Vol. 60: 289–297, 2010
doi: 10.3354/ame01426

AQUATIC MICROBIAL ECOLOGY
Aquat Microb Ecol

Published online August 3

Ecophysiology of a common unannotated gene transcript in surface water microbial assemblages of the oligotrophic open ocean

Julia M. Brown, Ian Hewson*

Department of Microbiology, Cornell University, Ithaca, New York 14853, USA

ABSTRACT: Metagenomic and metatranscriptomic surveys of open ocean microbial assemblages have revealed a vast number of nucleic acid sequences that are of unknown function or phylogenetic affiliation. We examined metatranscriptomic databases and found several highly expressed, yet unidentified, transcripts which were present across diverse locations in the open ocean. Comparison of the 5 most highly represented transcript contiguous sequences (hRCs) against the Global Ocean Survey (GOS) assembled metagenomic sequence database revealed that 4 of the 5 hRCs had been observed previously in open ocean microbial DNA. Amplification of the hRCs from pelagic microbial DNA yielded 1 amplicon (hRC93). To confirm that hRC93 was syntenous with the nucleotide sequence on a matching GOS contig (JCVI_SCAF_1096627329331), we amplified a 706 bp region between hRC93 and an adjacent gene from pelagic microbial DNA. We further characterized hRC93 by examining its relative frequency in metatranscriptomic libraries and performed quantitative PCR on microbial DNA collected from the equatorial Atlantic Ocean, the Gulf of Maine, and a tropical estuary (Bayboro Harbor). Our data demonstrate that hRC93 is in a rare (~0.1–1% of cells) component of productive tropical surface water microbial assemblages, and that its role in host physiology is not related to photosynthesis or to the expression of adjacent genes on the matching GOS genome fragment. However, the large investment in hRC93 transcription suggests an important role in host ecology. These results highlight the lack of genomic information on rare marine microorganisms, but also suggest that unannotated reads in metagenomic and metatranscriptomic surveys can provide useful information on the ecophysiology of uncultivated microorganisms.

KEY WORDS: Metatranscriptome · Bacterioplankton · Genomics · Riboswitch

INTRODUCTION

Several decades of research have highlighted the crucial role of marine microorganisms in the pelagic marine ecosystem (Williams 1981, Azam et al. 1983, Ducklow 1983). The marine microbial loop comprises *Bacteria*, *Archaea*, unicellular eukaryotes, and viruses which mediate the flux of carbon from atmosphere to deep ocean and form the basis of most marine food webs (Pomeroy et al. 2007). Applications of molecular tools in the past 2 decades have elucidated a wide diversity of microorganisms inhabiting marine habitats (Giovannoni et al. 1990, Delong 1992, Fuhrman et al. 1992, 1993, Fuhrman & Ouverney 1998, Giovannoni & Rappe 2000, Giovannoni 2004, DeLong et al. 2006), and recent applications of whole community genome (metagenomic) and transcript (metatranscriptomic) shotgun sequencing (Venter et al. 2004, DeLong et al. 2006, Rusch et al. 2007) have revealed the presence of previously unrecognized metabolic pathways and unexpectedly high transcription of genes of unknown function (Hewson et al. 2009b). These studies show the value of large-scale sequencing in its ability to reveal unanticipated and undiscovered processes occurring in complex microbial communities. The large number of unidentifiable sequences within these datasets, however, shows that there is still a great deal about genetic diversity and function in marine microbial communities to be extracted from such surveys.

In shotgun sequencing surveys of open ocean plankton, 33 to 87% of all sequences could be assigned to an organism or group of organisms (Venter et al. 2004,

*Corresponding author. Email: hewson@cornell.edu

Poretsky et al. 2005, DeLong et al. 2006, Rusch et al. 2007, Frias-Lopez et al. 2008, Gilbert et al. 2008, Hewson et al. 2009a, Poretsky et al. 2009) using traditional BLAST analyses comparing metatranscriptomes and metagenomes to databases of known sequences. In these analyses, the remainder of reads bore no identity or homology above an investigator-assigned annotation cutoff (E-value, bit score). The phylogenetic origin and function of the 13 to 67% remaining sequence reads is therefore unknown. Comparisons against assembled sequences from environmental genomic surveys provide further resolution of putative origin and ecophysiology, but taxonomic origin may be obscure (Hewson et al. 2009a). The presence of diverse, highly-expressed RNAs in metatranscriptomes (Frias-Lopez et al. 2008, Gilbert et al. 2008, Poretsky et al. 2009, Shi et al. 2009) represents a large genetic investment by microorganisms with streamlined genomes typical of open ocean bacteria (Rocap et al. 2003, Giovannoni et al. 2005). This highlights the probable importance of such molecules in the regulation of cellular activities (Poretsky et al. 2005, 2009). Analysis of transcripts at Stn ALOHA near Hawaii, USA revealed commonly occurring short (<100 bp) transcripts in microbial community RNA pools, which may represent putative small RNAs of potential regulatory function such as riboswitches or termination motifs (Shi et al. 2009). These regulatory RNAs made up the bulk of undefined sequences in metatranscriptomic surveys generated at Stn ALOHA (Frias-Lopez et al. 2008).

