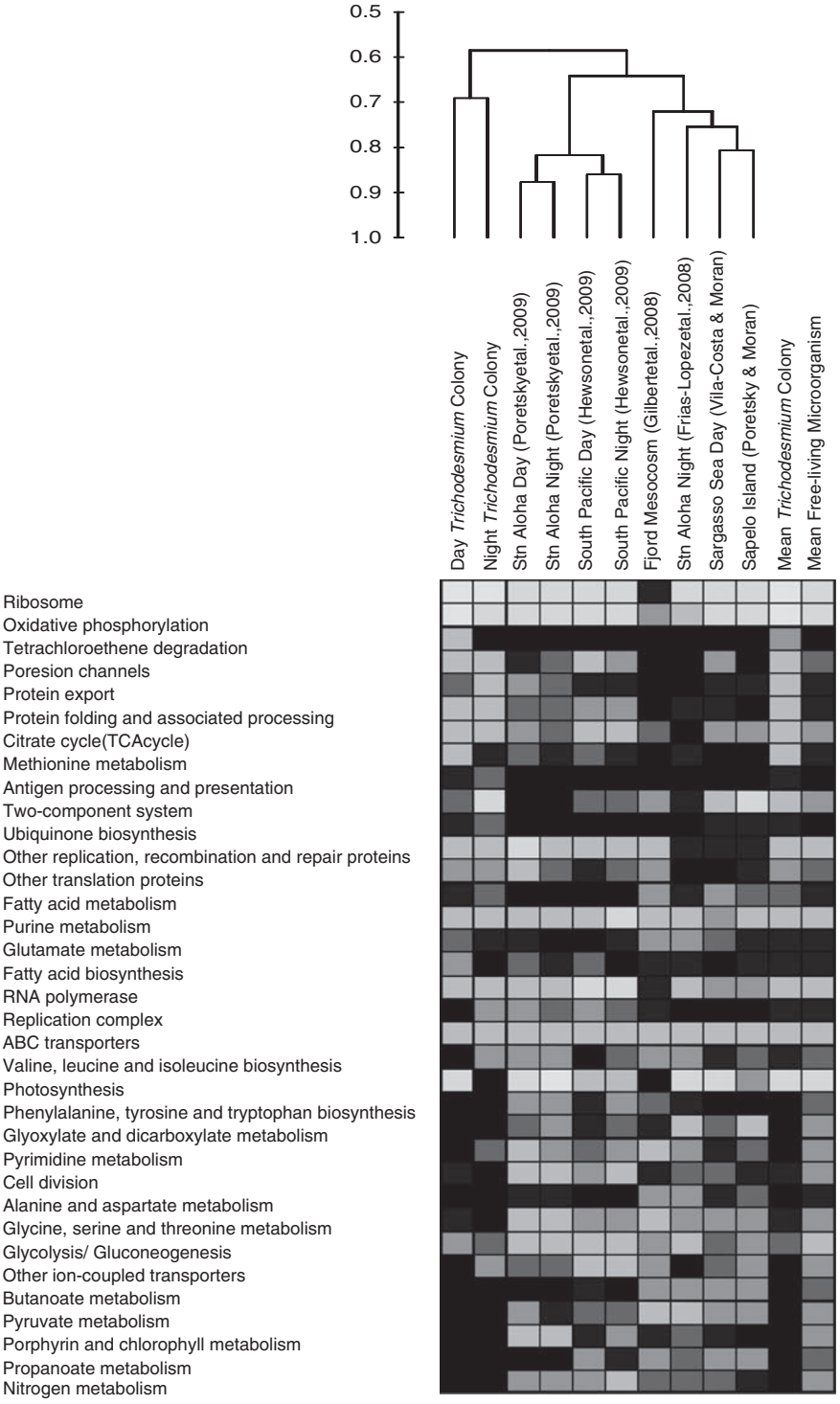

Figure 7 Heat map of dominant (>1%) transcript orthologs in *Trichodesmium* colony and free-living communities reported elsewhere. Columns in the heat map are clustered by Whittaker Index of similarity and unweighted pair-group-mean average (UPGMA) based on the frequency of transcript associated with each KEGG pathway. The total number of orthologs compared was different between samples ($n=508$ *Trichodesmium* colony day, $n=417$ *Trichodesmium* colony night, $n=14\,905$ Stn ALOHA day (Poretsky *et al.*, 2009a, b), $n=13\,841$ Stn ALOHA night (Poretsky *et al.*, 2009a, b), $n=1725$ South Pacific day (Hewson *et al.*, 2009), $n=1192$ South Pacific night (Hewson *et al.*, 2009), $n=4145$ Stn ALOHA night (Frias-Lopez *et al.*, 2008), $n=29\,144$ Fjord Mesocosm (Gilbert *et al.*, 2008), $n=45\,461$ Sargasso Sea day (Vila-Costa and Moran, unpublished data) and $n=21\,788$ Sapelo Island day (Poretsky and Moran, unpublished data). For the Sargasso Sea and Sapelo Island samples, which are as yet unpublished, 'control' samples were used for comparison.

observations show that although a large number of transcripts do not share homology with known proteins of sequenced microorganisms, they share similarity with genome fragment inventories of open ocean communities awaiting annotation of further sequencing of representative genomes.

