TRANSCRIPTOMIC AND METATRANSCRIPTOMIC ANALYSES OF MARINE MICROBIAL COMMUNITIES

by

RACHEL SUSAN PORETSKY

(Under the Direction of Mary Ann Moran)

ABSTRACT

Bacteria are abundant (10^4-10^7 cells/ml), diverse organisms and are responsible for much of the nutrient cycling in aquatic systems. Advances in molecular techniques have made it possible to examine the composition of marine microbial communities and there is a growing interest in understanding the linkage between microbial structure and function. The work described herein investigates gene expression by bacteria from a variety of aquatic ecosystems in order to assess their functional roles. Initially, a method for analyzing mRNA from environmental samples, i.e., metatranscriptomics was developed. Gene expression in bacterioplankton communities of the Sapelo Island and the Mono Lake Microbial Observatories was analyzed using this method. Transcripts were found for genes from a variety of microbial taxonomic groups. Many of the expressed genes were involved in central intermediary metabolism or were unclassified or unidentified. About 5% of the genes were responsible for ecologically important processes, such as sulfur oxidation and cellulose degradation. Improving upon the metatranscriptomics method and applying it to bacterioplankton at the Hawaii Ocean Time-Series, gene expression was examined over a day/night cycle. Taxonomic binning of mRNAs suggested that Cyanobacteria might represent the most metabolically active cells in

surface seawater. The composition of the transcriptome was consistent with models of prokaryotic gene expression. Statistical comparisons between the day vs. night transcriptomes revealed preferential biosynthesis of vitamins, membrane components and amino acids at night, and photosynthesis, heterotrophic C1 metabolism, and oxidative phosphorylation in the day. In a final study, bacterial expression patterns were characterized in response to dissolved organic matter from phytoplankton, using pure cultures of a model marine bacterium and a diatom in a microarray-based analysis. Several genes were upregulated in the presence of diatom DOM, including some involved in transport and utilization of amino acids, protocatechuate catabolism, and transcriptional regulation. These results provided a novel method for examining bacterial-phytoplankton associations on the level of gene expression and have implications for our understanding of phytoplankton/bacterial interactions. Together, the results of these gene expression characterizations contributed to our understanding of how microbial communities function, how microbial processes are regulated, and how microbes interact with each other and with their environment.

INDEX WORDS: marine bacteria, mRNA, gene expression, microbial ecology, microarray, dissolved organic matter, metatranscriptomics

TRANSCRIPTOMIC AND METATRANSCRIPTOMIC ANALYSES OF MARINE MICROBIAL COMMUNITIES

by

RACHEL SUSAN PORETSKY

B.S., Brandeis University, 1999

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

TRANSCRIPTOMIC AND METATRANSCRIPTOMIC ANALYSES OF MARINE MICROBIAL COMMUNITIES

by

RACHEL SUSAN PORETSKY

Major Professor: Mary Ann Moran

Committee: James T. Hollibaugh
William B. Whitman
Marc E. Frischer
Elizabeth Mann

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2008

DEDICATION

For their love and support, I dedicate this work to my parents, Allan and Esther Poretsky.

ACKNOWLEDGEMENTS

There are so many wonderful people without whom I could never have completed this dissertation. I consider myself incredibly fortunate to have worked with my major professor, Dr. Mary Ann Moran. To say that I greatly appreciate her encouragement, advice, mentorship, and support over the years is a gross understatement. I have learned so much from her enthusiasm, curiousity, and healthy dose of skepticism.

I also sincerely appreciate the support and advice of my committee members: Drs. Tim Hollibaugh, Barny Whitman, Marc Frischer, and Liz Mann.

I have enjoyed working in the Department of Marine Sciences, surrounded by so many intelligent and supportive students and faculty members. I am exceedingly grateful for the conversations, help, and company provided by current and former members of the Moran Lab. In addition to all of her assistance in the lab, Wendy Ye has been a good friend, source of information and, recently, supplier of coffee. Xie xie! For their help and friendship, I also sincerely thank Shulei Sun, Xiaozhen Mou, Justine Lyons, Alison Buchan, Maria Vila-Costa, Johanna Rinta-Kanto, Scott Gifford, Jennifer Edmonds, and Camille English. In addition, I have worked with some fantastic undergraduate students who helped me collect samples and conduct experiments: Jennifer Oliver, Whitney Pate, Jacob Shalack, and Claire Hierling. Mandy Joye and Beth! Orcutt have been my links to the world of geobiology and have been two great sources of inspiration, both within and outside of academia.

Over the years, I have received financial support from the National Science Foundation, The UGA Graduate School, The Department of Marine Sciences, and through NSF and Moore Foundation grants to Dr. Mary Ann Moran.

Numerous diversions outside the walls of UGA have helped keep me sane, reminding me that there other things to life than lab work and writing a dissertation. My cycling friends encouraged me to get outside and ride as often and as far as possible; three years at Canopy Studio have turned me into a trapeze dancer; Daily Co-op has served as an outlet for me to channel my activism and community interests; my swim buddies made sure I got my exercise/therapy in every day- or gave me a hard time if I didn't. There are so many talented, inspirational people in Athens and I'm thrilled to have been involved in a variety of aspects of life in this great community.

The support and love from my family has kept me going over the years. I am exceedingly grateful for their encouragement. That I can be a source of *nachas* simply by "playing in the dirt or water and making DNA" makes me incredibly happy. More recently, The McPhersons/Ratards/Lafleurs (my "other" family) have helped me keep my life in proper perspective and balance.

Finally, I could not have completed this without Matt, whose support- emotional, moral, scientific, and by making sure didn't starve myself or stay up *too* late- was truly unending.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The oceans cover more than 70% of the Earth's surface and are most likely where life on this planet originated. Direct observation of microbes in the ocean in the 1970's hinted at their abundance (19). It is now known that bacteria are the most abundant organisms in aquatic ecosystems (10^4-10^7 cells/ml) (44). The majority of these organisms, however, are not cultivable in the laboratory (1). To circumvent the difficulties of culturing and assessing community composition, molecular methods that exploit the wealth of taxonomic information provided by the small subunit (SSU) rRNA are frequently used (45). Concurrent with advances in molecular techniques and technologies is the sequencing of rRNAs, functional genes, and whole genomes of many important environmental microbes and communities, thus facilitating discoveries and providing insights into the roles of microbes in ecosystems. As more sequences become available, there is a growing interest in linking phylogeny to function in natural microbial communities. This has led to the development of new methods to target genomes, functional genes, and gene expression.

Genomics and metagenomics

Genomics, the sequencing and analysis of whole genomes, has been valuable for revealing clues about the functions of marine microbial communities. By sequencing and studying the genomes of those organisms that are amenable to culturing, it is possible to create hypotheses about their ecology as well the ecology of their close relatives in the environment.

Some important model marine organisms with full genome sequences are the cyanobacterium *Prochlorococcus marinus* (32), the SAR11 clade member *Pelagibacter ubique*, (15), and the marine Roseobacter *Silicibacter pomeroyi* (26). Recent efforts to sequence many more marine microbial genomes, particularly through large scale sequencing projects such as The Moore Foundation Marine Microbiology Initiative, have greatly expanded the number of existing full genome sequences. Genome sequences have yielded new information about how these organisms make a living in the environment and about how they carry out different processes. For example, the genome sequence of *P. marinus* SS120 has revealed what may be the minimal genome composition required for a photosynthetic organism (9). Comparison of multiple genome sequences demonstrated the prevalence of secretion system genes, resulting in the hypothesis that these are important in microbe-microbe interactions (46).Genome sequences of both *Synechococcus* sp. strain CC9311 and *S. pomeroyi* provided information about how these organisms adapt to the coastal environment (26, 27).

Metagenomics is another technique that has revolutionized the way microbial communities are sampled. This technique provides the ability to analyze the biology of any organism or community without cultivation or development of a unique genetic system while generating large datasets of environmental DNA in a fairly representative fashion. Large insert clone libraries or metagenomic sequencing efforts can provide information about the gene composition of microbial assemblages and identify potential biogeochemical processes (3, 33, 39, 41). These large insert clone libraries have proven useful for access to community microbial genomes and have revealed previously unknown genes such as proteorhodopsin, a bacterial light harvesting proton pump (2). Through sequencing innovations such as pyrosequencing, it is now possible to obtain millions of sequences in a matter of hours without cloning and for less than the

cost of traditional Sanger sequencing (24). Recent pyrosequencing efforts have greatly expanded existing sequence databases, revealed many new genes (38, 42, 47), and provided insight into the distribution of genes in the environment (7, 40).

Transcriptomics

In addition to detection, quantification, and characterization of marine microorganisms, there is interest in assessing gene expression in order to further understand the ecological roles and functions of microorganisms. Expression screening of metagenomic DNA is carried out by cloning the DNA fragments into a host organism, usually *Eschericia coli*, and screening the host for activity (34). Thus, detection of gene expression from metagenomic DNA is limited to that which can be expressed in the host and does not necessarily indicate natural expression in the environment. Recently, many studies have turned to the products of transcription for more detail on *in situ* patterns of gene expression. Gene expression in natural environments can be assessed with functional gene probes or primer sets applied to messenger RNA (mRNA) extracted directly from natural communities (6, 43).

Within the past several years, microarrays have become common and efficient tools for the study of gene expression, both for typical lab cultures such as *E. coli* (31) and for environmental isolates such as *Shewanella oneidensis* (12). Microarrays offer the opportunity to examine process-specific presence and expression of multiple genes in complex microbial assemblages simultaneously (48) and can be designed using whole genome sequences (5), various genotypes of a particular functional gene (8, 25), or a combination of sequences isolated directly from the environment (30). Several recent reviews discuss the feasibility of applying microarray technology to microbial communities in the environment (14, 48). Although

microarrays have been used increasingly to study microbial communities, there are still challenges associated with the current approaches (49). The challenges are primarily related to sensitivity, specificity and quantitation attributed to the difficulty of environmental probe design (and choice) and obtaining sufficient amounts of nucleic acids from many environments (49). Because target and probe sequences in the environment are diverse, microarray studies are often limited to known genes and functions, making it difficult to use these approaches for unbiased surveys of microbial gene expression in the environment. Until continued technological advances allow us to use microarrays for high-throughput and real-time analysis of gene expression *in situ*, culture-based array technologies that focus on ecologically-relevant model organisms can generate valuable hypotheses which can subsequently be tested and validated in the environment.

