Microbial community gene expression in ocean surface waters

Jorge Frias-Lopez*, Yanmei Shi*, Gene W. Tyson*, Maureen L. Coleman*, Stephan C. Schuster†, Sallie W. Chisholm*‡, and Edward F. DeLong*‡§

Departments of *Civil and Environmental Engineering and §Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; †Center for Comparative Genomics and Bioinformatics, Pennsylvania State University, University Park, PA 16802

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved January 22, 2008 (received for review September 19, 2007)

Metagenomics is expanding our knowledge of the gene content, functional significance, and genetic variability in natural microbial communities. Still, there exists limited information concerning the regulation and dynamics of genes in the environment. We report here global analysis of expressed genes in a naturally occurring microbial community. We first adapted RNA amplification technologies to produce large amounts of cDNA from small quantities of total microbial community RNA. The fidelity of the RNA amplification procedure was validated with *Prochlorococcus* cultures and then applied to a microbial assemblage collected in the oligotrophic Pacific Ocean. Microbial community cDNAs were analyzed by pyrosequencing and compared with microbial community genomic DNA sequences determined from the same sample. Pyrosequencing-based estimates of microbial community gene expression compared favorably to independent assessments of individual gene expression using quantitative PCR. Genes associated with key metabolic pathways in open ocean microbial species—including genes involved in photosynthesis, carbon fixation, and nitrogen acquisition—and a number of genes encoding hypothetical proteins were highly represented in the cDNA pool. Genes present in the variable regions of *Prochlorococcus* genomes were among the most highly expressed, suggesting these encode proteins central to cellular processes in specific genotypes. Although many transcripts detected were highly similar to genes previously detected in ocean metagenomic surveys, a significant fraction (≈50%) were unique. Thus, microbial community transcriptomic analyses revealed not only indigenous gene- and taxon-specific expression patterns but also gene categories undetected in previous DNA-based metagenomic surveys.

bacterial communities | metagenomics | metatranscriptomics | marine | cDNA

Cultivation-independent genomic approaches have greatly advanced our understanding of the ecology and diversity of microbial communities in the oceans (1, 2). Metagenomic methods applied in a variety of microbial habitats have led to the discovery and characterization of new genes and gene products from uncultivated microorganisms (3), assembly of whole genomes from community DNA sequence data (4), and comparisons of community gene content among diverse microbial assemblages (4–9). Recently, a very large metagenomic sampling survey was conducted in ocean surface waters, doubling the number of predicted protein sequences in public databases (10). All currently available data suggest that gene and protein "sequence space" still remain largely under sampled.

At the same time, studies of cultured members of the microbial community, such as *Prochlorococcus*, are helping to further link the ecology of genes and the ecology of organisms (11). From the considerable *Prochlorococcus* diversity observed in metagenomic datasets, clear structure has emerged, including clusters of sequence similarity and chromosomal hot spots for rearrangements (6, 8, 10). Meanwhile, laboratory studies have described physiological differentiation among isolates (12), and field surveys have documented the distribution of ecotypes in the oceans (13). These cross-scale comparisons provide a useful approach in which taxon-specific metagenomic information can be embedded and understood in the context of ecological and physiological data.

Given current research trends, it seems likely that metagenomic datasets will continue to grow rapidly and soon will dwarf whole-genome sequence datasets derived from cultivated microorganisms. The nature, size, and complexity of this information present formidable challenges to analyses and interpretation. In addition, although these data provide information about genome content, there is no clear indication of gene expression or expression dynamics. Although techniques like quantitative PCR (qPCR) can be used to quantify gene expression in natural samples, these are limited usually to measurement of a small number of known genes. Many questions remain to be answered. What fraction of the many new genes discovered in metagenomic datasets are actually expressed? Of the many hypothetical genes present, which are significantly expressed, and what is their function? What are the dynamics and time scales for gene expression in different microbial species, gene suites, and environments?

Measuring bacterial and archaeal gene expression in the wild has been challenging. The half-life of mRNA is short (14, 15), and mRNA in bacteria and archaea usually comprises only a small fraction of total RNA. Several approaches for overcoming these challenges recently have been developed. In one approach, ribosomal RNA (rRNA) subtraction was used in combination with randomly primed RT-PCR to generate microbial community cDNA for cloning and downstream sequence analysis (16). Although preliminary results were encouraging, relatively large sample volumes (≈10 liters) and long sample collecting times were required. Linear RNA amplification methods have been widely used to study gene expression in eukaryotic tissues (17, 18) but are not generally applicable to bacterial and archaeal mRNA because of the requirement of a poly(A) tail. Wendisch *et al.* (19) developed a method for the polyadenylation of bacterial messenger RNA using *Escherichia coli* poly(A) polymerase, which facilitates preferential isolation of bacterial mRNA from rRNA in crude extracts. This approach has been adapted in a commercially available kit (MessageAmp

Author contributions: J.F.-L. and Y.S. contributed equally to this work. J.F.-L., Y.S., G.W.T., S.W.C., and E.F.D. designed research; Y.S. performed research; S.C.S. contributed new reagents/analytic tools; J.F.-L., Y.S., G.W.T., M.L.C., S.W.C., and E.F.D. analyzed data; and J.F.-L., Y.S., G.W.T., M.L.C., S.W.C., and E.F.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: DNA and cDNA sequences reported in this paper have been deposited in the GenBank database (accession nos. SRA000262 and SRA000263).

