LETTERS

Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column

Yanmei Shi[1], Gene W. Tyson[1] & Edward F. DeLong[1,2]

Microbial gene expression in the environment has recently been assessed via pyrosequencing of total RNA extracted directly from natural microbial assemblages. Several such 'metatranscriptomic' studies[1,2] have reported that many complementary DNA sequences shared no significant homology with known peptide sequences, and so might represent transcripts from uncharacterized proteins. Here we report that a large fraction of cDNA sequences detected in microbial metatranscriptomic data sets are comprised of well-known small RNAs (sRNAs)[3], as well as new groups of previously unrecognized putative sRNAs (psRNAs). These psRNAs mapped specifically to intergenic regions of microbial genomes recovered from similar habitats, displayed characteristic conserved secondary structures and were frequently flanked by genes that indicated potential regulatory functions. Depth-dependent variation of psRNAs generally reflected known depth distributions of broad taxonomic groups[4], but fine-scale differences in the psRNAs within closely related populations indicated potential roles in niche adaptation. Genome-specific mapping of a subset of psRNAs derived from predominant planktonic species such as *Pelagibacter* revealed recently discovered as well as potentially new regulatory elements. Our analyses show that metatranscriptomic data sets can reveal new information about the diversity, taxonomic distribution and abundance of sRNAs in naturally occurring microbial communities, and indicate their involvement in environmentally relevant processes including carbon metabolism and nutrient acquisition.

Microbial sRNAs are untranslated short transcripts that generally reside within intergenic regions (IGRs) on microbial genomes, typically ranging from 50 to 500 nucleotides in length[3]. Most microbial sRNAs function as regulators, and many are known to regulate environmentally significant processes including amino acid and vitamin biosynthesis[5], quorum sensing[6] and photosynthesis[7]. Because the identification and characterization of microbial regulatory sRNAs has relied primarily on a few model microorganisms[8–10], relatively little is known about the broader diversity and ecological relevance of sRNAs in natural microbial communities.

During a microbial gene expression study comparing four metatranscriptomic data sets from a microbial community depth profile (25 m, 75 m, 125 m and 500 m at Hawaii Ocean Time-series (HOT) Station ALOHA[11]), we discovered that a large fraction of cDNA sequences could not be assigned to protein-coding genes or ribosomal RNAs (Fig. 1). However, >28% of these unassigned cDNA reads from each data set mapped with high nucleotide identity (≥85%) to IGRs on the genomes of marine planktonic microorganisms (Supplementary Fig. 1), indicating that they may be sRNAs. Consistent with the genomic location of known sRNAs[12], many of these reads mapped on IGRs distant from predicted open reading frames (ORFs), or were localized in clearly predicted 5′ and 3′ untranslated regions (UTRs).

A covariance-model-based algorithm[13] was used to search all unassigned cDNA reads for both sequence and structural similarity to known sRNA families archived in the RNA families database Rfam[14]. Thirteen known sRNA families were captured in the environmental transcriptomes, representing only ~16% of the total reads detected by IGR mapping. The most abundant sRNAs belonged to ubiquitous or highly conserved sRNA families including transfer-messenger RNA (tmRNA), RNase P RNA, signal recognition particle RNA (SRP RNA) and 6S RNA (SsrS RNA; Supplementary Table 1). In addition, a number of known riboswitches (*cis*-acting regulatory elements that regulate gene expression in response to ligand binding[15]) were detected in lower abundance, including glycine, thiamine pyrophosphate, cobalamin and *S*-adenosyl methionine riboswitches (Supplementary Table 1). The apparent taxonomic origins of the most abundant known sRNAs revealed depth-specific variation that was generally, but not always, consistent with known microbial depth distributions[4] (Supplementary Fig. 2). For example, although SRP RNAs are abundant in our data sets, very few *Pelagibacter*-like SRP

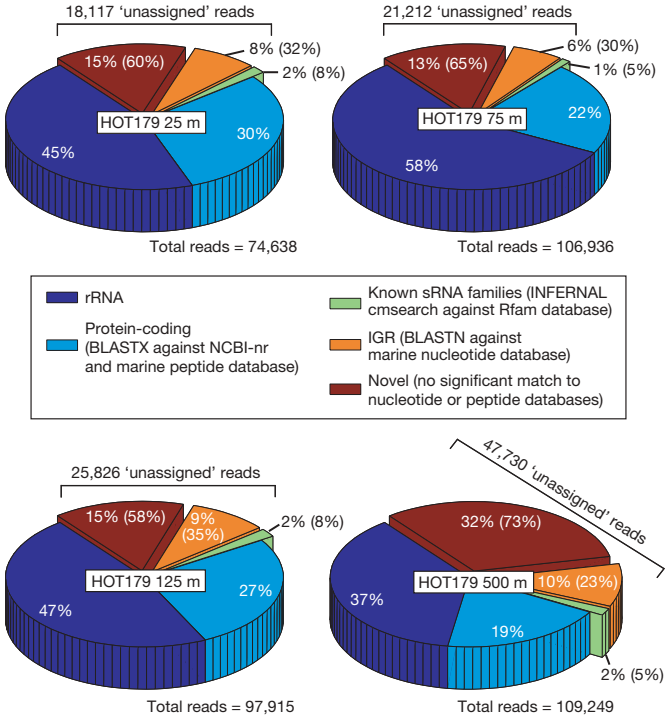


Figure 1 | Inventory of RNAs from each depth in the microbial metatranscriptomic datasets. The three offset slices represent reads that are not assigned to rRNA or known protein-coding genes, and are referred to as 'unassigned'. Numbers in parentheses represent the percentage of the total unassigned cDNA reads in each category.

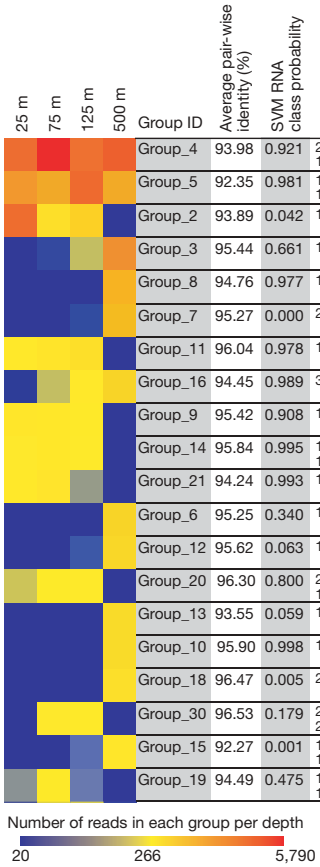
[1]Department of Civil and Environmental Engineering, and [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

RNA reads were detected, indicating that SRP-dependent protein recognition and transport may not be a dominant form of protein translocation in oceanic *Pelagibacter* populations.