Another approach to surveying microbial activity in the environment is through the direct retrieval of expressed genes (mRNAs) (10, 29). This is distinct from amplifying target genes using directed primers in that it provides access to the transcriptome of a microbial assemblage, analogous to accessing the metagenome of a community by shotgun cloning (42, 47). Despite the potential for bacterial mRNAs to provide a direct link between genetic potential and biogeochemical activity in natural environments, it is difficult to work with environmental mRNAs because they degrade quickly, they lack the convenient polyA tails of eukaryotic messages, and they are much less abundant than rRNAs, which can comprise > 80% of the RNA in total RNA extracts (20). Analysis of the mRNA pool in the environment, however, can provide one of the most effective ways of discovering connections between key activities and the organisms that mediate them.

Chapter overview

The research described in this dissertation involves a combination of environmental transcriptomics (i.e., obtaining mRNA profiles from environmental samples) and whole genome microarrays in order to examine expression of environmentally relevant functional genes in the environment. The following broad questions are addressed:

***Question 1*: Can complex microbial communities be analyzed on the level of gene expression (mRNA)?**

a) Can a representative mRNA pool be obtained from seawater?

b) What can environmental mRNA reveal about microbial activities in the environment?

***Question 2*: What can comparative metatranscriptomic studies reveal about the functions of microbial communities over a day/night cycle?**

a) Do daytime transcript pools conform to expectations of light-driven processes?

b) Do nighttime transcript pools differ from daytime pools?

***Question 3*: What is the functional gene response of a marine bacterium to phytoplankton DOM?**

a) Can gene expression studies of model organisms in pure cultures provide relevant information on microbial interactions?

b) What pathways for transport and metabolism of DOM are expressed when a model marine bacterium is exposed to phytoplankton DOM?

Question 1 is addressed in Chapter 2. A procedure was developed for analyzing environmental transcriptomes by creating clone libraries generated from environmental mRNAs. In this method, total RNA was first collected from the environment, rRNA was selectively removed, and cDNA synthesized from the enriched mRNA pool using random 10- 14 bp primers for reverse transcription and low-specificity PCR was cloned and sequenced with Sanger sequencing. Recovered sequences were annotated using bioinformatics techniques to identify the expressed genes. Two NSF Microbial Observatory sites served as the locations for this study. Mono Lake is a closed-basin, alkaline, hypersaline soda lake located east of the Sierra Nevada Mountains, approximately 160 km south of Lake Tahoe, California. Dean Creek (Sapelo Island, GA) is a tidal marsh creek that experiences a varying salinity range influenced by river discharge, ground water delivery, and tides. It is part of the Sapelo Island Microbial Observatory and the Georgia Coastal Ecosystems Long Term Ecological Research sampling area and is typical of southeastern U.S. salt marshes. In our first application of this method, analysis of the expressed genes in our transcript libraries revealed gene sequences of biogeochemical interest without constraints imposed by existing sequence data and with preference for those genes being actively expressed. This technique provided one of the first views of the composition and dynamics of the bacterial mRNA pool in a natural ecosystem

Chapter 3 describes a more efficient method for generating cDNA from environmental mRNA. This method relies on a linear amplification of mRNA by polyadenylating the mRNA and carrying out *in vitro* transcription followed by synthesis of double stranded cDNA using random hexamers.

Question 2 is addressed in Chapter 4. Bacterioplankton communities have been shown to vary temporally in coastal environments (35, 36). Among heterotrophic bacterioplankton,

populations fluctuations are expected to occur daily in response to substrate availability, predation, or other factors (11, 13, 18, 21, 28, 37). In order to examine diel gene expression, metratranscriptomics were carried using the technique described in Chapter 3 along with high throughput pyrosequencing. This research was conducted in the North Pacific subtropical gyre system, part of the Hawaiian Ocean Time-Series (23).

Chapter 5 examines Question 3. Unlike the previous chapters, this chapter explores microbial gene expression of a single, model organism in relation to an environmentally-relevant activity. Marine bacteria of the Roseobacter clade are known to associate with marine algae, both physically and in response to algal products (16, 17, 22). Indeed, Roseobacter organisms have been proposed to be ecologically successful in marine systems because of the competitive advantage gained by their ability to use algal-derived dissolved organic matter (DOM) (4, 26). The research presented in Chapter 5 exploited the observation that members of this lineage are often found in association with marine algae. Using whole genome microarray technology, gene expression of the marine α-Proteobacteria isolate *S. pomeroyi* was examined in response to complex DOM of known origin, i.e. exudate from an axenic culture of the marine diatom *Skeletonema costatum*, with the expectation that this model system could provide insights into similar interactions *in situ*.

REFERENCES

1. Amann, R. I., W. Ludwig, and K. H. Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59:143-69.
2. Béjà, O., E. N. Spudich, J. L. Spudich, M. Leclerc, and E. F. DeLong. 2001. Proteorhodopsin phototrophy in the ocean. Nature 411:786-789.
3. Béjà, O., M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong. 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ. Microbiol. 2:516-529.
4. Buchan, A., J. M. Gonzalez, and M. A. Moran. 2005. Overview of the marine Roseobacter lineage. Appl. Environ. Microbiol. 71:5665-5677.
5. Bürgmann, H., E. C. Howard, W. Ye, F. Sun, S. Sun, S. Napierala, and M. A. Moran. 2007. Transcriptional response of *Silicibacter pomeroyi* DSS-3 to dimethylsulfoniopropionate (DMSP), p. 2742-2755, Environ. Microbiol., vol. 9.
6. Bürgmann, H., F. Widmer, W. V. Sigler, and J. Zeyer. 2003. mRNA extraction and reverse transcription-PCR protocol for detection of *nifH* gene expression by *Azotobacter vinelandii* in soil. Appl. Environ. Microbiol. 69:1928-1935.
7. DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N.-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. 2006. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science 311:496-503.

8. Dennis, P., E. A. Edwards, S. N. Liss, and R. Fulthorpe. 2003. Monitoring gene expression in mixed microbial communities by using DNA microarrays. Appl. Environ. Microbiol. 69:769-778.

9. Dufresne, A., M. Salanoubat, F. Partensky, F. Artiguenave, I. M. Axmann, V. Barbe, S. Duprat, M. Y. Galperin, E. V. Koonin, F. Le Gall, K. S. Makarova, M. Ostrowski, S. Oztas, C. Robert, I. B. Rogozin, D. J. Scanlan, N. T. de Marsac, J. Weissenbach, P. Wincker, Y. I. Wolf, and W. R. Hess. 2003. Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome, p. 10020-10025, vol. 100.

10. 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.

11. Fuhrman, J. A., R. W. Eppley, A. Hagstrom, and F. Azam. 1985. Diel Variations in Bacterioplankton, Phytoplankton, and Related Parameters in the Southern-California Bight. Mar. Ecol. Prog. Ser. 27:9-20.

12. Gao, H. C., Y. Wang, X. D. Liu, T. F. Yan, L. Y. Wu, E. Alm, A. Arkin, D. K. Thompson, and J. Z. Zhou. 2004. Global transcriptome analysis of the heat shock response of *Shewanella oneidensis*. J. Bacteriol. 186:7796-7803.

13. Gasol, J. M., M. D. Doval, J. Pinhassi, J. I. Calderon-Paz, N. Guixa-Boixareu, D. Vaque, and C. Pedros-Alio. 1998. Diel variations in bacterial heterotrophic activity and growth in the northwestern Mediterranean Sea. Mar. Ecol. Prog. Ser. 164:107-124.

14. Gentry, T. J., G. S. Wickham, C. W. Schadt, Z. He, and J. Zhou. 2006. Microarray applications in microbial ecology research. Microb. Ecol. 52:159-175.

15. Giovannoni, S. J., H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappe, J. M. Short, J. C. Carrington, and E. J. Mathur. 2005. Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science 309:1242-1245.

16. Gonzalez, J. M., R. Simo, R. Massana, J. S. Covert, E. O. Casamayor, C. Pedros-Alio, and M. A. Moran. 2000. Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. Appl. Environ. Microbiol. 66:4237-46.

17. Grossart, H. P., F. Levold, M. Allgaier, M. Simon, and T. Brinkhoff. 2005. Marine diatom species harbour distinct bacterial communities. Environ. Microbiol. 7:860-873.

18. Hagstrom, A., J. Pinhassi, and U. L. Zweifel. 2001. Marine bacterioplankton show bursts of rapid growth induced by substrate shifts. Aquat. Microb. Ecol. 24:109-115.

19. Hobbie, J. E., R. J. Daley, and S. Jasper. 1977. Use of Nuclepore Filters for Counting Bacteria by Fluorescence Microscopy. Appl. Environ. Microbiol. 33:1225-1228.

20. Ingraham, J. L., O. Maaløe, and F. C. Neidhardt. 1983. Growth of the bacterial cell. Sinauer Associates, Sunderland, Mass.

21. Jacquet, S., J. F. Lennon, D. Marie, and D. Vaulot. 1998. Picoplankton population dynamics in coastal waters of the northwestern Mediterranean Sea. Limnol. Oceanogr. 43:1916-1931.

22. Jasti, S., M. E. Sieracki, N. J. Poulton, M. W. Giewat, and J. N. Rooney-Varga. 2005. Phylogenetic diversity and specificity of bacteria closely associated with *Alexandrium* spp. and other phytoplankton. Appl. Environ. Microbiol. 71:3483-3494.

23. Karl, D. M., and R. Lukas. 1996. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. Deep Sea Research Part II: Topical Studies in Oceanography 43:129-156.

24. Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

25. Moisander, P. H., A. E. Morrison, B. B. Ward, B. D. Jenkins, and J. P. Zehr. 2007. Spatial-temporal variability in diazotroph assemblages in Chesapeake Bay using an oligonucleotide nifH microarray. Environ. Microbiol. 9:1823-1835.

26. Moran, M. A., A. Buchan, J. M. Gonzalez, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B. Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Ye, T. R. Miller, J. Carlton, D. A. Rasko, I. T. Paulsen, Q. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S. Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut, and N. Ward. 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 432:910-913.

27. Palenik, B., Q. Ren, C. L. Dupont, G. S. Myers, J. F. Heidelberg, J. H. Badger, R. Madupu, W. C. Nelson, L. M. Brinkac, R. J. Dodson, A. S. Durkin, S. C. Daugherty, S. A. Sullivan, H. Khouri, Y. Mohamoud, R. Halpin, and I. T. Paulsen. 2006. Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. Proc. Natl. Acad. Sci. USA 103:13555-13559.

28. Pinhassi, J., F. Azam, J. Hemphala, R. A. Long, J. Martinez, U. L. Zweifel, and A. Hagstrom. 1999. Coupling between bacterioplankton species composition, population dynamics, and organic matter degradation. Aquat. Microb. Ecol. 17:13-26.