‡To whom correspondence may be addressed. E-mail: delong@mit.edu or chisholm@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0708897105/DC1.

© 2008 by The National Academy of Sciences of the USA

Table 1. Characterization of the pyrosequenced DNA and cDNA libraries from the microbial community analyzed in this study

	DNA library	cDNA library
Total number of reads	414,323	128,324
Average length, bp	110	114
Number of rRNA reads	5,877	67,859
Total base pairs, Mb	45.4	14.7
Number of NCBI-nr hits	205,747 (50% of reads)	7,275 (13% of reads)
Number of GOS peptide hits	290,741 (70% of reads)	23,203 (43% of reads)

Only sequences with bit score cutoffs ≥40 were considered hits.

II-Bacteria Kit, Ambion), which couples microbial RNA polyadenylation with a linear amplification step using T7 RNA polymerase (20). Polyadenylation-dependent RNA amplification approaches have been used in studies of cultured microbes using single-genome microarrays (21, 22). We adapted this approach to enable the synthesis of microbial community cDNA from small amounts of mixed population microbial RNA. Specifically, after polyadenylation of nanogram quantities of RNA (19), the RNA was linearly amplified with T7 RNA polymerase (20) and then converted to cDNA. The cDNA was directly sequenced by pyrosequencing, avoiding the need to prepare clone libraries and their associated biases (23, 24). By pyrosequencing both genomic DNA and cDNA from the same sample, the abundance of cDNA copies can be normalized to corresponding gene copy numbers in the community DNA pool.

We report here the application, validation, and field testing in the North Pacific Subtropical Gyre (25) of these methodologies for studying microbial community gene expression. We used the technique to analyze the expression of genes across the entire microbial community, to assess the taxonomic origins of the expressed genes, and to examine gene expression in *Prochlorococcus*, the dominant phototroph in the surface waters at this site. Genes from *Prochlorococcus* are highly represented in metagenomic databases (5, 8, 10), and extensive genomic and transcriptomic data exists from culture studies (6, 26–28) and so were useful in guiding the interpretation of field observations.

Results and Discussion

Assessing the Fidelity of Bacterial mRNA Amplification. We tested the fidelity of the RNA amplification technique using *Prochlorococcus* cultures and custom-designed Affymetrix arrays [see supporting information (SI) *Methods*] (27). Levels of gene expression measured from the amplified *Prochlorococcus* RNA compared favorably with those of unamplified RNA for protein coding genes (r^2 between 0.85 and 0.92; SI Fig. 4), and the results were highly reproducible (r^2 between 0.94 and 0.99 for biological replicates; SI Fig. 5). Linearly amplified RNA also revealed the same physiologically relevant changes in gene expression, as did unamplified RNA in an experiment designed to examine the response of strain MIT9313 to phosphate starvation (SI Fig. 6) (27). Both amplified and unamplified RNA identified the same four genes, all involved in phosphate acquisition, as highly up-regulated under phosphate starvation. In contrast to this high fidelity for mRNA, rRNA transcripts were consistently underrepresented in amplified versus unamplified RNAs (SI Fig. 7), reflecting a preferential polyadenylation of mRNA, consistent with previous reports of this polyadenylation bias in crude extracts (19) and with the known inefficiency of amplification of molecules with a high degree of secondary structure (29).

Field Testing Microbial Gene Expression Profiling in the Open Ocean. As a field test, we analyzed a picoplanktonic sample collected from 75-m depth at the well characterized Hawaii Ocean Time Series (HOT) station ALOHA in the North Pacific Subtropical Gyre (25). Because metagenomic analyses already have been performed at this site (5), and the cyanobacterium *Prochlorococcus* comprises a large fraction of its microbial communities (30), significant databases exist to facilitate the interpretation of our field results. The detection frequency for any given transcript in the community depends on the abundance of transcript-bearing cells and the average number of transcripts per cell. We recovered sequence data from both cDNA and genomic DNA in the same sample, which facilitated representation of specific cDNA classes relative to their occurrence in the genomic DNA pool.

