Metatranscriptomics: Eavesdropping on Complex Microbial Communities

Large-scale sequencing of mRNAs retrieved from natural communities provides insights into microbial activities and how they are regulated

Mary Ann Moran

At any moment, an estimated 10^{30} bacterial and archaeal genes are mediating essential ecological processes throughout the world. The new field of metatranscriptomics, using an approach that sequences microbial genes expressed within intact natural communities, allows us to understand microbial gene expression patterns. It is now feasible to deeply sequence the assortment of microbial community transcripts from a particular time and place, whether from bacteria, archaea, or small eukaryotes in the ocean, the soil, or the human gut. Putting aside the challenge of correctly assigning functions to the mRNAs being sequenced, the scientific promise of identifying all the processes simultaneously mediated by an undisturbed microbial assemblage is apparent. Essentially, we can eavesdrop on microbial ecology.

Summary

- Metatranscriptomics overcomes constraints inherent in the use of qPCR and microarrays to monitor gene expression in complex communities, avoiding limits on numbers of genes surveyed, a need to select what genes to target, and requirements for probes or primers.
- Current studies are providing microbial expression profiles from diverse ecosystems ranging from marine surface waters to grassland soils.
- Comparative and experimental metatranscriptomics hold enormous promise for understanding the regulation of microbial activities in response to environmental drivers and defined manipulations.

Beyond Metagenomics

Metagenomic approaches to inventory microbial genes in the ocean and soils have fueled a revolution in efforts to understand the genetic potential of uncultured bacteria and archaea. Metagenomics allows us to sequence genomes from a complex assemblage of microbes as a single unit in a culture-independent manner. Deeper sequencing and better annotation improves the quality of knowledge that comes from such analyses.

While it provides information on the possible activities of a microbial community, metagenomics cannot reveal the actual activities at a specific time and place, or how those activities change in response to environmental forces or biotic interactions. The challenge is to narrow down the suite of possible actions—the metagenome—to those that are ongoing at a particular time—the metatranscriptome—and ultimately to identify what is responsible for the difference.

Interest in describing in situ gene expression patterns for natural microbial communities is not new. Most of what we know comes from reverse-transcription PCR (RT-PCR), in which mRNAs extracted from microbial communities are converted to cDNAs and amplified to allow detection with primers targeting conserved regions of genes of interest. This can also be done quantitatively (RT-qPCR) if there is value in estimating the number, not just presence or absence, of transcripts in a community. However, this approach can target only a small number of genes at a time, and the vast nucleotide sequence diversity observed

Mary Ann Moran is a Professor in the Department of Marine Sciences at the University of Georgia, Athens.

for functional genes in nature makes it difficult to design primers that bind only to orthologs of a given gene. Moreover, since non-PCR-based sequence data are available for only a tiny fraction of natural microbial communities, we do not know how well PCR primers match target genes in particular microbial communities, and assessing the results is not easy. Nonetheless, much has been learned by tracking expression of well-conserved marker genes for critical microbial processes, such as those for carbon and nitrogen fixation.

Environmental microarrays overcome the gene number constraint by measuring expression levels of hundreds to thousands of genes at a time. However, designing microarray probes to encompass the full diversity of ortholog sequences encountered in nature is no less of a challenge than doing so for PCR primers. In addition, obtaining sufficient mRNA for replicated microarray studies with environmental samples is difficult. As a result, environmental microarrays have most often been used to survey DNA rather than mRNA, providing little information on community gene expression.

Why Metatranscriptomics?

Like metagenomics, metatranscriptomics (or environmental transcriptomics) involves random sequencing of microbial community mRNA (Fig. 1). Neither primers nor probes are needed, so there is no need to anticipate important genes beforehand and transcripts from microbial assemblages are sequenced with little bias. Further, paralagous sequences which might cross-hybridize on a microarray can be distinguished. The approach is particularly amenable to an experimental framework in which gene expression is monitored while a biotic or abiotic parameter is manipulated. Experimental metatranscriptomics is one of the most powerful tools for understanding the timing and regulation of complex microbial processes within communities and consortia, as well as microbial dexterity in response to changing conditions.