Recently, surface water metatranscriptomic libraries were prepared from both day and night at 7 stations in the tropical North Atlantic Ocean, Stn ALOHA, and in the Southwest Pacific Ocean (Hewson et al. 2009b, 2010, Poretsky et al. 2009), as well as a metatranscriptome prepared from plankton-net-collected *Trichodesmium* north of the Fijian Islands (Hewson et al. 2009c; see Table 1). The majority of transcript sequences obtained were unannotated, with few additional sequences matching environmental shotgun sequencing derived protein sequences (Hewson et al. 2009c, 2010). These reports were consistent with previous reports at other open-ocean and coastal locations (Frias-Lopez et al. 2008, Gilbert et al. 2008, Poretsky et al. 2009) and illustrated that the majority of gene transcripts in marine plankton were novel.

The aim of the present study was to investigate the ecophysiology of highly expressed, unannotated gene transcripts derived from surface water microbial communities of the open ocean. We describe the occurrence of a highly-expressed RNA sequence which bears no nucleotide identity or homology with known genomes or proteins, respectively, yet shares identity with a contiguous fragment assembled from environmental genome sequences generated in the Global Ocean Survey (GOS). Using molecular approaches and sequence database comparisons, we describe ecological information about the transcript which provides clues to its function in surface water microbial communities despite the absence of direct sequence identification. Moreover, the purpose of the present study was to describe an approach by which ecophysiologies of uncultivated microorganisms represented by numerous unannotated gene transcripts within metagenomic and metatranscriptomic datasets can be investigated in the absence of cultivation and expression studies.

MATERIALS AND METHODS

Sequence database alignment. Random transcript sequence databases prepared from 14 open ocean samples (Hewson et al. 2009b,c, 2010, Poretsky et al. 2009) were obtained from the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA), and were used to identify highly expressed yet unannotated sequences for further investigation (see Table 3). Because these transcript libraries were generated from RNA prepared using the same method, the resulting relative frequencies of transcripts within these 14 libraries are comparable. To identify unique, unannotated sequences in each library, several steps were performed (Fig. S1 in Supplement 1, available at www.int-res.com/articles/suppl/a060p289_supp.pdf). (1) Sequences matching proteins or genomes by BLASTx or BLASTn were removed from the raw sequence libraries. (2) Sequences <75 bp, sequences containing >60% of any single nucleotide, and sequences matching rRNA databases were removed, as described by Hewson et al. (2009b). (3) Sequences identical over the first 100 bp were removed to eliminate additional artifacts of 454 sequencing (Gomez-Alvarez et al. 2009), thus leaving 1 sequence to represent all sequencing replicates within the library.

The edited, unannotated sequences were then subjected to reciprocal BLASTn (nucleotide–nucleotide) alignment with an E-value cutoff of 0.001 to cluster similar sequences. Sequences matching >99% over at least 80% of the query length were considered the same in this comparison. The number of reads matching each query was then enumerated, and the 5 most highly repeated transcripts were selected for further analysis.

The 5 most highly expressed transcript sequences were assembled using the CAP3 program (Huang 1992). A minimum overlap of 20 bp and minimum identity on overlaps of 99% was used in the assembly. The assembly generated contiguous gene transcripts for each of the highly expressed unannotated reads (highly represented contiguous sequences [contigs], hRCs). The assembled hRCs were then verified to be unique by BLASTx against the non-redundant protein database at the National Center for Biotechnology Information (NCBI), and by BLASTn against the non-redundant nucleotide database at NCBI. hRCs were compared by BLASTx, and BLASTn against GOS assembled nucleotide sequences and proteins using CAMERA. The hRCs were further compared by BLASTn against eukaryotic microbial transcript sequence databases (J. Zehr unpubl. data), marine metaviriome databases accessed through CAMERA, and against metatranscriptomic sequence databases generated from other marine communities (Frias-Lopez et al. 2008, Gilbert et al. 2008, Poretsky et al. 2009, 2010) using an E-value cutoff score of 10^{-3} (Table 1).

Samples for ecophysiology investigation. Microbial samples were collected independently of those used to generate the metatranscriptomic libraries from surface waters and at depth at biogeochemically diverse locations (Table S1 in Supplement 1, available at www.int-res.com/articles/suppl/a060p289/_supp.pdf). Samples generally comprised cell collection on 0.2 µm Durapore or Sterivex filters, which were flash-frozen and transported to the laboratory at Cornell University.

Microbial DNA was extracted using the phenol:chloroform approach modified for pelagic microorganisms described by Fuhrman et al. (1988), with modifications (Hewson et al. 2006). Briefly, filters (which were either in Sterivex capsules or 25 mm membranes) were treated with 0.5 ml of boiling 10:1 salt-Tris-EDTA: sodium dodecyl sulfate (SDS) to heat-lyse bacterioplankton cells, followed by overnight precipitation with 0.95M ammonium acetate ($C_2H_3O_2NH_4$) and 60 % ethanol (EtOH) by volume (final concentration). Nucleic acids and proteins were pelleted by centrifugation, resuspended in 0.2 ml deionized water, and sequentially extracted with 0.2 ml of phenol, phenol: chloroform:isoamyl alcohol (24:1:0.1), and finally chloroform:isoamyl alcohol (10:1). The extracted nucleic acids were precipitated once again with $C_2H_3O_2NH_4$ and EtOH, pelleted by centrifugation, dried, and reconstituted in deionized water.