29. 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.

30. Rich, V. I., K. Konstantinidis, and E. F. DeLong. 2008. Design and testing of 'genome-proxy' microarrays to profile marine microbial communities. Environ. Microbiol. 10:506-521.

31. Richmond, C. S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner. 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. 27:3821-35.

32. Rocap, G., F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser, and S. W. Chisholm. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042-1047.

33. Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman,

M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman. 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66:2541-2547.

34. Rondon, M. R., S. J. Raffel, R. M. Goodman, and J. Handelsman. 1999. Toward functional genomics in bacteria: Analysis of gene expression in *Escherichia coli* from a bacterial artificial chromosome library of *Bacillus cereus*. Proc. Natl. Acad. Sci. USA 96:6451-6455.

35. Seymour, J. R., L. Seuront, and J. G. Mitchell. 2005. Microscale and small-scale temporal dynamics of a coastal planktonic microbial community. Mar. Ecol. Prog. Ser. 300:21-37.

36. Shiah, F. K., and H. W. Ducklow. 1995. Multiscale Variability in Bacterioplankton Abundance, Production, and Specific Growth-Rate in a Temperate Salt-Marsh Tidal Creek. Limnol. Oceanogr. 40:55-66.

37. Simon, M. 1994. Diel Variability of Bacterioplankton Biomass Production and Cell Multiplication in Lake Constance. Archiv Fur Hydrobiologie 130:283-302.

38. Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. USA 103:12115-12120.

39. Stein, J. L., T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong. 1996. Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment front a planktonic marine archaeon. J. Bacteriol. 178:591-599.

40. Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. 2005. Comparative Metagenomics of Microbial Communities. Science 308:554-557.

41. Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43.

42. Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. W. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.

43. Wawrik, B., J. H. Paul, and F. R. Tabita. 2002. Real-time PCR quantification of rbcL (ribulose-1,5-bisphosphate carboxylase/oxygenase) mRNA in diatoms and pelagophytes. Appl. Environ. Microbiol. 68:3771-3779.

44. Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. Prokaryotes: The unseen majority. Proc. Natl. Acad. Sci. USA 95:6578-6583.

45. Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. 51:221-271.

46. Worden, A. Z., M. L. Cuvelier, and D. H. Bartlett. 2006. In-depth analyses of marine microbial community genomics. Trends Microbiol. 14:331-336.

47. Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J.-M. Chandonia, D. A. Soergel,

Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter. 2007. The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. PLoS Biol. 5:e16.

48. Zhou, J. H. 2003. Microarrays for bacterial detection and microbial community analysis. Curr. Op. Microbiol. 6:288-294.

49. Zhou, J. Z., and D. K. Thompson. 2002. Challenges in applying microarrays to environmental studies. Curr. Opin. Biotechnol. 13:204-207.

CHAPTER 2

ANALYSIS OF MICROBIAL GENE TRANSCRIPTS IN ENVIRONMENTAL SAMPLES[1]

[1] Poretsky, R. S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M., Pate, W. M., Moran, M. A., & Hollibaugh, J. T. (2005) *Appl. Environ. Microbiol.* 71, 4121-4126.
Reprinted here with permission of publisher.

ABSTRACT

We analyzed gene expression in marine and freshwater bacterioplankton communities by the direct retrieval and analysis of microbial transcripts. Environmental mRNA, obtained from total RNA by subtractive hybridization of rRNA, was reverse transcribed, amplified with random primers, and cloned. Approximately 400 clones were analyzed, of which ~80% were unambiguously mRNA derived. mRNAs appeared to be from diverse taxonomic groups, including both Bacteria (mainly α- and γ- *Proteobacteria*) and Archaea (mainly *Euryarchaeota*). Many transcripts could be linked to environmentally important processes such as sulfur oxidation (*soxA*), assimilation of C1 compounds (*fdh1B*), and acquisition of nitrogen via polyamine degradation (*aphA*). Environmental transcriptomics is a means of exploring functional gene expression within natural microbial communities without bias toward known sequences, and provides a new approach for obtaining community-specific variants of key functional genes.

The technology of environmental genomics is based on sequence analysis of fragments of environmental DNA, and retrieves genes without any previous sequence information and with relatively little apparent bias (1, 18, 21). An analogous method for environmental mRNA (i.e., environmental transcriptomics) could similarly retrieve the transcriptome of a microbial assemblage without any prior information on what genes the community might be expressing. The prospect for using environmental transcriptomics to link genetic potential with biogeochemical activity of microbes has been hindered, however, by the difficulties of working with mRNAs. Prokaryotic transcripts generally lack the poly(A) tails that make isolation of most eukaryotic messages straightforward (12). Some mRNAs degrade quickly, with half lives as short as 30 seconds based on studies of cultured bacteria (2). Finally, mRNA molecules are much less abundant than rRNA molecules in total RNA extracts, so the mRNA signal is often overwhelmed by background.

We have developed a protocol to analyze partial environmental transcriptomes by collecting total RNA from the environment, enriching for mRNA by subtractive hybridization of rRNA, and using randomly primed reverse transcription (RT) to produce a cDNA template population. The templates are amplified by PCR and used to generate cDNA clone libraries. Here we report results from the analysis of approximately 400 environmental gene transcripts retrieved directly from bacterioplankton communities of Sapelo Island, GA and Mono Lake, CA.

Protocol for library generation. Water samples were collected from the Sapelo Island Microbial Observatory (SIMO; tidal salt marsh creek in the southeastern U.S.; simo.marsci.uga.edu) and the Mono Lake Microbial Observatory (MLMO; closed-basin, hypersaline soda lake near Lake Tahoe, CA; www.monolake.uga.edu). SIMO water samples (10 l) were collected in October 2002 and August 2003 and screened immediately after collection to

remove particles >3.0 µm, including most eukaryotic cells. Cells for RNA extraction were collected on a 0.2 µm pore size polycarbonate membrane filter. MLMO samples (~8 l) were collected in May 2003 at depths of 5 m (surface) and 23 m (chemocline). Because the dominant phytoplankter (*Picocystis salinarum*) is of similar size to the bacterioplankton, MLMO samples were not screened. MLMO samples were stored on ice during transport to the laboratory and then filtered onto a 0.2 µm pore size membrane filter.

The process from sample collection to RNA extraction was done as rapidly as possible to limit degradation of mRNA. RNA was extracted using the RNAqueous-Midi kit (Ambion, Austin, TX) with several modifications (see supplemental material for detailed protocol). For SIMO samples, the elapsed time between water collection and RNA extraction was less than 30 minutes. For MLMO samples, the elapsed time was ~2 h. Subtractive hybridization was used to selectively remove rRNA (MICROB*Express* Bacterial mRNA Enrichment kit, Ambion). DNase-treated mRNA preparations were amplified by RT-PCR using two of six possible random primers (Supplemental Table S2.1): 10-mer primers OPA04, OPA13, and OPA17 from a commercial primer stock (Operon Technologies, Inc., Alameda, CA), primer SD14 designed to target the Shine-Dalgarno region of bacterial mRNAs (5), and primers SES3-1 and SESRT-3 designed with low G+C content (MLMO only). Clone libraries of some PCR products were screened to eliminate sequences derived from contaminating rRNA using probes constructed by amplifying rRNA genes from DNA harvested from the same sample (see supplemental material for detailed protocol). Sequences of 347 SIMO clones (40 from October 2002 and 307 from August 2003) and 60 MLMO clones were analyzed using the BLASTX and BLASTN tools (http://www.ncbi.nlm.nih.gov/BLAST/). Additionally, 282 of the August 2003 SIMO clones were automatically annotated using the Annotation Engine service provided by The Institute for

Genomic Research (TIGR, Rockville, MD). The sequences were deposited in GenBank under the accession numbers AY793704-AY794012.

Environmental transcript libraries. We calculate that 2.4 x 10^{13} bacterial mRNAs were present in the 10 l SIMO water sample collected in August 2003, of which 80,000 were unique [calculated assuming a late summer population of 1.7 x 10^{6} bacterial cells ml^{-1} (http://gce-lter.marsci.uga.edu/lter/asp/db/data_catalog.asp) each with 1380 total mRNA molecules per cell (10), and 200 bacterial species represented in the community (http://simo.marsci.uga.edu/MainWeb/pages/database.htm) each with 400 unique mRNAs per cell (10)]. Thus the 342 SIMO clones and 60 MLMO clones analyzed here were a small fraction of the total transcript pool in each environment. Yet while these small libraries do not provide a quantitative inventory of bacterioplankton transcripts, they offer a novel glimpse of microbial activity that is unconstrained by existing sequence data and not restricted to previously characterized processes. Further, the standard cloning and sequencing methods used for these manually-assembled libraries can be readily adapted to high-throughput approaches, potentially allowing the sequencing of thousands of amplicons from a single community.

Sub-libraries were generated from a single sample using different primer combinations, with one primer chosen at random for the reverse transcription (RT) step and that primer used in combination with a second primer in the PCR step (Supplemental Table S2.2). When we compared transcript retrieval with different permutations of the random primers, the SD14 primer appeared to out-compete the others. Often, both ends of the amplicons were primed by SD14. No amplicons were generated for MLMO samples without SD14 in either the RT or PCR step, although several primer combinations without SD14 were used with success in the SIMO samples (Supplemental Table S2.2). The higher amplification efficiency with the SD14 primer is

not surprising, as primers designed to bind to the Shine-Dalgarno region (the ribosomal binding site) have been used in differential display analyses of mRNA transcripts in both pure cultures and in soils (5). When used as a PCR primer, it ostensibly biases amplification to the 5' end of mRNA transcripts for bacteria that possess a typical *Escherichia coli*-like Shine-Dalgarno region (e.g., AGGAGG) (10). When used as an RT primer, we expected to see the SD14 primer sequence only for polycistronic operons because it would target the Shine-Dalgarno site at the beginning of the gene downstream from the one that was reverse transcribed. Because the SD14 primer sequence was often identified at both the beginning and end of sequences following RT-PCR, we concluded that SD14 does not necessarily target the Shine-Dalgarno site exclusively when used under low-specificity PCR conditions.

Although we still observed a few rRNA generated cDNA sequences after repeated subtractive hybridizations, analysis of the clone libraries indicated that the protocol for removing rRNA worked efficiently, as typically fewer than 20% of the clones were derived from 16S, 23S, and 28S rRNA combined (Supplemental Table S2.2). Results from the colony hybridizations of the SIMO clone libraries indicated that perhaps a higher percentage of clones were rRNA generated, but this screening step reduced the number of rRNA clones sequenced. Even though they were not screened by hybridization, MLMO cDNA libraries contained few rRNA genes, suggesting that the subtractive hybridization alone worked efficiently for these samples.