For better characterization of sRNAs in our data sets, including previously unknown sRNA families (referred to as putative sRNAs (psRNAs) hereafter), we pooled all cDNA reads from each sample and used a self-clustering approach to group homologous cDNA reads (see Methods). On the basis of observations from the IGR mapping (Supplementary Fig. 1), the self-clustering approach would help identify potential sRNAs because they are likely to span short genomic regions and exhibit high abundance (in many cases orders of magnitude higher than transcripts of protein-coding genes found in the same data sets). A total of 66 groups that comprised at least 100 overlapping cDNA reads were identified (Fig. 2 and Supplementary Table 2). For several of these groups, the abundance and depth-dependent distribution detected by means of cDNA pyrosequencing was confirmed using reverse transcription–quantitative polymerase chain reaction (RT–qPCR) analyses (Supplementary Fig. 3). Among the 66 groups, 9 were identified as belonging to Rfam sRNA families (Supplementary Table 2), and most of the remaining psRNA groups mapped to IGRs on metagenomic fragments derived from marine planktonic microorganisms.

Although they bear no resemblance to known peptide sequences, the psRNA groups could potentially represent mRNA degradation products or small unannotated protein-coding regions. We applied several criteria to help rule out these possibilities, including location within IGRs, psRNA length, lack of coding potential and conserved secondary structure. First, the psRNAs ranged in size between 100 and 500 nucleotides (Supplementary Fig. 4 and Supplementary Table 2), and tended to have an increased GC content when located within an AT-rich genome context[16] (Fig. 3a). Second, we systematically screened multiple sequence alignments of all 66 groups for coding potential, as indicated by three-base periodicity in the nucleotide substitution patterns[17] (Methods). Only sequences in group 92 were identified as possibly encoding proteins (Fig. 3b), and these were subsequently mapped to a specific hypothetical protein (NCBI accession number: ABZ07689) from a recently described uncultured marine crenarchaeote[18]. Third, the psRNA groups encompassed relatively divergent sequences that internally shared conserved secondary structures (for example, Fig. 3a, inset), indicating evolutionary coherence of functional roles and mechanisms. The alignment of full-length psRNA sequences revealed clear nucleotide co-variation that preserved base pairing in the consensus secondary structure (for example, Supplementary Fig. 5). In a specific example (group 5), although three divergent *Pelagibacter*-like psRNA sequences (one from 4,000 m depth[18] and two from surface waters[19]) shared pairwise nucleotide identities of only 78% to 87%, predicted secondary structures were nearly identical (Supplementary Fig. 6). Although computational analyses alone cannot be completely definitive, these combined criteria support our hypothesis that most of the psRNA groups that we identified represent authentic microbial sRNAs.

Many of the psRNAs identified here may be derived from as-yet uncharacterized microorganisms. For example, nine self-clustered psRNA groups shared no obvious homology with known nucleotide



Group ID	Average pair-wise identity (%)	SVM RNA class probability	Average length	Location	Flanking ORF annotations	Putative taxonomy of environmental fragments
Group_4	93.98	0.921	212 bp 173 bp	5′ 3′	COG0469: pyruvate kinase COG2234: predicted aminopeptidases	Bacteria; Proteobacteria
Group_5	92.35	0.981	148 bp 155 bp	5′ 3′	COG0350: methylated DNA-protein cysteine methyltransferase COG0482: tRNA(5-methylaminomethyl-2-thiouridylate) methyltransferase	Bacteria; Proteobacteria; Alphaproteobacteria/Gammaproteobacteria
Group_2	93.89	0.042	189 bp	3′	Hypothetical protein	Bacteria; Cyanobacteria; Prochlorales (MIT9215 strain specific)
Group_3	95.44	0.661	170 bp	3′	COG3202: ATP/ADP translocase	Bacteria; Proteobacteria; Alphaproteobacteria/Gammaproteobacteria
Group_8	94.76	0.977	157 bp	5′	Hypothetical protein	Bacteria; Proteobacteria; Gammaproteobacteria
Group_7	95.27	0.000	205 bp	5′	COG1254: acylphosphatases	Archaea; Crenarchaeota
Group_11	96.04	0.978	187 bp	5′	COG0077: prephenate dehydratase	Bacteria; Proteobacteria; Deltaproteobacteria/Gammaproteobacteria
Group_16	94.45	0.989	362 bp	3′	COG0114: fumarase	Bacteria; Spirochaetes
Group_9	95.42	0.908	191 bp	5′	COG0571: dsRNA-specific ribonuclease	Bacteria; Cyanobacteria; Prochlorales
Group_14	95.84	0.995	198 bp 158 bp	5′ 3′	COG0116: predicted N6-adenine-specific DNA methylase Hypothetical protein	Bacteria; Cyanobacteria; Prochlorales
Group_21	94.24	0.993	139 bp	5′	COG0180: tryptophanyl-tRNA synthetase	Bacteria; Proteobacteria; Gammaproteobacteria
Group_6	95.25	0.340	138 bp*	Unknown	No hits	Unassigned
Group_12	95.62	0.063	135 bp	Unknown	No hits	Unassigned
Group_20	96.30	0.800	201 bp 157 bp	Unknown 3′	COG0822: NifU homolog involved in Fe-S cluster formation COG1804: predicted acyl-CoA transferases/carnitine dehydratase	Bacteria; Proteobacteria; Alphaproteobacteria/Gammaproteobacteria
Group_13	93.55	0.059	194 bp	Unknown	No hits	Archaea; Crenarchaeota
Group_10	95.90	0.998	133 bp*	Unknown	No hits	Unassigned
Group_18	96.47	0.005	275 bp	3′	COG0531: amino acid transporters	Archaea; Euryarchaeota
Group_30	96.53	0.179	222 bp 220 bp	5′ 3′	COG0596: predicted hydrolases or acyltransferases Conserved hypothetical protein	Bacteria; Cyanobacteria; Prochlorales
Group_15	92.27	0.001	107 bp 101 bp	5′ 3′	COG1245: predicted ATPase, RNase L inhibitor (RLI) homologue No hits	Archaea; Crenarchaeota
Group_19	94.49	0.475	173 bp 173 bp	3′ 5′	No hits COG1629: outer membrane receptor proteins, mostly iron transport	Bacteria; Proteobacteria; Gammaproteobacteria