PCR amplification of hRC sequences. PCR primers were designed around the 5 most abundant hRCs in metatranscriptomic libraries (Table 2; sequence

Table 1. Summary of unannotated sequence reads derived from sequenced metatranscriptomic and metagenomic libraries. The number of unannotated sequence reads was enumerated using the Metagenome-Rapid Annotation using Subsystems Technology (MG-RAST; Meyer et al. 2008) analysis of sequence read libraries. Reads were considered unannotated if the E-value was >0.001

Location	Total no. of sequences	Unannotated reads (%)	Library used	Source
South Pacific Ocean	18218	87	Day *Crocosphaera*	Hewson et al. (2009b)
South Pacific Ocean	5385	86.5	Night *Trichodesmium*	Hewson et al. (2009a)
South Pacific Ocean	12105	86	Night *Crocoshaera*	Hewson et al. (2009b)
South Pacific Ocean	5711	84	Day *Trichodesmium*	Hewson et al. (2009a)
Rauenfjord	173447	78.2	Present-day CO_2 initial	Gilbert et al. (2008)
Stn ALOHA	113205	75.1	70 m	Frias-Lopez et al. (2008)
Stn ALOHA	56206	59.2	Day	Poretsky et al. (2009)
Stn ALOHA	50797	59.1	Night	Poretsky et al. (2009)
Sapelo Island	141016	48		Poretsky et al. (2010)
Sargasso Sea	217223	44		M. Vila-Costa & M. A. Moran (unpubl. data)

Table 2. Oligonucleotide primers and probe used to attempt highly represented contiguous sequence (hRC) amplification and sequencing by PCR and quantitative PCR (qPCR)

Target	Forward primer (5′ - 3′)	Probe (5′ - 3′)	Reverse sequence (5′ - 3′)
hRC93	CCAACTGCAGACGGAGAACT		CGAGTTTCGAAGGGGTCTTA
hRC93 qPCR	GGCGAAAGCGTTATAAAGAG	CCCACCAAGAGGAAAAGAGAGAATCC	TTTCGAAGGGGTCTTAAGTG
GOS fragment hRC_93 to tonB	GGCGAAAGCGTTATAAAGAG		GAACCATAAATTGACGCAGC
hRC41	CGAGTTTCGAAGGGGTCTTA		CTCGATTAAGCTTGCCATCC
hRC195	TTTCCGTCTGCAGGAAACTT		ATCTCTTCATCCTTGGTGGGT
hRC182	GCTGAACCTTCATGAGATCG		TTTAATCAGCCCAATAAATCCC
hRC61	TGCAGTTGTTCCTGTGTCAGT		TCATCCTTGGTGGGTCGTA

information of hRCs is given in Fig. S2 in Supplement 1 available at www.int-res.com/articles/suppl/a060p289_supp.pdf). PCR reactions (50 µl) consisted of 1 µl of template DNA, 1× PCR Reaction Buffer (Invitrogen), 10 mM $MgCl_2$ (Invitrogen), 2 µM forward and reverse primers, and 2.5 U *Taq* polymerase (Invitrogen), with 1 µl of microbial DNA from samples collected from surface waters near Ferry Reach, Bermuda, as template. The reactions were subject to 30 cycles of amplification: 95°C denaturation for 30 s, 55°C annealing for 30 s, and 72°C elongation for 45 s in a BioRad MyCycler. Thermal cycling was preceded by a heating step at 95°C for 3 min, and followed by a final extension step at 72°C for 7 min. PCR products were electrophoresed on a 1.2% agarose gel at 8V cm^{-1} for 90 min, post-stained with SYBR Gold to visualize PCR products. PCR amplicons were gel extracted using the Zymo Gel Extraction kit and protocol, ligated into pGEM®-T Easy Vectors (Promega), and transformed into Z-competent JM109 *Escherichia coli* (Zymo Research). Plasmid extraction on overnight 5 ml cultures was accomplished using a Zymo Plasmid Miniprep Kit (Zymo Research), and cloned amplicons were sequenced by the Center for Life Sciences DNA Sequencing Facility at Cornell University.

Quantitative PCR. A quantitative PCR primer and probe set (TaqMan©) was developed to assay the abundance of hRC93 in the DNA of surface water bacterioplankton (Table 2). Quantitative PCR (qPCR) was conducted in 25 µl reaction volumes consisting of 1× qPCR Master Mix (Invitrogen), 5 pmol forward and reverse primers, 10 pmol probe, and 2 µl template DNA (comprising 0.1 to 0.4 ng extracted DNA). Standard curves were constructed for each qPCR run using plasmid DNA containing the target sequence ranging from 10^1 to 10^8 copies $reaction^{-1}$. Reactions were run in duplicate with an internal plasmid standard of 10^4 copies in a third replicate to assess possible inhibition of the PCR reaction by materials co-extracted with microbial community DNA. qPCR was carried out using an ABI 7300 qPCR machine with the following thermal cycling protocol: 50°C for 2 min, 95°C for 10 min, followed by 60 cycles of 95°C for 15 s and 60°C for 60 s.

