The ISME Journal (2011) 5, 1881–1895

www.nature.com/ismej

ORIGINAL ARTICLE

Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton

Elizabeth A Ottesen[1], Roman Marin III[2], Christina M Preston[2], Curtis R Young[1], John P Ryan[2], Christopher A Scholin[2] and Edward F DeLong[1]
[1]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA and* [2]*Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA*

Planktonic microbial activity and community structure is dynamic, and can change dramatically on time scales of hours to days. Yet for logistical reasons, this temporal scale is typically undersampled in the marine environment. In order to facilitate higher-resolution, long-term observation of microbial diversity and activity, we developed a protocol for automated collection and fixation of marine microbes using the Environmental Sample Processor (ESP) platform. The protocol applies a preservative (RNALater) to cells collected on filters, for long-term storage and preservation of total cellular RNA. Microbial samples preserved using this protocol yielded high-quality RNA after 30 days of storage at room temperature, or onboard the ESP at *in situ* temperatures. Pyrosequencing of complementary DNA libraries generated from ESP-collected and preserved samples yielded transcript abundance profiles nearly indistinguishable from those derived from conventionally treated replicate samples. To demonstrate the utility of the method, we used a moored ESP to remotely and autonomously collect Monterey Bay seawater for metatranscriptomic analysis. Community RNA was extracted and pyrosequenced from samples collected at four time points over the course of a single day. In all four samples, the oxygenic photoautotrophs were predominantly eukaryotic, while the bacterial community was dominated by *Polaribacter*-like *Flavobacteria* and a *Rhodobacterales* bacterium sharing high similarity with *Rhodobacterales* sp. HTCC2255. However, each time point was associated with distinct species abundance and gene transcript profiles. These laboratory and field tests confirmed that autonomous collection and preservation is a feasible and useful approach for characterizing the expressed genes and environmental responses of marine microbial communities.
The ISME Journal (2011) **5,** 1881–1895; doi:10.1038/ismej.2011.70; published online 30 June 2011
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: metatranscriptomics; gene expression; automated sampling; marine bacterioplankton; RNA preservation; Monterey Bay

Introduction

Community sequencing techniques have become a prominent tool in microbial ecology. Marine environments have been the focus of major metagenomic surveys, which have provided insight into microbial gene content and community structure (Venter *et al.*, 2004; DeLong *et al.*, 2006; Martin-Cuadrado *et al.*, 2007; Rusch *et al.*, 2007; Feingersch *et al.*, 2010). Metatranscriptomic analyses, with their focus on the transcriptional activity of the community, are also yielding insight into community function and gene regulation (Frias-Lopez *et al.*, 2008; Poretsky *et al.*, 2009; Shi *et al.*, 2009; Hewson *et al.*, 2010). However, current protocols for such studies require manual sample processing and shipboard collections, and as a result sampling schemes are often limited by ship availability, shipboard sampling logistics, and expense. Given the importance of episodic nutrient delivery events in modulating biogeochemical cycles (Fasham *et al.*, 2001; Karl *et al.*, 2001; Karl, 2002), new tools for observation and sampling of microbes are needed to facilitate observation of microbial processes at ecologically meaningful temporal and spatial scales.

The purpose of this study was to develop protocols for automated collection and preservation of samples for transcriptomic analysis using the Environmental Sample Processor (ESP), an automated platform for water sampling and molecular analysis. The ESP is an automated fluid handling system that collects and processes biological samples from seawater (Scholin *et al.*, 2009). Current real-time capabilities include array-based detection of target organisms including harmful algal species (Greenfield *et al.*, 2006, 2008; Haywood *et al.*, 2007), invertebrate larvae (Goffredi *et al.*, 2006; Jones *et al.*,

Correspondence: EF Delong, Division of Biological Engineering and Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Parsons Laboratory 48-427, 15 Vassar Street, Cambridge, MA 02139, USA.
E-mail: delong@MIT.edu
Received 10 January 2011; revised 12 April 2011; accepted 19 April 2011; published online 30 June 2011

2008) and major bacterial and archaeal clades (Preston *et al.*, 2009). Quantitative PCR capabilities are also currently in development (Scholin *et al.*, 2009). The ESP can return samples preserved in a saline–ethanol solution, but sample analysis has been primarily limited to *in situ* hybridization techniques (Goffredi *et al.*, 2006; Greenfield *et al.*, 2006, 2008; Jones *et al.*, 2008). To extend the scope of laboratory analysis of archived samples, new protocols for sample preservation were developed for gene expression analysis and transcriptomics.

In this study, we validated and field-tested protocols for automated collection and preservation of community mRNA from samples of marine bacterioplankton. Together with the ESP platform (Scholin *et al.*, 2009), these protocols enable the autonomous collection of samples and their return to the laboratory for transcriptional analysis. As the ESP also incorporates standard oceanographic instrumentation and the ability to transmit collected data and receive remote commands, environmental context monitoring and event response are also possible. Following successful laboratory tests, our protocols were applied in field deployments. We used the ESP and previously developed metatranscriptomic protocols to analyze microbial community gene expression in Monterey Bay, a coastal system that has been the focus of previous molecular microbial community analyses (Suzuki *et al.*, 2001, 2004; O'Mullan and Ward, 2005; Mincer *et al.*, 2007; Rich *et al.*, 2010). Transcriptomic analysis of ESP-collected surface water samples provided high-resolution sequence data useful in determining the identity, relative abundance, and expressed gene profiles of predominant marine bacterioplankton.

Materials and methods

ESP operation for sample archival

Only methods for ESP sample archival are presented here; for a full description of ESP operation, see Scholin *et al.* (2009) and Roman *et al.* (2007). Samples archived by the ESP for metatranscriptomic analysis were collected in titanium sample 'pucks' (filter holders) containing the following 25 mm diameter filters stacked from top to bottom: a 5 μm Durapore prefilter (Millipore, Billerica, MA, USA), a 0.22 μm Durapore sample filter, 0.45 μm Metricell backing filter (Pall Gelman Corporation, Port Washington, NY, USA) and a 10 μm sintered frit (Chand Eisenmann Metallurgical, Burlington, CT, USA); see Greenfield *et al.* (2006) for details. Pucks are stored in a rotating carousel in the ESP. A robotic mechanism transfers fresh pucks from the carousel to a collection position where they are immobilized and connected to the ESP's sample acquisition and reagent fluid handling system. Collection of samples was achieved by drawing seawater through the stacked filters with repeated pulls of a 25 cc syringe. Between pulls, filtrate was exhausted back to the environment. The instrument maintained a +10 p.s.i. pressure differential across the filter puck throughout sample collection. Filtration continued until the desired sample volume was reached (typically 40–60 min for 1 l seawater), or until the flow rate fell below 25 ml in 2.5 min, after which filtration was terminated and the filtered volume recorded. The material retained on the filter was then preserved with two, 20-min incubations with 1 ml of RNAlater (Ambion, Austin, TX, USA). Following sample collection and preservation, the puck was removed from the collection station and returned to the storage carousel. The sample intake line was then flushed with a 0.2% (v/v) sodium hypochlorite solution, which remained in the line until the next sampling event. Immediately before collection of the next sample, the sample intake line was flushed with a dilute Tween-20 solution (0.05%, v/v), a fresh puck was loaded from the carousel and the sample archival procedure repeated. Pucks remained onboard the instrument (at *in situ* temperatures and under an N_2 gas atmosphere) until the end of the full deployment. After deployment, the 5 and 0.22 μm filters were recovered and stored at −80 °C until use. Metagenomic and metatranscriptomic studies were performed only on the 0.22 μm filters.

To mimic *in situ* sampling conditions, seawater collected for processing in the laboratory was loaded into a dispensing pressure vessel (Millipore), attached to the intake and exhaust valves of the ESP and pressurized to 20 psi, to simulate conditions at ~18 m from the sea surface. Collection and processing then proceeded as described above.

Validation of fixation protocols

To examine the long-term stability of RNA following fixation, a near-shore sample was collected from the Monterey Bay Municipal Pier and prefiltered through a 10 μm Nitex mesh (Wildlife Supply Company, Yulee, FL, USA). The ESP then collected and preserved five replicate samples using the protocol described above. Following collection of all samples, filter pucks were removed from the instrument and stored at room temperature under conditions that mimic deployment (in a sealed container with an N_2 atmosphere and damp paper towels to generate humidity). Pucks were retrieved at 1-week intervals, and the sample filters removed and stored at −80 °C. Following completion of the time series, total RNA was extracted from all 0.22 μm filters simultaneously as described below. The integrity of the recovered RNA was evaluated using a Bioanalyzer high-sensitivity electrophoresis system and the RNA 6000 Pico mRNA protocol (Agilent, Santa Clara, CA, USA).

To evaluate the effect of long-term preservation on metatranscriptomic profiles, an ESP-collected and preserved sample was compared against a sample collected by vacuum filtration. Seawater was

collected from the Santa Cruz (CA, USA) municipal wharf with one aliquot dedicated for ESP processing (as above) and a second for collection using traditional laboratory vacuum filtration. The material collected using the latter method was immediately flash frozen in liquid nitrogen, and stored at −80 °C until extraction. The ESP-processed aliquot of that same sample remained on the instrument during a field test of the ESP at MBARI's station M0 (36.83°N, 121.90°W) from 6 April to 29 April 2009. Following recovery of the instrument, both ESP- and vacuum-collected samples were extracted and used for metatranscriptomic analysis as described below. To further validate collection and preservation protocols, samples collected *in situ* over the course of the deployment were also used for community RNA extraction and the RNA quality examined by size fractionation.