The diversity of sequences captured in the cDNA and DNA reads (Table 1) was determined by comparing all sequences to the National Center for Biotechnology Information nonredundant protein database (NCBI-nr; as of March 28, 2007) and to predicted peptides from the recent Global Ocean Sampling (GOS) metagenomic dataset (31). The number of cDNA and DNA reads with significant database matches (bit score >40; SI Fig. 8) was higher with GOS peptides than with the NCBI-nr database. A large number of GOS matches was expected because the GOS data are derived from similar microbial communities and contain a larger number of total protein sequences. The enrichment in GOS matches over NCBI-nr matches was much greater for the cDNA library (≈3-fold) compared with the DNA library (≈1.4-fold) (Table 1). The fraction of reads matched in the cDNA, however, still was relatively low (43% of total reads) compared with the DNA library (70% of reads). The large proportion of unmatched cDNA reads in part may reflect the presence of novel, rare genes, not detected in the GOS metagenomic survey, that nevertheless contribute significantly to the microbial community expression profile.

To corroborate the results, we selected a suite of genes and performed RT-qPCR and qPCR on the same RNA and DNA samples analyzed by pyrosequencing (*SI Methods*, SI Table 2, and SI Fig. 9). Three different gene expression classes were investigated: (*i*) genes shared in both genomic DNA and cDNA sequence datasets but with higher relative frequency in the cDNA pool, (*ii*) genes present in both genomic DNA and cDNA datasets but with lower relative frequency in the cDNA pool, and (*iii*) genes detected in the cDNA but not in the genomic DNA sequence dataset. The calculated RT-qPCR/qPCR ratios followed the same trends as gene expression patterns inferred from cDNA/DNA pyrosequencing analyses (SI Fig. 9). In some cases, the RT-qPCR/qPCR analysis appeared more sensitive for detecting a broader range of gene expression patterns. For example, genes found only in the cDNA sequence dataset were detected by qPCR in both RNA and DNA samples, which likely reflects the limited extent of sampling depth of the DNA pyrosequencing relative to indigenous genetic complexity.

To evaluate the protein family representation in our dataset and to functionally categorize genes, reads from both cDNA and DNA libraries were assigned to GOS protein clusters. DNA reads were assigned to 35,178 different GOS protein clusters, and cDNA reads were assigned to 4,376 clusters. There were

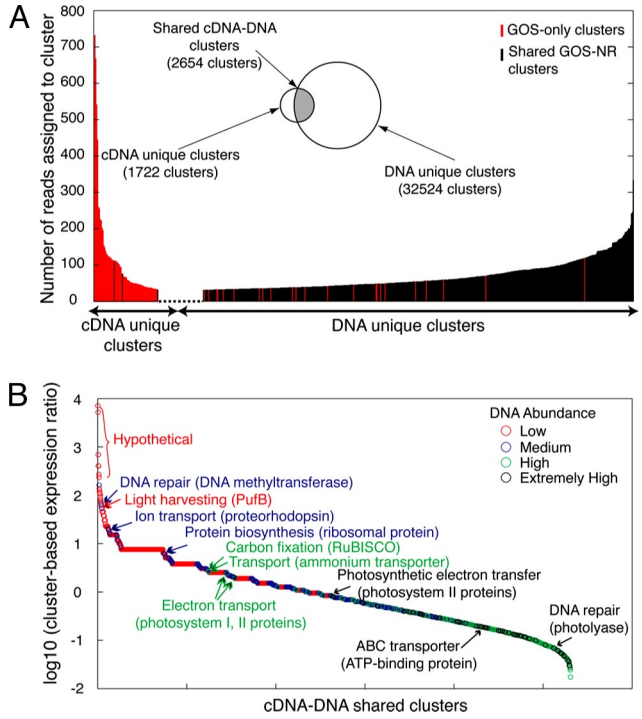

Fig. 1. Community-level gene expression profile based on GOS peptide database. (*A*) GOS protein clusters with DNA or cDNA matches at bit scores ≥40 are shown in the Venn diagram. Numbers of reads assigned to GOS protein clusters, when >70, are plotted for both cDNA-unique protein clusters and DNA-unique protein clusters. GOS protein clusters shared by DNA and cDNA libraries (shaded in gray) were further illustrated in *B*. (*B*) GOS protein clusters shared by cDNA and DNA libraries were ranked by their cluster-based expression ratio (representation of each cluster in the cDNA library normalized by its representation in the DNA library). Furthermore, each protein cluster was categorized (and color-coded) according to its abundance in the DNA library. Representative protein clusters were highlighted from each category and discussed in the text.

