The ISME Journal (2009) 3, 618–631
© 2009 International Society for Microbial Ecology All rights reserved 1751-7362/09 $32.00
www.nature.com/ismej

ORIGINAL ARTICLE

In situ transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*

Ian Hewson[1], Rachel S Poretsky[2], Roxanne A Beinart[1,5], Angelicque E White[3], Tuo Shi[1,6], Shellie R Bench[1], Pia H Moisander[1], Ryan W Paerl[1], H James Tripp[1], Joseph P Montoya[4], Mary Ann Moran[2] 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]*College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA and* [4]*Georgia Institute of Technology, School of Biology, Atlanta, GA, USA*

The diazotrophic cyanobacterium *Crocosphaera watsonii* supplies fixed nitrogen (N) to N-depleted surface waters of the tropical oceans, but the factors that determine its distribution and contribution to global N_2 fixation are not well constrained for natural populations. Despite the heterogeneity of the marine environment, the genome of *C. watsonii* is highly conserved in nucleotide sequence in contrast to sympatric planktonic cyanobacteria. We applied a whole assemblage shotgun transcript sequencing approach to samples collected from a bloom of *C. watsonii* observed in the South Pacific to understand the genomic mechanisms that may lead to high population densities. We obtained 999 *C. watsonii* transcript reads from two metatranscriptomes prepared from mixed assemblage RNA collected in the day and at night. The *C. watsonii* population had unexpectedly high transcription of hypothetical protein genes (31% of protein-encoding genes) and transposases (12%). Furthermore, genes were expressed that are necessary for living in the oligotrophic ocean, including the nitrogenase cluster and the iron-stress-induced protein A (*isiA*) that functions to protect photosystem I from high-light-induced damage. *C. watsonii* transcripts retrieved from metatranscriptomes at other locations in the southwest Pacific Ocean, station ALOHA and the equatorial Atlantic Ocean were similar in composition to those recovered in the enriched population. Quantitative PCR and quantitative reverse transcriptase PCR were used to confirm the high expression of these genes within the bloom, but transcription patterns varied at shallower and deeper horizons. These data represent the first transcript study of a rare individual microorganism *in situ* and provide insight into the mechanisms of genome diversification and the ecophysiology of natural populations of keystone organisms that are important in global nitrogen cycling.

The ISME Journal (2009) **3,** 618–631; doi:10.1038/ismej.2009.8; published online 19 February 2009
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: metatranscriptome; *Crocosphaera*; gene expression; transposase; diazotroph; oligotrophic

Introduction

Biological nitrogen (N_2) fixation is a major source of nitrogen in the oligotrophic oceans (Capone, 2000; Carpenter, 1983; Karl *et al.*, 2002). N_2 fixation by unicellular cyanobacteria, including *Crocosphaera watsonii*, can account for a large fraction (>50%) of total N_2 fixation in oligotrophic waters (Zehr *et al.*, 2001; Montoya *et al.*, 2004; Zehr *et al.*, 2007b). These microorganisms are keystone taxa in oligotrophic surface waters, because they are typically outnumbered by co-occurring nondiazotrophic microorganisms by 10^3- to 10^5-fold, yet support the nitrogen demand of nitrogen-starved microbial food webs (Zehr *et al.*, 2007b). *C. watsonii* is widely distributed in tropical surface waters of the northwestern Atlantic and Pacific oceans, where it is typically present at abundances $<10^3$ cells per ml (Church *et al.*, 2005; Hewson *et al.*, 2007; Zehr *et al.*, 2007b). Although blooms of large colonial (for example, *Trichodesmium*) (Capone *et al.*, 1997; Carpenter, 1983) and symbiotic (Carpenter *et al.*,

Correspondence. Current address: I Hewson, Department of Microbiology, Cornell University, Wing Hall 403, Ithaca, NY 14853, USA.
E-mail: hewson@cornell.edu
[5]Current address: Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.
[6]Current address: Algenol Biofuels, 1700 Union Ave, Baltimore, MD 21211, USA.

Received 27 October 2008; revised 12 January 2009; accepted 12 January 2009; published online 19 February 2009

1999) diazotrophs have been reported in tropical waters worldwide (Capone *et al.*, 1997), unicellular cyanobacteria are less easily observed (Zehr *et al.*, 2008), so little is known about the variability in abundance of these microorganisms.

Studies of *C. watsonii* gene expression in natural populations have been limited to a handful of proteins implicated in iron stress (iron-deficiency-induced protein A, *idiA*) (Webb *et al.*, 2001) and inorganic nutrient acquisition (Dyhrman and Haley, 2006; Zehr and Turner, 2001). Whole genome transcript profiling by RNA oligonucleotide microarrays (de Saizieu *et al.*, 1998) allows observation of genome-wide gene transcription patterns in cultivated microorganisms; however, it may be unsuitable for natural mixed-taxon assemblages.

Random community transcript (mRNA) sequencing (hereafter referred to as metatranscriptomics) offers promise to capture snapshots of active transcripts under *in situ* conditions. Recent applications of metatranscriptomics revealed differences in community gene expression in response to tidal and diel cycles (Frias-Lopez *et al.*, 2008; Poretsky *et al.*, 2005; Poretsky *et al.*, in press) and in response to elevated CO_2 concentrations (Gilbert *et al.*, 2008). These studies revealed a large component of sequences that are of unknown origin (that is, no hits to gene databases), including possibly intergenic spacer-encoded riboswitches and termination factors (Frias-Lopez *et al.*, 2008; Gilbert *et al.*, 2008; Poretsky *et al.*, 2005; Poretsky *et al.*, in press). Because sequencing is random, most sequences obtained in metagenomic and metatranscriptomic studies are from abundant taxa, and rarer organisms, such as *C. watsonii*, may be absent altogether (Johnston *et al.*, 2005). In most metagenomic sequencing efforts (Venter *et al.*, 2004), microdiversity of dominant prokaryotic groups (Brown and Fuhrman, 2005; Garcia-Martinez and Rodriguez-Valera, 2000) complicates assembly and taxonomic identification of short genomic fragments. *C. watsonii* has an unusually high degree of genome conservation, with similar nucleotide composition between strains isolated two decades ago and bacterial artificial chromosome-cloned genomic fragments obtained from station ALOHA in 2005 (Zehr *et al.*, 2007a). Because *C. watsonii* has a high degree of genome conservation, metagenomic and metatranscriptomic studies are relatively simple as sequence reads identical to the *C. watsonii* WH8501 genome can be isolated even among complex read libraries.

The objective of this study was to analyze the *in situ* gene expression to understand the ecophysiology of the N_2-fixing unicellular cyanobacterium *C. watsonii*. We applied metatranscriptomics to capture snapshots of microbial community expressed gene transcripts under *in situ* conditions at a sampling site in the South Pacific that had high *C. watsonii* abundances. We identified several *C. watsonii* transcripts that were present in large numbers in the metatranscriptome, and developed quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) assays to verify the *in situ* transcription pattern of these highly expressed genes.

Methods

Sampling location and station characteristics

Samples were collected on board the R/V Kilo Moana at station KM0703.025 (14°59.9′S, 175°0′E) on 13 April 2007. Additional samples were obtained at seven stations in the southwest Pacific and equatorial Atlantic oceans in June 2006 and April–May 2007 (Table 1). Station KM0703.025 seawater samples were collected from two CTD casts. The first cast, which sampled depths at 5, 30, 45, 75 and 106 m (deep chlorophyll maximum depth), was completed at 1246 hours (local time). The second cast to sample a subsurface bloom of *C. watsonii* with peak abundance at 37 m occurred at 1425 hours. Sunrise occurred at 0525 hours and sunset at 1838 hours. Sampling times for other stations are indicated in Table 1.