An important payoff of community expression profiling via direct sequencing is that the individual metatranscriptomic studies contribute to a growing community database that can be used to address intractable or unanticipated questions. The sequence of a gene transcript should always be the same, unlike microarray or RT-qPCR data, which are influenced by parameters such as array composition, primer design, and hybridization conditions. Moreover, environmental transcriptome data can be archived in one of the community metagenomic databases, such as CAMERA (http://camera.calit2.net/), MG-RAST (http://metagenomics.nmpdr.org/), or IMG/M (http://img.jgi.doe.gov/cgi-bin/m/main.cgi). Such data will become more valuable as better gene annotations become available.

Some of the first nontargeted glimpses of transcriptomes of natural microbial assemblages came from Rachel Poretsky and colleagues at the University of Georgia, who studied a coastal salt marsh and hypersaline lake, and from Lina Botero and colleagues at Montana State University and the University of Cincinnati College of Medicine, who studied a geothermal soil. Although these initial analyses were low-throughput (<400 sequences), low coverage, and potentially biased by priming methods, they laid the conceptual groundwork for why metatranscriptomics was worth undertaking. Soon after, environmental transcript analysis was coupled for the first time to a high-throughput sequencing technology (454 pyrosequencing) in a study of a grassland microbial assemblage by Sven Leininger and colleagues at the University of Bergen in Norway. This approach and the commercial availability of methods to amplify bacterial and archaeal mRNA, thereby providing sufficient template from small and rapidly processed samples, established the outline for current protocols (Fig. 1).

Challenges of Environmental Transcript Sequencing

Though metatranscriptomics is a small conceptual leap from metagenomics, practical considerations have slowed its development. One difficulty is that bacterial and archaeal mRNAs typically are not polyA tailed, so methods for specific capture of eukaryotic cDNAs are not applicable. This results in coextraction of the more abundant and stable rRNAs, which can lead to a disappointingly low yield of expressed gene sequences in a large-scale sequencing run, potentially as low as 10%. Selectively removing rRNA from the total RNA pool (Fig. 1) or embracing the rRNA sequences for their insight into community structure (if amplification and

Moran: Turning Point in a Hot Tub, Musing about a Marine Bacterium

After Mary Ann Moran moved to Georgia from upstate New York, she kept her skis in the living room "just in case I'd need them" for 10 full years, she says. Later she put them away and otherwise adjusted to the warmer temperatures and temperaments. "I've really come to love the easygoing and very polite southern culture," she says. "Now it takes me a few days to get my chip back on my shoulder when I return to New York."

Moran, 53, is distinguished research professor in marine sciences at the University of Georgia in Athens (UGA), where she studies the marine carbon and sulfur cycles, release of climate-relevant gases from ocean surfaces, ecological genomics, metagenomics, and metatranscriptomics, and the ecology and physiology of marine *Roseobacter* bacteria.

Moran was born in Boston, but soon moved with her family to Huntington, N.Y., on Long Island, where she found herself drawn to the ocean. "I spent lots of time at the beach during my childhood," she says. "And as strange as it sounds for a teenage girl, I actually preferred the beach in winter, when I would walk for miles. I also loved to garden as a kid, and spent a lot of time digging around in the dirt. Overall, I would do anything that would keep me outside, whether at the beach, on a hike, or in our garden."

After her parents came to America from Ireland in 1950, her father worked for United Airlines at Kennedy Airport in the food services division, while her mother worked in retail jobs after Moran, her sister, and brother began high school. Although neither parent attended college, "they had a deep love of reading—someone was always reading in my house—and a passion for education," she says. "They were amazingly supportive in my long trek through college and graduate school, although I know they didn't always quite understand what I was doing or why. But that was OK with them as long as it was what I wanted to do." Several years after Moran's father died, her mother moved to a nearby apartment in Athens. "I see her daily," Moran says.