RESULTS AND DISCUSSION

Initial bioinformatic analyses

We found through reciprocal BLAST analyses that 18.7 to 79.4% of unannotated transcripts within marine metatranscriptomes occur more than once within surface microbial communities of the oligotrophic open ocean (Table 1). Overlapping unannotated transcripts were assembled into contiguous sequences, denoted as hRCs. Five hRCs comprised on average 1.0 ± 0.2% of all sequence reads in the metatranscriptome surveyed (see Supplement Fig. S2). The most abundant hRC (hRC195, 215 bp) made up 0.4 ± 0.1%, and another prevalent unannotated contig (hRC93, 168 bp), made up 0.22 ± 0.04% of all reads within the 14 metatranscriptomes surveyed (Table 3).

Table 3. Metatranscriptomic datasets used to identify highly represented contiguous sequences (hRCs) within metatranscriptomic libraries. Frequency of hRCs in metatranscriptomes generated as part of the studies by Hewson et al. (2009a,b, 2010) and Poretsky et al. (2009). The occurrence of each hRC (see Fig. S2 in Supplement 1) was obtained by BLASTn comparison against each metatranscriptomic library, where matches at >99% nucleotide identity across at least 80% of the query sequence were considered the hRC fragment. The number of hRC matches was then converted to a percentage of the total mRNA library size (for details on how RNA libraries were edited for mRNAs, see Hewson et al. 2010)

CAMERA library name	Stn	Day/ night	Library size (reads)	hRC44 (%)	hRC61 (%)	hRC93 (%)	hRC182 (%)	hRC195 (%)
	ALOHA	Day	56206	0.01	0.00	0.01	0.00	0.02
	ALOHA	Night	50797	0.01	0.00	0.03	0.00	0.05
TA_20173	SJ060903	Day	44142	0.07	0.03	0.14	0.06	0.04
TA_20179	SJ060912	Night	25733	0.11	0.11	0.13	0.13	0.11
TA_20197	SJ060907	Day	24368	0.06	0.11	0.10	0.04	0.14
TA_20199	SJ060907	Night	34721	0.17	0.07	0.18	0.05	0.18
TA_34838	SJ060909	Day	31828	0.32	0.26	0.24	0.01	0.36
TA_34840	SJ060909	Night	26707	0.26	0.52	0.22	0.03	0.38
TA_34871	KM070304	Day	19487	0.25	0.37	0.24	0.01	0.94
TA_34877	KM070304	Night	23668	0.30	0.44	0.16	0.00	0.79
TA_34921	KM070310	Day	6753	0.56	0.43	0.31	0.00	0.37
TA_34960	KM070310	Night	21900	0.46	0.19	0.25	0.00	0.21
TA_35115	KM070325	Day	18218	0.14	0.33	0.52	0.05	0.74
TA_35117	KM070325	Night	12105	0.17	0.27	0.48	0.02	0.75
Mean				0.21	0.22	0.22	0.03	0.36
SE				0.04	0.05	0.04	0.01	0.08

In an attempt to assign a host identity to these hRCs, we first ensured that they were only found in metatranscriptomes originating from bacterioplankton samples. Comparison by BLASTn of the 5 hRCs against databases of eukaryotic transcripts (polyA-selected; J. Zehr & V. Armbrust unpubl. data) did not reveal any matches. In addition, no matches were found when the 5 hRCs were compared by BLASTn against the 'MarineViromes: All Metagenomic Sequence Reads' dataset using CAMERA (http://camera.calit2.net). Thus, we conclude that transcripts were likely bacterial or archaeal in origin. This is supported by their presence in metagenomes prepared from the 1.0 to 0.1 µm size fraction.

Four of the 5 hRCs matched multiple GOS contigs at nucleotide identities of >99% when clustered against GOS datasets through a BLASTn comparison. BLAST analysis of the nucleic acid sequence surrounding the hRCs on the larger GOS contigs revealed adjacent putative protein encoding regions that shed light on possible host organisms for these hRCs (Fig. 1). For example, hRC44 was found to be upstream of a gene resembling a short-chain dehydrogenase of a sequenced *Flavobacterium*. hRC93 matched several similar GOS contigs at 100% nucleotide identity, the longest being 4012 bp (JCVI_SCAF_1096627329331). The larger GOS contig has 2 weakly annotated putative protein encoding regions. At the 5′ end of the GOS contig there is a hypothetical protein gene most similar to marine *Gammaproteobacteria* HTCC2080 (MGP2080_0412), and at the 3′ end of the contig there is a tonB-dependent receptor most similar to that from the same gammaproteobacterial strain. The homology of the surrounding genes for all 4 hRCs is very low, with 22 to 60% homology to genes from sequenced representatives. Nonetheless, with these comparisons we may begin to assign putative hosts to these unannotated transcripts.