Sample collection and nucleic acid recovery

Station KM0703.025 seawater samples were collected using 8 × 10 l Niskin bottles that were triggered at the sampling depth and returned to the research vessel. For the first cast (deepest sample collected at 1226 hours and CTD retrieved at 1246 hours), small samples (3–4 l) for analysis of gene and transcript

Table 1 Locations and sampling dates and times for samples collected at stations away from the *Crocosphaera watsonii* bloom

Station	*Latitude*	*Longitude*	*Date*	*Daytime*	*Nighttime*	*Day*	*Night*	*Day library*	*Night library*
				(hours)	*(hours)*	*vol (l)*	*vol (l)*	*size (reads)*	*size (reads)*
SJ0609.03	12°17.1′N	56°7.5′W	27 June 2006	0830	NA	110	NA	98 359	NA
SJ0609.07	11°45.8′N	46°10.3′W	1 July 2006	1300	0120	20	110	111 638	114 431
SJ0609.09	12°0.4′N	40°6.8′W	4 July 2006	1330	0000	20	120	109 761	96 685
SJ0609.12	15°2.3′N	31°57.4′W	7 July 2006	NA	2030	NA	20	NA	113 423
KM0703.04	15°3.3′S	155°1.1′E	18 March 2007	1500	2230	120	40	102 227	114 751
KM0703.10	30°0′S	160°0′E	24 March 2007	1530	2300	120	40	30 771	70 770
KM0703.24	15°0′S	178°45′E	12 April 2007	1545	0300	[a]	[a]	98 096	107 916

The number of reads for day and night metatranscriptomes is also indicated.
[a] Sample was net-collected *Trichodesmium*.

abundance over a vertical profile were taken. Four samples were collected for daytime expression from each depth (5, 30, 45, 75 and 106 m) into acid-washed and seawater-rinsed 4.5 l polycarbonate bottles and processed by 1320 hours. Four additional 4.5 l bottles were filled with seawater from each depth for nighttime expression. The night expression bottles were placed immediately into an on-deck flow-through incubator where they remained at ambient temperature before sample processing at 0025 hours on 14 April. Water for DNA and RNA samples was filtered using peristaltic-pump-driven positive-pressure filtration serially through 10 μm polycarbonate (GE Osmonics, Trevose, PA, USA) and 0.2 μm Supor (Pall Gelman, Ann Arbor, MI, USA) filters. For RNA, 3 l were filtered, and the filters were placed immediately into microcentrifuge tubes containing 100 μl glass beads and 350 μl RLT buffer (RNEasy kit, Qiagen, Valencia, CA, USA). Samples for DNA (4 l) were collected from all depths, filtered per the RNA samples and placed into microcentrifuge tubes containing 200 μl TE buffer. The filters were then frozen in liquid nitrogen for transport to the laboratory at the University of California Santa Cruz.

For the second cast (samples collected at 1420 hours and CTD retrieved at 1425 hours; 37 m), water samples for metatranscriptome sampling were transferred from the Niskin bottles into acid-washed and seawater-rinsed 60 l polypropylene and 40 l polycarbonate carboys. The 40 l carboy was immediately transferred to an on-deck flow-through incubator for nighttime expression analysis. Water in the 60 l carboy was serially filtered immediately by air-driven positive-pressure filtration (<2 kPa) through a 142-mm-diameter 5 μm pore-size filter (Isopore; Millipore, Billerica, MA, USA) then through a 0.2 μm pore-size Durapore (Millipore) filter. The filters were housed in 142-mm-diameter polycarbonate filter holders (Geotech Environmental, Denver, CO, USA) and kept cool during the filtration, which took approximately 30 min. The incubated 40 l carboy sample was processed starting at 2145 hours in the same manner as the day sample. During filtration, the bottle and filter holders were kept in darkness. Filtration ended at 2205 hours. After filtration the filters were placed into 50 ml conical tubes (BD Falcon, Franklin Lakes, NJ, USA) and frozen immediately in liquid nitrogen for transport to the laboratory. Samples from the southwest Pacific and equatorial Atlantic oceans were processed similarly. Smaller samples (1.8 l for RNA and 4 l for DNA) were also collected during the 1403-hour cast and filtered for DNA and RNA as described above for the 1215-hour cast.

Particle concentrations and dissolved organic carbon and nitrogen measurements

Particle concentrations were inferred from particulate backscatter measurements at 650 nm (collected with a WET Labs Inc. (Philomath, OR, USA) hyperspectral absorbance and attenuation meter (ac-s)). All profiles were corrected for absorbance and attenuation of 0.2 μm filtered seawater profiles and $b_{bp}650$ was calculated as the difference of c_p650 and a_p650 as per Jonasz and Fournier (2007). Dissolved organic carbon and nitrogen samples collected from standard depths at each station were measured by high-temperature combustion methods described by Carlson *et al.* (1998).

Microscopy of Crocosphaera*-like cells*

Samples for microscopy of *Crocosphaera*-like cells were taken on both casts at the same depths as small-volume RNA samples. Two seawater-rinsed 50 ml conical tubes were filled directly from Niskin bottles, and preserved immediately with 2% (final concentration) 0.02-μm-filtered formaldehyde. The samples were kept at 4 °C before processing. Subsamples (40 ml each depth) were filtered through black 1.0 μm pore-size polycarbonate filters (Whatman) and mounted sample-side up on glass slides. The filters were then mounted with a drop of immersion oil and covered with a coverslip. Duplicate filters were prepared from independent samples at each depth. The slides were kept at −20 °C before microscopic examination.

Slides were examined under ×40 magnification using green light excitation with a Zeiss Axioplan epifluorescence microscope. *Crocosphaera*-like cells were identified as unicellular cyanobacteria between 2 and 5 μm in diameter and by phycoerythrin fluorescence. Two hundred cells were counted in 20–40 fields (depending on cell density) on each replicate slide.

Microscopy of virus-like particles and bacteria

Samples (50 ml) of seawater from 0, 37, 75 and 106 m were collected in sterile centrifuge tubes and immediately amended with 2%, 0.02-μm-filtered formalin. The samples were processed immediately for SYBR Green I microscopy following established protocols (Noble and Fuhrman, 1998; Patel *et al.*, 2007), using 1–3 ml of preserved sample in the slide preparation. Slides were observed using epifluorescence microscopy under blue excitation at ×100 magnification. More than 200 total virus-like particles and bacteria were counted in 20 microscopic fields for each slide.

DNA extraction

DNA was extracted from filters using a protocol modified from the DNEasy plant kit (Qiagen). DNA filters in 200 μl TE buffer were first submerged in a dry ice–ethanol bath to freeze. The tubes were then amended with buffer AP1 (Qiagen) and replaced in the dry ice–ethanol bath for 1 min before transferring to a heat block at 70 °C for 2 min. This freeze-thaw was repeated three times. After freeze-thaw, the tubes were bead-beaten in the BioSpec Products Mini Beadbeater for 2 min, followed by brief (pulse) centrifugation in a benchtop microcentrifuge. Afterward, 45 μl Proteinase K (supplied in the DNEasy plant kit; Qiagen) was added to each reaction and tubes were incubated at 55 °C for 1 h in a hybridization oven. The tubes were then amended with 4 μl RNase A (supplied in the DNEasy plant kit; Qiagen) and incubated at 65 °C for 10 min. After incubation, the filters were removed from the tubes using sterile steel needles, 130 μl of buffer AP2 (Qiagen) added and the samples vortexed and placed on ice. The tubes were then centrifuged at 14 000 *g* for 2 min, and the supernatant was processed per the manufacturer's protocol. The final extraction volume was 50 μl.

RNA extraction

Small-volume samples for quantitative PCR (qPCR) were extracted using RNEasy kits (Qiagen) according to manufacturer's protocols. Genomic DNA (gDNA) contamination in the small-volume extracts was eliminated by treatment with the Turbo DNA-free kit (Applied Biosystems, Austin, TX, USA) according to manufacturer's recommendations.

RNA for metatranscriptomic analysis was extracted from filters using the RNA Mini Isolation Kit II (Zymo Research, Orange, CA, USA), with the following modifications: 6 ml ZR buffer (Zymo Research) was added to the filters in centrifuge tubes while the filters remained frozen. Glass Beads (0.1 mm diameter, 100 μl total) were also added to the centrifuge tubes, and tubes were vortexed to ensure coverage of the filter by buffer. The tubes were then bead-beaten for 2 min, after which the filters were mechanically homogenized using sterile disposable serological pipettes for a further 2 min. After votexing the samples again, the tubes were centrifuged to remove filter debris from the buffer. The ZR buffer supernatant was then passed through the spin columns provided in the kit, and then the extraction followed manufacturer's recommendations. Contaminating gDNA was removed from extracts using the Zymo DNA-free RNA kit.

Metatranscriptomic sample processing

DNA-free total RNA for metatranscriptomic analysis was first subjected to terminator exonuclease using the mRNA-Only Kit (Epicentre, Madison, WI, USA), which was stopped by the addition of 1 μl kit-supplied stop solution. The resulting RNA mixture was subjected to a second mRNA purification protocol, using the MicrobExpress kit (Applied Biosystems), which employs capture hybridization by magnetic separation. At the conclusion of the mRNA-Only and MicrobExpress protocols, mRNA was depleted by 40–60% from total RNA as measured by spectrophotometry. The enriched mRNA sample was amplified by *in vitro* transcription using the MessageAmp II-Bacteria kit (Applied Biosystems) following the manufacturer's recommendation without amendment. The resulting amplified messenger RNA (aRNA) was 100–200 times the concentration of the mRNA after removing rRNA. The aRNA was converted to single-stranded (ss) cDNA using SuperScript III reverse transcriptase (Invitrogen, Carlsbad, CA, USA). The ss cDNA was then converted to double-stranded (ds) cDNA by treatment with *Escherichia coli* DNA polymerase I and RNase H. A total of 12 and 8 μg of ds cDNA was produced from the day and night samples, respectively, at the conclusion of processing.