Figure 2 | Abundance, distribution and features of the top twenty most abundant sRNA and psRNA groups identified in the metatranscriptomic data. The twenty groups were ranked based on total abundance. Each group's depth distribution is shown in the left panel, with the number of reads in each data set indicated by colour, from high (red) to low (blue). Each group's proximity (5′ or 3′) to the nearest gene, annotation and putative taxonomy for that gene (where possible) are shown. The RNA-class probability values were generated with an SVM learning algorithm using RNAz[29]. Group 9 is comprised of *Prochlorococcus*-like RNase P RNAs. Group 21 sRNAs probably mediate regulation (via transcription attenuation) of tryptophanyl tRNA synthetase. Group 30 contains overlapping sRNAs *Yfr8* and *Yfr9* indentified in *Prochlorococcus* MED4 in ref. 8. Lengths of putative sRNAs with no homology with known nucleotide sequences (each marked with an asterisk) were predicted through assembly of cDNAs from each group (average contig size, see Methods). A complete list of sRNA and psRNA groups containing >100 cDNA reads is provided in Supplementary Table 2.

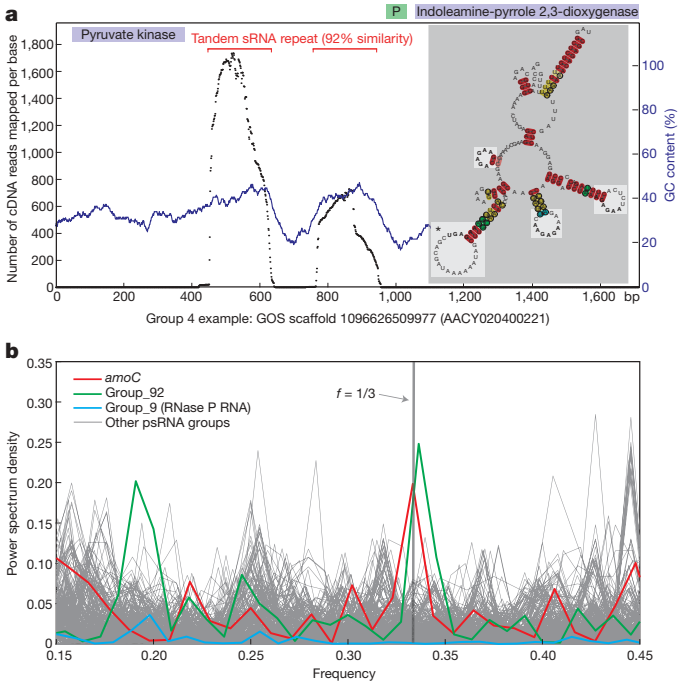


Figure 3 | Characteristics of psRNA groups consistent with known sRNAs. **a**, Genomic context and features of the most abundant psRNA group, group 4, mapped onto a Gammaproteobacteria-like contig from the Global Ocean Sampling (GOS) database. Sequence coverage (black dots, left axis) and reference GC content (blue dots, right axis) are shown. Gene annotations are indicated along the top of the panel (annotated ORFs shown outside and inside the panel are on the forward and reverse strand, respectively; P represents promoter). In the predicted structure (inset), loops containing conserved sequence motifs (in bold letters) are highlighted, and the loop marked with an asterisk contains sequences predicted to interact with 5′ translation start site of a flanking gene. **b**, Three-base periodicity analysis of multiple sequence alignments for the 66 self-clustered groups. A significant peak in the power spectrum density at the frequency of 1/3 indicates three-base periodicity in the nucleotide substitution patterns, indicating protein-coding potential[17]. See Methods.

sequences (for example, groups 6 and 10), and seem to represent completely new sRNA families. Most of these were found only in the 500 m sample (Fig. 2). The remaining psRNA groups mapped to IGRs on genomic and metagenomic sequences derived from planktonic marine microbes. Although identifying sRNA regulatory functions and their target genes is a major challenge even for model microorganisms[20], the conserved genomic context of these psRNAs has potential to provide insight into their functional roles[21,22]. The most predominant gene families flanking these psRNA groups included transporter genes involved in nutrient acquisition (inorganic nitrogen, amino acids, iron and carbohydrates) and genes involved in energy production and conversion (Supplementary Table 2). These results highlight the potential importance of sRNA regulation of nutrient acquisition and energy metabolism in free-living planktonic microbial communities.

The most populated psRNA cluster, group 4, appeared to be involved in the regulation of central carbon metabolism and energy production in Proteobacteria (predominantly Gammaproteobacteria). The psRNAs from this group were flanked by genes involved in pyruvate metabolism (for example, pyruvate kinase and malate synthase), glucose transport (sodium glucose symporter) and nitrogen acquisition (ammonia permease and aminopeptidase; Fig. 2 and Supplementary Table 2). In several cases, group 4 psRNAs occurred in tandem copies within the same IGR (Fig. 3a). Small RNAs that display stable secondary structure typically mediate regulation using sequences in loop domains to interact with specific target sequences[3,23]. Consistent with this mechanism, a conserved six-nucleotide sequence motif (AAGAGN) appeared in multiple loops within predicted hairpin structures for group 4 (Fig. 3a, inset). The six-nucleotide sequence AAGAGA was previously verified as a ribosomal binding site[24], and indicates that group 4 psRNAs may have a regulatory role at the translational level. Indeed, sequences in one of the loop domains of the consensus structure (Fig. 3a, inset) have potential to interact (by base pairing across 32 bp) with the flanking pyruvate kinase gene near the 5′ translation initiation site.