Apparent taxonomic representation. The putative taxonomic origin of the transcripts was used to assess diversity in relation to the known microbial composition of the two systems. Putative taxonomic origin was assigned based on the taxon of the most similar sequence by BLAST analysis (Supplemental Table 2.3; http://aem.asm.org/cgi/data/71/7/4121/DC1/1). The accuracy of this assignment is negatively affected by lateral gene transfer and positively

correlated with the taxonomic coverage of the database for any given gene. Given these caveats, the SIMO libraries appeared to be almost entirely bacterial-derived, although similarities to known genes were sometimes low (Supplemental Table S2.3; http://aem.asm.org/cgi/data/71/7/4121/DC1/1). Using only those sequences with E values $\leq \sim e^{-10}$, gene expression at the SIMO site was inferred for α-, β-, γ-, δ-, and ε- *Proteobacteria, Bacteroidetes, Chlorobi, Cyanobacteria, Firmicutes, Actinobacteria, Spirochaetes, Planctomycetes, Euryarchaeota,* and *Crenarchaeota.* Apparent archaeal sequences represented a significant portion of the August 2003 SIMO library (Fig 1A). Almost all of the putative archaeal transcripts were most similar to genes from *Sulfolobus tokodaii* or *Methanococcus voltae*, but identification may be skewed toward organisms for which a genome sequence is available.

A small-subunit rRNA database of SIMO bacterioplankton that was generated during a different year, but for the same season (summer) and the same size fraction (0.2 - 3.0 μm) (http://simo.marsci.uga.edu/), provided a comparison with the putative taxonomic assignment of the transcripts in the August 2003 mRNA library. The SIMO 16S rRNA libraries were dominated by sequences from α- and γ-*Proteobacteria* (18 and 16%) (Figure 2.1B). These two taxonomic groups were similarly represented in the mRNA library (16 and 19%, respectively). *Cyanobacteria* played a larger role in the 16S rRNA library (Figure 2.1B) than in the mRNA library while *Chlorobi,* ε-*Proteobacteria,* and *Spirochaetes* appeared to contribute to the mRNA pool but were not well represented in the 16S rRNA library. Overall, relatively similar distributions among apparent taxonomic groups existed between the two libraries. The transcripts in the August 2003 SIMO library were also compared to the genome of *Silicibacter pomeroyi,* a marine α-*Proteobacteria* isolated from coastal water near the SIMO site (http://www.marsci.uga.edu/s_pomeroyi/) (6, 15). Using BLASTX, almost 10% of the clones in

the SIMO transcript library matched predicted proteins in the *S. pomeroyi* genome with identities higher than other entries in GenBank, with E values between e^{-70} and e^{-97}in most cases.

At MLMO, 33% of the 60 transcripts appeared to be eukaryotic in origin, not surprising given that the spring phytoplankton bloom was underway during sample collection and eukaryotes were too small to be removed by size-selective screening. Prokaryotic MLMO transcripts matched genes from *Firmicutes, Cyanobacteria, Bacteroidetes, Spirochaetes, Actinobacteria, Planctomycetes* and α-, β-, δ-, and ε- *Proteobacteria.* Putative taxonomic affiliations of MLMO transcripts were consistent with a 16S rRNA gene library constructed in July 2000, as evidenced by the presence of γ-*Proteobacteria*-like sequences in surface and chemocline samples in both mRNA and 16S rRNA libraries, as well as a cyanobacterial-like mRNA and 16S rRNA sequences in the chemocline (9). Although there is uncertainty in the taxonomic assignment of mRNA sequences as discussed above, environmental transcripts appeared to be retrieved from a diversity of microorganisms at both the SIMO and MLMO sites.

Transcript annotation. Most of the sequences obtained were not full-length transcripts (~200-500 bp), although some amplicons were large (>1000 bp). In all cases, there was no amplification in controls that lacked the RT step. The mRNA sequences appeared to be transcribed from a range of housekeeping genes, components of transport systems, and genes for energy metabolism (Table 2.1). Like the taxonomic assignments, the identities of transcripts were inferred from the closest matches by BLASTX (Supplemental Table S2.3). These assignments are only as good as the existing database, and genes that are rare in genomes because they code for unusual or specialized traits are particularly susceptible to poor database coverage. For example, a MLMO transcript with a strong BLAST hit to an arsenite transporter (*arsA*) from *Arabidopsis thaliana* (Table 2.1) predicts a function expected in Mono Lake given

the high concentration of arsenite (200 μM; ref. (14) but predicts an organism quite distant from any lake plankton.

In all libraries, several instances of multiple mRNAs transcribed from homologous genes were seen. Some of the repeated mRNAs, such as those having sequence similarity to s*oxA* (sulfur oxidation; eight sequences) and *surE* (stationary-phase survival protein; four sequences) were found in different sub-libraries (i.e., libraries constructed from the same RNA sample but using different primer pairs). In all but one case, the homologous sequences were found in eight or fewer clones, with the exception being transcripts putatively encoding acetylpolyamine amidohydrolase (*aphA*) that accounted for 35 of 307 clones in the August 2003 sample.

Annotation of the clones from the SIMO libraries revealed that the majority (~80%) were found only once in the library. In contrast, nearly half (42%) of the 60 MLMO sequences were homologous to another sequence in the library. Four clones from the SIMO August 2003 sample and three clones from MLMO had no significant matches using an EXPECT threshold of 10 in homology searches (BLASTN and BLASTX), and thus are either transcripts of novel genes or are not transcripts.

The TIGR Annotation Engine organized the August 2003 SIMO sequences into role categories based on assigned functions of the highest matching gene sequences, including central intermediary metabolism (18.5% of the clones), cellular processes (5.5%), and protein synthesis (5.0%) (Table 2.2). Transcripts that appeared to code for transport/binding proteins (3.8%) were potentially involved in amino acid, carbohydrate, and organic acid and alcohol transport and metabolism (Table 2.2). The largest fraction of transcripts was categorized as hypothetical (35.3%), and 12% were “unclassified” (typically of known function but not readily placed in a

role category during autoannotation). As discussed above, inferred functional assignments of transcripts are subject to effects of database coverage and lateral gene transfer.

There are significant methodological obstacles in retrieving an environmental transcriptome that may result in the unequal capture of transcripts, including choice of primer, preferential targeting of transcripts, and bias toward the longest lived mRNAs. In assessing the issue of targeting bias, we found evidence for selective amplification of some targets by a given primer pair, such as *soxA* transcript amplification only when both OPA13 and OPA17 primers were used and *aphA* transcript amplification only if primer SD14 was used. In assessing the issue of mRNA lifetime, we examined three gene categories predicted to have longer half-lives based on studies of *E. coli* transcripts: cell envelope genes, energy metabolism genes, and transport/binding genes (3). The 307-member August 2003 SIMO environmental transcript library was not dominated by any of these functional categories, although evidence from organisms such as *Bacillus subtilis* indicates that there are both long and short half-life transcripts in almost all gene classes (7). For the MLMO transcript library, for which the time from collection to processing was several hours, potential biases related to mRNA half-life could not be evaluated.

Applications of environmental transcriptomics. A promising application of environmental transcriptomics is the retrieval of community-specific functional gene sequences with relevance for quantitative ecological studies. Functional gene discovery in natural environments is typically based on primer sets designed from a limited database that is heavily biased toward cultured organisms (17). Environmental transcript libraries can alleviate this problem by supplying site-specific functional gene sequences from active cells without the constraints of prior sequence information. For example, the eight putative s*oxA* sequences in the

SIMO library were similar to one another but distinct from those found in cultured bacteria (Figure 2.2). Quantitative PCR (qPCR) analysis of DNA from an August 2004 SIMO bacterioplankton community using a primer set designed to target only the SIMO clade *soxA* genes indicated that they were present at concentrations of ~4.6 x 10^6 l^{-1}, or in 1 of every 370 cells (assuming 1.7 x 10^6 cells l^{-1} and one gene copy per cell). Further, *soxA* transcripts were retrieved from four samples collected within an 11 h period in August 2004 using RT-qPCR (averaging 2.6 x 10^3 transcripts l^{-1}), suggesting that SIMO clade *soxA* genes are consistently transcribed within the bacterioplankton community. The putative chitinase transcript in the MLMO library (Table 2.1), which has low identity to known chitinase sequences (<27%), is also of significance because chitinase genes cannot be amplified from the Mono Lake ecosystem using existing *chi* primer sets (11). Nevertheless, the abundance of arthropod exoskeletons in the lake along with previous demonstrations of chitinase activity (11) suggest that chitin degradation is a major microbial process in this system. Environmental transcriptomics thus provides gene sequences of biogeochemical interest (Table 2.1) without constraints imposed by existing sequence data and with preference for those genes being actively expressed.

Environmental transcriptomics also has considerable potential for generating novel hypotheses about microbial processes. In the SIMO library, putative acetylpolyamine amidohydrolase (*aphA*) transcripts accounted for 11% of the sequences in the August 2003 SIMO library, represented by at least seven distinct sequences in four sub-libraries. The possible ecological relevance of these sequences is not immediately apparent because prokaryotic *aphA* genes are poorly characterized. However, they are suspected to be involved in the degradation of polyamines (19), a class of nitrogen-rich compounds including putrescine and spermidine that form complexes with DNA and RNA and act as important signaling compounds for cell growth

(20). Evidence in support of a hypothesis that the *aphA* transcripts reflect a role for polyamines as a nitrogen source for coastal bacterioplankton includes the facts that polyamines are produced by marine algae, plants, invertebrates, and microorganisms (8, 13, 16, 20), they reach concentrations of 30 nM in coastal seawater during algal blooms (16), and they are readily assimilated by coastal and open ocean bacterioplankton communities (8). Further, the genome sequence of marine bacterium *S. pomeroyi* contains an *aphA* homolog located in an apparent operon with a polyamine transporter (*potABCD*) (15) and candidate genes for a putrescine degradation pathway (putrescine transaminase, aminobutyraldehyde dehydrogenase). The *aphA* transcripts may be a response to unusual conditions caused by sample processing (e.g., a spike in polyamine concentrations due to eukaryotic cell breakage during filtration), but nonetheless indicate an ability to rapidly respond to the availability of these nitrogen-rich compounds in seawater. While polyamine assimilation by marine bacteria has been considered in the past (8), the SIMO transcript library forms the foundation of a hypothesis that these compounds are a more important source of dissolved organic nitrogen for coastal bacterioplankton than currently suspected.