Pyrosequencing

ds cDNA (5 μg) were subjected to picoliter reactor pyrosequencing at 454 Life Sciences, resulting in 123 686 individual reads. The average read length was 174 bp for day metatranscripts and 155 bp for night metatranscripts, representing 11.9 Mbp of day and 8.6 Mbp of night nucleotide sequence. Sequences from this study are deposited as libraries TA_35115 (day) and TA_35117 (night) under accession no CAM_P0000051 to the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA).

Sequence analysis

Pyrosequences were compared against the draft *C. watsonii* WH8501 genome by BLASTn analysis (Altschul *et al.*, 1997). Sequences with $<99\%$ similarity and *E*-values $>10^{-20}$ to the genome contigs were discarded. This resulted in an effective sequence length threshold of 105 bp. The sequences with $\geqslant 99\%$ similarity to the *C. watsonii* genome were then compared by BLASTx analysis against the complete nonredundant (nr) protein database. Sequences with matches $<95\%$ on the amino-acid level to gene products were discarded from further analyses. Similarly, rRNA sequences were removed. The remaining non-rRNA sequences matching $\geqslant 95\%$ to proteins were assigned to metabolic pathways by BLASTx analysis against KEGG (Kanehisa *et al.*, 2004).

Quantitative PCR and RT-PCR

Gene and transcript abundances of three previously unexamined genes (*isiA*, Transposase IS4 and Cwat_DRAFT_0061) and one previously examined gene (*nifH*) (Church *et al.*, 2005) were estimated by qPCR and qRT-PCR. Primers and TaqMan hybridization probes were designed for the three unexamined genes using Primer3 software (Rozen and Skaletsky, 2000), and checked against the nr and env_nr database by BLASTn to identify possible cross-amplification with environmental genes. The primer and probe sequences used are given in Table 2. qPCR and qRT-PCR were conducted as described elsewhere

Table 2 Primer and TaqMan probe sequences used in qRT-PCR analysis of highly abundant metatranscripts of *Crocosphaera watsonii* WH8501

Primer name	*Primer direction*	*Primer/probe sequence (5′-3′)*	*Source*
nifH	F	TGGTCCTGAGCCTGGAGTTG	Church *et al.*, 2005
	R	TCTTCTAGGAAGTTGATGGAGGTGT	
	Probe	TGTGCTGGTCGTGGTAT	
Iron-stress-induced protein A (*isiA*)	F	ATGGCTTCTCCTGAGAAAAT	—
	R	GCACATTGTTGTTCCTCCTA	—
	Probe	AGAGCTTTTCTTGCCCATGCC	—
Conserved hypothetical protein Cwat_DRAFT_0061	F	GCGATGTATTGGGAAGAAAT	—
	R	TTCGATCCTCATAACCCCTA	—
	Probe	TCCTCACTTGAATAAACACCACCACT	—
	Standard	TGCGATGTATTGGGAAGAAATA TCCTCACTTGAATAAACACCACCACTA TAGGGGTTATGAGGATCGAAA	—
Transposase IS4	F	GATCGCTTGTGAAGAATGTG	—
	R	CCAACTCTGAAGGTGTCCA	—
	Probe	AAAAATCAGACAGACCAGTGAGCCA	—
	Standard	CGATCGCTTGTGAAGAATG TGAAAAATCAGACAGACCAGTGAG CCATTGGACACCTTCAGAGTTGGC	—

(Church *et al.*, 2005) for the four genes, using oligonucleotide standards for IS4 and Cwat_DRAFT_0061 (Table 2) and plasmid standards for *isiA* and *nifH* over eight orders of magnitude. The copy numbers of genes and transcripts were calculated as described elsewhere (Church *et al.*, 2005). Extracted small-volume RNA was converted to ss cDNA using SuperScript III reverse transcriptase (Invitrogen) before use as template in qPCR (qRT-PCR), as described by Hewson *et al.* (2007). The PCR conditions for all four genes were identical to those used in previous studies of *nifH* (Church *et al.*, 2005).

Results and discussion

Characteristics of C. watsonii *population*

The mixed layer depth, as calculated per Lorbacher *et al.* (2006) was 25.9 m, where surface salinity was depressed by ~0.75 relative to waters immediately below. *Crocosphaera*-like cell abundances at station KM0703.025 reached 3.84–4.70 × 10^3 cells per ml between 37 and 45 m depth (the *C. watsonii*-enriched horizon), but were much less abundant at the surface, at 75 m, and were not detected by microscopy at the depth of the deep chlorophyll maximum (106 m, see Figure 1). This was the highest abundance of *C. watsonii* observed anywhere to date (Campbell *et al.*, 1997, 2005; Charpy, 2005; Church *et al.*, 2005; Foster *et al.*, 2007; Hewson *et al.*, 2007; Zehr *et al.*, 2007b). The *C. watsonii*-enriched horizon corresponded to distinct peaks of chlorophyll a fluorescence and total particle concentration. Thus, the observed bloom was not simply an increase in chlorophyll per cell, but was a localized enhancement of photosynthetic pigment and particles. This *C. watsonii* bloom feature was at the 27% light level just below the base of the mixed layer and was concomitant with the maximum percent oxygen saturation observed at this station. This indicates significant oxygen evolution rates coincident with the *C. watsonii* maxima, because photosynthetic picoeukaryotes and picocyanobacteria were at least an order of magnitude less abundant than *C. watsonii* (Figure 1).

Total prokaryote and viral abundances were much higher than the abundance of *C. watsonii* and typical of oligotrophic open ocean conditions (Hewson and Fuhrman, 2006, 2007). Bacterial abundance was highest at the enriched depth, but the range of abundance at the depths sampled was relatively narrow (2.63–5.88 × 10^5 cells per ml). Viral abundance was highest at the *C. watsonii*-enriched horizon (1.29 × 10^7 viruses per ml), but was about an order of magnitude less abundant at the surface and deep chlorophyll maximum. The ratio of viruses to bacteria was higher in the enriched horizon (ratio 22) compared to surface waters (ratio 12) and the deep chlorophyll maximum (ratio 14).

Total prokaryote and viral abundance is typically greater in more productive conditions (Hewson *et al.*, 2001, 2006). The large abundance of both prokaryotes and viruses in the *C. watsonii*-enriched horizon is consistent with an environment in which a greater amount of dissolved organic matter is present due to photosynthetic activity. In fact, measured dissolved organic carbon and nitrogen were enhanced at the *C. watsonii* maximum (Figure 1). The higher virus:prokaryote ratio within the bloom may be a consequence of either higher viral production by prokaryotes under a productive regime (Hewson *et al.*, 2001) or production of viruses by the abundant *C. watsonii*. To date there have been no reports of viruses specific to *C. watsonii*; however, viral lysis has been described in other pelagic nondiazotrophic and diazotrophic cyanobacteria including *Trichodesmium* (Hewson *et al.*, 2004; Ohki, 1999).