Acknowledgements

We thank E Orchard, J Landrum, R Horak, B Van Gessel, R Beinart, R Paerl, K London, P Moisander, I Biegala and the crew of the R/V Kilo Moana for field assistance; S Schuster, J Qi and L Tomsho for sequencing at Pennsylvania State University; B Carter, T Shi, S Bench, S Sun, P Gilna and the Community Cyberinfrastructure for Advance Marine Microbial Ecology Research and Analysis (CAMERA) for bioinformatics support; A Yocum, M Hogan and E Mondragon for laboratory assistance; and JH Paul, D John and MA Moran for comments on a draft of the paper. This work was supported by the Gordon and Betty Moore Foundation Marine Microbiology Initiative; NSF Grant EF0424599 for the Center for Microbial Oceanography: Research and Education (CMORE); and NSF Grant OCE0425583 awarded to JPM and JPZ.

References

Allen AE, Booth MG, Frischer ME, Verity PG, Zehr JP, Zani S. (2001). Diversity and detection of nitrate assimilation genes in marine bacteria. *Appl Environ Microbiol* **67**: 5343–5348.

Allers E, Niesner C, Wild C, Pernthaler J. (2008). Microbes enriched in seawater after addition of coral mucus. *Appl Environ Microbiol* **74**: 3274–3278.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl Acid Res* **25**: 3389–3402.

Bowman TE, Lancaster LJ. (1967). Bloom of the planktonic blue-green alga *Trichodesmium erythraeum* in the Tonga Islands. *Limnol Oceanogr* **10**: 291–293.

Capone DG. (2000). The marine microbial nitrogen cycle. In: Kirchman DL, (eds). *Microbial Ecology of The Oceans*. Wiley-Liss: New York. pp 455–493.

Capone DG, Carpenter EJ. (1999). Nitrogen fixation by marine cyanobacteria: Historical and global perspectives. *Bull Inst Oceanogr Monaco* **19**: 235–256.

Capone DG, Ferrier MD, Carpenter EJ. (1994). Amino acid cycling in colonies of the planktonic marine cyanobacterium *Trichodesmium thiebautii*. *Appl Environ Microbiol* **60**: 3989–3995.

Capone DG, Zehr JP, Paerl HW, Bergman B, Carpenter EJ. (1997). *Trichodesmium*, a globally significant marine cyanobacterium. *Science* **276**: 1221–1229.

Carpenter EJ. (1983). Estimate of global marine nitrogen fixation by Oscillatoria (Trichodesmium). In: Carpenter EJ, Capone DG (eds). *Nitrogen in the Marine Environment*. Academic Press: New York. p 920.

Carpenter EJ, Romans K. (1991). Major role of the cyanobacterium *Trichodesmium* in nutrient cycling in the North Atlantic Ocean. *Science* **254**: 1356–1358.

Codispoti LA, Brandes JA, Christensen JP, Devol AH, Naqvi SWA, Paerl HW *et al.* (2001). The oceanic fixed nitorgen and nitrous oxide budgets: moving targets as we enter the anthropocene. *Scientia Marina* **65**: 85–105.

Cole J, Chai B, Farris R, Wang Q, Kulam-Syed-Mohideen A, McGarrell D *et al.* (2007). The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acid Res* **35**: D169–D172.

Culley AI, Lang AS, Suttle CA. (2006). Metagenomic analysis of coastal RNA virus communities. *Science.* **312**: 1795–1798.

Cutter GA, Cutter LS. (2006). Biogeochemistry of arsenic and antimony in the North Pacific Ocean. Geochemistry Geophysics Geosystems **7**: Q05M08 (doi:10.1029/2005GC001159).

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al.* (2008). Functional metagenomic profiling of nine biomes. *Nature* **452**: 629–632.

Dugdale R, Menzel D, Ryther J. (1961). Nitrogen fixation in the Sargasso Sea. *Deep Sea Res* **7**: 297–300.

Dyhrman S, Haley S. (2006). Phosphorus scavenging in the unicellular marine diazotroph *Crocosphaera watsonii*. *Appl Environ Microbiol* **72**: 1452–1458.

Dyhrman S, Webb E, Anderson D, Moffett J, Waterbury J. (2002). Cell-specific detection of phosphorus stress in *Trichodesmium* from the western north Atlantic. *Limnol Oceanogr* **47**: 1832–1836.