Our environmental transcriptomics protocol was used successfully to survey two very different types of aquatic communities for microbial gene expression, without the constraints of targeting specific organisms, phylogenetic groups, or metabolic pathways. While the libraries analyzed here were small, this approach can be readily adapted for high-throughput processing and automated annotation, and can be coupled to environmental genomics methods to assess genetic potential and patterns of activity in natural microbial assemblages.

ACKNOWLEDGEMENTS

We thank J. R. Henriksen for help with sequence analysis, E. C. Hierling and J. Shalack for assistance with sample collection and analysis, and three anonymous reviewers for constructive comments. This work was supported by NSF grants MCB-0084164 to the Sapelo Island Microbial Observatory and MCB-9977886 to the Mono Lake Microbial Observatory, and by the Gordon and Betty Moore Foundation. R.S.P. was funded by an NSF graduate research fellowship, and M.P. and W.M.P. by NSF REU fellowships. The Annotation Engine service was provided by The Institute for Genomic Research as a result of funding from the DOE and the NSF.

REFERENCES

1. **Béjà, O., M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong.** 2000. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ. Microbiol. **2:**516-529.

2. **Belasco, J. G., and G. Brawerman.** 1993. Control of messenger RNA stability. Academic Press, San Diego.

3. **Bernstein, J. A., A. B. Khodursky, P. H. Lin, S. Lin-Chao, and S. N. Cohen.** 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. U S A **99:**9697-702.

4. **Felsenstein, J.** 1989. PHYLIP, phylogenetic inference package (version 3.2). Cladistics **5:**164-166.

5. **Fleming, J. T., W. H. Yao, and G. S. Sayler.** 1998. Optimization of differential display of prokaryotic mRNA: application to pure culture and soil microcosms. Appl. Environ. Microbiol. **64:**3698-706.

6. **González, J. M., J. S. Covert, W. B. Whitman, J. R. Henriksen, F. Mayer, B. Scharf, R. Schmitt, A. Buchan, J. A. Fuhrman, R. P. Kiene, and M. A. Moran.** 2003. *Silicibacter pomeroyi* sp. nov. and *Roseovarius nubinhibens* sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. Int. J. Syst. Evol. Microbiol. **53:**1261-9.

7. **Hambraeus, G., C. von Wachenfeldt, and L. Hederstedt.** 2003. Genome-wide survey of mRNA half-lives in *Bacillus subtilis* identifies extremely stable mRNAs. Mol. Genet. Genomics. **269:**706-14.

8. **Höfle, M. G.** 1984. Degradation of putrescine and cadaverine in seawater cultures by marine bacteria. Appl. Environ. Microbiol. **47:**843-849.

9. **Humayoun, S. B., N. Bano, and J. T. Hollibaugh.** 2003. Depth distribution of microbial diversity in Mono Lake, a meromictic soda lake in California. Appl. Environ. Microbiol. **69:**1030-1042.

10. **Ingraham, J. L., O. Maaløe, and F. C. Neidhardt.** 1983. Growth of the bacterial cell. Sinauer Associates, Sunderland, Mass.

11. **LeCleir, G. R., A. Buchan, and J.T.Hollibaugh.** 2004. Chitinase gene sequences from diverse aquatic habitats reveal complex patterns of diversity. Appl. Environ. Microbiol. **70:**6977-6983.

12. **Liang, P., and A. B. Pardee.** 1992. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science **257:**967-71.

13. **Lu, Y. H., and D.-F. Hwang.** 2002. Polyamine profile in the paralytic shellfish poison-producing alga *Alexandrium minutum*. J. Plankton Res. **24:**275-279.

14. **Maest, A. S., S. P. Pasilis, L. G. Miller, and D. K. Nordstrom.** 1992. Redox geochemistry of arsenic and iron in Mono Lake, California, p. 507-511. *In* Y. K. Kharaka and A. S. Maest (ed.), Water-Rock Interaction. A.A. Balkema.

15. **Moran, M. A., A. Buchan, J. M. González, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B. Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Ye, T. R. Miller,**

J. Carlton, D. A. Rasko, I. T. Paulsen, Q. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S. Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut, and N. Ward. 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature., in press.

16. **Nishibori, N., A. Yuasa, M. Sakai, S. Fujihara, and S. Nishio.** 2001. Free polyamine concentrations in coastal seawater during phytoplankton bloom. Fish. Sci. **67:**79-83.

17. **Rodríguez-Valera, F.** 2004. Environmental genomics, the big picture? FEMS Microbiol. Lett. **231:**153-8.

18. **Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman.** 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. **66:**2541-2547.

19. **Sakurada, K., T. Ohta, K. Fujishiro, M. Hasegawa, and K. Aisaka.** 1996. Acetylpolyamine amidohydrolase from *Mycoplana ramosa*: Gene cloning and characterization of the metal-substituted enzyme. J. Bacteriol. **178:**5781-5786.

20. **Tabor, C. W., and H. Tabor.** 1984. Polyamines. Annu. Rev. Biochem. **53:**749-790.

21. **Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37-43.

Table 2.1. Selected mRNA sequences with inferred functions of ecological or geochemical relevance in cDNA libraries constructed from SIMO (top) and MLMO (bottom) samples. A complete list of all transcript annotations is provided in Supplemental Table 2.3 (http://aem.asm.org/cgi/data/71/7/4121/DC1/1).

Gene	**Putative function**	**Closest Match Source**	**Accession number**	**Percent Identity**[a]
soxA	Inorganic sulfur oxidation	*Chlorobium tepidum*	NP_661911	51%
fdh1B	Formate dehydrogenase beta-subunit (C1 metabolism)	*Methanococcus voltae*	AAK57554	58%
trkA	Potassium uptake	*Bacteroides thetaiotaomicron*	NP_813009	63%
kefA	Potassium efflux	*Pseudomonas syringae*	AAO58452	48%
psbA2	Photosystem II protein	*Synechococcus sp. WH 8102*	NP_897076	78%
surE	Stationary-phase survival protein	*Coxiella burnetii*	NP_820653	58%
proV	Proline/glycine betaine/DMSP transport system	*Streptomyces coelicolor*	AAD29279.1	63%
mexE	Multi-drug efflux membrane protein	*Pseudomonas syringae*	NP_792891	85%
chi	Chitinase	*Bacillus thuringiensis*	AAM88400	27%
arsA	Arsenite-transporting ATPase	*Arabidopsis thaliana*	NP_563640	68%

[a]Based on deduced amino acid sequences

Table 2.2. SIMO transcript identities and assigned role categories (August 2003 sample) as determined by the TIGR Annotation Engine.

Main Role	Subrole	Number	%[a]
Amino acid biosynthesis		**6**	2.5
	Aspartate	2	
	Pyruvate	4	
Biosynthesis of cofactors		**1**	0.4
Cell envelope		**1**	0.4
Cellular processes		**13**	5.5
	Chemotaxis and motility	1	
	Detoxification	8	
	DNA transformation	2	
	Toxin production and resistance	2	
Central intermediary metabolism		**44**	18.5
DNA metabolism		**5**	2.1
	Replication, recombination, repair	5	
Energy metabolism		**10**	4.2
	Aerobic	1	
	Amino acids and amines	3	
	ATP proton motive force	1	
	Electron transport	1	
	Glycolysis/gluconeogenesis	1	
	Pentose phosphate pathway	1	
	Photosynthesis	1	
	TCA cycle	1	
Protein fates		**3**	1.3
	Degradation of proteins, peptides & glycolpeptides	1	
	Protein and peptide secretion	2	
Protein synthesis		**12**	5.0
	Ribosomal proteins	10	
	Translation factors	1	
	tRNA aminoacylation	1	
Regulatory functions		**10**	4.2
Transport and binding proteins		**9**	3.8
	Amino acids, peptides, amines	2	
	Carbohydrates, alcohols and organic acids	2	
	Unknowns	5	
Unknown functions		**2**	0.8
Unclassified[b]		**28**	11.8
Conserved hypothetical proteins[c]		**10**	4.2
Hypothetical proteins		**84**	35.3
rRNA		**44**	
Total number of clones		**282**	

[a] Calculation of percent representation in the library does not include the 44 rRNA sequences.
[b] Unclassified proteins have a known function, but have not been assigned to a role category.
[c] Conserved hypothetical proteins have homologs in other organisms, but none of the homologs have known functions.

Figure 2.1. Taxonomic assignment of SIMO mRNA (A) and 16S rRNA (B) sequences from the 0.2- 3.0 μm size fraction of SIMO bacterioplankton. *Archaea,* viruses and fungi were not captured by the 16S rRNA primers used in the rRNA library. No putative taxonomic assignment could be made for 37% of the mRNA clones because they had highest similarity to genes from unclassified organisms. Twenty-three percent of the mRNA clones did not match any database sequence well using a BLASTX E-value cutoff of ~e^{-10}, and were not included in this figure. Approximately 17% of the rRNA clones could not be readily assigned to a phylum using a cut-off value of 80% similarity to described organisms, while 20% could be classified as *Proteobacteria* but not assigned a class within this phylum.