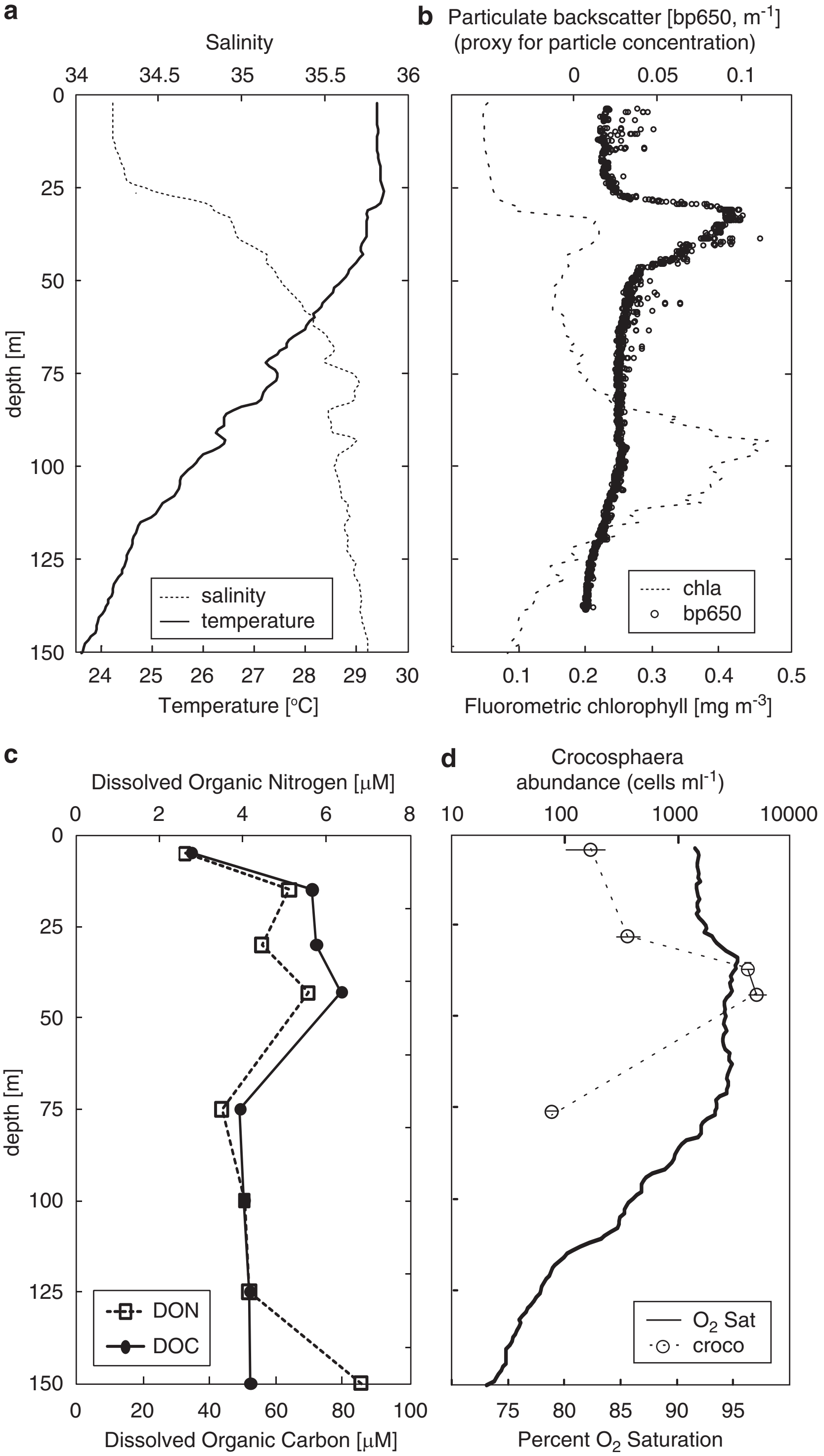

Figure 1 (**a**) Salinity and temperature profiles, (**b**) particle concentrations as estimated by particulate backscatter at 650 nm relative to fluorometric chlorophyll a measured by a CTD mounted Seapoint fluorometer, (**c**) dissolved organic carbon and dissolved organic nitrogen profiles and (**d**) the microscopically determined abundance of *Crocosphaera watsonii*-like cells relative to percent O_2 saturation at station KM0703.025 (14°59.9′S, 175°0′E) on 13 April 2008. (**d**) The bloom depth horizon fluctuated between 37 and 45 m over the course of station occupation. Note that microscopy counts for 37 m were taken approximately 2 h after the other depths. The depth horizon of the *Crocosphaera* maximum coincides with dissolved organic carbon and nitrogen enrichments. Error bars on abundance = s.e.

Metatranscriptome characteristics

At the station with a high abundance of *C. watsonii* (the ‘bloom’ station), 68 405 day and 55 281 night total RNA sequences were obtained, including 18 217 day and 12 105 night putative mRNAs that did not match rRNA libraries. The mean day mRNA read length was 173 bp and night 153 bp, with mean G+C contents of 46%. Of putative mRNAs, 672 day and 327 night sequences were >105 bp and were ⩾99% identical at the nucleotide level to the draft *C. watsonii* WH8501 genome (AADV02000000). An additional 53 day and 53 night sequences >105 bp were between 97% and 99% identical to the *C. watsonii* WH8501 genome sequence and were more similar to *C. watsonii* WH8501 than to other microorganisms; however, we did not include these in further analyses. Of ⩾99% *C. watsonii* WH8501 nucleotide sequence matches, 450 day and 148 night sequences (67% and 45%, respectively) shared homology with annotated nonribosomal genes of *C. watsonii* WH8501. The remainder of transcripts with *C. watsonii* WH8501 nucleotide matches did not share homology with protein-encoding regions. The largest number of nonprotein coding mRNAs (47 day and 29 night metatranscripts) matched *C. watsonii* contig 288 (of the September 2006 draft annotation) in a region between conserved hypothetical genes. The function of these intergenic region transcripts is unknown, however it is possible they represent rho-independent termination factors or riboswitches that are transcribed from intergenic regions beyond the protein-encoding region. These results are in line with previous studies showing large numbers of unidentified transcripts in surface seawater (Frias-Lopez *et al*., 2008; Gilbert *et al*., 2008).

Of the 5698 total protein-encoding genes in the *C. watsonii* WH8501 genome, 226 (4%) were detected in the combined day and night metatranscriptomes. The ratio of *C. watsonii* WH8501 nucleotide matches to total sequences (1.17% and 0.69% in day and night, respectively) is similar to the ratio of *C. watsonii* cell abundance to total prokaryote cell abundance. Gene length normalized frequencies (Frias-Lopez *et al*., 2008) demonstrated genome-wide diel expression of *C. watsonii* WH8501 genes, indicating that transcription of genes was not restricted to one part of the genome, or our sampling biases recovery of particular genome regions (Figure 2).

The low proportion of recovered *C. watsonii* WH8501 genes in the metatranscriptome indicates that sampling was undersaturated, which was a consequence of their low abundance relative to other microorganisms or due to differential expression of genes. A previous study of metatranscriptomes at station ALOHA indicated differential expression of *Prochlorococcus* genes (Frias-Lopez *et al*., 2008). Because *Prochlorococcus* is a dominant component of microbial assemblages in the open ocean, comparisons with a complementary metagenome were possible. However, for less-abundant components like *C. watsonii* WH8501, which comprise a small fraction of cell abundance and thus would have poor metagenomic coverage, such comparisons would require deep

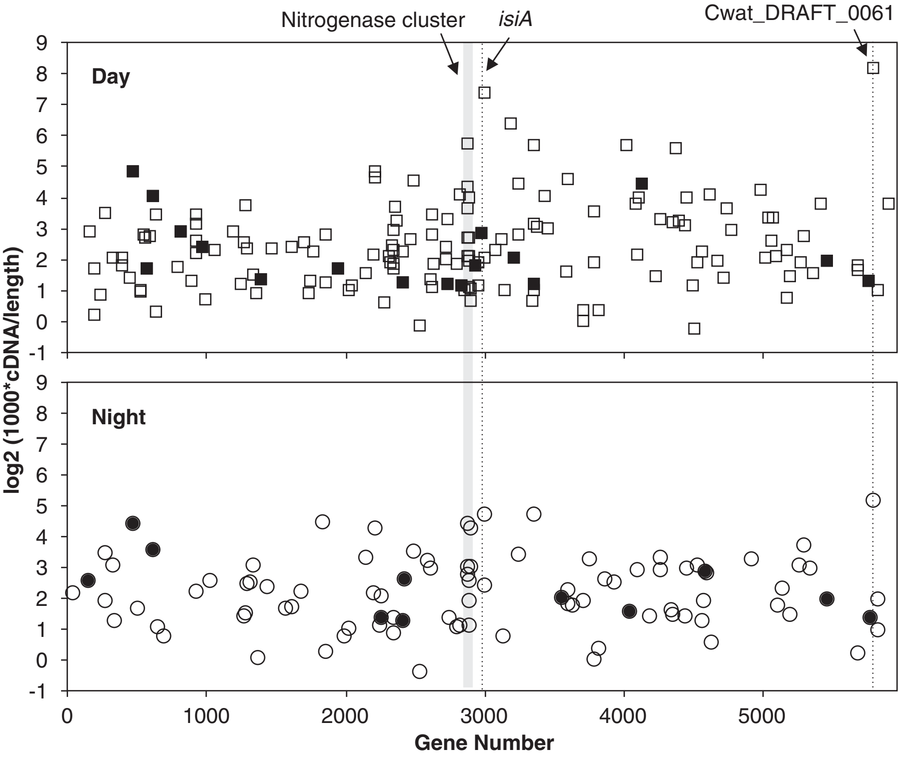

Figure 2 Length normalized frequency of detection of *Crocosphaera watsonii* WH8501 genes. Each point represents the number of transcripts detected for each of the 5958 genes in the draft genome assembly. Solid symbols indicate transposases. Sequence reads that share >95% homology to *C. watsonii* proteins by BLASTx analysis are indicated.

sequencing to obtain complimentary RNA and DNA frequency. Because we used a random survey approach, genes have equal probability of recovery. Hence, unless the genes are present in multiple copies in the *C. watsonii* WH8501 genome, comparison of transcript frequency in the random survey, in the absence of complementary gene frequency, allows identification of the most actively expressed genes.