In contrast to the broad taxonomic affiliations of group 4 psRNAs, the other highly abundant psRNA group, group 5, appeared almost exclusively on *Pelagibacter*-like genomic fragments recovered from both open ocean surface waters[19] and abyssal (4,000 m) depth[18], but did not map to the genomes of currently cultivated *Pelagibacter* strains (Fig. 2 and Supplementary Table 2). Group 5 psRNAs mapped onto 203 different metagenomic fragments, predominantly in the 5′ UTR of 6-*O*-methylguanine DNA methyltransferase (6-*O*-MGMT, COG0350; involved in DNA repair) and the 3′ UTR of tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase (*trmU*, COG0482; involved in tRNA modification). A predicted promoter and Rho-independent terminator flanked group 5 psRNAs upstream of 6-*O*-MGMT, and attenuator/riboswitch characteristics were identifiable in the 5′ UTR by secondary structure prediction (Supplementary Fig. 6). Indeed, the presence of riboswitch-like elements upstream of 6-*O*-MGMT genes was previously predicted by comparing 223 complete bacterial genomes[25].

Unlike group 4 and 5 psRNAs, the remaining self-clustered sRNA and psRNA groups showed depth-variable distributions (Fig. 2). Group 7 psRNAs were enriched at 500 m and were highly conserved in marine crenarchaeal genomes. Similarly, cyanobacteria-like psRNAs were enriched in the photic zone (for example, groups 2, 30, 48 and 17; Supplementary Table 2). One of these groups (group 30) includes two experimentally validated sRNAs (*Yfr8* and *Yfr9*), which were found antisense to one another and were hypothesized to be involved in a toxin–antitoxin system in *Prochlorococcus marinus* MED4 (ref. 8). Intriguingly, a few *Prochlorococcus*-like psRNA groups mapped to some but not all coexisting members of the *Prochlorococcus* population, indicating that such sRNAs may provide niche-specific regulation. Group 2 psRNAs, for example, were detected only in the genome of *P. marinus* strain MIT9215 and in a highly similar genomic fragment from the environment (NCBI accession number: DQ366713). Group 2 psRNAs are located in a hyper-variable region adjacent to phosphate transporter genes, and share a 14-bp exact match with the 5′ translation initiation site of the phosphate ABC transporter gene (*pstC*). In *Prochlorococcus* strains lacking the phosphate regulon two-component response regulator (*phoB*) and signalling kinase (*phoR*)[26], such as MIT9215, it is possible that sRNAs represent an alternative mechanism for regulating phosphorus assimilation.

To examine sRNA representation in specific abundant microbial groups, we aligned the psRNA reads to the genome of an abundant planktonic bacterium, *Candidatus Pelagibacter ubique* HTCC7211. Eleven IGRs on the *P. ubique* HTCC7211 genome coincided with the psRNAs identified in our samples (Fig. 4), 6 of which were also independently predicted to be sRNA-containing IGRs (support vector machine, SVM, RNA-class probability >0.9) by comparative analysis of three *P. ubique* genomes (Methods and Supplementary Table 3). Genes flanking these expressed psRNAs included DNA-directed DNA polymerase gamma/tau subunit (*dnaX*), *carD*-like transcriptional regulator family and alternative thymidylate synthase (Supplementary Table 3). Notably, covariance-model-based searches identified cDNAs mapping to glycine riboswitch motifs in two *Pelagibacter* IGRs (Fig. 4 and Supplementary Table 3). Recently, it was experimentally verified that *P. ubique* HTCC1062 uses one of these two glycine riboswitches to sense the intracellular glycine level and to regulate its carbon usage for biosynthesis and energy[27].

The diversity and abundance of sRNAs in microbial metatranscriptomic data sets indicates that natural microbial assemblages use a wide variety of sRNAs for regulating gene expression in response to variable environmental conditions. The data and analyses

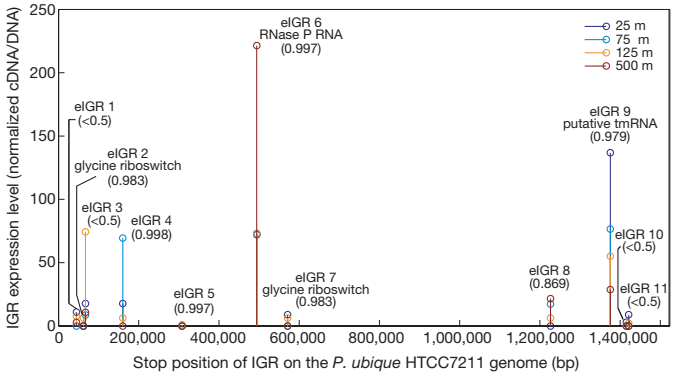


Figure 4 | Normalized cDNA/DNA ratios of expressed IGRs (eIGRs) on the *P. ubique* HTCC7211 genome at all four depths. Because a manually curated HTCC7211 genome annotation is not yet publicly available, the genomic regions that recruited psRNAs were manually inspected and confirmed as IGRs. The values in the parentheses are RNA-class probability values generated with a SVM learning algorithm using RNAz[29].

described here provide a culture-independent tool to expand our knowledge of the sequence motifs, structural diversity and genomic distributions of microbial sRNAs that are expressed under specific environmental conditions. Although the exact regulatory functions of many of the psRNAs remain to be experimentally verified, their *in situ* expression, structural features and genomic context all provide a solid foundation for future studies. These data, in conjunction with metatranscriptomic field experiments linking environmental variation with changes in RNA pools, have potential to provide new insights into environmental sensing and response in natural microbial communities.

METHODS SUMMARY

Bacterioplankton samples were collected from the HOT Station ALOHA (22° 45′ N, 158° W) in March 2006 at four different depths (25 m, 75 m, 125 m and 500 m), and immediately frozen and stored at −80 °C until processing. Nucleic acid extraction, RNA amplification, cDNA synthesis and pyrosequencing were performed as previously described[1]. Ribosomal RNA sequences were identified by querying against a comprehensive rRNA database using BLASTN, and were excluded from the subsequent sRNA analysis. Protein-coding genes were recognized by querying with BLASTX against published peptide databases as well as a custom marine-specific peptide database (Methods). A covariance-model-based program (INFERNAL)[13] was used to search for known sRNA elements in the data sets. The self-clustering approach (see Methods) to identify abundant psRNAs in the environment was based on sRNA reads spanning across a short genomic region in high abundance. Self-clustered groups that contained more than 100 cDNA reads were further characterized in detail, including secondary structure prediction using RNAalifold[28], coding potential evaluation, genomic context examination and sRNA-class probability calculation using RNAz[29] (see Methods). The genome sequences of an oceanic *Pelagibacter* strain (HTCC7211) were used to recruit psRNA reads to examine possible regulatory sRNAs related to oceanic *Pelagibacter* populations.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 31 October 2008; accepted 9 April 2009.

1. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl Acad. Sci. USA* **105,** 3805–3810 (2008).
2. Gilbert, J. A. *et al.* Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* **3,** e3042 (2008).
3. Storz, G. & Haas, D. A guide to small RNAs in microorganisms. *Curr. Opin. Microbiol.* **10,** 93–95 (2007).
4. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311,** 496–503 (2006).
5. Gottesman, S. Stealth regulation: biological circuits with small RNA switches. *Genes Dev.* **16,** 2829–2842 (2002).
6. Lenz, D. H. *et al.* The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell* **118,** 69–82 (2004).
7. Duehring, U., Axmann, I. M., Hess, W. R. & Wilde, A. An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proc. Natl Acad. Sci. USA* **103,** 7054–7058 (2006).
8. Steglich, C. *et al.* The challenge of regulation in a minimal photoautotroph: non-coding RNAs in *Prochlorococcus*. *PLoS Genet.* **4,** e1000173 (2008).
9. Vogel, J. *et al.* RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.* **31,** 6435–6443 (2003).
10. Silvaggi, J. M., Perkins, J. B. & Losick, R. Genes for small, noncoding RNAs under sporulation control in *Bacillus subtilis*. *J. Bacteriol.* **188,** 532–541 (2006).
11. Karl, D. M. & Lukas, R. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep-Sea Res. II* **43,** 129–156 (1996).
12. Kawano, M., Reynolds, A. A., Miranda-Rios, J. & Storz, G. Detection of 5′- and 3′-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. *Nucleic Acids Res.* **33,** 1040–1050 (2005).
13. Eddy, S. *INFERNAL User's Guide, Version 0. 72* ⟨http://infernal.janelia.org/⟩ (2007).
14. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33,** D121–D124 (2005).
15. Brantl, S. Bacterial gene regulation: from transcription attenuation to riboswitches and ribozymes. *Trends Microbiol.* **12,** 473–475 (2004).
16. Schattner, P. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.* **30,** 2076–2082 (2002).
17. Ré, M. & Pavesi, G. in *Applications of Fuzzy Sets Theory* (eds Masulli, F., Mitra, S. & Pasi, G.) 544–550 (Springer, 2007).
18. Konstantinidis, K. T. & DeLong, E. F. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J.* **2,** 1052–1065 (2008).
19. Rusch, D. B. *et al.* The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5,** 398–431 (2007).
20. Vogel, J. & Wagner, E. G. H. Target identification of small noncoding RNAs in bacteria. *Curr. Opin. Microbiol.* **10,** 262–270 (2007).
21. Hershberg, R., Altuvia, S. & Margalit, H. A survey of small RNA-encoding genes in *Escherichia coli*. *Nucleic Acids Res.* **31,** 1813–1820 (2003).
22. Yao, Z. *et al.* A computational pipeline for high-throughput discovery of *cis*-regulatory noncoding RNA in prokaryotes. *PLoS Comp. Biol.* **3,** 1212–1223 (2007).
23. Trotochaud, A. E. & Wassarman, K. M. A highly conserved 6S RNA structure is required for regulation of transcription. *Nature Struct. Mol. Biol.* **12,** 313–319 (2005).
24. Bruttin, A. & Brüssow, H. Site-specific spontaneous deletions in three genome regions of a temperate *Streptococcus thermophilus* phage. *Virology* **219,** 96–104 (1996).
25. Abreu-Goodger, C. & Merino, E. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res.* **33,** W690–W692 (2005).
26. Martiny, A. C., Coleman, M. L. & Chisholm, S. W. Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc. Natl Acad. Sci. USA* **103,** 12552–12557 (2006).
27. Tripp, H. J. *et al.* Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. *Environ. Microbiol.* **11,** 230–238 (2008).
28. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31,** 3429–3431 (2003).
29. Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA* **102,** 2454–2459 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the University of Hawaii HOT team, and the captain and crew of the RV *Kilo Moana* for their expert assistance at sea. We also thank S. Schuster for collaboration and advice on pyrosequencing, J. Eppley for help with computational analyses and discussion, and J. Maresca, A. Martinez, J. McCarren and V. Rich for their comments on this manuscript. We thank S. Giovannoni, J. Tripp and M. Schwalbach for sharing their in press manuscript on *Pelagibacter* riboswitches, and S. Giovannoni, U. Stingl, the J. Craig Venter Institute and the Gordon and Betty Moore Foundation for the genome sequence of *Pelagibacter* strain HTCC7211. This work was supported by the Gordon and Betty Moore Foundation, National Science Foundation Microbial Observatory Award MCB-0348001, the Department of Energy Genomics GTL Program, the Department of Energy Microbial Genomics Program, and an NSF Science and Technology award, C-MORE. This article is a contribution from the NSF Science and Technology Center for Microbial Oceanography: Research and Education (C-MORE).

Author Contributions E.F.D. conceived and directed the research, coordinated the sequencing effort and collected the samples. Y.S. prepared samples for sequencing, and made the initial observation of sRNA sequences. E.F.D., Y.S. and G.W.T. developed the concept of the paper together. Y.S. and G.W.T. performed the data analysis. Y.S. wrote the first draft of the paper, which was completed by G.W.T. and E.F.D. together.

Author Information The sequences reported here have been deposited in GenBank under accession numbers SRA007802.3, SRA000263, SRA007804.3 and SRA007806.3 corresponding to cDNA sequences, and SRA007801.5, SRA000262, SRA007803.3 and SRA007805.4 corresponding to DNA sequences, for 25 m, 75 m, 125 m and 500 m samples, respectively. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to E.F.D. (delong@mit.edu).

METHODS

Sample collection and RNA/DNA extraction. Bacterioplankton samples from the photic zone (25 m, 75 m, 125 m) and the mesopelagic zone (500 m) were collected from the HOT Station ALOHA site in March 2006, as described previously[1]. In brief, four replicate 1-l seawater samples were prefiltered through 1.6-mm GF/A filters (Whatman) and then filtered onto 0.22-μm Durapore filters (25 mm diameter, Millipore) using a four-head peristaltic pump system. Each Durapore filter was immediately transferred to screw-cap tubes containing 1 ml of RNAlater (Ambion Inc.), and frozen at −80 °C aboard the RV *Kilo Moana*. Samples were transported frozen to the laboratory in a dry shipper and stored at −80 °C until RNA extraction. Total sampling time, from arrival on deck to fixation in RNAlater, was less than 20 min.