To further evaluate the effect of automated collection and preservation on metatranscriptomic profiles, ESP-collected and preserved samples were compared with a replicate sample processed by peristaltic pump filtration. Water was collected in Monterey Bay (36.7173°N, 122.1147°W, ~9 km from MBARI station M1) at 0013 local time on 8 June 2009 by rosette sampler at 30 m depth. In all, 0.5 l was processed by ship-board ESP with 20-min filtration time and ~30 min from seawater collection to start of fixation. A replicate sample of 0.5 l was processed for RNA extraction using a standard peristaltic pump filtration protocol, as described previously (Frias-Lopez *et al.*, 2008). Following peristaltic pump filtration (10-min filtration time), the 0.22 μm filter was immediately submersed in 300 μl of RNA Later and stored at −80 °C, with a total time of 20 min from seawater collection to preservation. To provide DNA template for synthesis of sample-specific ribosomal RNA (rRNA) subtractive hybridization probes, an additional 9.8 l of water from the same sample was collected for DNA extraction with the same prefilter but using a 0.22 μm pore size Sterivex filter (Millipore). The Sterivex filter was subsequently filled with 2 ml lysis buffer (50 mM Tris-Hcl, 40 mM EDTA and 0.75 M sucrose) and stored at −80 °C.

Field tests in Monterey Bay

Monterey Bay time series samples were collected during a deployment of the ESP at MBARI's station M0 from 14 May to 11 June 2009. Sample pucks from that deployment were recovered and processed (filters separated and placed in sterile tubes at −80 °C) on 12 June. The autonomous underwater vehicle (AUV) *Dorado* was repeatedly deployed from 2 to 4 June 2009, to survey water masses and phytoplankton variability in an area surrounding the ESP mooring. Details of AUV *Dorado* sensors, operation and data processing can be found in Ryan *et al.* (2010b).

DNA for metagenomic sequencing and rRNA probe synthesis was extracted from seawater collected by rosette sampler from a ship in close proximity to the ESP on 2 June 2009 at 0830 local time. A seawater sample collected by the AUV *Dorodo* using the *Gulper* water sampling system (Ryan *et al.*, 2010a) on 4 June was used for DNA extraction and synthesis of probes for rRNA subtraction, but was not sequenced. Both DNA samples were filtered using Sterivex filters and stored in lysis buffer as described in the previous section.

Nucleic acid extraction and subtractive hybridization

Total RNA was extracted from filters as described previously (Frias-Lopez *et al.*, 2008). Briefly, community RNA was extracted using the *mir*Vana kit (Ambion). Turbo DNase (Ambion) was used to remove genomic DNA and the resulting samples purified and concentrated using the RNeasy MinElute cleanup kit (Qiagen, Valencia, CA, USA). RNA extraction yields for all samples are summarized in Table 1.

DNA was extracted and purified using the QuickGene 610 l system (Fujifilm, Tokyo, Japan) and DNA Tissue Kit L with a modified lysis protocol. In all, 2 mg of lysozyme in lysis buffer (described above) was added to thawed Sterivex filters, which were incubated with rotation to mix at 37 °C for 45 min. In total, 100 μl each of the kit buffers DET and MDT were added, and the sample incubated at 55 °C for 2 h with rotation. The lysate was decanted from the filter, 2 ml LDT solution was added, and incubated at 55 °C for a further 15 min. Finally, 2.7 ml EtOH was added, and the sample loaded onto the QuickGene instrument for purification according to the DNA Tissue protocol.

Antisense rRNA probes for subtractive hybridization were prepared as described previously (Stewart *et al.*, 2010). In brief, universal bacterial, archaeal and eukaryotic small subunit (SSU) and large

Table 1 Samples and RNA extraction efficiencies

Sample	*Vol (ml)*[a]	*Yield (μg)*
Monterey wharf—time 0	500	3.2
Monterey wharf—1 week	500	5.0
Monterey wharf—2 weeks	500	4.7
Monterey wharf—3 weeks	500	5.1
Monterey wharf—4 weeks	500	4.8
Station M0—4/7/09	1000	1.4
Station M0—4/9/09	1000	2.5
Station M0—4/12/09	1000	1.9
Station M0—4/16/09	1000	2.1
Station M0—4/20/09	1000	2.2
Station M0—4/25/09	1000	2.4
Santa Cruz wharf—vacuum	300	4.0
Santa Cruz wharf—ESP	1000	11.8
Station M1—peristaltic pump	500	0.7
Station M1—ESP	500	0.9
Station M0—6/2/09 0500 hours	1000	0.7
Station M0—6/2/09 1000 hours	1000	2.5
Station M0—6/2/09 1800 hours	1000	0.6
Station M0—6/2/09 2200 hours	1000	1.2

Abbreviation: ESP, Environmental Sample Processor.
[a]Total volume of seawater filtered.

subunit (LSU) primers with attached T7 promoters were used in PCR reactions with Herculase II Fusion DNA polymerase (Agilent) to generate templates for antisense-rRNA probe synthesis. Biotin-labeled antisense rRNA probes were generated from the PCR products using the MegaScript T7 kit (Ambion). The Santa Cruz municipal pier samples used for the vacuum/ESP comparison lacked a paired DNA sample, so PCR was instead performed on first-strand complementary DNA prepared with the SuperScript III kit (Invitrogen, Carlsbad, CA, USA) with random primers and 40 ng of the ESP-collected total RNA sample.

Subtractive hybridization was carried out using published protocols (Stewart *et al*., 2010). Hybridization reactions were carried out on 200 ng of total RNA and sample-specific antisense rRNA probe mixtures. For the Santa Cruz wharf vacuum/ESP comparison, the original, two-step hybridization protocol was followed, using 200 ng total RNA and 250 ng each of the SSU and LSU rRNA bacterial probes. For the remaining samples, the amended protocol presented in the Supplementary materials of Stewart *et al*. (2010) was utilized. Station M1 samples utilized probes synthesized from the paired DNA sample, with 400 ng each bacterial SSU and LSU, 200 ng each archaeal SSU and LSU, and 300 ng each eukaryotic SSU and LSU. Station M0 samples utilized probes generated from both the 2 June (rosette-collected) and 4 June (AUV *Dorado*-collected) DNA samples at 0.75 × concentration (for example, 300 ng 2 June bacterial SSU + 300 ng 4 June bacterial SSU). For the station M0 samples, the archaeal SSU rRNA primers exhibited nonspecific amplification, and as a result no archaeal SSU rRNA probes were included. rRNA probe duplexes were subsequently bound to Strepavidin-coated magnetic beads, and removed from the total RNA preparation. Following this procedure, samples were purified and concentrated using the RNeasy MinElute cleanup kit.

rRNA-subtracted (and unsubtracted total RNA) samples were amplified as described previously (Frias-Lopez *et al*., 2008). In brief, RNA was amplified using the MessageAmp II Bacteria kit (Ambion) and a poly-T primer with an additional 5′ *Bpm*I restriction site. First-strand complementary DNA was synthesized from the amplified RNA using random primers and SuperScript III (Invitrogen), second-strand complementary DNA synthesized using DNA pol I, *Escherichia coli* ligase, and T4 DNA polymerase (Invitrogen) and remaining poly-A tails removed by digestion with *Bpm*I (New England Biolabs, Ipswich, MA, USA).

All samples were sequenced using the 454 Genome Sequencer (Roche Applied Science, Indianapolis, IN, USA). Metatranscriptomic samples were prepared and sequenced using the GS FLX protocol, and the metagenomic DNA sample using the GS FLX Titanium protocol. Library preparation and sequencing was carried out according to the manufacturer's protocols. Sequences have been submitted to the NCBI Sequence Read Archive under sample accession numbers SRS167142 (Santa Cruz wharf samples), SRS183627 (Station M1 samples) and SRS010641 (Station M0 samples).