2,654 clusters that had both DNA and cDNA reads (Fig. 1). The smaller number of cDNA assignments is in part because the total number of cDNA reads was only one-eighth the number of DNA reads after removing rRNA sequences. Another factor likely responsible for the decreased number of high-quality sequence reads in the cDNA relative to genomic DNA includes the inefficient enzymatic removal of the poly(A) tail produced during the amplification of the mRNA. These homopolymers cause a significant number of sequences to be filtered out during processing because of lower-quality scores, low flow counts, and carry forward (premature incorporation of bases caused by incomplete flushing) (see *Materials and Methods* and ref. 24). Nevertheless, 40% of the cDNA-containing GOS clusters (referred to as "cDNA-unique clusters" hereafter) did not overlap with those in the DNA library, suggesting that the full diversity of sequences was undersampled in both the DNA and cDNA pools. This difference in representation is supported by rarefaction analysis, showing a near linear increase in the rate of recovery of GOS protein clusters with increasing number of sequence reads for both cDNA and DNA (SI Fig. 10). This finding is consistent with other large-scale metagenomic surveys that showed no sign of sequencing saturation for similar marine microbial communities (31, 32).

To maximize functional genomic information drawn from the data, the 2,654 GOS protein clusters (protein families) that were represented in both the DNA and cDNA libraries were analyzed further, calculating the number of cDNA reads matching a given GOS protein cluster, divided by the number of corresponding DNA reads in the same cluster (see *Materials and Methods*)—the "cluster-based expression ratio." This approach allowed us to bypass the difficulties associated with traditional annotation of short pyrosequencing reads (average trimmed length of ≈96 bp), which would have segmented the reads into many apparently unrelated, nonoverlapping clusters, even though they were potentially derived from the same gene. This level of analysis allows us to look at the expression profile of the microbial community at the level of protein family, without losing the resolution inherent in the data.

The 2,654 shared GOS protein clusters were categorized based on their abundance in the DNA library (low, medium, high, and extremely high; SI Fig. 11). Protein clusters with the highest cluster-based expression ratios (up to 10^3 higher than the average ratio) tended to fall into the low DNA abundance category (Fig. 1*B*). This observation, together with apparent high expression levels in cDNA-unique clusters, suggested the presence of actively transcribed genes that are relatively low in abundance in the total community. Interestingly, these highly expressed protein clusters consist mostly of hypothetical proteins that are found only in the GOS peptide database (Fig. 1 and SI Table 3). The high degree of sequence similarity (up to 100%; average 89.5%) between these GOS-only hypothetical protein matches and the cDNA reads supports the GOS gene predictions and confirms that these genes are actively expressed *in situ*. Conversely, the DNA-unique clusters are composed of protein families that are well represented in current protein databases (e.g., NCBI-nr and fully sequenced microbial genomes; Fig. 1 and SI Table 4). This finding indicates that cDNA analysis captures novel genes, with potentially important functions, that have escaped detection even in the largest metagenomic DNA survey conducted to date.

Highly Expressed Gene Categories in Known Metabolic Pathways. Expression patterns of environmentally diagnostic genes can provide significant insight into microbial processes active in the environment. For example, genes involved in microbial phototrophy—e.g., oxygenic and anoxygenic photosynthesis and photoheterotrophy—were among the most highly expressed classes in cluster-based expression ratios (Fig. 1*B*), even though the sample was collected 3 h before sunrise.

In the case of genes related to oxygenic photosynthesis, ribulose bisphosphate carboxylase (RuBisCo) large subunit (*rbc*L) homologs, encoding subunits of the key enzyme in the Calvin Cycle carbon fixation enzyme, were among the highly expressed genes in the sample (Fig. 1*B*). Expression levels of this gene were on a par with those of glutamine synthase (GS), suggesting high expression levels of this key enzyme in nitrogen metabolism that is found in all microorganisms. RuBisCo and GS gene copies were present in comparable numbers in the microbial genomic DNA of our sample, in contrast to the recently reported GOS datasets, where relatively low numbers of the *rbc*L gene were identified relative to GS (31). With respect to alternative forms of phototrophy, several protein clusters associated with aerobic, anoxygenic phototrophy showed extremely high cluster-based expression ratios (Fig. 1*B*). These proteins include light-harvesting protein β-chain (PufB), photosynthetic reaction center cytochrome *C* subunit (PufC), and chlorophyllide reductase subunit Y (BchY), which all appear to be derived from Alphaproteobacteria closely related to *Roseobacter* species (33). Although these correspond to relatively low abundances in the DNA libraries, their high expression levels support the potential ecological importance of aerobic anoxygenic phototrophy to microbial species in the open ocean.