Total RNA was extracted as previously described[1], using the mirVana RNA isolation kit (Ambion), with several modifications as follows. Samples were thawed on ice, and then 1-ml RNAlater was loaded onto two Microcon YM-50 columns (Millipore) to concentrate and desalt each sample. The resulting 50 μl of RNAlater was added back to the sample tubes, and total RNA extraction was performed following the mirVana manual. Genomic DNA was removed using a Turbo DNA-free kit (Ambion). Finally, extracted RNA (DNase-treated) from four replicate filters was combined, purified and concentrated using the MinElute PCR Purification Kit (Qiagen).

Bacterioplankton sampling for DNA extraction was performed as previously described[1].

Complementary DNA synthesis and sequencing. The synthesis of microbial community cDNA from small amounts of mixed-population microbial RNA was performed as previously described[1]. In brief, nanogram quantities of total RNA were polyadenylated using *Escherichia coli* poly(A) polymerase I (E-PAP)[30]. First-strand cDNA was then synthesized using ArrayScript (Ambion) with an oligo(dT) primer containing a T7 promoter sequence and a restriction enzyme (BpmI) recognition site sequence, followed by the second-strand cDNA synthesis[1]. The double-stranded cDNA templates were transcribed *in vitro* using T7 RNA polymerase at 37 °C for 6 h[31], yielding a large amount of antisense RNA. The SuperScript double-stranded cDNA synthesis kit (Invitrogen) was used to convert antisense RNA to microgram quantities of cDNA, which was then digested with BmpI to remove poly(A) tails. Purified cDNA was then directly sequenced by pyrosequencing[32].

Removal of low-quality and rRNA GS20 cDNA sequences. Low-quality cDNA reads were removed as previously described[1].

Reads encoding rRNA were identified and removed from the cDNA data sets by comparing them to a combined 5S, 16S, 18S, 23S and 28S rRNA database derived from available microbial genomes and sequences from the ARB SILVA LSU and SSU databases (http://www.arb-silva.de). BLASTN[33] matches with bit score ≥50 were considered significant and deemed rRNA sequences. In test simulations, this bit score cutoff resulted in <1.7% false positives against a database of all non-rRNA microbial genes from available microbial genomes.

Identification of protein-coding genes. Protein-coding cDNA reads were identified by translating nucleotide sequences in all 6 frames and comparing each to Global Ocean Sampling peptides, the NCBI-nr protein database and a custom peptide database using BLASTX[33]. The custom peptide database contained marine-specific ORF sequences predicted from four sources: the Moore Microbial Genome Project genomes (http://www.moore.org/microgenome/strain-list.aspx), large genome fragments (~40 kb) from a variety of marine habitats (Rich *et al.*, in preparation), and both fosmid end sequences and shotgun library sequences generated from depth profile bacterioplankton samples collected in multiple HOT cruises (E.F.D. *et al.*, in preparation). Unpublished databases are available on request.

After rRNA sequences were removed, each cDNA data set contained between 40,000 and 70,000 pyrosequence reads. Of these cDNA reads, a large fraction (~50% of those from photic-zone samples; ~70% from the mesopelagic sample) showed no significant homology to either the non-redundant peptide database from NCBI or marine microbial peptide sequences, using the bit score of 40 that has been previously validated as a cutoff for calling homology in short pyrosequencing reads[1].

Assignment of cDNA reads to known non-coding RNA families. We searched the Rfam database[14] to investigate the representation and diversity of known sRNA families in our data sets. Rfam is a collection of non-coding RNA families, represented by multiple sequence alignments and covariance models, including those from 400 complete genomes including 233 bacterial and 24 archaeal genomes (June 2008 version). The INFERNAL program (http://infernal.janelia.org/) was used to search for RNA structure and sequence similarities based on covariance models (also called profile stochastic context-free grammars)[34]. The reference database was a collection of covariance models for all non-coding RNA families downloaded from the Rfam (version 8.1) ftp site (http://www.sanger.ac.uk/Software/Rfam/ftp.shtml). A perl wrapper named Rfamscan.pl (http://www.sanger.ac.uk/Software/Rfam/help/software.shtml), written by Sam Griffiths-Jones, was used to run batch queries (>200,000 cDNA reads) on a local machine.

To test the specificity and sensitivity of the INFERNAL Rfam-seeded search of our cDNA reads, two data sets were created from the *E. coli* strain K12 substrain MG1655, in which sRNAs have been well defined[35]. The two test data sets were protein-coding sequences and known sRNA sequences, each with the same length distributions as our cDNA data set (that is, 206,418 sequence fragments with mean sequence length 97 bp). The INFERNAL Rfam-seeded search of the *E. coli* MG1655 protein-coding test data set yielded no significant hits, indicating high specificity and a false-positive rate below detection. However, the INFERNAL Rfam-seeded search did not identify all *E. coli* MG1655 sRNA fragments, probably owing to the short lengths of the query sRNA fragments. To compensate for the decreased search sensitivity due to shorter read length, we queried all cDNA reads against all full-length sRNA sequences in the Rfam database by BLASTN. Reads that did not meet the default cutoffs defined by Rfamscan, but shared good homology with Rfam member sequences by BLASTN (alignment length ≥90% of sequence length; sequence identity ≥85%), were also assigned to the corresponding sRNA families.

Putative taxonomic assignment of cDNA reads in known sRNA families. Potential taxonomic origins of the known sRNAs were investigated by searching against NCBI-nt (4 July 2008) using BLASTN (word size of 7, default e-value cutoff, low complexity filter off, and the ten best hits retained). The BLASTN results were then parsed using MEGAN[36] using default parameters—that is, the congruent taxonomy of the hits that were within 10% below the best hit was assigned to the cDNA read.