Sequence processing and annotation

Sequencing and annotation statistics for each sample are summarized in Table 2. Sequences derived from rRNA were identified using BLASTN with a bit score cutoff of 50 against a database composed of 5S, 16S, 18S, 23S and 28S rRNA sequences from

Table 2 Read numbers and statistics

Sample	*Reads*[a]	*% rRNA*[b]	*Non-rRNA reads*		
			Replicates[c]	*nr Hits*[d]	*KEGG hits*[d]
Santa Cruz wharf: vacuum	179 643	80%	20%	17 988	15 869
Santa Cruz wharf: ESP	160 364	82%	18%	14 702	12 850
Station M1: peristaltic pump	118 595	41%	26%	20 981	18 677
Station M1: ESP	203 574	42%	11%	41 295	37 086
0500 hours rRNA-subtracted	248 016	33%	4%	82 387	69 157
0500 hours unsubtracted	298 380	91%	7%	11 802	10 089
1000 hours rRNA-subtracted	102 024	40%	17%	25 197	22 250
1000 hours unsubtracted	149 186	82%	27%	9979	8612
1800 hours rRNA-subtracted	238 635	38%	17%	54 040	46 253
1800 hours Unsubtracted	232 248	83%	12%	15 694	13 156
2200 hours rRNA-subtracted	235 339	35%	13%	52 069	43 956
2200 hours unsubtracted	202 650	82%	24%	10 701	8890
DNA[e]	1 535 834	0.4%	0.97%	1 035 676	956 510

Abbreviations: KEGG, Kyoto Encyclopedia of Genes and Genomes; nr, non-redundant; rRNA, ribosomal RNA.
[a]Total number of sequence reads passing quality filters.
[b]Percentage of total pyrosequencing reads with significant (bitscore > 50) BLASTN hits to prokaryotic and eukaryotic rRNA (16S, 18S, 23S, 28S, 5S).
[c]Percentage on non-rRNA reads identified as artificial replicates (99% identity, 1-bp length difference) and removed.
[d]Non-replicate, non-rRNA reads with significant (bitscore > 50) BLASTX hits to proteins in the NCBI nr or KEGG Genes databases.
[e]Metagenomic data set, sequenced using GS FLX Titanium chemistry rather than GS FLX.

microbial genomes and the SILVA LSU and SSU databases (http://www.arb-silva.de). Non-rRNA sequences with identical start sites (first 3 bp), 99% identity and <1-bp length difference were identified as probable artificially duplicated sequences (Stewart *et al.*, 2010) and removed using the cd-hit program (Li and Godzik, 2006) and scripts developed by Gomez-Alvarez *et al.* (2009). Non-rRNA sequences were compared with the 3 November 2009 version of NCBI's non-redundant (nr) protein database reference databases using BLASTX. Unless otherwise specified, a bit score cutoff of 50 was used to identify significant matches to the database.

For pairwise comparisons of metatranscriptomic profiles, each sequence was assigned to a single reference gene in the NCBI-nr database based on BLASTX alignment bit score. When a single sequence aligned equally well to multiple potential reference genes, it was assigned to the reference gene that was most frequently identified in the data set. Reference gene abundances were compared using the cumulative form of the AC Test (Audic and Claverie, 1997) and a false discovery rate correction for multiple comparisons (Benjamini and Hochberg, 1995); details of how these tests were conducted are in the Supplementary online materials.

The MEGAN program (Huson *et al.*, 2007) was used to assign sequences to a higher-order taxonomy. All analyses used a bit score cutoff of 50 and database matches with bit scores within 10% of the top-scoring hit. Unique non-rRNA sequences from both subtracted and unsubtracted sequence data sets were pooled and assigned to the NCBI taxonomy based on the results of a BLASTX search of the NCBI-nr database. rRNA genes were assigned to the NCBI taxonomy using manually curated rRNA databases constructed based on the approach used by Urich *et al.* (2008) as described in the Supplementary online materials. Only rRNA-unsubtracted samples were utilized in taxonomic analysis of putative rRNA sequences. Owing to the higher copy number and lower genetic diversity of rRNA genes compared with mRNA genes, rRNA taxonomies were constructed without removal of duplicates.

Analyses of gene expression in the *Rhodobacterales* sp. HTCC2255 and *Polaribacter* taxonomic groups used all sequences for which the taxon in question was among the top-scoring database matches (all matches with bit scores equal to the highest-scoring alignment were considered). For composite analyses of *Polaribacter* expression, read counts and annotations for genes shared by both *Polaribacter irgensii* 23-P and *Polaribacter* sp. MED152 were combined. Shared genes were defined as reciprocal best BLASTP hits with e-value $<1\times10^{-5}$, and at least 80% alignment coverage for both genes. The draft version of the *Rhodobacterales* sp. HTCC2255 genome has a large number of contigs that were annotated as contamination and removed from the genome scaffolds. These contigs (and the 2267 coding sequence identified within them) were not identified in surveys of Monterey Bay and were excluded from our genome analyses.

Functional comparison using KEGG gene categories

For comparisons of differences in community function, sequences were assigned to functional categories based on Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology groups (Kanehisa and Goto, 2000). For bulk community-level analyses, sequence reads were assigned a single reference gene in the 7 November 2009 version of the KEGG database as described above for the NCBI-nr database, with the additional weighting factor that proteins that were assigned to a KEGG ortholog category were preferred when choosing between multiple matches with identical alignment scores and frequencies in the data set. For analyses focused on *Polaribacter* and HTCC2255, the KAAS automated annotation pipeline (Moriya *et al.*, 2007) was used to annotate each reference genome. KEGG ortholog counts for each taxon were then compiled using all sequences for which the taxon in question is among the top-scoring hits by BLASTX against NCBI-nr database (all hits with bit scores equal to the highest-scoring alignment).

KEGG Pathway counts were generated based on the total number of sequences assigned to KO annotations within that pathway (because of functional overlap, some orthologs were represented in multiple pathways). All comparisons used KEGG Pathway rather than BRITE hierarchies, and pathways within the 'Human Diseases' or 'Organismal Systems' hierarchies were not analyzed. Both ortholog and pathway counts for each sample were normalized to the total number of non-rRNA sequences with significant hits to the KEGG database (for bulk community analyses) or to the total number of sequences assigned to the taxon in question (analyses focused on HTCC2255 or *Polaribacter*). Statistical evaluation of KEGG pathway abundances used in-house R scripts utilizing a methodology explained in detail in the Supplementary online materials.

Results and Discussion

Validation of ESP preservation protocols

Following collection and preservation on the ESP, marine bacterioplankton samples were found to be stable for at least 4 weeks (Figure 1). In one experiment, replicate samples of a near-shore surface water sample were filtered and preserved using an ESP in the laboratory and stored under ESP-like conditions at room temperature (high humidity, N_2 atmosphere). Each of the five samples that were processed at weekly intervals had similar yields (Table 1), and gel electrophoresis indicated comparable RNA quality (discrete 16S and 23S rRNA peaks with no obvious degradation) (Figure 1a). In a

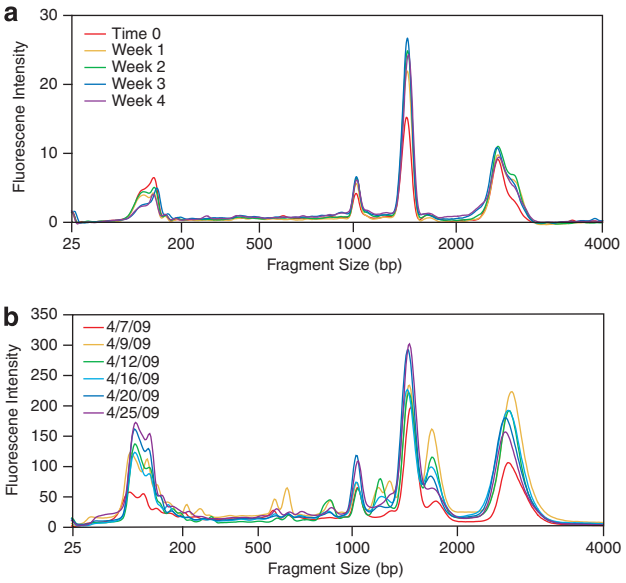

Figure 1 Size fractionation of total RNA extracted from ESP-collected and preserved samples. All samples were diluted to approximately equal concentrations before analysis to facilitate comparison of RNA quality. (**a**) Total RNA extracted from replicate surface water samples collected and preserved using the ESP and stored at room temperature under conditions that simulate a deployed instrument (high humidity, N_2 atmosphere). (**b**) Total RNA extracted from samples collected throughout a single deployment of the ESP at Monterey Bay station M0.

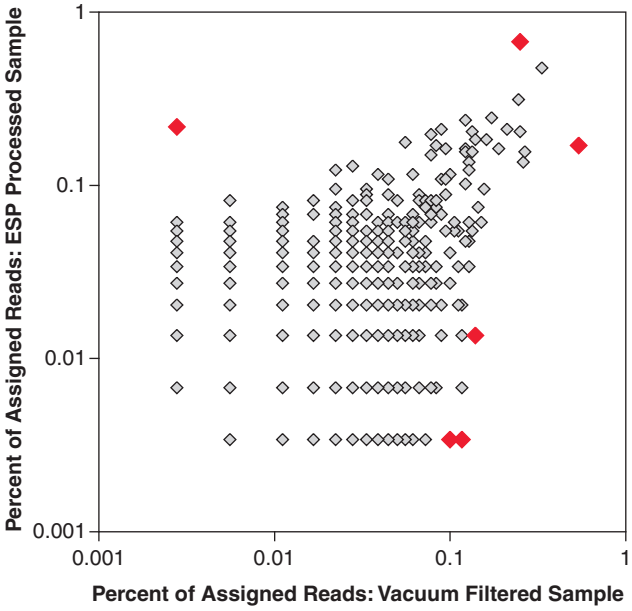

Figure 2 Metatranscriptomic analysis of ESP-collected and preserved samples. The abundance of NCBI-nr reference genes is shown for a sample collected by ESP and retained on the instrument for a 29-day deployment in Monterey Bay and a replicate sample collected by vacuum filtration and flash frozen. For visualization purposes, reference genes with 0 assigned sequences in one of the two data sets are shown as 0.5 sequence reads. Reference genes with significantly different abundances in the two data sets (false discovery rate-corrected *P*-value <0.05) are marked in red. Accession numbers and *P*-values of significantly different reference genes are listed in Table 3.

separate field test, RNA was extracted from cells that had been filtered and fixed by the ESP over the course of a 29-day deployment in Monterey Bay. These *in situ* filtered and fixed cells also yielded high yields of RNA (Table 1) with comparable high quality over all time points sampled (Figure 1b). We conclude that the ESP sampling and preservation protocol provides material that is sufficient for downstream extraction of high quality, high yield total cellular RNA.