The largest fraction of *C. watsonii* WH8501 transcript sequences obtained from the bloom station in the day (30%) and night (18%) was conserved hypothetical proteins or proteins of unknown function, which is in line with *C. watsonii* WH8501 transcripts recovered from other stations in the southwest Pacific and equatorial Atlantic oceans (30.7%), and previous metagenomic studies of microorganisms (Tyson *et al*., 2004; Venter *et al*., 2004; Rusch *et al*., 2007) and viruses (Angly *et al*., 2006; Breitbart *et al*., 2002, 2003). The most highly expressed *C. watsonii* hypothetical protein (Cwat_DRAFT_0061; ZP_00519306.1) was present in the bloom day (15% of transcripts) and night (5% of transcripts) metatranscriptomes. Cwat_DRAFT_0061 was also recovered at station ALOHA (Poretsky *et al*., in press) and at a station in the southwest Pacific (KM070304) (Supplementary Table 1). These highly expressed genes of unknown function may be attractive targets for expression studies because they appear to be important in the ecophysiology of *C. watsonii*.

The transcripts from the bloom with assigned KEGG pathways were predominately associated with genetic information processing and energy metabolism, with fewer sequences involved in carbohydrate, amino acid, lipid and cofactors and vitamins metabolism, and cell information processing and signaling (Figure 3). The *C. watsonii* WH8501 transcripts from energy metabolic genes included those associated with photosynthesis and N_2 fixation.

Surprisingly, transposase transcripts (predominately IS605 and IS4) were among the most highly represented *C. watsonii* WH8501 transcripts of known function, representing 10.5% and 15.9% of *C. watsonii* WH8501 sequences in the day and night, respectively. This corresponded with a greater number of *C. watsonii* WH8501 recombination and repair orthologs (7.1–10.7% of all transcripts) than in the dominant sympatric cyanobacterium, *Prochlorococcus marinus* MIT9301 (2.0–2.7% of all transcripts), which lacks transposase genes altogether. Our results are consistent with a metatranscriptome sequencing effort from station ALOHA in the North Pacific Ocean, where 3 transposase sequences were recovered from 50 total *C. watsonii* WH8501 transcripts (Zehr *et al*., 2007a). Transposases were also recovered at other locations in the southwest Pacific Ocean and equatorial North Atlantic Ocean, but represented a smaller percentage of total *C. watsonii* WH8501 transcripts

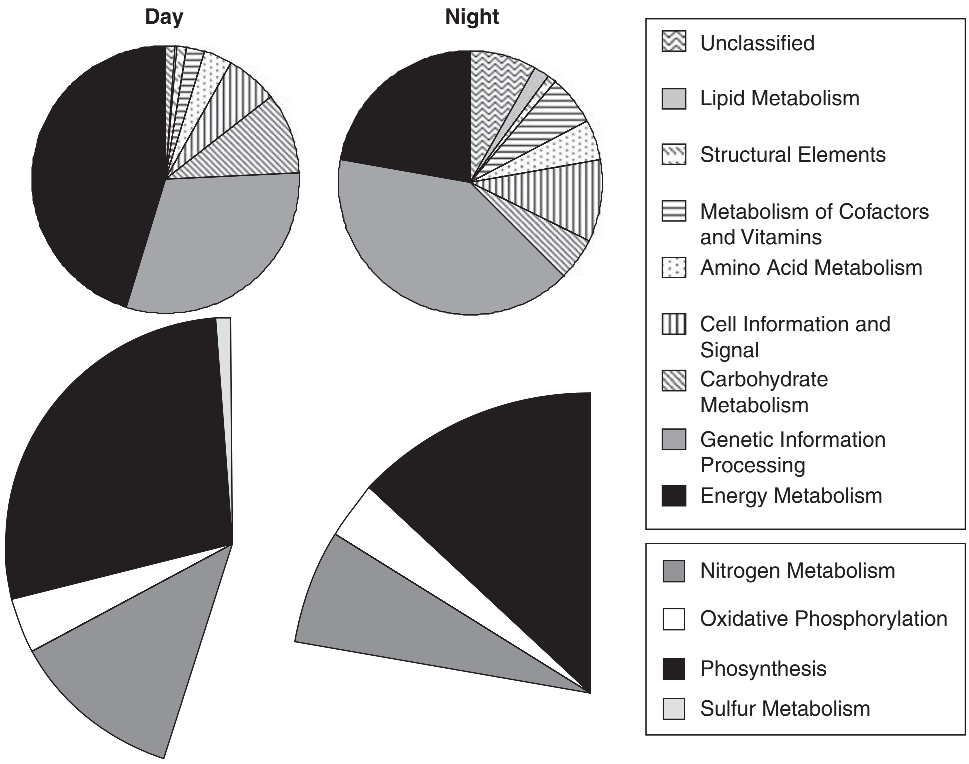

Figure 3 Metabolic KEGG pathway categories of *Crocosphaera watsonii* WH8501 genes recovered at station KM0703.025 at 37 m on 13 April 2008. The pie charts represent 442 day and 144 night sequences. The lower pie chart fragments show different energy metabolic pathways.

(Supplementary Table 1). At station KM0703.04 in the southwest Pacific Ocean, transposases represented 5.5% and 2.2% of *C. watsonii* WH8501 metatranscripts, respectively, whereas at station SJ0603.09 in the eastern portion of the equatorial North Atlantic Ocean, one transposase sequence was recovered among four total *C. watsonii* WH8501 sequences.

The abundance of *C. watsonii* transposase transcripts was surprising, as their function in *C. watsonii* ecology is unclear. It has been hypothesized that genetic diversity in *C. watsonii* may be maintained by genome rearrangement, rather than by gene sequence divergence because the diazotroph has low levels of genome DNA sequence variability between habitats and over time (Zehr *et al*., 2007a). High transposase activity in *C. watsonii* is similar to that found in populations of the archaeon *Ferroplasma acidarmanus* (Allen *et al*., 2007), and positive selection of transposases has been linked to increased fitness or invasion of new environments (Mes and Doeleman, 2006; Allen *et al*., 2007; Franguel *et al*., 2008). The *C. watsonii* WH8501 genome has a large number of transposases (419 transposases in the ∼6.3 Mb genome), some of which interrupt reading frames of genes (Zehr *et al*., 2007a) and may result in phenotypic diversity. The activity of transposases may be linked to gene modification or silencing. In other cyanobacteria, the activity of transposases is linked to gene modification or perhaps even gene silencing through insertions within protein-encoding regions (Mlouka *et al*., 2004). The biosynthetic investment in transposase expression suggests a vital function in the cyanobacterium for population fitness in oligotrophic environments.

Another possible function of transposase activity in *C. watsonii* is defense against lysogenic conversion, because enhanced genomic diversity results in lower frequency of insertion sites for phage integration. The effect of transposases may be similar to clustered regularly interspaced short palindromic repeats (CRISPRs) (Jansen *et al*., 2002), the frequency of which corresponds to phage attack frequency (Barrangou *et al*., 2007). CRISPRs have been implicated both as gene silencers in prokaryotes (similar to siRNAs in eukaryotes) and as a mechanism of viral resistance (Barrangou *et al*., 2007; Tyson and Banfield, 2008). The *C. watsonii* WH8501 genome contigs contain 1 confirmed (lying within a gene encoding for peptide synthetase; NP_486684.1) and 12 questionable CRISPR-like structures, more than found in common cultivated marine microorganisms (Grissa *et al*., 2007). One of the questionable structures shares sequence homology with heat-shock protein *DnaJ*, three sequences of which were retrieved as metatranscripts in the day sample. *C. watsonii* are typically rare and therefore are unlikely to have a multiplicity of infection congruent with lytic phage attack under non-bloom conditions (Wilcox and Fuhrman, 1994). However, *Synechococcus* in Tampa Bay demonstrate greater frequency of lysogenic conversion in summer months when abundances are greater (McDaniel *et al*., 2002). Thus, when *C. watsonii* abundance is elevated, transposase activity may be enhanced to defend against phage attack.