Moran attended graduate school at Cornell University in Ithaca, N.Y., where she met her husband Ken, an insect evolutionary biologist. "We moved to UGA for his two-year postdoc," she says, adding: "Ha! That was 25 years ago." They have two teenage daughters.

Earlier, Moran received her B.A. in 1977 from Colgate University in Hamilton, N.Y., her M.S. in 1982 from Cornell, and her doctorate in 1987 from the University of Georgia. "Yes, I broke the rule of staying at the same institution where I got my Ph.D., but Ken and I were thrilled to get two faculty positions in the same institution," she says.

Two teachers at Colgate greatly influenced her, according to Moran. Robert Goodwin taught plant taxonomy, and Ronald Hoham taught algal taxonomy and ecology. "I wonder if they remember me," she says. "They were the first two 'naturalists' I had met, with insatiable curiosities about the details and patterns of nature. I remember many lab sessions out and about the Chenango Valley as I came to truly admire their detailed knowledge of things that most people would not even notice. That was probably when ecology got into my blood."

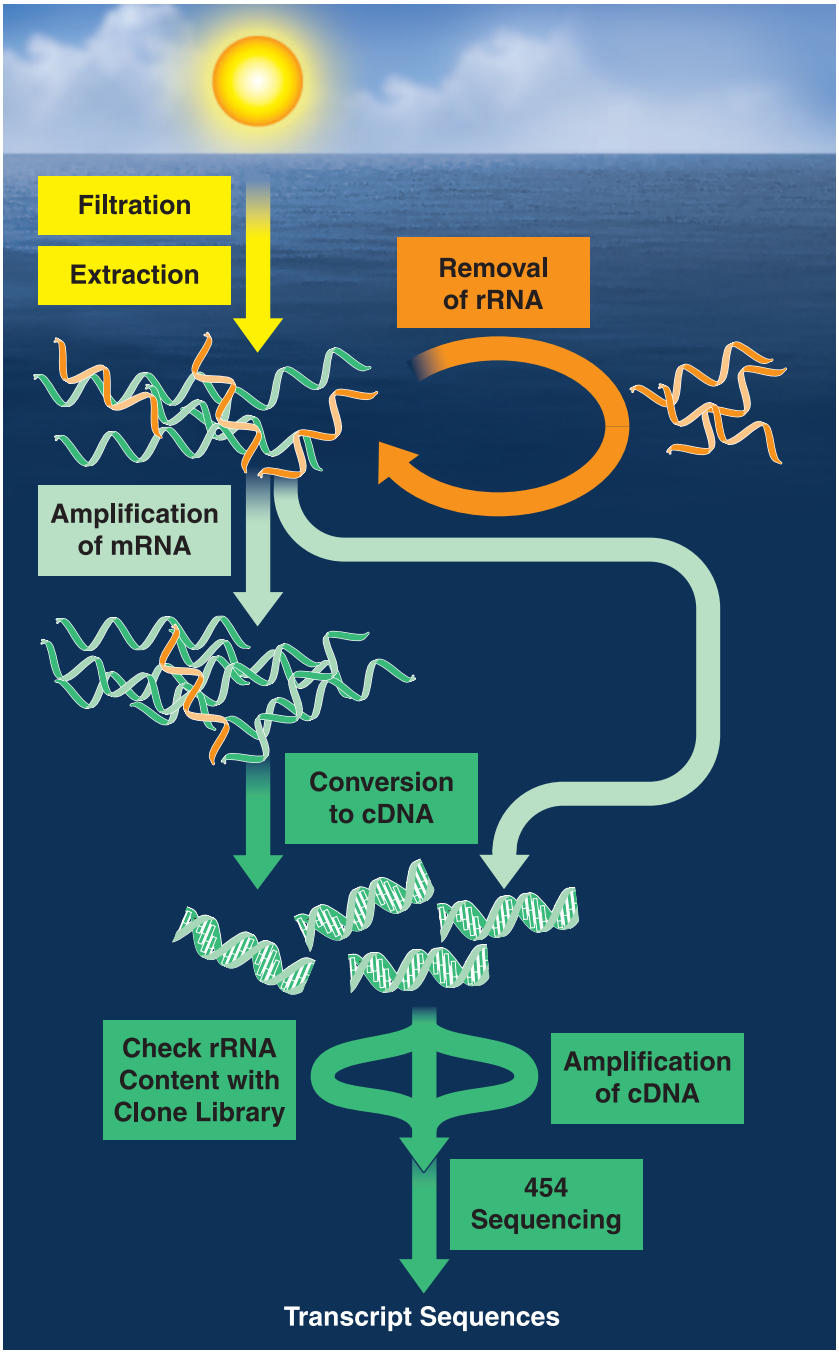
Later, at UGA, Robert Hodson, her doctoral adviser, "taught me how to think big, to find the synthesis," while Lawrence Pomeroy, "who still scares me a little," taught her not to worry about how fashionable or popular her work might be, "but just do good science."