Dyhrman ST, Chappell PD, Haley ST, Moffett JW, Orchard ED, Waterbury JB *et al.* (2006). Phosphonate utilization by the globally important marine diazotroph trichodesmium. *Nature* **439**: 68–71.

El-Shehawy R, Lugomela C, Ernst A, Bergman B. (2003). Diurnal expression of *hetR* and diazocyte development in the filamentous non-heterocystous cyanobacterium *Trichodesmium erythraeum*. *Microbiol-SGM* **149**: 1139–1146.

Evans KM, Wortley AH, Mann DG. (2007). An assessment of potential diatom 'barcode' genes (*cox1*, *rbcL*, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist* **158**: 349–364.

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 Nat Acad Sci USA* **105**: 3805–3810.

Fuhrman JA, Ouverney CC. (1998). Marine microbial diversity studied via 16S rRNA sequences: cloning results from coastal waters and counting of native archaea with fluorescent single cell probes. *Aquat Ecol* **32**: 3–15.

Gilbert JA, Field D, Huang Y, Edwards R, Li WKW, Gilna P *et al.* (2008). Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PloS ONE* **3**: e3042.

Giovannoni SJ, Rappe M. (2000). Evolution, diversity and molecular ecology of marine prokaryotes. In: Kirchman DL (eds). *Microbial Ecology of The Oceans*. Wiley-Liss: New York. pp 47–84.

Hewson I, Govil SR, Capone DG, Carpenter EJ, Fuhrman JA. (2004). Evidence for *Trichodesmium* viral lysis and potential significance for biogeochemical cycling in the oligotrophic ocean. *Aquat Microb Ecol* **36**: 1–8.

Hewson I, Poretsky R, Beinart R, 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.

John D, Zielinksi B, Paul J. (2009). Creation of a pilot metatranscriptome library from eukaryotic plankton of a eutrophic bay (Tampa Bay, Florida). *Limnol Oceanogr Methods* **7**: 249–259.

Karl D, Letelier R, Tupas L, Dore J, Christian J, Hebel D. (1997). The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature* **388**: 533–538.

Koshiishi C, Kato A, Yama S, Crozier A, Ashihara H. (2001). A new caffeine biosynthetic pathway in tea leaves: utilisation of adenosine released from the S-adenosyl-L-methionine cycle. *FEBS Lett* **499**: 50–54.

Kramer J, Wyman M, Zehr J, Capone D. (1996). Diel variability in transcription of the structural gene for glutamine synthetase (glnA) in natural populations of the marine diazotrophic cyanobacterium *Trichodesmium thiebautii*. *FEMS Microbiol Ecol* **21**: 187–196.

Lin S, Zhang H, Hou Y, Zhuang Y, Miranda L. (2009). High-level diversity of dinoflagellates in the natural environment, revealed by assessment of mitochondrial *cox1* and *cob* genes for dinoflagellate DNA barbcoding. *Appl Environ Microbiol* **75**: 1279–1290.

Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–380.

McGrath KC, Thomas-Hall SR, Cheng CT, Leo L, Alexa A, Schmidt S *et al.* (2008). Isolation and analysis of mRNA from environmental microbial communities. *J Microbiol Methods* **75**: 172–176.

Michaels A, Olson D, Sarmiento J, Ammerman J, Fanning K, Jahnke R *et al.* (1996). Inputs, losses and transformations of nitrogen and phosphorus in the pelagic North Atlantic Ocean. *Biogeochemistry* **35**: 181–226.

Mulholland M, Bernhardt P, Heil C, Bronk D, O'Neil J. (2006). Nitrogen fixation and release of fixed nitrogen by *Trichodesmium* spp. in the Gulf of Mexico. *Limnol Oceanogr* **51**: 1762–1776.

Nausch M. (1996). Microbial activities on *Trichodesmium* colonies. *Mar Ecol Prog Ser* **141**: 173–181.

Noble RT, Fuhrman JA. (1998). Use of SYBR Green I rapid epifluoresence counts of marine viruses and bacteria. *Aquat Microb Ecol* **14**: 113–118.

O'Neil JM, Metzler PM, Glibert PM. (1996). Ingestion of $^{15}N_2$-labelled *Trichodesmium* spp. and ammonia regeneration by the harpactacoid copepod *Macrosetella gracilis*. *Mar Biol* **125**: 89–96.

O'Neil JM, Roman MR. (1994). Ingestion of the cyanobacterium *Trichodesmium* spp. by pelagic harpacticoid copepods *Macrosetella, Miracia* and *Oculosetella*. *Hydrobiol* **292/293**: 235–240.

Ohki K. (1999). A possible role of temperate phage in the regulation of *Trichodesmium* biomass. *Bull Inst Oceanogr Monaco* **19**: 287–292.