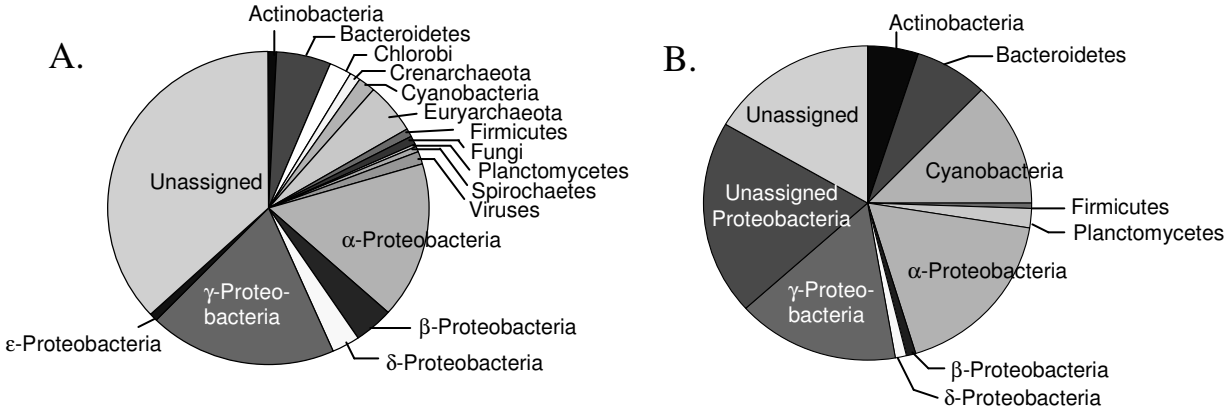
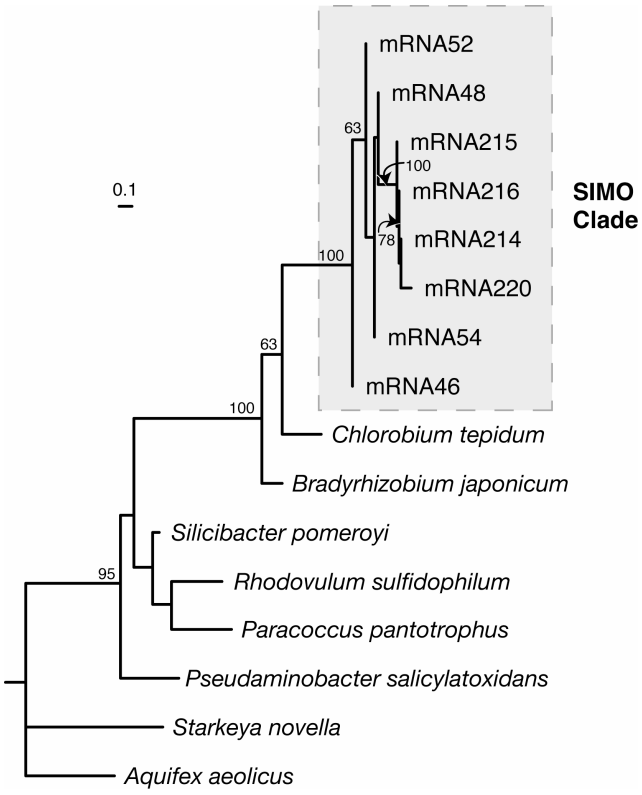

A.
Actinobacteria
Bacteroidetes
Chlorobi
Crenarchaeota
Cyanobacteria
Euryarchaeota
Firmicutes
Fungi
Planctomycetes
Spirochaetes
Viruses
Unassigned
α-Proteobacteria
γ-Proteo-
bacteria
β-Proteobacteria
ε-Proteobacteria
δ-Proteobacteria
B.
Actinobacteria
Bacteroidetes
Unassigned
Cyanobacteria
Unassigned
Proteobacteria
Firmicutes
Planctomycetes
α-Proteobacteria
γ-Proteo-
bacteria
β-Proteobacteria
δ-Proteobacteria

Figure 2.2. Phylogenetic tree of SIMO-specific *soxA* sequences and those from representative cultured organisms constructed using the neighbor-joining method of the PHYLIP package (4). The tree is based on the deduced amino acids encoded by the *soxA* transcripts or genes (positions 8 to 247; *C. tempidum* numbering system) and is unrooted, with *soxA* from *A. aeolicus* (AE000757) as the outgroup. Bootstrap values ≥ 50% are indicated at branch nodes. The scale bar indicates Dayhoff PAM distance.


mRNA52
mRNA48
63
mRNA215
100
0.1
mRNA216
SIMO
Clade
78
mRNA214
100
mRNA220
63
mRNA54
mRNA46
100
Chlorobium tepidum
Bradyrhizobium japonicum
Silicibacter pomeroyi
95
Rhodovulum sulfidophilum
Paracoccus pantotrophus
Pseudaminobacter salicylatoxidans
Starkeya novella
Aquifex aeolicus

Supplemental Methods

Environmental sample collection. Two National Science Foundation (NSF) Microbial Observatory sites served as the locations for this study. The Sapelo Island, GA (SIMO) site is a tidal marsh creek typical of southeastern U.S. salt marshes and supports a bacterioplankton abundance averaging 1.7 x 10^6 cells ml^{-1} (http://gce-lter.marsci.uga.edu/lter/). Mono Lake (MLMO) is a closed-basin, alkaline (pH 9.8), hypersaline (~90 g l^{-1}) soda lake located east of the Sierra Nevada Mountains, approximately 160 km south of Lake Tahoe, California. The lake is naturally eutrophic and bacterioplankton abundance is typically > 10^7 cells ml^{-1} (3). The Mono Lake Microbial Observatory web site (www.monolake.uga.edu) has further information on lake biogeochemistry.

SIMO water samples (10 l) were collected in October 2002 and August 2003. Immediately after collection, the samples were screened to remove particles rich in eukaryotic nucleic acid by filtration through an 8.0 µm pore size, 293 mm diameter Poretics polycarbonate membrane filter (Osmonics) (October sample) or a 5.0 µm pore size polypropylene cartridge filter (USFilter, Warrendale, PA) (August sample) followed by a 3.0 µm pore size polycarbonate filter. Cells for RNA extraction were collected on a 0.2 µm pore size, 293 mm diameter polycarbonate membrane filter. MLMO water samples (~8 l) were collected in May 2003 at depths of 5 m (surface) and 23 m (chemocline) in the center of the lake using a Niskin water sampler. These samples could not be screened to remove eukaryotes prior to collecting bacteria because a dominant phytoplankter (*Picocystis salinarum*) and the bacteria in Mono Lake are of the same size, especially in chemocline samples where bacteria are large and filamentous forms are common (www.monolake.uga.edu). An on-site laboratory was not available Mono Lake, and thus these samples were stored on ice for 2 hours until arrival at a lab, at which point they were

filtered onto a 0.2 µm pore size, 293 mm diameter polycarbonate membrane filter (Osmonics, Minnetonka, MN).

RNA isolation and mRNA enrichment. The entire process from sample collection to RNA extraction was done as rapidly as possible to limit degradation of mRNA. We used standard precautions to minimize nuclease contamination. For SIMO, where samples were processed on-site, the elapsed time between water collection and RNA extraction was less than 30 minutes. RNA was extracted using the RNAqueous-Midi kit (Ambion, Austin, TX) with modifications to the manufacturer's instructions. Immediately after filtration, the filter was cut into small pieces with a sterile, RNase-free razor and forceps and homogenized by vortexing for 10 min with beads from a soil DNA extraction kit (MoBio, Carlsbad, CA) and the lysis/binding solution provided by the RNAqueous-Midi kit. Following centrifugation at ~10,000 rpm for 10 min to clarify the lysate, the supernatants were poured into sterile petri dishes to which 5 ml of the RNAqueous-Midi kit ethanol solution was added. The solution was mixed gently and then passed several times through an 18 or 22 gauge needle in order to shear genomic DNA. While the mixture was in the syringe, the needle was removed and a glass fiber filter (provided with the kit) was placed on the end of the syringe. The mixture was passed through the filter unit and then washed and eluted according to the manufacturer's instructions. Immediately after the extraction, the RNA was frozen on dry ice and stored at -70°C. The same procedure was followed for MLMO except that the water samples were first transported to a lab on ice (~2 h elapsed time between water collection and RNA extraction).

To enrich for mRNA, rRNA was removed by subtractive hybridization with capture oligonucleotides hybridized to magnetic beads using the MICROB*Express* Bacterial mRNA Enrichment kit (Ambion). The enrichment process was repeated a second time for some samples

to ensure removal of as much rRNA as possible. To remove any contaminant DNA, the mRNA was treated with DNase I using the DNA-free kit (Ambion). RNA quantity was determined by measuring absorbance at 260 nm and purity was checked using the ratio of absorbance at 260 nm to 280 nm and found to be consistently >1.7.

RT-PCR. RT-PCR was performed on 1- 6.5 µl of SIMO mRNA using either 10-mer primers randomly chosen from a commercial primer stock (OPA04, OPA13, and OPA17; Operon Technologies, Inc., Alameda, CA) or with primer SD14 (Table S1.1). SD14 was designed by Fleming et al. (2) to target the Shine-Dalgarno region of bacterial mRNAs. One primer was used in the RT reaction, and this primer was combined with a second primer for PCR (Table S1.2). The RT reaction was performed using the Omniscript RT kit (Qiagen, Valencia, CA) in 10 µl volumes containing 1X RT buffer, 0.5 µl of 10 µM primer, 1 µl of 5 mM dNTPs, 2 U of reverse transcriptase, and 20 U of RNase inhibitor (Promega, Madison, WI) at 37°C for 1 h, followed by inactivation of the reverse transcriptase at 95°C for 2 min. PCR amplification of the resulting SIMO cDNA was performed in 25-µl volumes containing 12.5 µl of MasterAmp 2X PCR Premix F (Epicentre, Madison, WI), 0.5 µl of each 10 µM primer, 1.25 U of *Taq* DNA polymerase (Qiagen), and 5 µl cDNA from the RT reaction. PCR conditions included a preliminary denaturation for 5 min at 94°C, followed by 40 cycles of denaturation at 94°C for 45 sec, annealing at 37-45°C for 45 s, primer extension at 72°C for 1 min, and a final extension at 72°C for 10 min. The low annealing temperatures were used in order to decrease the specificity of the reaction so as to maximize the number of resulting amplicons. A PCR control without an initial RT step was included with every set of reactions to ensure that there was no DNA contamination in the mRNA extracts. MLMO mRNA samples were similarly processed, but with the use of two additional arbitrary primers (SES3-1 and SESRT-3, Table S1.1) with lower G+C

content (40%) than the OPA primers (60-70%). In addition, MLMO RT reactions were carried out in 20 µl reaction volumes and used the *Taq* PCR Core Kit (Qiagen) in 100 µl PCR volumes.

Cloning, colony hybridizations, and sequencing. SIMO PCR products were cleaned using the Ultra-Clean PCR Clean-up Kit (Mo-Bio) and cloned into the pCR2.1 vector using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA). The presence of an insert was verified by restriction digest for 1 h at 37°C with *Eco*RI. Because some rRNA was present even after subtractive hybridization, the SIMO libraries were screened by colony hybridizations to identify clones containing rRNA fragments. rRNA probes were constructed by amplifying rRNA genes (PCR Ready-To-Go beads; Amersham Pharmacia Biotech, Piscataway, NJ) from DNA harvested from the same sample (MoBio Soil DNA extraction kit) using primers specific for 16S, 23S, 28S, and 18S rRNA gene sequences (Table S1.1). PCR products were then labeled with DIG-High Prime (Roche, Indianapolis, IN).

Colonies were grown on LB plates for 1 h at 37°C and then transferred to nylon transfer membranes (Amersham Pharmacia Biotech) that were placed on fresh LB plates and grown overnight at 37°C. Following immobilization of colony DNA on the membrane by treatment at 42°C in 5X SSC buffer (50% deionized formamide, 0.1% sodium-lauroylsarcosine, 0.02% SDS, and 2% blocking reagent (Roche)), the labeled probes were hybridized to colony DNA by incubating the membrane with all of the probes simultaneously overnight in the 5X SSC hybridization buffer at 42°C. The blots were washed according to the manufacturer's instructions. Probe binding was detected by conjugation to Anti-DIG-alkaline phosphatase and exposure to film.