To evaluate preservation of mRNA, a near-shore surface water sample taken from the municipal wharf in Santa Cruz was processed via the ESP for filtration and fixation, and then subjected to 29 days of storage on board the instrument during a deployment in Monterey Bay. This sample, and a control sample that was filtered in parallel by vacuum filtration and immediately flash frozen, were used to prepare a community transcriptome pyrosequencing library. Transcript abundance profiles for the conventionally processed flash-frozen sample and the 29-day ESP-preserved sample were highly similar, with only 6 out of 17 284 transcripts showing significantly different abundances in the two samples (Figure 2, Table 3). An ESP-processed and preserved sample collected near MBARI station M1 in Monterey Bay was also compared with a replicate sample processed by rapid peristaltic pump filtration. This pair of samples again yielded similar transcript abundance profiles, with significant differences for 28 out of 35 036 reference genes (Supplementary Figure S1, Supplementary Table S1).

Both of our comparisons of ESP- and manually processed samples yielded numbers of significantly different references, and percentages of sequence reads mapping to significantly different transcripts, which were within the range typically observed for technically replicated metatranscriptomic profiles (Stewart *et al.*, 2010; see Supplementary Table S2 for comparisons recalculated using our updated database and cutoffs). In fact, the sample subjected to long-term storage on board the instrument (Santa Cruz wharf ESP) was more similar to its manually processed control (in terms of taxonomic composition and significantly different transcripts), than were the paired ship-board samples (Station M1 PP and ESP), in which the ESP-collected sample was immediately removed from the instrument and stored at $-80\,^{\circ}C$ (Table 4). This suggests that the water filtration method used may have more impact on the observed metatranscriptomic profile than the length of time between preservation and removal to low-temperature, permanent storage.

Metatranscriptomic analysis of Monterey Bay surface water samples

Four time points during a deployment of the ESP in Monterey Bay surface waters were chosen for transcriptomic sequencing. These time points represented transcriptomic profiles of microbial

Table 3 NCBI-nr reference genes with significantly different abundances in metatranscriptomes generated from vacuum filtered vs ESP-filtered and preserved replicate samples[a]

NCBI-nr reference gene	*Vacuum*[b]	*ESP*[b]	P-*value*
AAM48736 antenna pigment protein, alpha chain (uncultured marine proteobacterium)	ND	0.22%	6.1×10^{-08}
ZP_01224258 hypothetical protein (gamma proteobacterium HTCC2207)	0.25%	0.67%	4.1×10^{-05}
ZP_01447883 branched-chain amino acid ABC transporter, periplasmic substrate-binding protein (*Rhodobacterales* sp. HTCC2255)	0.54%	0.17%	5.4×10^{-05}
ZP_03559919 50S ribosomal subunit protein L3 (*Glaciecola* sp. HTCC2999)	0.12%	ND	8.5×10^{-03}
ZP_03559922 50S ribosomal protein L2 (*Glaciecola* sp. HTCC2999)	0.10%	ND	0.041
ZP_01447418 glutamate synthase large subunit (*Rhodobacterales* sp. HTCC2255)	0.14%	0.01%	0.043

Abbreviations: ESP, Environmental Sample Processor; FDR, false discovery rate; ND, not determined; nr, non-redundant.
[a]Reference genes with FDR-corrected *P*-value >0.05.
[b]The percentage of unique, non-rRNA reads with significant NCBI-nr database hits mapping to each reference gene in vacuum-filtered and the ESP-processed and preserved samples.

Table 4 Comparison of metatranscriptomic profiles from manually processed and ESP-collected and preserved samples

DS compared[a]		*NCBI-nr taxa*[b]			*WI*[c]	*NCBI-nr references*[d]			*Sig. diff. Refs*[e]	*% hits in sig. diff. refs*[f]	
DS 1	*DS 2*	*DS1*	*DS2*	*Both*		*DS1*	*DS2*	*Total*		*DS1*	*DS2*
SC Vac	SC ESP	785	666	448	0.89	10 723	9386	17 284	6	1.2%	1.1%
M1 PP	M1 ESP	1280	1613	1028	0.82	13 754	25 211	35 036	28	1.6%	1.3%

Abbreviations: ESP, Environmental Sample Processor; nr, non-redundant.
[a]Data sets used in pairwise comparisons (SC Vac and ESP: Santa Cruz wharf, vacuum-filtered and flash frozen sample and ESP-processed and preserved sample, stored on board ESP for a 30-day deployment in Monterey Bay; M1 PP and ESP: Monterey Bay water collected by rosette sampler, filtered immediately by peristaltic pump or processed and preserved by ESP. Additional pairwise comparisons are listed in Supplementaary Table S2.
[b]Number of taxa (NCBI-nr taxonomy ID's) with one or more uniquely assigned sequences (reads with exactly one top-scoring database match).
[c]Whittaker's index of association (Whittaker 1952) for NCBI-nr taxon counts.
[d]Number of NCBI-nr reference genes with one or more mapped reads.
[e]NCBI-nr reference genes with significantly different abundances in the two data sets.
[f]Percentage of sequences with NCBI-nr hits that map to reference genes with significantly different abundances.

communities at dawn (0500–0600 hours), late morning (1000–1100 hours), dusk (1800–1900 hours) and night (2200–2300 hours) on 2 June 2009. As the ESP was deployed in a moored configuration, these samples represented distinct microbial populations from water masses with different physical and chemical conditions (Supplementary Figures S2 and S3, Supplementary Table S3). To provide genomic context and additional information on population structure, community DNA was extracted and sequenced from a 10 l water sample collected near the ESP on 2 June 2009.

Taxonomic composition of station M0 metagenomic and metatranscriptomic samples

Metagenomic and metatranscriptomic sequences were assigned to taxonomic groups using MEGAN (Huson *et al*., 2007) to analyze both protein-coding regions and rRNA SSU and LSU sequences (Figure 3). The metagenomic data set showed a similar taxonomic composition among all transcript types, with a primarily bacterial community dominated by flavobacteria and alpha- and gamma-proteobacteria. Cyanobacterial sequences represented only a small proportion of metagenomic and metatranscriptomic libraries, suggesting that primary production in these samples was dominated by eukaryotic phytoplankton. A diversity of sequences from eukaryotic nano- and picoplankton capable of passing through the 5 μm prefilter were detected in the metatranscriptomic samples. SSU rRNA-based analysis showed a higher representation of eukaryotic organisms in metatranscriptomic samples than in the metagenome, consistent with previous observations that eukaryotic picoplankton exhibit higher transcriptional activity relative to genomic abundance than bacteria (Man-Aharonovich *et al*., 2010). In taxonomic analyses of LSU rRNA and coding sequences, eukaryotes represented a much smaller proportion of assigned sequences than expected based on the SSU rRNA results, but this may be due to limited database coverage of marine picoeukaryotes.

The annotated protein-coding transcript pool contained two particularly abundant organisms/groups. The first was an alpha proteobacterium closely related to *Rhodobacterales* sp. HTCC2255. HTCC2255 is a proteorhodopsin-containing *Roseobacter* of the NAC11-7 clade (Newton *et al*., 2010; Yooseph *et al*., 2010) isolated by dilution-to-extinction near the coast of Oregon as part of the high-throughput culture collection (Connon and Giovannoni, 2002). HTCC2255 was also the most frequently identified taxon in both of the water samples used for validation studies, constituting $>25\%$ of