Another dominant transcript recovered was the photosystem I-associated iron-stress-induced protein A (*isiA*; annotated in genome as the photosystem II *psbC*-encoding COP43 homolog, but with highest amino-acid sequence similarity to *isiA* in *Cyanothece* and *Trichodesmium*). *isiA* is formed under prolonged iron stress in cultures of cyanobacteria (Laudenbach *et al*., 1988; Laudenbach and Straus, 1988) as a mechanism to prevent high-light-induced oxidative damage by increasing cyclical electron flow around the photosynthetic reaction center photosystem I (Michel and Pistorius, 2004; Ohkawa *et al*., 2000). *isiA* sequences were recovered from transcriptomes at other stations in the southwest Pacific Ocean and the eastern portion of the equatorial North Atlantic Ocean (Supplementary Table 1), suggesting that this transcript may be a useful marker for the physiological status of natural populations of *C. watsonii* and may be more sensitive than the *idiA* (Webb *et al*., 2001), transcripts of which were not detected in our libraries. The energy metabolic gene expression pattern of *C. watsonii* was different from the non-N_2-fixing cyanobacterium *P. marinus* WH9301*,* which had a lower percentage of transcripts for nitrogen metabolism (16.7- to 27.6-fold higher in *C. watsonii*), but greater number of transcripts of genes in the glycolysis and gluconeogenesis pathway (1.3- to 2.3-fold lower in *C. watsonii*).

The large number of *C. watsonii* N_2 fixation gene transcripts at the bloom station is consistent with the environmental conditions at the sampling location, where N_2 fixation rates at the bloom depth horizon as measured by average $^{15}N_2$ uptake over 24 h were 0.5–1.5 nmol l^{-1} h^{-1}, accounting for 60% of the total areal rate at the station (Montoya JP, unpublished data). We recovered transcripts for most components of the *C. watsonii* nitrogenase gene cluster in the day and some at night in the bloom sequence libraries (Figure 4). Only a single nitrogenase transcript was found at another southwest Pacific station (KM0703.04) and none were found elsewhere. Nitrogenase activity has been measured predominately at night in cultures of *C. watsonii* WH8501, and *C. watsonii* WH8501-related cyanobacterial diazotroph *nifH* transcripts are in low abundance during the light phase (Church *et al*., 2005; Hewson *et al*., 2007). The disproportionate number of sequences observed between day and night metatranscriptomes prevents direct comparison of day and night expression levels. The abundance of nitrogenase gene cluster transcripts in the day may reflect constitutive expression of some parts of the cluster that do not directly lead to N_2 fixation activity. In *Cyanothece* sp. ATCC 51142,

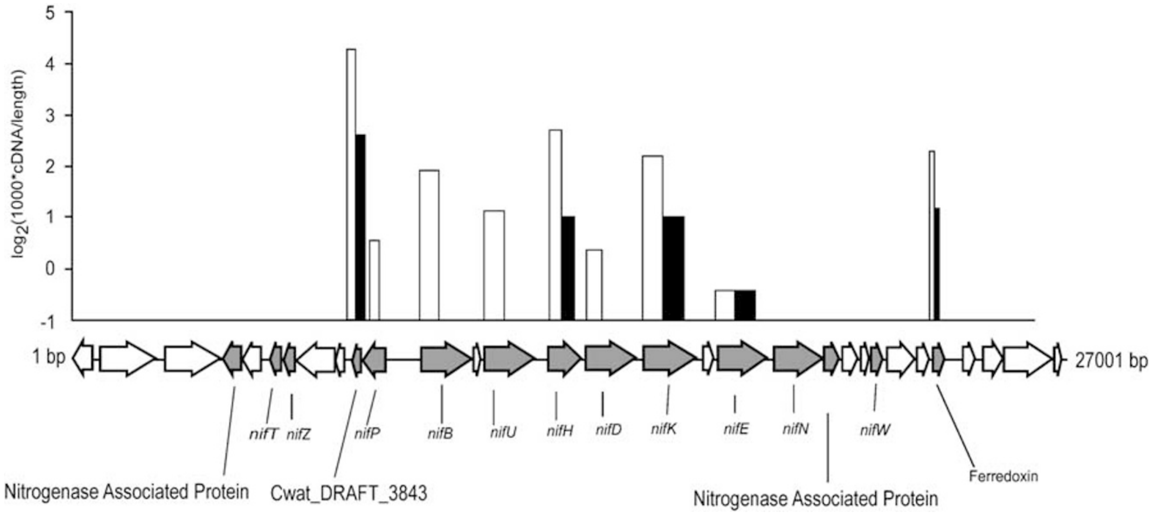

Figure 4 Graphical interpretation of *Crocosphaera watsonii* WH8501 nitrogenase cluster gene length normalized metatranscript frequency in day (light bars) and night (dark bars). The nitrogenase cluster is on a single contig of the unfinished genome; the direction of arrows indicates the orientation of the gene in the cluster. The frequency of metatranscripts is indicated by the height of the bars above the cluster representation.

differential expression of nitrogenase cluster genes occurs over diel cycles, with expression of some genes during the light, even though most genes are more highly expressed in the dark (Colon-Lopez and Sherman, 1998; Stoeckel *et al.*, 2008; Toepel *et al.*, 2008). The Fe–Mo structural protein *nifK*, which is encoded for by the *nifHDK* polycistronic mRNA (Chen *et al.*, 1998), had almost as many transcripts represented as *nifH*, which is the gene commonly used to examine diazotroph abundance and nitrogenase expression in environmental samples (Church *et al.*, 2005; Short and Zehr, 2007; Zehr *et al.*, 2007b). The *nifB* gene, which forms part of the *nifBQ* cluster and is necessary for the processing of the Fe–Mo cofactor encoded by *nifHDK* (Roberts *et al.*, 1978), was also highly represented among the daytime transcripts, in line with expression observed in *Cyanothece* sp. ATCC 51142 (Toepel *et al.*, 2008). Finally, a hypothetical protein flanked by *nifZ* and *nifP* (Cwat_DRAFT_3843) had the greatest number of transcripts; however, its function in nitrogenase is unknown and it may be an attractive target for expression studies. These nitrogenase genes may provide more sensitive targets for study of environmental nitrogenase transcription in future studies. Because *C. watsonii* WH8501 also has the genomic potential to assimilate NH_4^+, and reduce NO_3^- and NO_2^-, the biosynthetic cost to *C. watsonii* of gene transcription for N_2 acquisition at this station appears necessary to maintain high abundances under otherwise favorable but oligotrophic environmental conditions.

Quantitative PCR and RT-PCR

Based on metatranscriptomic sequence frequency within libraries, we developed qPCR and qRT-PCR assays for Transposase IS4, iron-inducible stress protein A (isiA), and the conserved hypothetical protein similar to COP23 in *Synechococcus* (Cwat_DRAFT_0061). We applied these and the previously developed group B cyanobacterial *nifH* assay (of which *C. watsonii* WH8501 is a representative) (Church *et al.*, 2005) to vertical profiles of microbial DNA and RNA collected during day and night phases at the bloom location. These genes were targeted because their transcript abundance suggested a key function in the physiology of *C. watsonii* in natural populations; however, there was no information on their expression levels over diel cycles.

The abundance of *Crocosphaera*-like cells at the bloom depth determined microscopically (Figure 1) and confirmed by flow-cytometric analysis (data not shown) was several orders of magnitude higher than gene abundance (Table 3) at each depth, which may be a result of microbial DNA extraction efficiency or the identification of cells as *C. watsonii* that were actually morphologically similar cyanobacteria. The qRT-PCR results of transcript abundance were consistent with observations of metatranscriptomic sequence frequency (Figure 5). Transcript abundance data normalized to cell abundance (determined microscopically) indicated that most transcripts were present as multiple copies per cell at the depth sampled for metatranscriptomics; however, at shallower and deeper depths the number of transcripts was less, often <1 transcript per cell (Figure 5). This is likely due to low RNA extraction efficiency, but may also indicate the presence of inactive cells within the assemblage, because inactive and active cells are not distinguished by microscopy. As expected, but consistent with the observed metatranscript frequency, the

Table 3 Cell abundance—normalized gene abundance of Fe–Mo protein of nitrogenase (*nifH*), conserved hypothetical protein Cwat_DRAFT_0061 (most similar to COP23 circadian-related protein in *Synechococcus* sp. PCC 8801), iron-stress-induced protein A (*isiA*), and Transposase IS4 at bloom station KM0703.25

Depth (m)	*Time (hours)*	*nifH*		*Cwat_DRAFT_0061*		isiA		*Tranposase IS4*	
		Transcript per cell	*s.e.*	*Transcript per cell*	*s.e.*	*Transcript per cell*	*s.e.*	*Transcript per cell*	*s.e.*
15	1244	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01
30	1244	0.00	0.00	0.00	0.00	0.03	0.01	0.02	0.00
37	1403	1.07	0.15	1.37	0.22	36.34	5.94	0.59	0.02
45	1244	0.11	0.01	0.05	0.01	2.38	0.52	0.72	0.07
75	1244	0.01	0.00	0.00	0.00	0.01	0.00	0.08	0.01
15	0025	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
37	0025	2.31	0.08	0.01	0.01	0.32	0.11	1.90	0.37
45	0025	0.58	0.02	0.00	0.00	0.01	0.00	0.22	0.04
75	0025	0.63	0.07	0.00	0.00	0.01	0.00	0.04	0.00

Gene abundances were measured by qRT-PCR, where mean and standard error of three analytical replicates are indicated. The data were normalized to *C. watsonii* WH8501 cell abundance as determined by epifluorescence microscopy. Note that 15 m gene abundance data were normalized to 5 m microscopic counts.

ratio of expression between day and night transcripts (Figure 5) was the least for *nifH*. This gene was previously shown to have strong circadian rhythm, with night expression typically 2–3 orders of magnitude greater than day expression (Church *et al.*, 2005; Hewson *et al.*, 2007; Zehr *et al.*, 2007b).