Clones that did not hybridize to the rDNA probes were selected for sequencing. All of the 60 MLMO clones and 65 of the SIMO clones (40 from October 2002 and 25 from August 2003)

were sequenced at the University of Georgia on the ABI PRISM 310 or 9600 Genetic Analyzer (Applied Biosystems, Foster City, CA). Sequences were analyzed using the BLASTX and BLASTN tools (http://www.ncbi.nlm.nih.gov/BLAST/). Additionally, 282 SIMO clones (from August 2003) were sequenced by a commercial firm (Seqwright, Inc., Houston TX). The sequences were analyzed by two different approaches: they were manually submitted to NCBI for BLASTN and BLASTX searches and they were automatically annotated using the Annotation Engine service provided by The Institute for Genomic Research (TIGR, Rockville, MD). BLASTN analysis served to identify rRNA sequences, while functional gene assignments were based on BLASTX.

Methods for the MLMO samples were identical except that PCR products were cleaned using the QIAquick PCR Purification Kit (Qiagen) and cloned into the pGEM-T Easy Vector (Promega), and clones were not screened for rRNA inserts.

The sequences obtained in this study were deposited in GenBank under the accession numbers AY793704-AY794012.

Real-time PCR quantification. Genes and transcripts of putative *soxA* sequences were quantified using DNA and RNA extracted from 10 l samples of SIMO water collected in August 2004. Quantitative PCR (qPCR) primers were designed based on the *soxA* sequences retrieved in the August 2003 SIMO library (Table S2.1). RT reactions were carried out as described above, but in 20 or 25 µl reaction volumes using either the SoxA-F or SoxA-R primer. *soxA* sequences in DNA or cDNA were quantitatively amplified on the iCycler iQ real-time PCR detection system (Bio-Rad, Hercules, CA) in a 20 µl reaction volume containing 10 µl of iQ SYBR Green Supermix (Bio-Rad), 0.4 µl each of 10 µM SoxA-F and SoxA-R primers, and either DNA or cDNA as template. PCR conditions included a preliminary denaturation at 95°C for 3 min

followed by 45 cycles of 95°C for 15 s, annealing at 50°C for 1.5 s, 95°C for 1 min, and 55°C for 1 min. A melt curve was generated following the PCR, beginning with 55°C and increasing 0.4°C every 10 s until 95°C. A standard curve was generated using a cDNA clone of *soxA* that had been excised from the cloning vector and quantified on a Hoefer DyNA Quant 200 fluorometer (Amersham Biosciences, Piscataway, NJ).

Supplemental Table S2.1. Sequences of primers used in this study.

Primer	Sequence (5'-3')	Purpose	Reference
OPA13	CAGCACCCAC	RT-PCR	
OPA17	GACCGCTTGT	RT-PCR	
OPA04	AATCGGGCTG	RT-PCR	
SD14	GGGGAACGACGATG	RT-PCR	(2)
SES3-1	CTAAACTCACTCTTACGGATCA	RT-PCR	
SESRT-3	AATGCGTACA	RT-PCR	
23S F	GCGATTTCYGAAYGGGGRAACCC	rRNA screen	(1)
23S R	TTCGCCTTTCCCTCACGGTACT	rRNA screen	(1)
1492R	GGTTACCTTGTTACGACTT	rRNA screen	(5)
6F	GGAGAGTTAGATCTTGGCTCA	rRNA screen	(5)
LR0R-28S	ACCCGCTGAACTTAAGC	rRNA screen	(6)
LR3R-28S	GTCTTGAAACACGGACC	rRNA screen	(6)
LR7-28S	TACTACCACCAAGATCT	rRNA screen	(6)
NS1-18S	GTAGTCATATGCTTGTCTC	rRNA screen	(4)
NS2-18S	GGCTGCTGGCACCAGACTTGC	rRNA screen	(4)
NS3-18S	GCAAGTCTGGTGCCAGCAGCC	rRNA screen	(4)
NS4-18S	CTTCCGTCAATTCCTTTAAG	rRNA screen	(4)
SoxA-F	CTGGGAGGAGTCAATAAATG	qPCR	(this study)
SoxA-R	CAACAGAAGATGCTGAGAAAG	qPCR	(this study)

Supplemental Table S2.2. Summary of the sequences in cDNA libraries constructed from SIMO (top) and MLMO (bottom) samples.

			Number of transcripts encoding for					
Sample ID	**Primer**[a]	**Clones**	**Protein**	**16S rRNA**	**18S rRNA**	**23S rRNA**	**28S rRNA**	**5S rRNA**
SIMO October 2002	2 OPA primers	40	8 (20%)	9		17	6	
SIMO August 2003	SD14+OPA17	47	47 (100%)					
SIMO August 2003	SD14+OPA13	9	6 (67%)	3				
SIMO August 2003	SD14+SD14	57	53 (93%)	3			1	
SIMO August 2003	OPA13+OPA17	36	13 (36%)				23	
SIMO August 2003	OPA13+SD14	60	53 (88%)	4		2	1	
SIMO August 2003	OPA17+OPA17	2	1				1	
SIMO August 2003	OPA17+OPA13	36	30 (83%)			6		
SIMO August 2003	OPA17+SD14	60	55 (92%)	4		1		
MLMO Surface	SD14+SESRT	13	10 (77%)	1				
MLMO Surface	SD14+OPA17	11	6 (55%)	2	1	2		
MLMO Surface	SD14+OPA13	7	5	1				
MLMO Chemocline	SD14+SESRT	15	11 (73%)	2		2		
MLMO Chemocline	SD14+OPA17	10	6 (60%)	1				3
MLMO Chemocline	SD14+OPA13	4	4 (100%)					

[a] The first primer listed was used in the RT step and both primers were used in the PCR step

REFERENCES

1. **Anthony, R. M., T. J. Brown, and G. L. French.** 2000. Rapid diagnosis of bacteremia by universal amplification of 23S ribosomal DNA followed by hybridization to an oligonucleotide array. J. Clin. Microbiol. **38:**781-788.
2. **Fleming, J. T., W. H. Yao, and G. S. Sayler.** 1998. Optimization of differential display of prokaryotic mRNA: application to pure culture and soil microcosms. Appl. Environ. Microbiol. **64:**3698-706.
3. **Humayoun, S. B., N. Bano, and J. T. Hollibaugh.** 2003. Depth distribution of microbial diversity in Mono Lake, a meromictic soda lake in California. Appl. Environ. Microbiol. **69:**1030-1042.
4. **Innis, M. A.** 1990. PCR protocols: a guide to methods and applications. Academic Press, San Diego.
5. **Lane, D. J., B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace.** 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. USA **82:**6955-9.
6. **Vilgalys, R., and M. Hester.** 1990. Rapid genetic identification and mapping of enzymatically amplified ribosomal DNA from several *Cryptococcus* species. J. Bacteriol. **172:**4238-46.

CHAPTER 3

ENVIRONMENTAL TRANSCRIPTOMICS: A METHOD TO ACCESS EXPRESSED GENES IN COMPLEX MICROBIAL COMMUNITIES[1]

[1] Poretsky, R. S., N. Bano, A. Buchan, M. A. Moran, and J. T. Hollibaugh (2008) in *Molecular Microbial Ecology Manual*, eds. Kowalchuk, G. A., Bruijn, F. J. d., Head, I. M., Akkermans, A. D. L., & Elsas, a. J. D. v. (Springer Netherlands), pp. 1892-1904.

INTRODUCTION

As new molecular methods become available and applicable to environmental systems, information on microbial diversity, community structure, and ecological function continues to expand. Of particular interest to assessing microbial functions *in situ* have been advances in the use of functional gene probes or primer sets to examine messenger RNA (mRNA) extracted directly from natural communities [4, 12]. Novel approaches involving microarrays offer the opportunity to examine the presence and potentially expression of multiple genes in complex microbial assemblages simultaneously [13]. These exciting methods provide valuable information on gene expression in natural communities, but they are limited to genes for which sequence information is already available.

The technology of environmental genomics is based on sequence analysis of clone libraries containing inserts of environmental DNA, and retrieves genes without any previous sequence information in a fairly representative fashion [1, 9, 11]. In this chapter, we present an analogous method for environmental mRNA that similarly retrieves the transcriptome of a microbial assemblage without prior knowledge of what genes the community might be expressing. We refer to this strategy as environmental transcriptomics.

Despite the prospect for environmental transcriptomics to provide a direct link between the genetic potential and biogeochemical activity of microbes, mRNAs are inherently difficult to work with. Some degrade quickly, with half lives as short as 30 seconds based on studies of mRNAs of cultured bacteria [2]. Prokaryotic mRNAs generally lack the poly(A) tails that make isolation of eukaryotic messages relatively straightforward [6]. Finally, mRNAs are much less abundant than rRNAs in total RNA extracts, thus an rRNA background often overwhelms

mRNA signals. However, techniques for overcoming some of the difficulties of working with environmental microbial mRNA have been recently developed and have facilitated novel discoveries.

A procedure for analyzing environmental transcriptomes by creating clone libraries generated from environmental mRNAs is described in this chapter. Environmental transcriptomics provides a way to survey an intact community for gene expression without the constraints of targeting a specific organism, phylogenetic group, or metabolic pathway. In general, total RNA is first collected from the environment, rRNA is selectively removed, and cDNA synthesized from the enriched mRNA pool is cloned and sequenced. Recovered sequences can be annotated using standard bioinformatics techniques to identify the expressed genes. We present two alternative methods for generating cDNA from environmental mRNA (Figure 3.1). One method uses random 10- 14 bp primers for reverse transcription and low-specificity PCR. The second method relies on a linear amplification of mRNA by polyadenylating the mRNA and carrying out *in vitro* transcription followed by synthesis of double stranded cDNA using random hexamers. We have successfully employed both of the methods described here to water samples from several different environments including a hypersaline lake in California [7], a salt marsh off the coast of Georgia [7], and the North Pacific ocean. It should be noted that the protocols described here are optimized for aquatic systems, and appropriate modifications may be necessary for analysis of other types of samples (e.g. sediment, soil, etc.).

PROCEDURES

Working with RNA

Because RNases are ubiquitous and mRNAs degrade rapidly, standard precautions for working in a ribonuclease-free environment must be followed and samples should be processed or preserved as soon as possible following collection.