Self-clustering approach to identify sRNA and psRNA groups. A self-clustering approach allowed related cDNA reads to form distinct groups that could be separated from other transcripts based on sequence similarity and overall abundance. Combined cDNA reads (206,418 reads after the removal of rRNAs) from all 4 depths were locally aligned to each other (that is, all sequences served both as queries and subjects) using BLASTN with the following settings different from default: $W = 7$, $F =$ F, $m = 8$, $v = 206418$, $b = 206418$, $e = 1 \times 10^{-5}$. A perl script was used to group similar cDNA reads based on the BLASTN output. In brief, for each cDNA query, all matches that met a minimum cutoff of 85% sequence identity over 90% average sequence length were considered significant and stored into a hash. The hash then was ranked on the basis of the number of matches stored for each hash key (query). The cDNA read with the most matches served as a seed sequence of the first cluster. After all matches of the seed sequence were recruited, the script looped over each one of the matches and gathered all subsequent matches until the chain disconnected and a new cluster started to form.

The self-clustering approach was successful in identifying a number of highly abundant psRNA groups. These psRNAs were clearly distinct from protein-coding clusters as they were found in much higher copy number than most mRNAs, and the typical length of psRNAs was ~100–500 nucleotides. The sequence identity cutoff (85%) was chosen because it allowed known RNase P RNAs from closely related microbial populations (for example, all *Prochlorococcus* RNase P RNAs) to form a distinct sequence group. However, because sRNA species by nature differ in their primary sequence divergence, clustering based on one sequence identity cutoff inevitably yields psRNA groups with different within-group diversity, which either represent homologues from closely related microbial populations or highly conserved elements from diverse microbial taxa.

Systematic screening for coding potentials of the self-clustered groups. We identified a total of 66 groups that contained more than 100 cDNA reads (a file named 'H179_sRNA_groups.tgz', containing all sequences from these 66 groups, and a file named 'H179_sRNA_groups_CLUSTAL.tgz', containing multiple sequence alignments of subsets of sequences from these 66 groups, can be downloaded from http://web.mit.edu/ymshi/Public/). To assess the possibility that some groups represent unannotated small proteins, we systematically screened multiple sequence alignments of these 66 groups for coding potentials based on three-base periodicity in nucleotide substitution patterns. The rationale of detecting three-base periodicity in coding regions is that codons encoding the same amino acid often differ only in a single nucleotide located in the third position of the codon. As a direct consequence, in coding sequences under selective evolutionary pressure, substitutions are more often tolerated if they occur at the third position of codons. Therefore, if aligned sequences are protein-coding, the spectral signal of the mismatches along the alignment is expected to be maximal at frequency 1/3 (three-base periodicity)[17].

We generated a pipeline for multiple sequence alignment, nucleotide diversity calculation (conversion of DNA sequence alignments to numerical sequences) and Fourier transform and power spectrum analysis of the numerical sequences for all 66 groups (including known sRNAs and psRNAs). Specifically, 100

sequences were randomly sampled from a subset of overlapping sequences in each group, and aligned using MUSCLE 3.6 (ref. 37). The random sampling and alignment was repeated multiple times proportional to the number of sequences in the group. For each alignment, average nucleotide diversity was calculated for each column of the alignment as following:

$$D_{\text{average}} = \sum D_{\text{pair-wise}} \Big/ N(N-1)/2$$

where D_{average} represents average nucleotide diversity, $D_{\text{pair-wise}}$ represents pair-wise nucleotide diversity (a pair of identical nucleotides was given a value of 0, and a pair of different nucleotides was given a value of 1) and $N(N-1)/2$ represents the total number of pairs in the column of the alignment. Owing to high insertion/deletion error rate of pyrosequencing[32], any alignment column where greater than 75% of sequences had a gap resulted in that column being ignored in the subsequent calculation. After the multiple sequence alignments were converted to numerical sequences, a Fourier transform and power spectrum analysis[38] of the numerical sequences were performed using MATLAB (http://www.mathworks.com/) to find significant frequencies of periodicity.

RT–qPCR analysis of psRNA group 7 and sRNA group 9. The apparent abundance and depth-dependant distribution of group 7 and group 9 in our metatranscriptomic data sets were validated using RT–qPCR. Owing to lack of absolute quantification standards for these groups, we calculated their relative abundance to the crenarchaeal ammonia monooxygenase subunit A (*amoA*) transcript in the 500 m sample. Primers for these groups were designed using the Invitrogen web-based OligoPefect primer designer. The primer sequences are: G7_Primer1 (5′-AGCTCTGCTGGTTCYAGACT-3′) and G7_Primer2 (5′-TCGAACATTCACGCTTCCT-3′); G9_Primer1 (5′-TAAGCCGGGTTCTG TTCATC-3′) and G9_Primer2 (5′-GCCGCTTGAGACTGTGAAGT-3′). The primer set for the crenarchaeal *amoA* transcript was the same as previously published[39]: CrenAmoAQ-F (5′-GCARGTMGGWAARTTCTAYAA-3′) and CrenAmoAModR (5′-AAGCGGCCATCCATCTGTA-3′). All primers were blasted against NCBI-nt database to avoid potential matches to unwanted regions.

Possible traces of DNA were removed from all RNA samples using the Turbo DNA-free kit (Ambion) following the manufacturer's instructions. For each reverse transcription reaction, 1 μl of RNA (4–7.5 ng) was reverse transcribed using gene-specific primer and Superscript III reverse transcriptase (Invitrogen). Reverse transcription was performed at 50 °C for 50 min, after an initial incubation step of 5 min at 65 °C. The reverse transcription reactions were terminated at 85 °C for 5 min, and 1 μl RNase H was added to each reverse transcription reaction, followed by incubation at 37 °C for 20 min. Subsequently, SYBR Green qPCR reactions were performed on LC480 (Roche Applied Science) using the specific primer set for each gene of interest. We used the $2^{-\Delta\Delta CT}$ method[40] to compare the relative abundance of group 7 and group 9 transcripts in all 4 samples (25 m, 75 m, 125 m and 500 m) to the crenarcheal *amoA* transcript in the 500 m sample.

Characterizing psRNA groups. The psRNA groups were further characterized to determine the approximate psRNA length, proximity to (5′ or 3′ or unknown (when the psRNA is not flanked by one ORF on each side)) and annotation of nearest flanking ORF on available genome/metagenome fragments, putative taxonomy and SVM-based RNA class probability. Pooled cDNA reads (not including rRNA reads) from each transcriptomic data set were queried against a custom database of nucleotide sequences from available genome and metagenomic projects (see above) using BLASTN. Metagenomic fragments in this database were run through Metagene[41] to identify predicted ORFs (coding) and intergenic (non-coding) regions.