Another important family of proteins involved in phototrophy are the proteorhodopsins (PRs), a group of membrane proteins that function as a light-driven proton pump (3). PR genes were not only abundant in community genomic DNA but also were

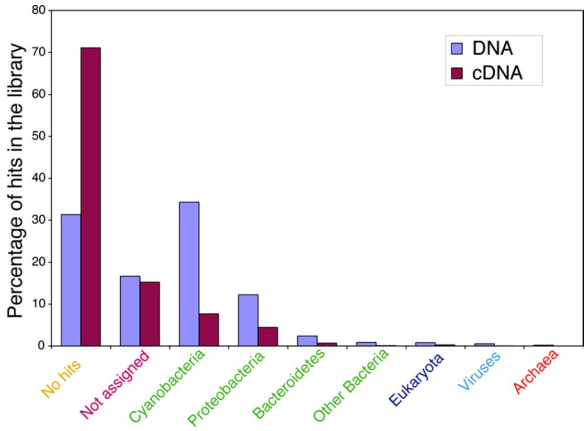

Fig. 2. Distribution of different phylogenetic groups in DNA and cDNA libraries. Percentages of the different phylogenetic groups were calculated from the MEGAN analysis results at the phylum level cutoff (SI Table 5 shows a detailed list of the distribution of number of hits and percentages for all phyla). Not assigned reads are sequences with an NR hit but a bit score <40.

among the most highly expressed genes in the cDNA pool (Fig. 1*B*). Preliminary taxonomic assignments suggest that the expressed PR genes were derived from diverse microbial taxa, supporting their general ecological significance in planktonic microbial communities (3, 34). Heterologous expression experiments have confirmed the ability of PR to function as a proton pump and enable photophosphorylation in *E. coli* (3, 35). Moreover, some (but not all) PR-containing bacteria display enhanced growth rates and cell yields in the presence of light (36, 37).

Putative Taxonomic Origins of Expressed Genes. Metatranscriptomic analyses, in principle, can be used to associate specific microbial taxa with *in situ* expression dynamics. However, phylogenetic inference based on protein-coding genes is highly dependent on a given gene's conservation across taxa, the depth of taxonomic sampling, taxon richness and evenness in the sample, and sequence read length. Further, taxonomic inferences also have the potential to be confused by horizontal gene transfer events (38). With these caveats in mind, we performed a preliminary taxonomic assessment of DNA and cDNA reads using MEGAN (39), software that assigns putative taxonomic origins based on BLAST outputs and NCBI taxonomic hierarchy. Not surprisingly, based on their known abundance in the wild and their abundance in the genomic databases, the genus *Prochlorococcus* and Alphaproteobacteria (genus *Pelagibacter*) were the two most highly represented taxonomic groups in both DNA and cDNA libraries (Fig. 2 and SI Table 5). Another noteworthy observation was the detection of expressed genes of viral origin, suggesting there was active viral infection occurring in cells *in situ* in the sample we analyzed (SI Table 5). The most common viral transcripts were related to the major capsid protein of myoviridae. Previous metagenomic analyses reported a high viral abundance in the cellular fraction from the same depth and site (5). For the most abundant groups, there was general agreement between the taxonomic origins of sequence reads in the DNA and cDNA datasets.

Evaluating Gene Expression in a Naturally Occurring *Prochlorococcus* Assemblage. The vast majority (>90%) of putative *Prochlorococcus* reads shared highest sequence similarity with strains MIT9301, AS9601, and MIT9312, all representatives of the high light-adapted eMIT9312 ecotype (40). This result (data not shown) is consistent with depth-specific ecotype abundance data based on qPCR analysis of the rRNA internally transcribed

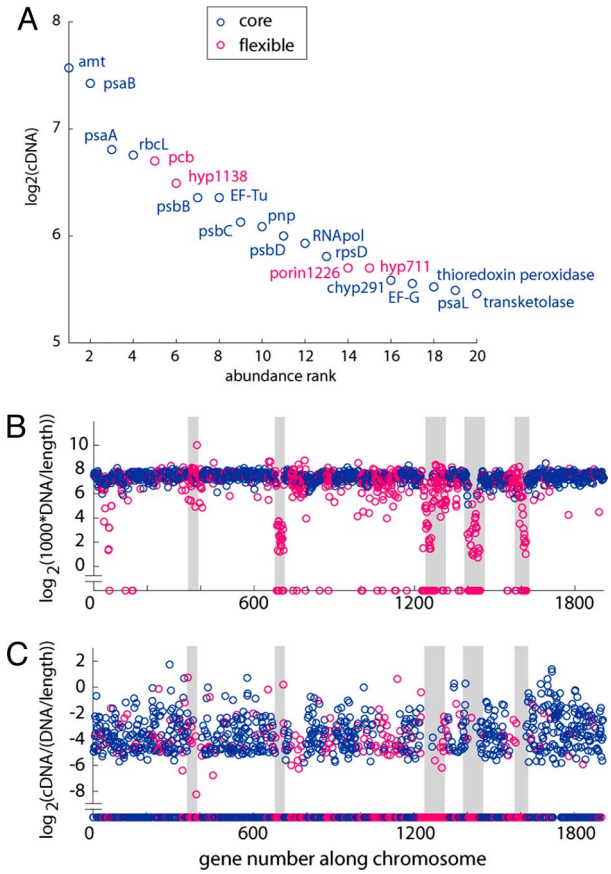

Fig. 3. *Prochlorococcus* gene and transcript abundance using strain MIT9301 as a reference genome. (*A*) Rank abundance of the 20 genes with highest frequency in the raw cDNA, reflecting transcription of the entire *Prochlorococcus* population. (*B*) Frequency of DNA hits from the natural sample along the genome of MIT9301 normalized to gene length. (*C*) Frequency of cDNA hits from the natural sample normalized to the DNA values in *B*. Gray bars indicate the location of genomic islands identified through whole-genome analysis of cultured isolates (6). Core genes, genes present in all genomes of *Prochlorococcus* sequenced, are shown in blue. Flexible genes, genes not present in all genomes of *Prochlorococcus* sequenced, are shown in pink.