One of the defining moments of her career came in 2001 while she was soaking in a hot tub during a family vacation. She had just received a message from the National Science Foundation, notifying her that her group would receive funding with which to sequence the genome of the marine bacterium *Silicibacter pomeroyi*. "We cultured it from southeastern U.S. seawater," she says. "It was a new species, so we . . . named it after Larry [Pomeroy]. This was back in the day when sequencing was far from routine, and I was terrified. I sat out in the hot tub staring off into the fog-shrouded trees around the cabin and knew that it was a major turning point in my career. Incorporating genomics and functional genomics into my research in marine microbial ecology would be a huge and powerful step toward answering the many questions I was chasing."

Today she has little time for leisure activities because of "the kids and an elderly parent I am caring for," she says. "But if I can sneak away for a few minutes, I love to garden, knit, play with the dog, or read a good book."

Marlene Cimons
Marlene Cimons is a freelance writer in Bethesda, Md

FIGURE 1



Protocol for metatranscriptomic sequencing of microbial communities. Rapid collection and stabilization of RNA reduces degradation and artifacts; in aquatic systems, filter pore size allows different microbial size fractions to be targeted; selective removal of rRNA can increase the final proportion of mRNA sequences; transcripts can be linearly amplified before or after conversion to cDNA; pyrosequencing is cost effective and eliminates biases associated with cloning. This protocol summarizes methods presented in Frias-Lopez et al. Proc. Natl. Acad. Sci. USA **105**:3805–3810, 2008; Gilbert et al. PLoS ONE **3**:e3042, 2008; and Poretsky et al. Environ. Microbiol. doi: 10.1111/j.1462–2920.2008.01863.x., 2009.

other steps that could bias the rRNA pool are avoided) help mitigate this issue. Another technical challenge of working with RNA is a half-life on the order of minutes even under optimal conditions.

On a conceptual level, there is not always a predictable relationship between mRNA abundance and protein activity, since genes can be constitutively expressed and enzyme activities can be regulated posttranscriptionally. Metaproteomics, a promising complementary technique, offers a better link to metabolic function but a less-resolved view of instantaneous regulatory responses. Technical issues of protein extraction, separation, and identification make metaproteomics more onerous, at least for now, than metatranscriptomics.

The most difficult challenge for all "meta" approaches may be that only a small percentage of the vast number of ecologically important genes has been correctly annotated. Sequence datasets contain only the most abundant genes from a very limited number of natural microbial communities; even so, most of these cannot be confidently assigned a function. Many sequences have no close matches in existing databases. In a recent metatranscriptomics study by Poretsky and colleagues, only 33% of the possible protein-encoding sequences had matches to annotated proteins in the NCBI RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/), only 24% had matches to a KEGG pathway (http://www.genome.ad.jp/kegg/kegg2.html), and only 16% to a COG category (http://www.ncbi.nlm.nih.gov/COG/).