Using the BLASTN and Metagene results, cDNA reads were mapped to each genome/metagenome fragment based on sequence similarity (≥85% identity over 90% of the read length), which could be used to calculate coverage values for each coding and intergenic region on each genomic/metagenomic fragment. Two groups were identified as highly expressed protein-coding genes (group 35, ammonia monooxygenase subunit C; and group 42, ammonia permease) and were excluded from further analyses. In most cases, reads belonging to putative sRNA groups mapped with high coverage to IGRs on genomic/metagenomic fragments. In these cases, we estimated the size of psRNAs in each group by defining the psRNAs as the sequence region in intergenic space having minimum sequence coverage of greater than ten times. In addition, it was also possible to determine the location of these psRNAs with respect to coding sequences. psRNAs were labelled as either 3′ or 5′ based on their position relative to the nearest flanking gene. Functional annotation for each of the genes flanking psRNA groups was obtained by comparing the amino acid sequences against the KEGG[42], COG[43] and the NCBI-nr databases from NCBI using BLASTP. Putative taxonomic origins of each fragment were assigned based on the NCBI taxonomy of matches in the NCBI-nr database.

Only 9 psRNA groups had no homology to sequences in the currently available database. To estimate the size of each of these psRNA groups, reads from each were assembled using PHRAP (−minmatch 15, −minscore 20, revise_greedy) and the average length of contigs (<10 contigs) formed used to infer sequence space spanned by the sRNA group.

To calculate the RNA class probability for each group, the first twenty cDNA reads recruited to each psRNA group were extracted from the data set and placed in the same sequence orientation. Multiple sequence alignments were performed using MUSCLE 3.6 (ref. 37). The sequence alignment for each psRNA groups (CLUSTALW format) was then used to predict consensus structure and the thermodynamic stability using RNAz[29], and an RNA-class probability was calculated based on the SVM regression analysis.

Secondary structure prediction. The minimum free energy structure was predicted based on the multiple sequence alignment of full-length psRNA sequences extracted from metagenomic sequence reads. The RNAalifold program from the Vienna RNA package[28,44] was used to produce consensus secondary structure and sequence alignment colour-coded based on nucleotide variations. The colour hue indicates how many of the six possible types of base-pairs (GC, CG, AU, UA, GU, UG) occur in at least one of the sequences. Pairs without sequence covariation are shown in red. Ochre, green, turquoise, blue and violet mark pairs that occur in two, three, four, five and six types of pairs, respectively. Pale colours mark pairs that cannot be formed by all sequences (that is, inconsistent base changes occur in some sequences). Attenuator-like structure was predicted using RibEx program[25].

Mapping cDNA reads to the genome of *P. ubique* HTCC7211. *Candidatus Pelagibacter ubique* HTCC7211 genome sequences were downloaded from the Moore Microbial Genome Project (http://www.moore.org/microgenome/strain-list.aspx). Based on the genome annotations, all IGR sequences greater than 50 bp (excluding rRNA and tRNA) were extracted and used to create BLASTN database. Both DNA and cDNA reads from each sample were then queried (BLASTN) against the database and parsed using same criteria as above (alignment length ≥90% of sequence length; identity ≥85%). For each IGR an expression ratio was calculated as the percentage of cDNA reads assigned to the IGR, relative to that in the DNA library. If there were cDNA hits but no DNA hits, the number of DNA hits was considered to be 1. This normalization compensates for the IGR length differences, and differences in DNA and cDNA library sizes.

Prediction of sRNA-containing IGRs in *Pelagibacter* genomes. Three *Pelagibacter* genomes (*Pelagibacter ubique* HTCC1062, HTCC1002 and HTCC7211) were used in the comparative genome analysis to predict possible sRNAs in the IGRs based on conserved secondary structure among closely related genomes[45]. A total of 1,113 IGRs were extracted from these three genomes (again only IGRs ≥50 bp and excluding tRNAs and rRNAs), and locally aligned to pooled ORFs and IGRs (5,398) from the three genomes using BLASTN with the following settings changed from default: $W = 7$, $F =$ F, $v = 5398$, $b = 5398$. ORFs were included so that *cis*-acting regulatory elements of mRNA were also examined. A total of 1,848 IGR sequences were extracted from all the high-scoring segment pairs with bit scores greater than 50, using Bioperl[46]. Self-clustering of this subset of *Pelagibacter* IGR sequences was then performed, as described above. Sequences in each cluster were aligned using MUSCLE 3.6 (ref. 37) and the alignments were scored for their secondary structure conservation and thermodynamic stability using RNAz 1.0 (ref. 29). SVM-based RNA-class probability values from the RNAz pipeline were gathered for each cluster and ranked from high to low.

30. Wendisch, V. F. *et al.* Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Anal. Biochem.* **290,** 205–213 (2001).
31. Vangelder, R. N. *et al.* Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl Acad. Sci. USA* **87,** 1663–1667 (1990).
32. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
33. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
34. Eddy, S. R. & Durbin, R. RNA sequence-analysis using covariance-models. *Nucleic Acids Res.* **22,** 2079–2088 (1994).
35. Rudd, K. E. EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.* **28,** 60–64 (2000).
36. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–386 (2007).
37. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

38. Holste, D., Weiss, O., Grosse, I. & Herzel, H. Are noncoding sequences of *Rickettsia prowazekii* remnants of "neutralized" genes? *J. Mol. Evol.* **51,** 353–362 (2000).
39. Mincer, T. J. *et al.* Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific subtropical gyre. *Environ. Microbiol.* **9,** 1162–1175 (2007).
40. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-[Delta][Delta]CT Method. *Methods* **25,** 402–408 (2001).
41. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34,** 5623–5630 (2006).
42. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
43. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28,** 33–36 (2000).
44. Hofacker, I. L., Fekete, M. & Stadler, P. F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319,** 1059–1066 (2002).
45. Axmann, I. M. *et al.* Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol.* **6,** R73 (2005).
46. Jason, S. & Ewan, B. The Bioperl project: motivation and usage. *SIGBIO Newsl.* **20,** 13–14 (2000).