The *C. watsonii* photosystem I-binding *isiA* iron-stress protein had the greatest variation from day to night at the bloom depth (Figure 5), with most transcripts found during the day. The opposite pattern was observed in shallower and deeper depths, suggesting a key function for *isiA* at the *C. watsonii*-enriched horizon. *isiA* protects the photosynthetic complex from superoxide damage during photosynthesis, hence the strong diel change is consistent with photosynthetic patterns. However, the observation of *isiA* genes at high copy number at the bloom depth was somewhat unexpected because the photosynthetically active radiation at 37 m was only 27% of surface irradiance and *isiA* transcripts were also observed in the night metatranscriptome. Furthermore, this expression of *isiA* was surprising because Fe concentrations in the region are typically higher than in oligotrophic waters elsewhere due to undercurrent flux originating near Papua New Guinea (Wells *et al.*, 1999). This suggests that although *isiA* has strong diel cycle, it may have a function other than photosynthetic activity or Fe starvation in *C. watsonii* cells at the bloom depth.

Transposase IS4 demonstrated an ~3-fold increase in transcript abundance determined by qRT-PCR at night at the bloom depth (Figure 5). In contrast, transposase transcript abundance at shallower and deeper depths was higher during day than night. While consistent with metatranscript frequency, the diel cycling of the gene is unclear. Most phage infection occurs at night in surface waters of the ocean (Winter *et al.*, 2004), hence it is

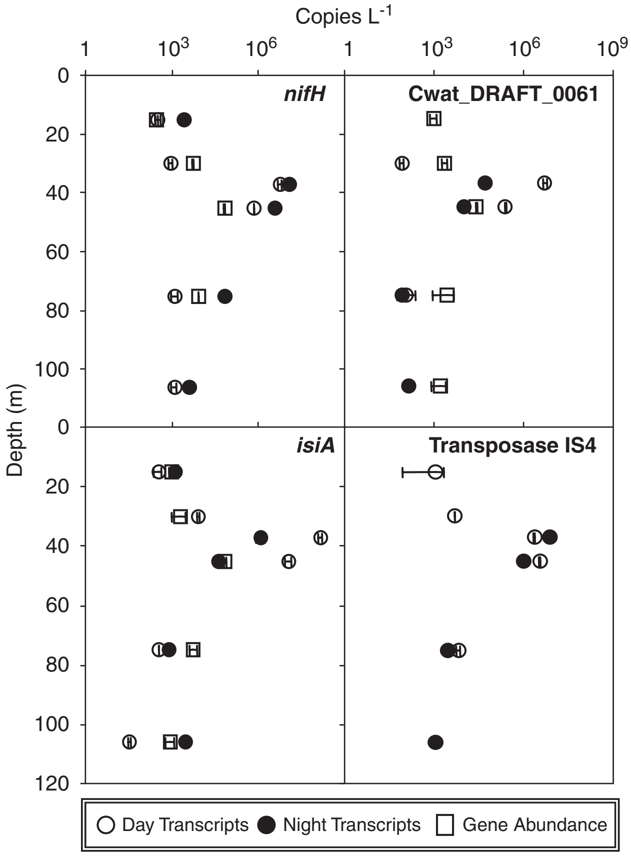

Figure 5 Gene and transcript abundance of Fe–Mo protein of nitrogenase (*nifH*) (Church *et al.*, 2005), iron-stress-induced protein A (*isiA*), conserved hypothetical protein Cwat_DRAFT_0061 and Transposase IS4, as measured by qPCR and qRT-PCR; Church *et al.*, 2005). Each data point represents the mean of three analytical replicates; s.e., standard error. Cwat_DRAFT_0061 transcripts were not detected in day and night at 15 m, transcripts of Transposase IS4 were not detected at night at 15 m and *nifH* genes were not detected at 106 m. No night transcript samples were obtained from 30 m. Note that samples for 37 m were taken approximately 2 h after the other sample depths.

possible that some transposase activity may be related to phage defense. Because most DNA damage in prokaryotic cells occurs as a consequence of UV irradiation (Jeffrey and Mitchell, 1997), it is also possible that transposase activity may be related to genome maintenance during darkness. No information is available on the growth characteristics of *C. watsonii* during day–night cycles; however, it is possible that enhanced cell division at night may cause higher transposase activity. Regardless of the mechanism by which it occurs, the increased transcript number per cell ratio at depths with higher abundance indicates that transposase activity is linked to enhanced fitness and enables *C. watsonii* cells to form dense concentrations.

The strong diel changes in transcripts of Cwat_DRAFT_0061 at all depths (Figure 5) suggest that this protein is linked to diel cycles. Cwat_DRAFT_0061 is similar (39% identical on the amino-acid level) to a circadian oscillating polypeptide COP23 precursor of *Synechococcus* sp. PCC 8801/RF-1, a cell membrane protein that is stable in arrhythmic cells but degrades rapidly under light–dark cycles (Chen *et al.*, 1996). Our observation of a higher number of day transcripts than night transcripts of Cwat_DRAFT_0061 in the bloom suggests that it may have similar function to *Synechococcus* sp. PCC 8801 COP23.

The time between sample collection and processing (∼45 min) may have resulted in degradation of some mRNAs (Andersson *et al.*, 2006). Energy metabolic transcripts may be overrepresented, and putative enzymes and translation transcripts underrepresented in degraded metatranscriptomes (Selinger *et al.*, 2003). Hence, we cannot discount the possibility that the highly expressed *isiA* and *nifH* genes are a consequence of mRNA degradation during sample processing.

Conclusion

The different metabolic pathway gene transcription pattern of *C. watsonii* compared to co-occurring cyanobacteria may indicate a unique strategy for maintaining its niche within the open ocean ecosystem. *C. watsonii* has genome expression patterns that reflect the nutrient-limited environment of the oligotrophic ocean that provides insight into highly expressed genes important for the microorganism under favorable (that is, bloom) conditions. Whereas the function of highly expressed transposase and conserved hypothetical proteins observed in the bloom and elsewhere in the open ocean is unknown, our results demonstrate that the transcription (that is, transcripts per gene copy) of these genes were enhanced in the bloom of *C. watsonii* (Figure 5). The function of transposases in the population biology and physiology of *C. watsonii* needs further study, and these genes may provide markers for cell growth or physiological stress in the environment. The heavy investment of *C. watsonii* cells in genes that respond directly to ambient nutrient availability suggests that nutrient acquisition is a dominant activity of the diazotroph. Finally, the genes observed under ambient conditions may only be observed using our *in situ* gene expression approach, because cultivation studies may demonstrate different gene expression in confinement and highlight the need for more targeted transcriptomic studies to understand the autecology of individual key microorganisms in nature.

Acknowledgements

We thank J Landrum, R Horak, B Van Gessel, T Peterson, K London, I Biegala, J Becker, D Hayakawa, E Orchard, T McGovern, B Issler, E Speicher, T Chen and the crew of the R/V Kilo Moana for assistance in the field. We also thank B Carter, M Hogan, the MEGAMER facility, S Schuster, J Qi, L Tomsho, R Martone, P Gilna and the Community Cyberinfrastructure for Advance Marine Microbial Ecology Research and Analysis (CAMERA) for sequencing and bioinformatics support. This work was supported by the Gordon and Betty Moore Foundation Marine Microbiology Initiative, and NSF grant EF0424599 for the Center for Microbial Oceanography: Research and Education (CMORE).

References

Allen E, Tyson G, Whitaker R, Detter J, Richardson P, Banfield J. (2007). Genome dynamics in a natural archaeal population. *Proc Natl Acad Sci USA* **104**: 1883–1888.

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. *Nucleic Acids Res* **25**: 3389–3402.

Andersson A, Lundgren M, Eriksson S, Rosenlund M, Bernander R, Nilsson P. (2006). Global analysis of mRNA stability in the archaeon Sulfolobus. *Genome Biol* **7**: R99.