Alternatively, sequences may be assigned to gene families with no known function or for which characterized members provide an incomplete picture of the diversity of functions. As an example, automated annotation placed a gene now known to encode an enzyme that demethylates a ubiquitous marine organic sulfur compound in a family whose canonical members mediate gly-

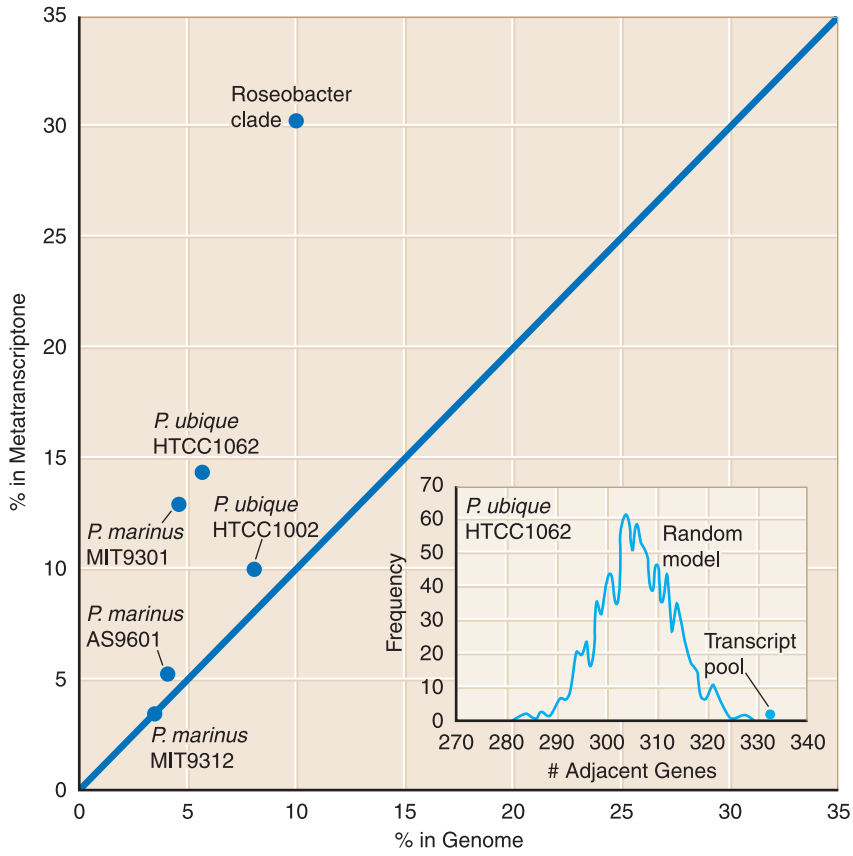
cine degradation, as shown recently by Erinn Howard and colleagues at the University of Georgia. Functional characterization of an enormous number of ecologically relevant genes is a genuinely daunting task for postgenomic environmental microbiology.

Do metatranscriptomic protocols provide a reasonably unbiased sample of the mRNA pool of a microbial community? First, compositions of environmental transcript pools are consistent with models of bacterial and archaeal gene expression. These include operon-based transcription patterns, a bias toward transcripts from highly expressed genes (Fig. 2), and a metatranscriptome that is a simpler subset of the metagenome.

Second, transcript frequency deduced from sequence libraries, including those derived from mRNA or cDNA that has been linearly amplified (Fig. 1), correlates well with measures of transcript frequency based on qPCR analysis. This finding suggests minor biases in the post-mRNA extraction stages of the technique.