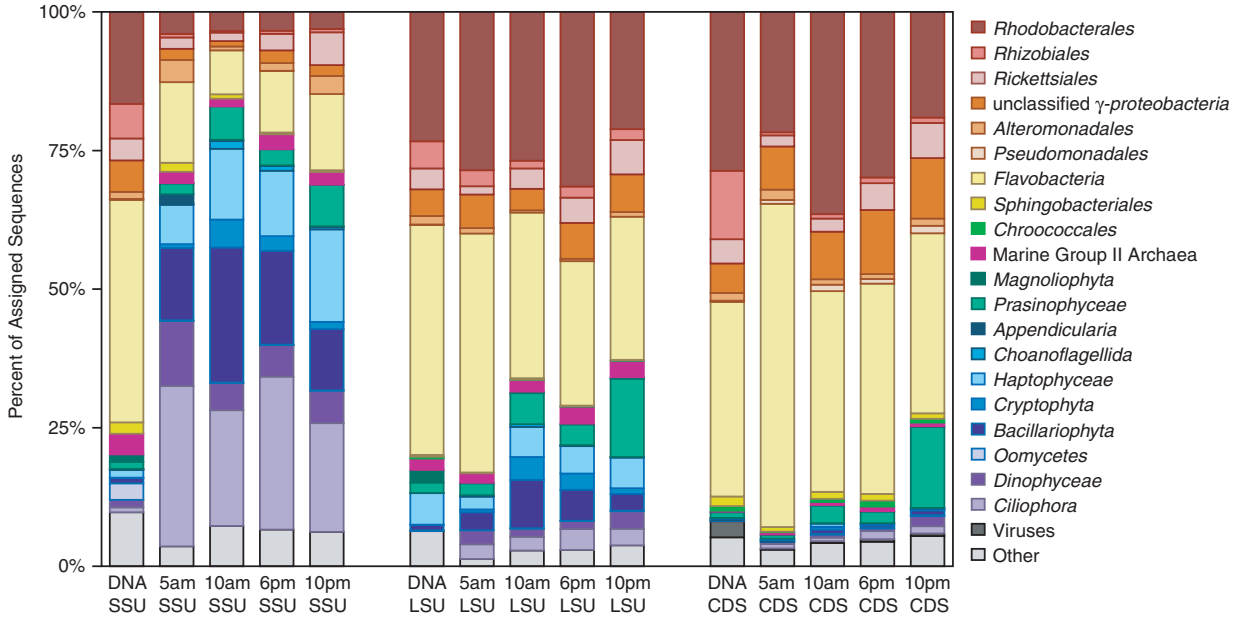

Figure 3 Relative abundance of major taxonomic groups in metatranscriptomic and metagenomic samples. Sequences were assigned to the NCBI taxonomy using the MEGAN program (Huson *et al.*, 2007), (bitscore >50, top 10% of hits). Taxonomic analyses of SSU and LSU rRNA sequences are based on unsubtracted RNA samples only. Coding sequence taxonomy generated from the combined non-replicate, non-rRNA fraction of both unsubtracted and subtracted RNA. Groups representing >1% of assignable sequences in one or more samples are shown, those representing <1% of sequences in all samples are included in the 'other' category, and those assigned at lower taxonomic levels are not shown.

assignable non-rRNA sequences in the two Santa Cruz wharf transcriptomes and ~10% of assignable sequences from the Station M1 transcriptomes (Supplementary Figure S4). The other highly abundant group of sequences appears to represent one or more *Flavobacteria*, which are most closely related to the two sequenced *Polaribacter* species, *Polaribacter* sp. MED152 (González *et al.*, 2008) and *P. irgensii* 23-P (Gosink *et al.*, 1998). In all, 8–9% of sequences from the Santa Cruz wharf, but <2% of sequences from the Station M1 samples, were also assigned to one of these two *Polaribacter* genomes.

Other bacteria representing significant proportions of the community included gamma proteobacterium HTCC2207 (Stingl *et al.*, 2007), a diversity of *Flavobacteria* including *Flavobacteria* sp. MS024-2A (Woyke *et al.*, 2009) and eubacterium SCB49 (Yooseph *et al.*, 2010). *Ostreococcus* (Derelle *et al.*, 2006; Palenik *et al.*, 2007) and *Micromonas* (Worden *et al.*, 2009) were among the most abundant picoeukaryotes identified in annotated protein-coding transcripts, but rRNA-based analysis suggests a large diversity of eukaryotes without sequenced genomes were present in the samples.

Across the four time points, the community composition remained broadly similar, with Whittaker's index of association (Whittaker, 1952) values between 0.68–0.85 for NCBI taxon counts (Supplementary Table S2). The five most abundant bacterial taxa dominated in all four samples (Supplementary Figure S5), although the picoeukaryote *Ostreococcus lucimarinus* showed more dramatic shifts in abundance between the different samples. All taxa present at >0.02% of uniquely assignable transcripts in at least one sample were detected in all four samples. However, the relative RNA abundances of these organisms varied dramatically. One of the largest changes was in *O. lucimarinus*-like sequences, which ranged from 0.2% to 6.2% of transcripts, while MED152-like sequences ranged from 5.5% to 16.8% and HTCC2255-like sequences from 9.12% to 16.7%. This variability is greater than the differences seen in comparisons of abundant taxa in transcriptomic profiles at different depths in the euphotic zone at station ALOHA (Shi *et al.*, 2010) or of day and night samples collected 2 days apart at a single station in the North Pacific Subtropical gyre (Poretsky *et al.*, 2009). However, these comparisons cannot independently discriminate between changes in organism abundance and changes in activity level (cellular RNA content). For example, in incubation experiments where dimethylsulfoniopropionate was added to oligotrophic waters, noticeable shifts in transcript taxonomic composition were observed within 30 min, presumably too short an interval for extensive growth (Vila-Costa *et al.*, 2010).

Rhodobacterales sp. HTCC2255

HTCC2255-like transcripts were abundant at all four time points. HTCC2255 was among the top-scoring alignments in the NCBI-nr database, accounting for a total of 12%, 14%, 15% and 8% of assignable

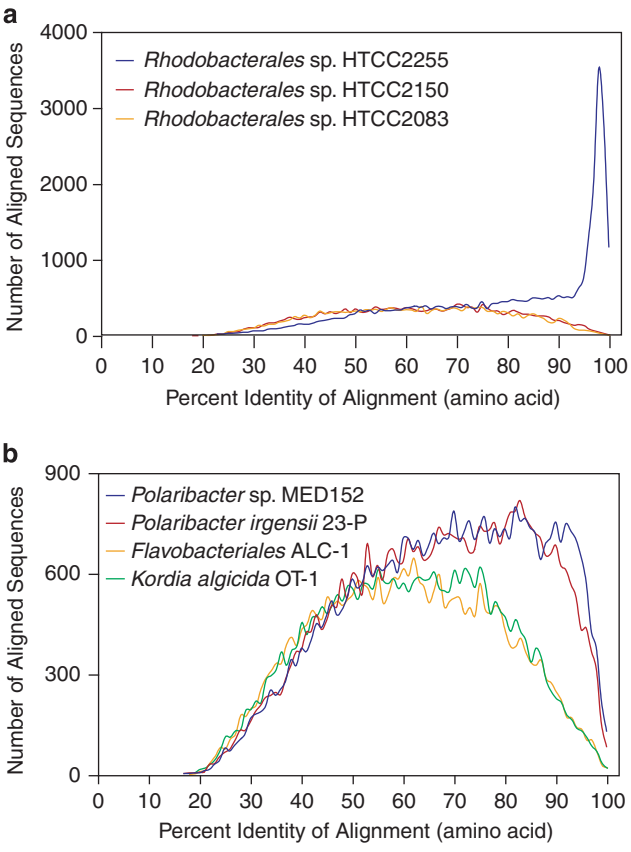

Figure 4 Percent identity histograms for sequences assigned to *Rhodobacterales* sp. HTCC2255 or *Polaribacter*. A global percent identity (percent amino acid similarityfraction of read covered) was calculated for all significant (bitscore >50) BLASTX hits in the NCBI-nr database for metatranscriptomic libraries from Station M0, and the number of sequences aligning at each percent identity determined. (**a**) Sequences for which *Rhodobacterales* sp. HTCC2255 is among the top-scoring hits in the NCBI-nr database. To show the specificity with which these sequences were mapped, additional database hits passing the bitscore threshold were examined, and percent identity histograms generated for the two most frequently identified taxa, *Rhodobacterales* spp. HTCC2150 and HTCC2083. (**b**) Sequences with at least one top hit to either of the two *Polaribacter* reference genomes. Significant alignments of *Polaribacter*-assigned reads to *Flavobacteriales* sp. ALC-1 and *Kordia algicida* OT-1 are also shown.

sequences in the 0500, 1000, 1800 and 2200 hours metatranscriptomic data sets, and a corresponding 9%, 10%, 11% and 6% of sequences could be unambiguously assigned to HTCC2255 by the Megan LCA algorithm. HTCC2255-like protein coding genes averaged 95% identity to the sequenced strain at the amino acid level (Figure 4). In all, 2092 out of 2240 of HTCC2255 coding sequences were identified in at least one of the four Station M0 transcriptomes, including proteorhodopsin, the full citrate cycle, and genes involved in dimethylsulfoniopropionate degradation and sulfur oxidation. The draft genome sequence of HTCC2255 has two scaffolds, NZ_DS022282 and NZ_DS022288, of which the smaller, NZ_DS022288 has much lower coverage in both the metagenome and the transcriptomes (Supplementary Figure S6). This may represent a less conserved genomic island, or a plasmid that is missing from the Monterey Bay genotype. Alternatively, the metadata associated with the draft genome notes that contaminating *gammaproteobacteria* sequences were present in the raw sequence data and removed from the scaffolds during draft assembly; NZ_DS022288 may represent a scaffold that was inappropriately assigned to HTCC2255.

HTCC2255 appears to be a very common component of the Monterey Bay microbial community. Several bacterial artificial chromosome sequences previously isolated from Monterey Bay (Rich *et al.*, 2008, 2010) share a high percent identity and synteny with HTCC2255, and probes targeting these bacterial artificial chromosome clones and the HTCC2255 reference genome identified these organisms in 97–100% of surface water samples spanning a 4-year time series at Monterey Bay station M1 (Rich *et al.*, 2010). Additionally, the HTCC2255-like bacterial artificial chromosome EB000-55B11 was also detected in near-shore samples from Woods Hole, MA, USA during experiments with the prototype genome-proxy array (Rich *et al.*, 2008). In general, HTCC2255-like organisms appear to be widely present in marine communities (Yooseph *et al.*, 2010), but may be particularly abundant in near-shore waters. The reference strain was isolated off the Oregon coast, relatives have been detected off both the California and Massachusetts coasts, and HTCC2255-like sequences were reported as the most abundant sequence type in a proteorhodopsin library from the North Sea (Riedel *et al.*, 2010).