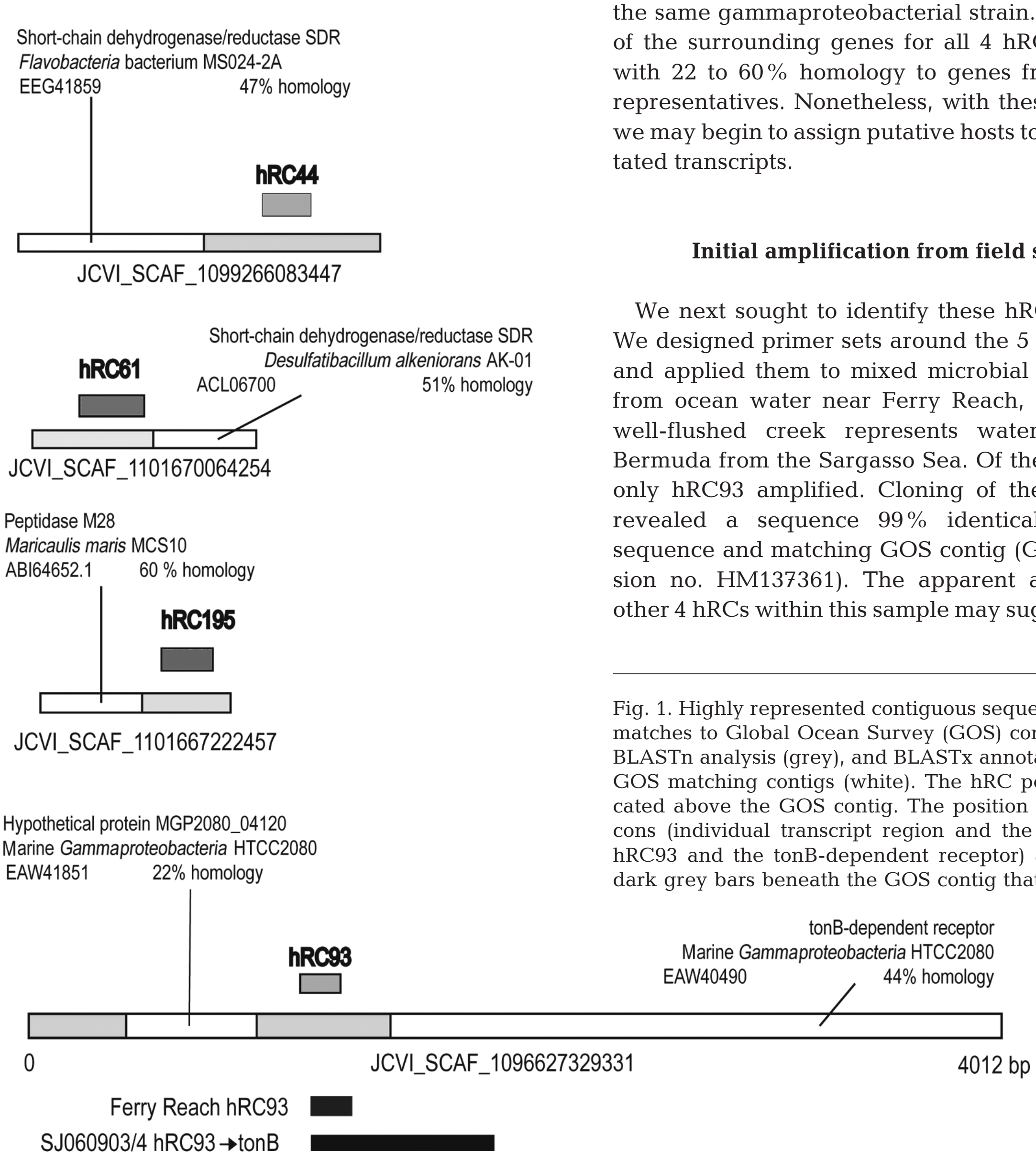

Fig. 1. Highly represented contiguous sequences (hRCs) best matches to Global Ocean Survey (GOS) contigs based upon BLASTn analysis (grey), and BLASTx annotation of genes on GOS matching contigs (white). The hRC positions are indicated above the GOS contig. The position of hRC93 amplicons (individual transcript region and the region between hRC93 and the tonB-dependent receptor) are indicated by dark grey bars beneath the GOS contig that matches hRC93

Initial amplification from field samples

We next sought to identify these hRCs in the field. We designed primer sets around the 5 hRCs (Table 2) and applied them to mixed microbial DNA obtained from ocean water near Ferry Reach, Bermuda. This well-flushed creek represents waters adjacent to Bermuda from the Sargasso Sea. Of the 5 primer sets, only hRC93 amplified. Cloning of the PCR product revealed a sequence 99% identical to the hRC sequence and matching GOS contig (GenBank accession no. HM137361). The apparent absence of the other 4 hRCs within this sample may suggest that these sequences are contaminants, or that these sequences are simply not found at this location or are present at such low genomic levels that they could not be amplified using our method. We chose to focus on hRC93 for further investigation for 3 reasons. (1) The successful amplification of the hRC93 fragment from open ocean DNA was attractive for further quantitative and gene synteny confirmation. (2) The longest GOS contig on which hRC93 recruited contained the most genetic information (4012 bp, including 2 putative protein encoding regions). (3) Aside from hRC195, hRC93 comprised the largest proportion of as-yet undefined metatranscripts with surrounding protein encoding regions on GOS contig matches.

Confirmation of genomic synteny of hRC93 with a tonB-like gene

We investigated whether hRC93 was syntenous with the tonB-like gene in open ocean microbial assemblages observed in larger matching GOS contigs (Fig. 1), by PCR amplification and sequencing between these 2 genes. Amplification using a reverse primer designed on the GOS tonB sequence and forward primer on hRC93 (Table 2) using open ocean microbial DNA yielded amplicons of predicted size (706 bp) that shared 90 and 92% nucleotide identity at Stns SJ060903 and SJ060904 in the equatorial Atlantic Ocean (GenBank accession no. HM137362).