Third, taxonomic binning of transcript sequences is consistent with the potential cellular sources of the transcripts in the sample, as determined from comparisons to companion metagenomic libraries, 16S rRNA clone libraries, or cosequenced rRNAs. That is, most transcript sequences have highest identity to genes from taxa known to be present in the microbial community.

FIGURE 2



The composition of metatranscriptomes is consistent with models of gene expression in Bacteria and Archaea. Main figure: Evidence of preferential representation in metatranscriptomes of sequences from genes predicted to be highly expressed. The proportion of predicted highly expressed genes in six taxonomic bins from a North Pacific surface ocean metatranscriptome was compared to the proportion in the reference genomes most similar to them. *P. marinus* = *Prochlorococcus marinus*; *P. ubique* = *Pelagibacter ubique*. Inset figure: Evidence of operon-based expression from a comparison of the number of adjacent genes in 1000 random samples of the *P. ubique* HTCC1062 genome (solid line; mean = 306 genes) to the number of adjacent genes in the North Pacific metatranscriptome assigned to the *P. ubique* taxonomic bin (332 transcripts). Each random genome sample and the *P. ubique* HTCC1062 transcript bin contained the same number of genes or transcripts (1,546). (Redrawn from Poretsky et al. Environ. Microbiol. doi: 10.1111/j.1462–2920.2008.01863.x., 2009).

Insights from Metatranscriptomics

Not surprisingly, genes necessary for the maintenance of basic cellular machinery enabling growth and metabolism dominate microbial transcript pools. Studies by Jorge Frias-Lopez and colleagues at the Massachusetts Institute of Technology in Cambridge; Jack Gilbert and colleagues at the Plymouth Marine Laboratory in the United Kingdom; Tim Urich and colleagues at the Universities of Bergen, Norway, Vienna, Austria, and Pennsylvania State University; and Rachel Poretsky and colleagues at the University of Georgia and the University of California Santa Cruz all indicate that the genes most frequently expressed by bacteria and archaea in marine and soil ecosystems are those involved in transcription and translation (e.g., synthesis of ribosomes, t-RNAs, and associated initiation, elongation, and termination factors); those for protein folding and export; and those for DNA replication and repair.

Genes required for energy generation are also highly expressed. Transcripts encoding proteins

that build machinery for light-driven processes, such as those required for oxygenic photosynthesis, anoxygenic photosynthesis, and proteorhodopsin-based proton pumping, are abundant in marine surface waters. Genes for carbon fixation are likewise well represented in community transcript pools from many microbial communities, as are those for amino acid metabolism and carbohydrate metabolism.

The variety of approaches used in metatranscriptomic studies thus far provides a hint of future applications. In at least two cases, random transcript sequencing has been used to capture organism-specific wild transcriptomes, obtaining insights into activities of ecologically important microbes whose biology is poorly known. These microorganisms include uncultured soil Crenarcheota, which were actively oxidizing ammonia and fixing carbon, according to Urich and colleagues, and the marine cyanobacterium *Crocosphaera watsonii*, for which ongoing nitrogen fixation in a rarely observed bloom was documented by Ian Hewson and colleagues at the University of California Santa Cruz.

Metatranscriptomics is tailor-made for comparative studies. Bacterioplankton communities in oligotrophic North Pacific surface water show greater investments in energy acquisition and metabolism (photosynthesis, oxidative phosphorylation, and C1 compound metabolism) during the day and in biosynthesis (of membranes, amino acids, and vitamins) at night, according to an analysis of RNA transcripts from these periods by Poretsky and colleagues. A powerful application of metatranscriptomics is in controlled experimental studies, in which microbial community gene expression can be measured in direct response to a defined manipulation. For instance, Gilbert and colleagues are assessing how microbial communities respond to ocean acidification.