spacer (ITS) region (13). Our current analysis using short pyrosequencing sequence reads from both DNA and cDNA therefore supports ecotype distributions inferred from independent analyses using a single taxonomic marker, the ITS.

Observed frequencies of the putative *Prochlorococcus* cDNA sequences reflect which genes are the most highly expressed in the *Prochlorococcus* assemblage sampled. These highly expressed genes include ammonium uptake (*amt*), photosynthesis (*psa*AB), and carbon fixation (*rbc*L) genes, pointing to key biogeochemical processes being driven, in part, by *Prochlorococcus* (Fig. 3*A* and SI Table 6). Two of the top 20 most highly expressed *Prochlorococcus* genes were hypothetical proteins: P9301_11381, which has orthologs only in the other MIT9312-like genomes (AS9601, MIT9312, and MIT9215), and P9301_07111, which has no orthologs in other *Prochlorococcus* genomes (but is paralogous to P9301_04361) (SI Table 6). High-level expression of hypothetical proteins has previously been observed in *Prochlorococcus* under nutrient limitation in laboratory experiments (27, 28). The current data indicate the potential relevance of these proteins to *Prochlorococcus* in its native environment. When a gene-length correction is applied (see *Materials and Methods*, SI Fig. 12, and SI Table 6), additional hypothetical proteins (P9301_03541 and P9301_02451) with high

per-copy transcript abundance appear to be rare in the population but are highly expressed.

The *Prochlorococcus* core genome (i.e., those genes shared by all sequenced *Prochlorococcus* isolates) consists of ≈1,250 genes (41). The "flexible" genome represents the remaining genes found in one or more genomes, and many of these variable genes are concentrated in genomic islands (6). Using strain MIT9301 as a reference, we calculated the abundance of genes belonging to the core and flexible genomes in both the DNA and cDNA libraries. In the DNA library, all *Prochlorococcus* core genes were represented with roughly equal abundance, supporting the idea that these genes are conserved and present in single copy in virtually every *Prochlorococcus* cell (Fig. 3*B*). In contrast, genes belonging to the MIT9301 flexible genome had highly variable occurrence in the DNA library, suggesting that the natural population likely harbors a different suite of such genes. In the cDNA library, core genes involved in photosynthesis and carbon fixation, for instance, were highly represented, but, surprisingly, a number of genes belonging to the flexible genome, some of which are located in genomic islands in MIT9301, also were highly represented (Fig. 3 *A* and *C*). Thus, some of these island genes appear to be highly expressed, corroborating laboratory evidence and suggesting that they are likely functionally important to naturally occurring *Prochlorococcus*. Furthermore, the majority of "flexible" genes, and hypothetical genes, were found in the cDNA pool and expressed at levels comparable to most other core genes, further indicating their significance in the biology and ecology of *Prochlorococcus.*

Microbial Community Transcriptomics: Prospects and Challenges. Many challenges are associated with the interpretation of microbial gene expression patterns at the community level. These arise in part from the remarkable diversity and complexity of microbial communities in the ocean environment, logistics associated with field sampling, and the lack of comprehensive representation in metagenomic databases. Rapid collection and processing of samples for gene expression studies, for example, still presents significant challenges. Although our approach used relatively small volumes (1 liter) and short filtration times (<15 min), there still remains significant room for improvement. Other factors that will influence community transcriptomic analyses include the specifics of mRNA synthesis and degradation rates, environmental conditions at the time of sampling (time of day, for example), sequence read size and target gene size, and the specific method used for gene identification and annotation. Some of these variables can be controlled or improved, and others are inherent to the specific environment or community being sampled.

It is well accepted that longer sequence reads are generally more informative, allowing more robust annotation. Side-by-side comparisons of Sanger dideoxy sequences versus pyrosequences derived from the same metagenomic samples, however, have been generally consistent and comparable (9, 42). The sequence reads in our dataset have an average size of ≈96 bp, sufficient for general functional annotation and, in the case of *Prochlorococcus*, for assignment of reads to specific genes and ecotypes. For as-yet-uncultivated microorganisms, 100 bp is not always sufficient for specific gene assignment. Improvements in pyrosequencing, however, now produce >230-bp length reads. Further improvements in pyrosequencing read lengths are anticipated, which will improve accuracy and precision in future microbial community transcriptomics studies.