After confirming the genomic synteny of hRC93 with the tonB-like gene showing that hRC93 is upstream to the tonB-like gene, we hypothesized that the hRC93 region of the gene may act as a riboswitch. Analysis of the matching GOS contig using the RibEx website (http://132.248.32.45/cgi-bin/ribex.cgi; Abreu-Goodger & Merino 2005) identified 1 riboswitch-like element (RLE) and transcriptional attenuator in hRC93, at the 3′ end. The closest cluster of orthologous group assignment of the detected RLE is to predicted kinases related to dihydroxyacetone kinases (COG 1461). Comparisons of the tonB-like gene by tBLASTx against the metatranscript libraries, however, did not yield any significant alignments. Since there were no detected transcripts of the tonB-like gene in metatranscriptome libraries, the highly expressed hRC93 is unlikely to be directly related to expression of the genomically adjacent tonB-like gene; however, the genomic potential for both elements is present in open ocean microbial assemblages.

Presence and ecophysiology of hRC93

hRC93 had the greatest frequency in metatranscriptomic libraries prepared from open ocean environments (Hewson et al. 2010) at Stn KM070325, which harbored an intense bloom of the diazotrophic microorganism *Crocosphaera watsonii* WH8501 (Fig. 2) (Hewson et al. 2009b). At this location, hRC93 made up 0.5% of all putative mRNAs. Comparison of the hRC93 sequence by BLASTn against metatranscriptomes prepared from other locations revealed a large number (187) of match-

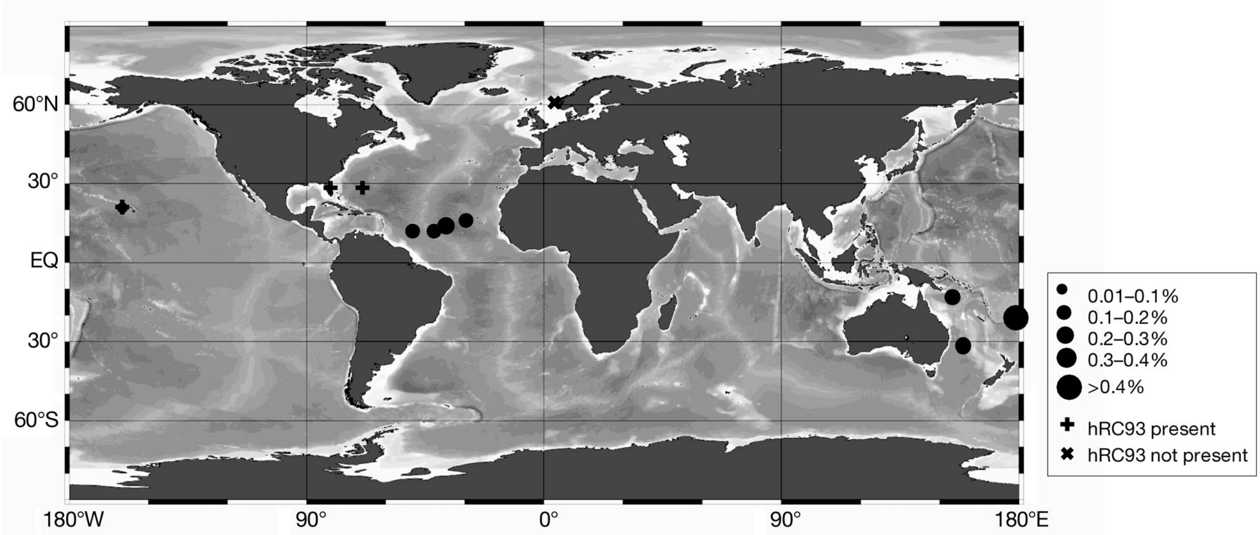

Fig. 2. Relative transcript abundance of hCR93 within metatranscriptomes of Hewson et al. (2010). Circles represent percent representation of hRC93 in analyzed metatranscriptomes at the noted locations. These numbers and respective day/night analyses can be found in Table 3. Additional metatranscriptomes were analyzed for the presence of hRC93. For the scanned metatranscriptomes, (✚) represents locations of metatranscriptomes where hRC93 was found to be present, and (✖) near Rauenfjord denotes that hRC93 was not found within metatranscriptomes at this location

ing sequence reads at Stn ALOHA (Frias-Lopez et al. 2008), at the Sapelo Island observatory (93 matching reads; Poretsky et al. 2010), with fewer matches to transcripts of the Sargasso Sea (5 matching sequences; M. Vila-Costa & M. A. Moran unpubl. data) and no matches to the Raunefjord mesocosm experiment (Gilbert et al. 2008). This pattern of occurrence highlights the tendency of the transcript to be found in warmer, oligotrophic surface waters. There was no consistent pattern in library-size normalized hRC93 frequency between samples taken during light and dark phases, suggesting that the transcript is likely not linked directly to photosynthetic physiology (Table 3).