Polaribacter-like sequences

Polaribacter-like transcripts were identified at all four time points, but were most abundant in the 0500 hours sample. A *Polaribacter*-derived sequence was among the top-scoring alignments in the NCBI-nr database for 27%, 12%, 12% and 9% of assignable sequences in the 0500, 1000, 1800 and 2200 hours metatranscriptomic data sets, and a corresponding 15%, 6%, 5% and 5% were unambiguously assigned to this genus by the Megan LCA algorithm. These sequences appear to represent one or more unsequenced organisms related to *Polaribacter*, as sequence reads with top database hits to a *Polaribacter* sp. averaged 82% amino acid identity to *P. irgensii* 23P and 83% identity to *Polaribacter* sp. MED152 (Figure 4). The two sequenced organisms average 72% amino acid identity among their shared genes (defined as reciprocal best hits with e-value $<1\times10^{-5}$ and 80% of the gene aligned). In all, 1519 out of 1636 shared *Polaribacter* genes were identified in one or more of the four transcriptomic samples, while 361 of 974 MED152-specific and 259 of 920 23P-specific genes were identified (Supplementary Figures S7 and S8).

There is less previous evidence for *Polaribacter* as a common component of the Monterey Bay bacterioplankton community than there is for HTCC2255.

Polaribacter-like sequences were not detected in experiments using genome proxy arrays to examine community structure at Monterey Bay station M1, despite inclusion of both sequenced *Polaribacter* genotypes in the array (Rich *et al.*, 2010). However, the sequences recovered in this study averaged ~82% amino acid identity to the reference strains, which is too genetically dissimilar to show strong hybridization signal on the array (Rich *et al.*, 2008, 2010). Additionally, our study used a 5 μm prefilter during sample collection, while the array experiments used a 1.6 μm prefilter, which may change the representation of larger and/or particle-attached bacterial cells. *Polaribacter*-like sequences were identified in 16S rRNA libraries prepared from samples collected at station M0 in September–October 2004, during development of their sandwich hybridization assay (Preston *et al.*, 2009). Other studies of Monterey Bay have not specifically examined abundance of *Flavobacteria* or *Polaribacter*, although Suzuki *et al.* (2004) identified the *Cytophaga-Flavobacteria-Bacteroides* group as representing 8.5% of bacteria in a surface water sample. However, *Flavobacteria* and the *Bacteroidetes* as a whole is thought to have a major role in degradation of particulate and high molecular weight dissolved organic matter in the ocean (Kirchman, 2002), and *Polaribacter* was found to be the most abundant *Flavobacterial* group across a transect in the North Atlantic (Gómez-Pereira *et al.*, 2010).

Nutrient acquisition strategies of Monterey Bay Rhodobacterales and Polaribacter

Although both HTCC2255 and the two *Polaribacter* reference strains are proteorhodopsin-bearing heterotrophs, their genome characteristics and transcriptional activity are consistent with distinctly different nutrient acquisition strategies. Among the most highly expressed HTCC2255-like genes were substrate binding proteins associated with tripartite ATP-independent periplasmic (TRAP) and ATP-binding cassette (ABC) transporters of amino acids, sugars and sugar alcohols (Supplementary Tables S4 and S5). In contrast, González *et al.* (2008) found that *Polaribacter* sp. MED152 carries relatively few transporters for free amino acids or sugars, and no sugar-specific ABC transporters. Both *Polaribacter* genomes appear to lack TRAP and TRAP-T transporters, and few transcripts mapped to those ABC transporters they do carry (<1% of *Polaribacter*-like sequences in contrast to 10–14% HTCC2255-like sequences; details in Supplementary Table S6). Similarly, in an examination of transporters in a coastal transcriptome, Poretsky *et al.* (2010) found an abundance of *Rhodobacterales* and SAR11 associated ABC- and TRAP-related transcripts in coastal environmental transcriptomes, and relatively few *Flavobacterales* associated sequences. A metaproteomic investigation of SAR11 in the Sargasso Sea (Sowell *et al.*, 2009) found that transport functions similarly dominated the proteome of that alpha proteobacterium, with the most abundant proteins being ABC and TRAP transporters. This is consistent with previous studies showing that *Alphaproteobacteria* dominate uptake of amino acids and monomers, while *Bacteroidetes* specialize in utilization of polymers (Cottrell and Kirchman, 2000; Kirchman, 2002).

The most abundant group of transport-related transcripts within the *Polaribacter*-associated sequences are TonB-dependent/ligand gated channels (Supplementary Table S7). TonB-dependent channels were also among the most highly expressed proteins from gamma proteobacterium HTCC2207 and flavobacterium MS024-2A. *Rhodobacterales* sp. HTCC2255 appears to lack a TonB system; the draft genome contained no significant hits to the pfam profiles of either TonB or the two TonB-dependent receptor domains. TonB-dependent transporters from a variety of taxonomic groups were the most abundant family of membrane proteins identified in a metaproteomic analysis of samples from the South Atlantic (Morris *et al.*, 2010). TonB-related proteins were also identified as among the most abundant transcripts assigned to *Idiomarinaceae* and *Alteromonadacea* in a marine microcosm enriched with high molecular weight dissolved organic matter (McCarren *et al.*, 2010). As mentioned above, ABC transporters from *Flavobacterales* were not found to be abundant in a coastal metatranscriptome (Poretsky *et al.*, 2010). However, these investigators did mention an abundance of *Flavobacterales* transporters for inorganic compounds, and COG1629, which includes some TonB transporters, is included in this group under the COG classification scheme. TonB-dependent channels were originally identified in the context of iron transport, but have since been associated with the transport of a large variety of compounds (Schauer *et al.*, 2008). Of particular interest is their newly recognized association with degradation of polymers and complex carbohydrates (Blanvillain *et al.*, 2007; Martens *et al.*, 2009). Several of the TonB-dependent channels in *Polaribacter* sp, MED152 were associated with predicted peptidases and glycosyl hydrolases (González *et al.*, 2008), suggesting they may be involved in utilization of high molecular weight substrates in this organism.

Differences in functional and metabolic profiles of metatranscriptomic samples

In order to examine transcript abundance dynamics, we used KEGG pathways to functionally profile the four transcriptomic samples. KEGG profiles were generated for bulk community data, and individually for *Rhodobacterales* sp. HTCC2255-like and *Polaribacter*-like transcripts. Metabolic genes represented the most abundant class of annotated transcripts, although they were out-numbered by unassigned transcripts for both the bulk community

Table 5 Percentage of transcripts assigned to different KEGG functional categories

Category	*Bulk community*[a]				*HTCC2255*[b]				*Polaribacter*[c]			
	0500 hours	*1000 hours*	*1800 hours*	*2200 hours*	*0500 hours*	*1000 hours*	*1800 hours*	*2200 hours*	*0500 hours*	*1000 hours*	*1800 hours*	*2200 hours*
Unassigned	51	49	48	54	31	30	33	40	56	54	51	60
Metabolism	30	35	32	29	35	41	37	31	27	32	29	27
Genetic information processing	18	13	17	14	23	20	20	15	17	15	21	13
Environmental information processing	4.4	4.5	4.9	4.7	15	14	16	19	3.0	2.3	3.2	2.6
Cellular processes	2.3	2.6	2.6	3.3	1.9	2.4	2.4	2.5	2.0	2.3	1.9	2.4

Abbreviation: KEGG, Kyoto Encyclopedia of Genes and Genomes.
[a]Percentage of sequences with significant hits in the KEGG genes database.
[b]Percentage of total sequences assigned to *Rhodobacterales* sp. HTCC2255.
[c]Percentage of total sequences assigned to either of the sequenced *Polaribacter* genomes.