Despite the caveats and potential improvements, we have shown metatranscriptomic sequencing and characterization is sufficient to identify many expressed biological signatures in complex biological samples such as seawater. Whole-community analysis relying on gene family clustering for analyses of pyrosequencing reads revealed clear patterns in community gene expression for individual taxa, specific genes, and within protein families. Taxon-specific analyses focusing on *Prochlorococcus* provided deep insight into the most highly expressed genes among these populations. Interestingly, both in the case of the whole community as well as in the case of *Prochlorococcus*, hypothetical genes were among the most highly expressed, underlining the potential importance of these unidentified proteins. The fact that a large fraction of cDNA reads were not present in the available databases, including the GOS database, indicates that we have just scratched the surface of the microbial metabolic diversity present in the ocean.

Metatranscriptomics (ref. 16 and this report) and proteomics (43, 44) represent two approaches in microbial ecology that have potential to significantly leverage, apply, and extend existing microbial metagenomic datasets. The two approaches each measure a different component and dynamic of the macromolecular pool, reflecting the different regulatory controls, expression rates, and turnover kinetics of mRNAs and proteins. Although transcriptomics has potential to reveal the near-instantaneous responses to environmental fluctuation, proteomics more directly reflects the immediate catalytic potential of the microbial community. In conjunction with metagenomic data, these approaches offer significant promise to advance measurement and prediction of *in situ* microbial responses and activities in complex, naturally occurring, or engineered microbial communities.

Materials and Methods

Sampling. Seawater was collected at the HOT station ALOHA (22°44′N, 158°2′W), 75-m depth, on March 9, 2006, 03:30 a.m. local time. Hydrocasts for sampling and hydrographic profiling were conducted by using a conductivity-temperature-depth (CTD) rosette water sampler equipped with 24 Scripps 12-liter sampling bottles aboard the R/V Kilo Moana. Continuous vertical profiles of physical and chemical parameters thus were recorded. DNA and RNA extraction, processing, and sequencing are detailed in the *SI Methods*.

RNA Amplification and cDNA Synthesis. Approximately 5 μl of RNA (≈100 ng total) was amplified by using MessageAmp II-Bacteria Kit (Ambion) following the manufacturer's instructions. Briefly, the method is based on polyadenylation of the 3′ end of total RNA. The A-tailed RNA is reverse-transcribed primed with an oligo(dT) primer containing a T7 promoter sequence and a restriction enzyme (BpmI) recognition site sequence [T7-BpmI-$(dT)_{16}$VN], then double-stranded cDNA is synthesized. Finally, the cDNA templates are transcribed *in vitro* (37°C for 6 h), yielding large amounts of antisense RNA (aRNA; ≈1,000-fold amplification). The aRNA is polyadenylated and further reverse-transcribed to cDNA with the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen). Finally, ≈2 μg of cDNA is digested with BpmI, purified, and used for pyrosequencing.

Pyrosequencing. DNA and cDNA libraries were constructed as previously described (23, 45) and sequenced with a Roche GS20 DNA sequencer. A full run of the sequencer yielded 45,380,301 bp from 414,323 reads (110-bp average length) from the DNA library, and 14,675,424 bp from 128,324 reads (114-bp average length) from the cDNA library (Table 1). The lower number of cDNA library reads may be because of shorter cDNA fragments and highly polymeric sequences resulting from inefficient removal of poly(A) tails introduced during mRNA amplification. To pass GS20 quality filters, flow-grams for each read require at least 84 flows (21 cycles or ≈50 bp) and <5% of flows with ambiguous bases (N) and <3% of flow-gram values between 0.5 and 0.7.

Analysis of Metagenomic GS20 DNA and cDNA Data. DNA and trimmed non-RNA cDNA reads were compared with the NCBI-nr (as of March 28, 2007) and GOS peptide databases using BLASTX (46). Top BLASTX hits with bit score >40 were used to assign DNA and cDNA reads to GOS peptides and NCBI-nr proteins (Table 1). Reads assigned to GOS peptides were linked to GOS protein clusters and associated GO, Pfam, and TIGRfam annotations (if available). Additional details are provided in *SI Methods*.

ACKNOWLEDGMENTS. We thank the HOT team, the captain and crew of the R/V Kilo Moana for the expert assistance at sea, and Chon Martinez for preparing the sample DNA. This work was supported by the Gordon and Betty Moore Foundation (E.F.D. and S.W.C.), the National Science Foundation (S.W.C.), National Science Foundation Microbial Observatory Award MCB-0348001 (to E.F.D.), the Department of Energy Genomics GTL Program (E.F.D. and S.W.C.), and the Department of Energy Microbial Genomics Program (E.F.D. and S.W.C.). This article is a contribution from the National Science Foundation Science and Technology Center for Microbial Oceanography: Research and Education (C-MORE).