To further investigate the ecophysiology of the hRC93 fragment, we examined geographic and depth-dependent variation in gene abundance using qPCR. The highest abundance of hRC93, examined using an hRC93-specific qPCR primer/probe set (Table 2), was observed in microbial DNA collected from surface waters of tropical regions of the Atlantic spanning from the Amazon River plume to Cape Verde in the eastern Atlantic. Concentrations ranged from 10^4 to 10^7 copies l^{-1}, reflecting presence as a single copy in ca. 0.1 to 1.0% of cells, and consistent with abundances seen in metatranscriptomic data from these regions (0.22 ± 0.4% of all metatranscripts). hRC93 was also detected in Bayboro Harbor, Florida, USA, an area with conditions similar to that of the Amazon River plume (water temperature ~26°C). In contrast, hRC93 was not detected in deeper waters of the eastern Atlantic (water temperatures <17°C) and could not be amplified from microbial DNA collected adjacent to Appledore Island in the Gulf of Maine (Fig. 3A, water temperature = 14.6°C). hRC93 was found predominantly in surface waters. The exception to this observation was at Stn SJ060904, where the highest surface abundances were observed. At this station, hRC93 was also detected at 75 m depth (Fig. 3B). This station and Stn SJ060903 are located in the Amazon plume, where enhanced productivity driven by allochthonous inputs and nitrogen fixation (Foster et al. 2007) leads to enriched diversity of microorganisms (Hewson et al. 2006). hRC93 was not detected in 0.7 µm prefiltered samples from surface waters at Hydrostation S, located near Ferry Reach, Bermuda, where the hRC93 gene fragment was amplified and sequenced. Because all other samples were not prefiltered and we would expect to see the transcript at this location, the microorganism that harbors this sequence may be larger than 0.7 µm. The presence of hRC93 in open ocean metagenomic libraries, metatranscript libraries, and PCR amplification of hRC93 fragments directly from marine samples indicates that this element is likely not a contaminant and originates from an open ocean marine microorganism.

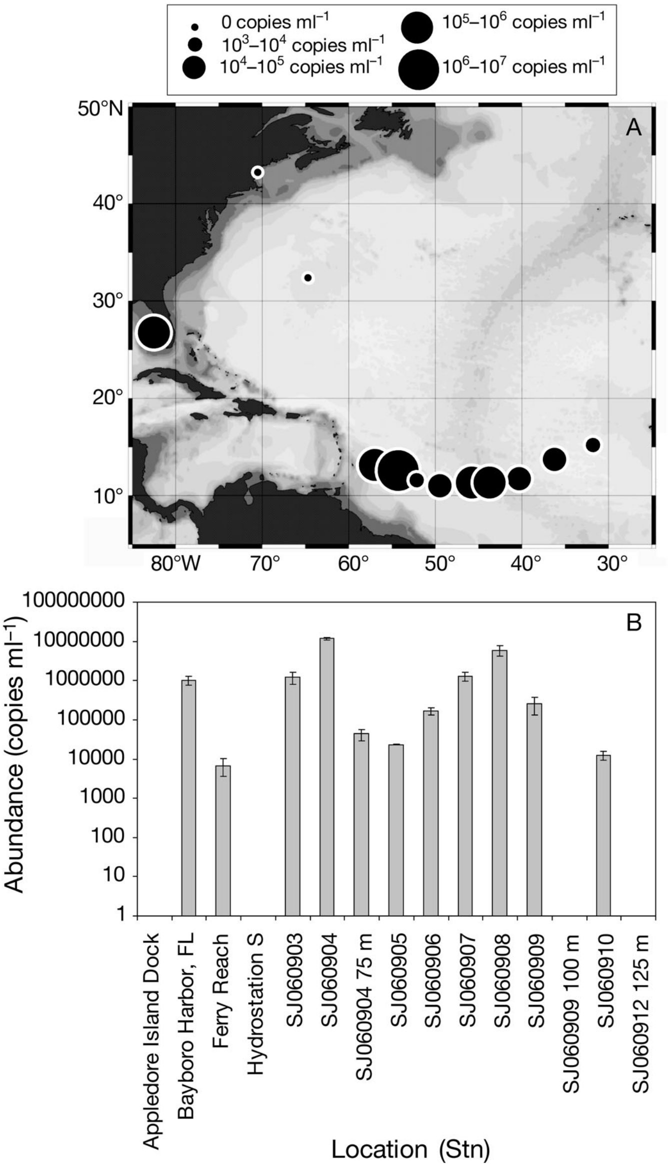

Fig. 3. (A) Geographic distribution of hRC93, examined using quantitative PCR (qPCR) in surface water microbial DNA from the Atlantic Ocean. hRC93 was most abundant in tropical surface waters of the Amazon River plume and absent from samples collected in the Gulf of Maine. (B) qPCR results based on a TaqMan primer/probe designed on the hRC93 sequence. Data indicate the abundance in mixed microbial DNA collected at various stations and depths (indicated on *x*-axis). Error bars = SE of analytical replicates, except at Stn SJ060903 the SE of 3 sample replicates, and at Stn SJ060909 the SE of 2 sample replicates

CONCLUSION

Aggregate analysis of the 5 hRCs, and hRC93 in particular, provides novel insight into the ecophysiology of these as-yet unannotated elements in open ocean microbial assemblages. Our data suggest that hRC93 is likely unrelated to photosynthetic physiology, that the host microorganism is rare, potentially related to *Gammaproteobacteria*, and a warm-water (tropical) microorganism located predominantly in surface assemblages. While it was not possible to definitively assign phylogenetic information to hRC93 based upon BLASTx or BLASTn comparisons, the high transcript abundance in metatranscriptomic libraries compiled from locations around the globe and genomic abundance (via qPCR) of hRC93 within Amazon plume waters and in the tropical Tampa Bay (Bayboro Harbor) lends argument to the significance of this transcript and its association with areas of enhanced tropical activity.