Orcutt KM, Lipschultz F, Gundersen K, Arimoto R, Michaels AF, Knap AH *et al.* (2001). A seasonal study of the significance of N2 fixation by *Trichodesmium* spp. at the Bermuda Atlantic Time-Series Study (BATS) site. *Deep Sea Res II* **48**: 1583–1608.

Paerl H, Bebout B, Prufert L. (1989). Bacterial associations with marine *Oscillatoria* sp (*Trichodesmium* sp) populations—ecophysiological implications. *J Phycol* **25**: 773–784.

Patel A, Noble R, Steele J, Schwalbach M, Hewson I, Fuhrman J. (2007). Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nat Protoc* **2**: 269–276.

Poretsky R, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M *et al.* (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* **71**: 4121–4126.

Poretsky R, Gifford S, Rinta-Kanto J, Vila-Costa M, Moran MA. (2009a). Analyzing gene expression from marine microbial communities using environmental transcriptomics. *J Visualized Exper* **24**: http://www.jove.com/index/Details.stp?ID=108 (doi:10.3791/1086).

Poretsky R, Hewson I, Sun S, Allen AE, Moran MA, Zehr JP. (2009b). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* (doi: 10.1111/j.1462-2920.2008.01863.x).

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T *et al.* (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064–1071.

Robba L, Russell SJ, Barker GL, Brodie J. (2006). Assessing the use of the mitochondrial cox1 marker for use in DNA barcoding of red algae (Rhodophyta). *Am J Bot* **93**: 1101–1108.

Rusch D, Halpern A, Sutton G, Heidelberg K, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: 398–431.

Seifert KA, Samson RA, Dewaard JR, Houbraken J, Levesque CA, Moncalvo JM *et al.* (2007). Prospects for fungus identification using C01 DNA barcodes, with *Penicillium* as a test case. *Proc Nat Acad Sci USA* **104**: 3901–3906.

Sheridan CC, Steinberg DK, Kling GW. (2002). The microbial and metazoan community associated with colonies of *Trichodesmium* spp.:a quantitative survey. *J Plankton Res* **24**: 913–922.

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

Siddiqui P, Bergman B, Carpenter E. (1992). Filamentous cyanobacterial associates of the marine planktonic cyanobacterium. *Trichiodesmium Phycologia* **31**: 326–337.

Sowell SM, Wilhelm LJ, Norbeck AD, Lipton M, Nicora CD, Barofsky DF *et al.* (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J* **3**: 93–105.

Stoeckel J, Welsh E, Liberton M, Kunnvakkam R, Aurora R, Pakrasi H. (2008). Global transcriptomic analysis of *Cyanothece* 51142 reveals robust diurnal oscillation of central metabolic processes. *Proc Nat Acad Sci USA* **105**: 6156–6161.

Subramaniam A, Carpenter EJ, Falkowski PG. (1999). Bio-optical properties of the marine diazotorphic cyanobacteria *Trichodesmium* spp. II. A reflectance model for remote sensing. *Limnol Oceanogr* **44**: 618–627.

Thurber RLV, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C *et al.* (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proc Nat Acad Sci USA* **105**: 18413–18418.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

Villareal TA. (1990). Laboratory culture and preliminary characterisation of the nitrogen-fixing *Rhizosolenia—Richelia* symbiosis. *Mar Ecol* **11**: 117–132.

Webb E, Moffett J, Waterbury J. (2001). Iron stress in open-ocean cyanobacteria (*Synechococcus*, *Trichodesmium*, and *Crocosphaera* spp.): Identification of the *IdiA* protein. *Appl Environ Microbiol* **67**: 5444–5452.

Westberry T, Siegel D. (2006). Spatial and temporal distribution of *Trichodesmium* blooms in the world's oceans. *Global Biogeochem Cycles* **20**.

Wommack KE, Bhavsar J, Ravel J. (2008). Metagenomics: Read length matters. *Appl Environ Microbiol* **74**: 1453–1463.

Zehr J, Ohki K, Fujita Y, Landry D. (1991). Unique modification of adenine in genomic DNA of the marine cyanobacterium *Trichodesmium* sp. strain NIBB-1067. *J Bacteriol* **173**: 7059–7062.

Zehr JP. (1995). Marine nitrogen fixation: Why only *Trichodesmium*? In: Joint I (ed). *Molecular Ecology of Aquatic Microbes*. Springer-Verlag: Berlin. pp 335–364.

Zehr JP, Jenkins BD, Short SM, Steward G. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ Microbiol* **5**: 539–554.

Zehr JP, McReynolds LA. (1989). Use of degenerate oligonucleotides for amplification of the *nifH* gene from the marine cyanobacterium *Trichodesmium thiebautii*. *Appl Environ Microbiol* **55**: 2522–2526.