1. Giovannoni SJ, Stingl U (2005) Molecular diversity and ecology of microbial plankton. *Nature* 437:343–348.
2. DeLong EF, Karl DM (2005) Genomic perspectives in microbial oceanography. *Nature* 437:336–342.
3. Beja O, Spudich EN, Spudich JL, Leclerc M, DeLong EF (2001) Proteorhodopsin phototrophy in the ocean. *Nature* 411:786–789.
4. Tyson GW, *et al*. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.
5. DeLong EF, *et al*. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311:496–503.
6. Coleman ML, *et al*. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768–1770.
7. Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet* 6:805–814.
8. Venter JC, *et al*. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304:66–74.
9. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444:1027–1031.
10. Rusch DB, *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PloS Biol* 5:398–431.
11. Coleman ML, Chisholm SW (2007) Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* 15:398–407.
12. Moore LR, Ostrowski M, Scanlan DJ, Feren K, Sweetsir T (2005) Ecotypic variation in phosphorus acquisition mechanisms within marine picocyanobacteria. *Aquatic Microb Ecol* 39:257–269.
13. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* 311:1737–1740.
14. Selinger DW, Saxena RM, Cheung KJ, Church GM, Rosenow C (2003) Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res* 13:216–223.
15. Andersson AF, Lundgren M, Eriksson S, Rosenlund M, Bernander R, Nilsson P (2006) Global analysis of mRNA stability in the archaeon Sulfolobus. *Genome Biol* 7:R99.
16. Poretsky RS, *et al*. (2005) Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* 71:4121–4126.
17. Feldman AL, *et al*. (2002) Advantages of mRNA amplification for microarray analysis. *Biotechniques* 33:906–912, 914.
18. Moll PR, Duschl J, Richter K (2004) Optimized RNA amplification using T7-RNA-polymerase based *in vitro* transcription. *Anal Biochem* 334:164–174.
19. Wendisch VF, Zimmer DP, Khodursky A, Peter B, Cozzarelli N, Kustu S (2001) Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Anal Biochem* 290:205–213.
20. Vangelder RN, Vonzastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc Natl Acad Sci USA* 87:1663–1667.
21. Moreno-Paz M, Parro V (2006) Amplification of low quantity bacterial RNA for microarray studies: time-course analysis of *Leptospirillum ferrooxidans* under nitrogen-fixing conditions. *Environ Microbiol* 8:1064–1073.
22. Rachman H, Lee JS, Angermann J, Kowall J, Kaufmann SH (2006) Reliable amplification method for bacterial RNA. *J Biotechnol* 126:61–68.
23. Margulies M, *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
24. Huse SM, Huber JA, Morrison HG, Sogin ML , Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 8:R143.
25. Karl DM, *et al*. (1996) Seasonal and interannual variability in primary production and particle flux at Station ALOHA. *Deep Sea Res Part II Topical Stud Oceanogr* 43:539–568.
26. Rocap G, *et al*. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042–1047.
27. Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* 103:12552–12557.
28. Tolonen AC, *et al*. (2006) Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol Syst Biol* 2:53.
29. von Wintzingerode F, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* 21:213–229.
30. Campbell L, Nolla HA, Vaulot D (1994) The importance of *Prochlorococcus* to community structure in the central North Pacific Ocean. *Limnol Oceanogr* 39:954–961.
31. Yooseph S, *et al*. (2007) The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PloS Biol* 5:432–466.
32. Sogin ML, *et al*. (2006) Microbial diversity in the deep sea and the underexplored ''rare biosphere.'' *Proc Natl Acad Sci USA* 103:12115–12120.
33. Oz A, Sabehi G, Koblizek M, Massana R, Beja O (2005) Roseobacter-like bacteria in Red and Mediterranean Sea aerobic anoxygenic photosynthetic populations. *Appl Environ Microbiol* 71:344–353.
34. Sabehi G, *et al*. (2005) New insights into metabolic properties of marine bacteria encoding proteorhodopsins. *PloS Biol* 3:1409–1417.
35. Martinez A, Bradley AS, Waldbauer JR, Summons RE, DeLong EF (2007) Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. *Proc Natl Acad Sci USA* 104:5590–5595.
36. Gomez-Consarnau L, *et al*. (2007) Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* 445:210–213.
37. Giovannoni SJ, *et al*. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* 438:82–85.
38. Boucher Y, *et al*. Lateral gene transfer and the origins of prokaryotic groups. *Ann Rev Genet* 37:283–328.
39. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
40. Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S–23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* 68:1180–1191.
41. Kettler GC, *et al*. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:e231.
42. Gill SR, *et al*. (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312:1355–1359.
43. Ram RJ, *et al*. (2005) Community proteomics of a natural microbial biofilm. *Science* 308:1915–1920.
44. Lo I, *et al*. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* 446:537–541.
45. Poinar HN, *et al*. (2006) Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* 311:392–394.
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.