Methods for representing and analyzing comparative metatranscriptomic data can be readily borrowed from the microarray domain, since both techniques provide relative measures of expression of many genes simultaneously, and from the metagenomics field, since both generate large, random sequence datasets from mixed microbial communities. A critical requirement for future studies is that sampling designs meet accepted standards for independent replication, since this will allow robust statistical analyses of changes or differences in expression patterns.

Increasing the coverage depth for transcript pools extracted from natural communities is also essential, since critical genes in an ecological or biogeochemical sense may not be among the most abundant community transcripts. For example, transcripts from bacterial transporter genes expressed in response to a myriad of compounds in the environment may each be poorly represented compared to transcripts from genes for central metabolism and protein synthesis machinery. Thus, identifying ecological signals amid the abundant cellular metabolic noise of ribosomal proteins, polymerases, elongation factors, and DNA replication enzymes requires deep sequencing of the community transcriptome. Alternatively, methods for subtracting shared sequences that are abundant in all transcriptomes, either physically prior to sequencing or in silico after sequencing, will help focus analysis on ecologically relevant transcripts. Advances in sequencing technology will make it easier to achieve both better replication and better coverage of informative mRNA sequences.

Metatranscriptomics enables us to identify activities and investigate gene regulation in complex microbial communities, both for descriptive studies of baseline gene expression and for experimental studies of manipulated systems, with no need to presuppose which genes should be targeted. For ecosystems as diverse as the ocean and the human microbiome, understanding and predicting in situ gene expression patterns is a major goal for the coming decade.

ACKNOWLEDGMENTS

Financial support was provided by grants from the National Science Foundation (MCB-0702125 and OCE-0724017) and the Gordon and Betty Moore Foundation.

I thank current and former members of my lab, particularly R. Poretsky, S. Gifford, J. Rinta-Kanto, and M. Vila-Costa, for ideas and discussions. This article is based on the D. C. White Research and Mentoring Award Lecture at the 2008 ASM General Meeting.

SUGGESTED READING

Bailly, J., L. Fraissinet-Tachet, M.-C. Verner, J.-C. Debaud, M. Lemaire, M. Wésolowski-Louvel, and R. Marmeisse. 2007. Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J. **1:**632–642.

Botero, L. M., S. D'Imperio, M. Burr, T. R. McDermott, M. Young, and D. J. Hassett. 2005. Poly(A) polymerase modification and reverse transcriptase PCR amplification of environmental RNA. Appl. Environ. Microbiol. **71:**1267–1275.

Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. DeLong. 2008. Microbial community gene expression in ocean surface waters. Proc. Natl. Acad. Sci. USA **105:**3805–3810.

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

Hewson, I., R. S. Poretsky, R. A. Beinart, A. E. White, T. Shi, S. R. Bench, P. H. Moisander, R. W. Paerl, H. J. Tripp, J. P.Montoya, M. A. Moran, and J. P. Zehr. 2009. In situ transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*. ISMEJ. doi: 10.1038/ismej.2009.8.

Leininger, S., T. Urich, M. Schloter, L. Schwark, J. Qi, G. W. Nicol, J. I. Prosser, S. C. Schuster, and C. Schleper. 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature **442:**806–809.

Poretsky, R. S., N. Bano, A. Buchan, G. LeCleir, J. Kleikemper, M. Pickering, W. M. Pate, M. A. Moran, and J. T. Hollibaugh. 2005. Analysis of microbial gene transcripts in environmental samples. Appl. Environ. Microbiol. **71:**4121–4126.

Poretsky, R. S., I. Hewson, S. Sun, A. E. Allen, J. P. Zehr, and M. A. Moran. 2009. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific Subtropical Gyre. Environ. Microbiol. doi: 10.1111/j.1462–2920.2008.01863.x.

Shendure, J. 2008. The beginning of the end for microarrays? Nature Methods **5:**585–587.

Urich, T., A. Lanzén, J. Qi, D. H. Huson, C. Schleper, and S. C. Schuster. 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS ONE **3:**e2527.