Angly F, Felts B, Breitbart M, Salamon P, Edwards R, Carlson C *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4**: 2121–2131.

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S *et al.* (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709–1712.

Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P *et al.* (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**: 6220–6223.

Breitbart M, Salamon P, Andresen B, Mahaffy J, Segall A, Mead D *et al.* (2002). Genomic analysis of an uncultured marine viral community. *Proc Natl Acad Sci USA* **99**: 14250–14255.

Brown MV, Fuhrman JA. (2005). Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. *Aquat Microb Ecol* **41**: 15–23.

Campbell L, Carpenter E, Montoya J, Kustka A, Capone D. (2005). Picoplankton community structure within and outside a *Trichodesmium* bloom in the southwestern Pacific Ocean. *Vie et Milieu* **55**: 185–195.

Campbell L, Liu H, Nolla H, Vaulot D. (1997). Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991–1994 ENSO event. *Deep-Sea Res I* **44**: 167–192.

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

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

Carlson CA, Ducklow HW, Hansell DA, Smith WO. (1998). Organic carbon partitioning during spring phytoplankton blooms in the Ross Sea Polynya and the Sargasso Sea. *Limnol Oceanogr* **43**: 375–386.

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

Charpy L. (2005). Importance of photosynthetic picoplankton in coral reef ecosystems. *Vie et Milieu* **55**: 217–223.

Chen H, Chien C, Huang T. (1996). Regulation and molecular structure of a circadian oscillating protein located in the cell membrane of the prokaryote *Synechococcus* RF-1. *Planta* **199**: 520–527.

Chen Y, Dominic B, Mellon M, Zehr J. (1998). Circadian rhythm of nitrogenase gene expression in the diazotrophic filamentous nonheterocystous Cyanobacterium *Trichodesmium* sp strain IMS101. *J Bacteriol* **180**: 3598–3605.

Church MJ, Jenkins BD, Karl DM, Zehr JP. (2005). Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat Microb Ecol* **38**: 3–14.

Colon-Lopez M, Sherman L. (1998). Transcriptional and translational regulation of photosystem I and II genes in light-dark- and continuous-light-grown cultures of the unicellular cyanobacterium *Cyanothece* sp. strain ATCC 51142. *J Bacteriol* **180**: 519–526.

de Saizieu A, Certa U, Warrington J, Gray C, Keck W, Mous J. (1998). Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nat Biotechnol* **16**: 45–48.

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

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.

Franguel L, Quillardet P, Castets A, Humbert J, Matthijs H, Cortez D *et al.* (2008). Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* **9**: 274.

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 USA* **105**: 3805–3810.

Garcia-Martinez J, Rodriguez-Valera F. (2000). Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group 1. *Mol Ecol* **9**: 935–948.

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.

Grissa I, Vergnaud G, Pourcel C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**: 172.

Hewson I, Capone DG, Steele JA, Fuhrman JA. (2006). Influence of the Amazon and Orinoco offshore surface water plumes on oligotrophic bacterioplankton diversity in the West Tropical Atlantic. *Aquat Microb Ecol* **43**: 11–22.

Hewson I, Fuhrman JA. (2006). Viral impacts upon marine bacterioplankton assemblage composition. *J Mar Biol Assoc UK* **86**: 577–589.

Hewson I, Fuhrman JA. (2007). Covariation between viral parameters and bacterial assemblage richness and diversity in the water column and sediments. *Deep-Sea Res I* **54**: 811–830.

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, Moisander PH, Morrison AE, Zehr JP. (2007). Diazotrophic bacterioplankton in a coral reef lagoon: phylogeny, diel nitrogenase expression and response to phosphate enrichment. *ISME J* **1**: 78–91.

Hewson I, O'Neil JM, Fuhrman JA, Dennison WC. (2001). Virus-like particle distribution and abundance in sediments and overlying waters along eutrophication gradients in two subtropical estuaries. *Limnol Oceanogr* **47**: 1734–1746.

Jansen R, van Embden J, Gaastra W, Schouls L. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**: 1565–1575.

Jeffrey W, Mitchell D. (1997). Mechanisms of UV-induced DNA damage and response in marine microorganisms. *Photochem Photobiol* **65**: 260–263.

Johnston AWB, Li Y, Ogilvie L. (2005). Metagenomic marine nitrogen fixation—feast or famine? *Trend Microbiol* **13**: 416–420.

Jonasz M, Fournier GR. (2007). *Light Scattering by Particles in Water: Theoretical and Experimental Foundations*. Elsevier: London.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**: D277–D280.

Karl DM, Michaels AF, Bergman B, Capone DG, Carpenter EJ, Letelier R *et al.* (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58**: 47–98.

Laudenbach D, Reith M, Straus N. (1988). Isolation, sequence analysis, and transcriptional studies of the flavodoxin gene from *Anacystic nudulans* R2. *J Bacteriol* **170**: 258–265.

Laudenbach D, Straus N. (1988). Characterization of a cyanobacterial iron stress induced gene similar to PsbC. *J Bacteriol* **170**: 5018–5026.

Lorbacher K, Dommenget D, Niiler PP. (2006). Ocean mixed layer depth: a subsurface proxy for ocean-atmosphere variability. *J Geophys Res* **111**: C07010.

McDaniel L, Houchin LA, Williamson SJ, Paul JH. (2002). Lysogeny in marine *Synechococcus*. *Nature* **415**: 496.

Mes T, Doeleman M. (2006). Positive selection on transposase genes of insertion sequences in the *Crocosphaera watsonii* genome. *J Bacteriol* **188**: 7176–7185.

Michel K, Pistorius E. (2004). Adaptation of the photosynthetic electron transport chain in cyanobacteria to iron deficiency: the function of IdiA and IsiA. *Physiol Plant* **120**: 36–50.

Mlouka A, Comte K, de Marsac N. (2004). Mobile DNA elements in the gas vesicle gene cluster of the planktonic cyanobacteria *Microcystis aeruginosa*. *FEMS Microbiol Lett* **237**: 27–34.

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. (2004). High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**: 1027–1031.

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

Ohkawa H, Pakrasi H, Ogawa T. (2000). Two types of functionally distinct NAD(P)H dehydrogenases in *Synechocystis* sp strain PCC6803. *J Biol Chem* **275**: 31630–31634.

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

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, Hewson I, Sun S, Allen AE, Moran MA, Zehr JP. (in press). 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).

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

Rozen S, Skaletsky H. (2000). Primer3 on the WWW for general users and for biologist programmers. In: Krawertz S and Misener S (eds). *Bioinformatics Methods and Protocols*. Humana Press: Totowa, NJ, pp 365–386.

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.

Selinger D, Saxena R, Cheung K, Church G, Rosenow C. (2003). Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Gen Res* **13**: 216–223.

Short SM, Zehr JP. (2007). Nitrogenase gene expression in the Chesapeake Bay Estuary. *Environ Microbiol* **9**: 1591–1596.

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 Natl Acad Sci USA* **105**: 6156–6161.

Toepel J, Welsh E, Summerfield T, Pakrasi H, Sherman L. (2008). Differential transcriptional analysis of the cyanobacterium *Cyanothece* sp strain ATCC 51142 during light-dark and continuous-light growth. *J Bacteriol* **190**: 3904–3913.

Tyson GW, Banfield JF. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **10**: 200–207.

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.

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.

Wells M, Vallis G, Silver E. (1999). Tectonic processes in Papua New Guinea and past productivity in the eastern equatorial Pacific Ocean. *Nature* **398**: 601–604.

Wilcox RM, Fuhrman JA. (1994). Bacterial viruses in coastal seawater: lytic rather than lysogenic production. *Mar Ecol Prog Ser* **114**: 35–45.

Winter C, Herndl GJ, Weinbauer MG. (2004). Diel cycles in viral infection of bacterioplankton in the North Sea. *Aquat Microb Ecol* **35**: 207–216.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T *et al.* (2008). Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**: 1110–1112.

Zehr JP, Bench SR, Mondragon EA, McCarren J, DeLong EF. (2007a). Low genomic diversity in tropical oceanic N_2-fixing cyanobacteria. *Proc Natl Acad Sci USA* **104**: 17807–17812.

Zehr JP, Montoya JP, Hewson I, Mondragon E, Short CM, Hansen A *et al.* (2007b). Nitrogenase gene expression and N_2 fixation in the North Pacific Subtropical Gyre. *Limnol Oceanogr* **52**: 169–183.

Zehr JP, Turner PJ. (2001). Nitrogen fixation: nitrogenase genes and gene expression. *Methods Microbiol* **30**: 271–286.

Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E, Steward G *et al.* (2001). Unicellular cyanobacteria fix N2 in the subtropical North Pacific Ocean. *Nature* **412**: 635–638.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)