It is striking that the increased number of microbial genomes sequenced in recent years has not paralleled a large increase in the fraction of gene information retrieved from metagenomic or metatranscriptomic surveys. This reveals the vast diversity of microorganisms in the ocean, most of whose existence we are only aware of through low-homology blast hits and great 'unknown' portions of metagenomic and metatranscriptomic data. In the present study, we investigated 5 highly expressed unannotated small RNA sequences that are found in abundance in several metatranscriptomic datasets. Our results highlight the low sequence coverage of representative strains of marine microorganisms and provide an autecological method that will help further characterize unknown portions of sequencing data from environmental samples.

Acknowledgements. We thank J. Zehr (University of California Santa Cruz) and J. Montoya (University of Georgia) for sample provision and access to metatranscriptomic libraries; P. Gilna, S. Sun (CAMERA) for bioinformatic support; R. Poretsky, J. Sohm, and R. Paerl for assistance in sample collection; and E. Madsen, R. Ley, J. Peters, J. Helmann (Cornell), M. Saito (Woods Hole Oceanographic Institution) and R. Kodner (University of Washington) for helpful comments on interpretation. Bioinformatic analysis was also supported by B. Carter (UCSC) and the MEGAMER Facility. Sample collection and metatranscriptome library preparation was supported by the Gordon and Betty Moore Foundation Marine Microbiology Initiative (Investigator grant to J. Zehr) and National Science Foundation Grants OCE0425583 and OCE0425363 awarded to J. Montoya and J. Zehr, respectively. This work contributes in part to a PhD by J.M.B.

LITERATURE CITED

Abreu-Goodger C, Merino E (2005) RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. Nucleic Acids Res 33:W690–W692

Azam F, Fenchel T, Field JG, Gray JS, Meyerreil LA, Thingstad F (1983) The ecological role of water-column microbes in the sea. Mar Ecol Prog Ser 10:257–263

Delong EF (1992) Archaea in coastal marine environments. Proc Natl Acad Sci USA 89:5685–5689

DeLong EF, Preston CM, Mincer T, Rich V and others (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503

Ducklow HW (1983) Production and fate of bacteria in the oceans. Bioscience 33:494–501

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

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF (2008) Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci USA 105:3805–3810

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

Fuhrman JA, Comeau DE, Hagstrom A, Chan AM (1988) Extraction from natural planktonic microorganisms of DNA suitable for molecular biological studies. Appl Environ Microbiol 54:1426–1429

Fuhrman JA, McCallum K, Davis AA (1992) Novel major archaebacterial group from marine plankton. Nature 356: 148–149

Fuhrman JA, McCallum K, Davis AA (1993) Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. Appl Environ Microbiol 59:1294–1302 doi: 10.137/journal.pone.003092

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

Giovannoni S (2004) Evolutionary biology—oceans of bacteria. Nature 430:515–516

Giovannoni S, Rappe M (2000) Evolution, diversity and molecular ecology of marine prokaryotes. In: Kirchman D (ed) Microbial ecology of the oceans. Wiley-Liss, New York, NY, p 47–84

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60–63

Giovannoni SJ, Tripp HJ, Givan S, Podar M and others (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245

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

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

Hewson I, Paerl RW, Tripp HJ, Zehr JP, Karl DM (2009a) Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. Limnol Oceanogr 54: 1981–1994

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

Hewson I, Poretsky RS, Dyhrman ST, Zielinski B and others (2009c) Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. ISME J 3:1286–1300

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 7:1940–1956

Huang X (1992) A contig assembly program based on sensitive detection of fragment overlaps. Genomics 14:18–25

Meyer F, Paarmann D, D'Souza M, Olson R and others (2008) The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386

Pomeroy LR, Williams PJI, Azam F, Hobbie JE (2007) The microbial loop. Oceanography 20:28–33

Poretsky RS, Bano N, Buchan A, LeCleir G and others (2005) Analysis of microbial gene transcripts in environmental samples. Appl Environ Microbiol 71:4121–4126

Poretsky RS, Hewson I, Sun SL, 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

Rocap G, Larimer FW, Lamerdin J, Malfatti S and others (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047

Rusch DB, Halpern AL, Sutton G, Heidelberg KB and others (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol 5:e77 doi: 10.1371/journal.pbio.0050077

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

Venter JC, Remington K, Heidelberg JF, Halpern AL and others (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74

Williams PJI (1981) Incorporation of microheterotrophic processes into the classical paradigm of the planktonic food web. Kieler Meeresforsch:1–28

Editorial responsibility: Dittmar Hahn,
San Marcos, Texas, USA

Submitted: January 7, 2010; Accepted: May, 2010
Proofs received from author(s): July 16, 2010