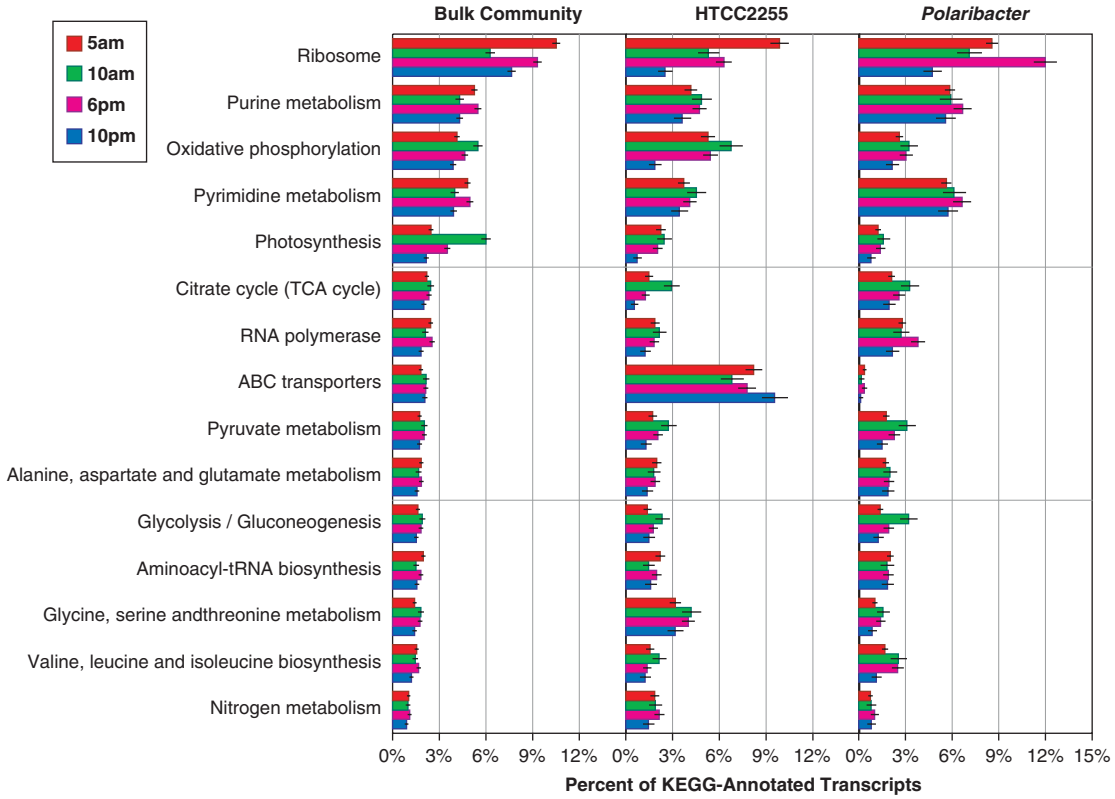

Figure 5 Relative abundances of KEGG pathways in metatranscriptomic data sets. The 10 most abundant KEGG pathways in the bulk community, plus pathways within the 10 most abundant pathways for either *Rhodobacterales* sp. HTCC2255 or *Polaribacter* are shown in order of descending abundance in the total community. Percentage of sequences with significant hits to the KEGG database (community) or percentage of total sequences assigned to specific taxa shown. Error bars represent 95% confidence limits. Photosynthesis signal in the (non-photosynthetic) HTCC2255 and *Polaribacter* bins is due to the assignment of F0F1 ATP synthase genes to this category.

and *Polaribacter*-like sequences (Table 5). Transcripts involved in environmental information processing were more abundant in the HTCC2255 fraction compared with the bulk community (14–19% vs 4.4–4.9% of transcripts); this signal was dominated by ABC transporters.

KEGG-annotated genes could be assigned to 188 (bulk community), 127 (HTCC2255) and 129 (*Polaribacter*) pathways, excluding BRITE hierarchies and pathways associated with human diseases and organismal systems. Of these, 133, 76 and 69 had significantly different levels of expression between samples (Supplementary Tables S8 and S9). Many central metabolic pathways, including oxidative phosphorylation, photosynthesis, the citrate cycle, pyruvate metabolism and glycolysis were overrepresented in the 1000 hours sample (Figure 5), both at the bulk community and taxon-specific levels. In

contrast, ribosomal proteins peaked in morning and evening samples, and had different maxima and minima for the bulk community, HTCC2255 and *Polaribacter* groups. RNA polymerase also displayed different trends among the three groups. Given the changes in taxonomic composition discussed above, we cannot rule out population-level effects in comparisons of relative transcript abundance. However, the synchronous changes in central carbohydrate metabolism and energy metabolism suggests that these pathways may be tuned to broader environmental factors, while the dynamics of translation and transcription suggest more population-specific controls.

Although complicating population-specific factors cannot be ruled out, these transcriptional profiles suggest that light may have a role in the metabolism of *Rhodobacterales* sp. HTCC2255 and *Polaribacter.* In recent years, a number of potential mechanisms by which light might influence the metabolism of heterotrophs in the ocean have been discovered (Béjà *et al.*, 2000, 2002; Kolber *et al.*, 2001; Venter *et al.*, 2004). Proteorhodopsin made up 0.18–0.82% of HTCC2255-like and 0.42–1.18% of *Polaribacter*-like transcripts. Surprisingly, in both groups the representation of proteorhodopsin transcripts was highest in the nighttime (2200 hours) sample. *Dokdonia* sp. MED134, another proteorhodopin-carrying flavobacterium, has been demonstrated to have higher levels of proteorhodopsin in light-incubated vs dark-incubated cultures, but these changes were examined at time scales of 3–13 days, not hours as in this study (Gómez-Consarnau *et al.*, 2007). Proteorhodopsin was one of the most abundant transcripts associated with gamma proteobacterium HTCC2207 (0.55–0.91% of HTCC2207 transcripts) and flavobacterium MS024-2A (0.86–3.9% of MS024-2A transcripts). HTCC2207-like proteorhodopsin transcripts appeared most abundant at night, but MS024-2A-like proteorhodopsin transcripts had the highest relative abundance at 1000 hours. However, while the daytime increase in proteorhodopsin expression for MS024-2A was significant in the context of the total number of transcripts assigned to this organism, the relatively low coverage (1610–5210 assigned sequences) precludes rigorous transcriptional analysis.

Even in the absence of photo-regulation of proteorhodopsin expression, the HTCC2255-like transcripts did exhibit potential light-dependent changes in energy metabolism. In particular, HTCC2255-like F-type adenosine triphosphate (ATP) synthase transcripts appeared to be down-regulated at night; 5 out of 9 ATP synthase-associated proteins had significant differences in abundance, and all were least abundant in the 2200 hours sample. This may indicate light-dependent changes in the cross-membrane proton gradient. In contrast, although ATP synthase as a whole (photosynthesis pathway in Figure 5) showed a slight but significant decrease in transcript abundance at night for *Polaribacter*-like sequences, only one subunit exhibited a significant change in expression at the transcript level. One explanation for this difference in light-dependence of ATP synthase expression is that HTCC2255-like organisms may be more dependent on ATP to power transport than *Polaribacter*, as a result of the expanded use of ABC transporters in this organism (TonB-dependent transporters utilize the proton gradient directly). More broadly, the transcriptomic profile of HTCC2255-like microbes at 2200 hours showed a larger decrease in many metabolic activities than seen in co-occurring *Polaribacter*-like bacteria, suggesting potentially greater light regulation in this organism. Alternatively, the particular population of HTCC2255-like organisms sampled during the 2200 hours timepoint could have been in a lower-energy metabolic state, for reasons independent of the time of day.

Implications

Although a number of metatranscriptomic studies of marine microbial communities have been conducted, most have represented single or relatively few time points that were spatially segregated. Although these studies have proven useful for general surveys of expressed genes and non-coding RNAs, the utility of such comparisons is limited in the absence of data on the spatial and temporal scales of natural environmental variability. In this study, we demonstrate the feasibility of *in situ* autonomous collection of metatranscriptomic samples using the ESP platform, along with synoptic data on environmental conditions. A distinct advantage of this approach is that it allowed longer term deployments and continuous monitoring of environmental fluctuations over the full time course of multiple sample collections. Consistent with the known dynamic variability in coastal systems, our observations reflected continuously changing conditions, consistent with high current velocity at the sample site. Each of the four metatranscriptomic data sets thus represents a different water mass, and a distinct microbial community. Although similar taxa were present in each time point, these taxa showed different bulk activity levels (as reflected in rRNA and mRNA abundances) and expression profiles in each of the four samples. As a result, although differences in gene expression levels could be observed, it was difficult to differentiate changes that reflect the specific water masses and microbial populations sampled vs discrete organismal responses to broader environmental parameters (such as time of day). Our study demonstrates that such effects are large enough to require serious consideration, even when a fixed location is sampled across relatively short (24 h) time scales.

Automated sample collection has the potential to greatly reduce the costs associated with long-term environmental monitoring, allowing longer duration and/or higher frequency sampling schemes.

Different deployment schemes for the sampler, for example, on 'drifters' or autonomous vehicles, may also facilitate short-term temporal sampling within coherent water masses. This may prove to be important for developing a picture of the temporal and spatial scales of natural variability in microbial populations. In this study, we found that the identities of the most abundant microbial populations did not shift dramatically in samples collected over the course of a day, but their relative abundances did. With ESP technology, it will be possible to examine such differences over longer time scales, and using different sampling modes (for example, Lagrangian vs Eularian) to better correlate changes in environmental conditions with shifts in microbial community composition and activity.

Acknowledgements

We thank the engineering technicians and machinists at MBARI for their invaluable help and dedication toward instrument development, and the crew of *R/V Zephyr* for their support and expertise during field operations. This work was supported by a grant from the Gordon and Betty Moore Foundation (EFD), the Office of Science (BER), US Department of Energy (EFD) and NSF Science and Technology Center Award EF0424599. Development and application of ESP has been funded in part by grants from the David and Lucille Packard Foundation (to CS) through funds allocated by the Monterey Bay Aquarium Research Institute (MBARI), NSF (OCE-0314222), NASA Astrobiology (NNG06GB34G, NNX09AB78G to CS) and the Gordon and Betty Moore Foundation (ERG731 to CS). This work is a contribution of the Center for Microbial Oceanography: Research and Education (C-MORE).

References

Audic S, Claverie JM. (1997). The significance of digital gene expression profiles. *Genome Res* **7**: 986–995.

Béjà O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP *et al.* (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902–1906.

Béjà O, Suzuki MT, Heidelberg JF, Nelson WC, Preston CM, Hamada T *et al.* (2002). Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**: 630–633.

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodo* **57**: 289–300.

Blanvillain S, Meyer D, Boulanger A, Lautier M, Guynet C, Denancé N *et al.* (2007). Plant carbohydrate scavenging through tonB-dependent receptors: a feature shared by phytopathogenic and aquatic bacteria. *PLoS One* **2**: e224.

Connon SA, Giovannoni SJ. (2002). High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. *Appl Environ Microbiol* **68**: 3878–3885.

Cottrell MT, Kirchman DL. (2000). Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. *Appl Environ Microbiol* **66**: 1692–1697.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496–503.

Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, Robbens S *et al.* (2006). Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci USA* **103**: 11647–11652.

Fasham MJR, Baliño BM, Bowles MC, Anderson R, Archer D, Bathmann U *et al.* (2001). A new vision of ocean biogeochemistry after a decade of the Joint Global Ocean Flux Study (JGOFS). *AMBIO* (Sp. Iss. 10): 4–31.

Feingersch R, Suzuki MT, Shmoish M, Sharon I, Sabehi G, Partensky F *et al.* (2010). Microbial community genomics in eastern Mediterranean Sea surface waters. *ISME J* **4**: 78–87.

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.

Goffredi SK, Jones WJ, Scholin CA, Marin III R, Vrijenhoek RC. (2006). Molecular detection of marine invertebrate larvae. *Mar Biotechnol (NY)* **8**: 149–160.

Gomez-Alvarez V, Teal TK, Schmidt TM. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**: 1314–1317.

Gómez-Consarnau L, González JM, Coll-Lladó M, Gourdon P, Pascher T, Neutze R *et al.* (2007). Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445**: 210–213.

Gómez-Pereira PR, Fuchs BM, Alonso C, Oliver MJ, van Beusekom JE, Amann R. (2010). Distinct flavobacterial communities in contrasting water masses of the north Atlantic Ocean. *ISME J* **4**: 472–487.

González JM, Fernández-Gómez B, Fernàndez-Guerra A, Gómez-Consarnau L, Sánchez O, Coll-Lladó M *et al.* (2008). Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (Flavobacteria). *Proc Natl Acad Sci USA* **105**: 8724–8729.

Gosink JJ, Woese CR, Staley JT. (1998). *Polaribacter* gen. nov., with three new species, *P. irgensii* sp. nov., *P. franzmannii* sp. nov. and *P. filamentus* sp. nov., gas vacuolate polar marine bacteria of the *Cytophaga-Flavobacterium-Bacteroides* group and reclassification of '*Flectobacillus glomeratus*' as *Polaribacter glomeratus* comb. nov. *Int J Syst Bacteriol* **48**(Part 1): 223–235.

Greenfield DI, Marin R, Jensen S, Massion E, Roman B, Feldman J *et al.* (2006). Application of environmental sample processor (ESP) methodology for quantifying *Pseudo-nitzschia australis* using ribosomal RNA-targeted probes in sandwich and fluorescent *in situ* hybridization formats. *Limnol Oceanog Methods* **4**: 426–435.

Greenfield DI, Marin III R, Doucette G, Mikulski C, Jones K, Jensen S *et al.* (2008). Field applications of the second-generation Environmental Sample Processor (ESP) for remote detection of harmful algae: 2006-2007. *Limnol Oceanog Methods* **6**: 667–679.

Haywood AJ, Scholin CA, Marin R, Steidinger KA, Heil C, Ray J. (2007). Molecular detection of the brevetoxin-producing dinoflagellate *Karenia brevis* and closely related species using rRNA-targeted probes and a semiautomated sandwich hybridization assay. *J Phycol* **43**: 1271–1286.

Hewson I, Poretsky RS, Tripp HJ, Montoya JP, Zehr JP. (2010). Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ Microbiol* **12**: 1940–1956.

Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Jones WJ, Preston CM, Marin R, Scholin CA, Vrijenhoek RC. (2008). A robotic molecular method for *in situ* detection of marine invertebrate larvae. *Mol Ecol Resour* **8**: 540–550.

Kanehisa M, Goto S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27–30.

Karl DM. (2002). Nutrient dynamics in the deep blue sea. *Trends Microbiol* **10**: 410–418.

Karl DM, Dore JE, Lukas R, Michaels AF, Bates NR, Knap AH. (2001). Building the long-term picture: the US JGOFS time-series programs. *Oceanography* **14**: 6–17.

Kirchman DL. (2002). The ecology of *Cytophaga-Flavobacteria* in aquatic environments. *FEMS Microbiol Ecol* **39**: 91–100.

Kolber ZS, Plumley FG, Lang AS, Beatty JT, Blankenship RE, VanDover CL *et al.* (2001). Contribution of aerobic photoheterotrophic bacteria to the carbon cycle in the ocean. *Science* **292**: 2492–2495.

Li W, Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**: 1658–1659.

Man-Aharonovich D, Philosof A, Kirkup BC, Le Gall F, Yogev T, Berman-Frank I *et al.* (2010). Diversity of active marine picoeukaryotes in the Eastern Mediterranean Sea unveiled using photosystem-II *psbA* transcripts. *ISME J* **4**: 1044–1052.

Martens EC, Koropatkin NM, Smith TJ, Gordon JI. (2009). Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem* **284**: 24673–24677.

Martin-Cuadrado AB, Lopez-Garcia P, Alba JC, Moreira D, Monticelli L, Strittmatter A *et al.* (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One* **2**: e914.

McCarren J, Becker JW, Repeta DJ, Shi Y, Young CR, Malmstrom RR *et al.* (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc Natl Acad Sci USA* **107**: 16420–16427.

Mincer TJ, Church MJ, Taylor LT, Preston C, Karl DM, DeLong EF. (2007). Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. *Environ Microbiol* **9**: 1162–1175.

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

Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J* **4**: 673–685.

Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE *et al.* (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4**: 784–798.

O'Mullan GD, Ward BB. (2005). Relationship of temporal and spatial variabilities of ammonia-oxidizing bacteria to nitrification rates in Monterey Bay, California. *Appl Environ Microbiol* **71**: 697–705.

Palenik B, Grimwood J, Aerts A, Rouzé P, Salamov A, Putnam N *et al.* (2007). The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci USA* **104**: 7705–7710.

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

Poretsky RS, Sun S, Mou X, Moran MA. (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* **12**: 616–627.

Preston CM, Marin III R, Jensen SD, Feldman J, Birch JM, Massion EI *et al.* (2009). Near real-time, autonomous detection of marine bacterioplankton on a coastal mooring in Monterey Bay, California, using rRNA-targeted DNA probes. *Environ Microbiol* **11**: 1168–1180.

Rich V, Pham V, Eppley J, Shi Y, Delong EF. (2010). Time-series analyses of Monterey Bay coastal microbial picoplankton using a 'genome proxy' microarray. *Environ Microbiol* **13**: 116–134.

Rich VI, Konstantinidis K, DeLong EF. (2008). Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. *Environ Microbiol* **10**: 506–521.

Riedel T, Tomasch J, Buchholz I, Jacobs J, Kollenberg M, Gerdts G *et al.* (2010). Constitutive expression of the proteorhodopsin gene by a flavobacterium strain representative of the proteorhodopsin-producing microbial community in the North Sea. *Appl Environ Microbiol* **76**: 3187–3197.

Roman B, Scholin C, Jensen S, Massion E, Marin III R, Preston C *et al.* (2007). Controlling a robotic marine environmental sampler with the Ruby scripting language. *JALA* **12**: 56–61.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Ryan JP, Johnson SB, Sherman A, Rajan K, Py F, Thomas H *et al.* (2010a). Mobile autonomous process sampling within coastal ocean observing systems. *Limnol Oceanog Methods* **8**: 394–402.

Ryan JP, McManus MA, Sullivan JM. (2010b). Interacting physical, chemical and biological forcing of phytoplankton thin-layer variability in Monterey Bay, California. *Continental Shelf Res* **30**: 7–16.

Schauer K, Rodionov DA, de Reuse H. (2008). New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci* **33**: 330–338.

Scholin C, Doucette G, Jensen S, Roman B, Pargett D, Marin R *et al.* (2009). Remote detection of marine microbes, small invertebrates, harmful algae, and biotoxins using the Environmental Sample Processor (ESP). *Oceanography* **22**: 158–167.

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

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

Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, 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.

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

Stingl U, Desiderio RA, Cho JC, Vergin KL, Giovannoni SJ. (2007). The SAR92 clade: an abundant coastal clade of culturable marine bacteria possessing proteorhodopsin. *Appl Environ Microbiol* **73**: 2290–2296.

Suzuki MT, Béjà O, Taylor LT, DeLong EF. (2001). Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ Microbiol* **3**: 323–331.

Suzuki MT, Preston CM, Béjà O, de la Torre JR, Steward GF, DeLong EF. (2004). Phylogenetic screening of ribosomal RNA gene-containing clones in bacterial artificial chromosome (BAC) libraries from different depths in Monterey Bay. *Microb Ecol* **48**: 473–488.

Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC. (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**: e2527.

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.

Vila-Costa M, Rinta-Kanto JM, Sun S, Sharma S, Poretsky R, Moran MA. (2010). Transcriptomic analysis of a marine bacterial community enriched with dimethylsulfoniopropionate. *ISME J* **4**: 1410–1420.

Whittaker RH. (1952). A study of summer foliage insect communities in the great smoky mountains. *Ecol Monographs* **22**: 1–44.

Worden AZ, Lee JH, Mock T, Rouzé P, Simmons MP, Aerts AL *et al.* (2009). Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**: 268–272.

Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H *et al.* (2009). Assembling the marine metagenome, one cell at a time. *Plos One* **4**.

Yooseph S, Nealson KH, Rusch DB, McCrow JP, Dupont CL, Kim M *et al.* (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**: 60–66.

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