Total RNA extraction

Several options are available for isolating total RNA from environmental samples. At present, the most efficient and effective methods are based on commercially available kits and use proprietary reagents, thus this protocol relies heavily on the availability of these, or similar, products. Filter-based extraction systems such as the RNAqueous Kit (Ambion, Austin, TX) or the RNeasy Kit (Qiagen, Valencia, CA) yield high quality RNA. Depending on the environment from which RNA is being isolated, phenol/guanidine-based methods [e.g. using RNAwiz (Ambion) or TRI-Reagent (Molecular Research Center, Cincinnati, OH)] founded on an approach developed by Chomczynski and Sacchi [5] may be preferred. Such methods often yield more RNA per volume sample and are more flexible with regards to sample size. An essential step in any isolation approach is the empirical determination of the most efficient method of RNA extraction for a given environment. Below, we present modifications to kit-based techniques for whole water samples where microbial assemblages were collected by filtration onto 0.2 μm polycarbonate filters in the field.

1. If field conditions preclude extracting RNA on-site, preserve filtered samples in the field by adding 2 ml of Buffer RLT (containing β-mercaptoethanol) from the RNeasy Kit (Qiagen) or Stop Solution (95% ethanol and 5% water saturated phenol) in an RNase-free tube. Shake

well and flash freeze in liquid nitrogen or dry ice until processing further. If Stop Solution is used, it is recommended that the solution be removed prior to freezing by centrifugation and decanting.

2. Thaw frozen samples slightly for 2 min in a 40-50°C water bath prior to extraction.
3. If Stop Solution was employed, add an appropriate extraction buffer, such as the first buffer from an RNA extraction kit, RNAwiz, or TRI-Reagent.
4. Vortex sample for 10 min (or bead-beat for 2 min) with 0.1 mm zirconia/silica beads or equivalent RNase-free beads, such as those supplied with the Mo-Bio Soil DNA kit (Carlsbad, CA). These beads are reusable after ashing.
5. Centrifuge 5 min at 3,000- 5,000 x *g* and transfer supernatant to a new tube.
6. Add 1 volume of 70% ethanol to the lysate. In order to sheer large molecular weight nucleic acids, draw the lysate up through an 18-22 gauge needle and pass it back out several (~5) times.
7. RNA extraction can be continued with any standard kit or technique at this point.
8. Remove any contaminating DNA by DNase I treatment.

mRNA enrichment

We have employed two different techniques for removing rRNA:

1. rRNA can be removed by subtractive hybridization with capture oligonucleotides hybridized to magnetic beads using the MICROB*Express* kit (Ambion) according to the manufacturer's instructions. This enrichment process can be repeated multiple times if rRNA contamination is substantial. This approach is efficient, but biased towards organisms whose rRNAs hybridize to the oligonucleotide probe. A list of known organisms compatible with the

oligonucleotide probe used in this kit is provided on the Ambion webiste http://ambion.com/techlib/misc/microbe.html

2. The mRNA-ONLY Prokaryotic mRNA Isolation Kit (Epicentre Biotechnologies, Madison, WI) is an enzymatic method that uses a 5'-phosphate-dependant exonuclease to cleave all prokaryotic rRNAs, all of which have a 5'-monophosphate. It is important to use RNA of high integrity, as the 5' -OH termini of partially degraded RNAs are not recognized by the exonuclease. Because some degree of RNA degradation can be expected if time elapses between sample collection and processing (which may be inevitable when working in remote environments), we have had limited success solely using the enzymatic rRNA removal approach on environmental samples. However, when used in conjunction with the oligonucleotide-based approach, we are able to remove as much as 90% of the rRNA (Figure 3.2).

RNA concentration can be determined with any standard spectrophotometer by measuring absorbance at 260 nm or use of specialized instruments with low sample volume requirements (e.g. NanoDrop Spectrophotometer; Wilmington, DE). We also recommend assessing RNA quality electrophoretically with an Experion Automated Electrophoresis System (Bio-Rad, Hercules, CA) or equivalent RNA analyzer.

Generating ds cDNA from environmental mRNA option 1: RT-PCR with random oligonucleotides

Random oligonucleotide primers can be used to reverse transcribe and amplify mRNA. This technique has been successfully used to create environmental transcript libraries from two

different aquatic environments, but it should be noted that the true random nature of the primers is debatable [7]. Despite this potential bias, an informative transcript library can be generated when random primers are combined with a low specificity amplification procedure. We used 10-mer primers randomly chosen from a commercial primer stock (Operon Technologies, Inc., Alameda, CA)

1. Reverse transcription (RT) can be performed on ca. 50 ng- 2 µg mRNA using 0.5 µM of a random oligonucleotide. We employ the Omniscript reverse transcriptase (Qiagen, Valencia, CA) in 10 µl volumes containing:

   1X RT buffer

   0.5 µl of 10 µM primer

   1 µl of 5 mM dNTPs

   2 U of reverse transcriptase

   20 U of RNasin Ribonuclease Inhibitor (Promega, Madison, WI)

   The RT reaction is carried out at 37°C for 1 h, followed by inactivation of the reverse transcriptase at 95°C for 2 min.

2. The random primer used in the RT step is combined with a second random primer for PCR amplification. Low annealing temperatures (37-45°C) are used in order to decrease the specificity of the reaction so as to maximize the number of resulting amplicons. PCR conditions likely need to be optimized for different samples. We typically perform PCR amplifications in 25-µl volumes containing:

   12.5 µl of MasterAmp 2× PCR Premix F (Epicentre, Madison, WI)

   0.5 µl of each 10 µM primer

   1.25 U of *Taq* DNA polymerase (Qiagen)

1-5 µl cDNA from the RT reaction

PCR conditions include a preliminary denaturation for 5 min at 94°C, followed by 40 cycles of denaturation at 94°C for 45 sec, annealing at 37-45°C for 45 s, primer extension at 72°C for 1 min, and a final extension at 72°C for 10 min.

3. A PCR control without an initial RT step should be included with every set of reactions to ensure that there is no DNA contamination in the mRNA extracts.
4. RT-PCR-generated cDNA products are cleaned up with a kit such as the Ultra-Clean PCR Clean-up Kit (Mo-Bio).
5. Clone libraries can then be constructed with the resulting cDNA. We clone the cDNA into the pCR2.1 vector using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA).

Generating ds cDNA from environmental mRNA option 2: Linear amplification of mRNA and double stranded cDNA synthesis

Recent advances have provided tools for dealing with limitations of the above approach such as primer bias, selective PCR amplification and systematic bias against short-lived transcripts. Amplification of mRNA makes it possible use less starting material. This decreases the collection and processing time of samples and thereby minimizes RNA degradation. *In vitro* transcription methods for amplifying mRNA involve polyadenylating the mRNA and incorporating a T7 promoter onto the 3' end of the transcript. These modifications obviate the need for random primers in the amplification step. mRNA amplification does not appear to substantially change the expression profile of the RNA and is often used for analysis of gene expression using microarrays [14]. Amplified RNA (aRNA) can then be converted to double stranded cDNA using random hexamers [10]. Random oligonucleotides are widely accepted for

first-strand synthesis of cDNA [3]. Although one must consider that the multiple conversions between RNA and DNA could introduce error, this technique is thought to reduce several of the biases associated with alternative ways of accessing the environmental transcriptome.

mRNA amplification

As of this printing, the only commercially available kit for amplification of bacterial mRNA is the MessageAmp II-Bacteria Kit (Ambion). We used this kit according to the protocol provided by the manufacturer:

1. Bacterial mRNA is polyadenylated using oligo(dA) and Poly(A) Polymerase.
2. Polyadenylated mRNA is reverse transcribed with a T7 oligo(dT) primer (provided by the kit) to generate first-strand cDNA.
3. Second-strand cDNA is synthesized using a T7 promoter primer with DNA polymerase. RNase H should be added to this reaction to degrade any contaminating RNA template.
4. Following cDNA purification, *in vitro* transcription with T7 RNA polymerase is used to generate many copies of amplified RNA (aRNA).

Starting with 10- 20 ng of mRNA, we typically obtain 50-100 μg of aRNA using this procedure.

cDNA generation for cloning

Several commercial kits are available for double-stranded cDNA synthesis, including the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Carlsbad, CA) and the Universal RiboClone cDNA Synthesis System (Promega, Madison, WI). We have used both of these kits with success. However, the ds cDNA syntheis can also be carried out in the absence of a commercial kit. The procedure is outlined below.

1. The first step of converting aRNA to single-stranded cDNA uses an RNA-dependent DNA polymerase reverse transcriptase with random oligonucleotide primers (typically a hexamer mixture). Higher concentrations of random primers typically increase the cDNA yield but decrease the average cDNA size. Typically, 1- 5 μg of RNA can be converted to cDNA using 50 ng-1 μg of random hexamers. However, the initial aRNA concentration and the amount of primer should be determined empirically for each sample.
2. aRNA templates are removed with RNase H treatment while DNA polymerase I extends the primers.
3. DNA ligase is added to repair the nicks formed by RNase H.
4. Finally, the ds cDNA overhang ends are filled using T4 DNA polymerase or *Pfu* and excess dNTPs.
5. Following inactivation of the above reaction (by addition of EDTA to a final concentration of 10mM) and purification (by ethanol precipitation), the blunt-ended ds cDNA can be ligated into an appropriate cloning vector, such as the pCR-Blunt vector from the Zero Blunt PCR Cloning kit (Invitrogen). When run on an agarose gel, the ds cDNA should appear as a smear, representing a range of fragment sizes, typically between 50- 1000 bp (Figure 3.3).

Analyzing sequences

Sequences can be analyzed using the BLASTX and BLASTN tools at the National Centers for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/). BLASTN analyses serve to identify contaminating rRNA sequences, while functional gene assignments are based on BLASTX homology searches. The Annotation Engine service provided by The Institute for

Genomic Research (TIGR, Rockville, MD) is one useful tool available for assigning functional roles to sequences (Table 3.1).

Application of the Method

The applications of environmental transcriptomics are vast and include: 1) as a tool for the discovery of previously unknown genes or functions, 2) as a means for examining gene expression in microbial communities in the context of potentially influential environmental factors, 3) to gain a broad understanding of how microorganisms function within their environments, 4) to elucidate controls on biogeochemical processes, and 5) to provide novel material for more quantitative investigations of gene expression, such as environmental microarrays and quantitative PCR primer sets.

Functional gene discovery in natural environments is typically based on primer sets designed from a limited database that is heavily biased toward cultured organisms [8]. However, the first environmental transcriptomic libraries illustrate the potential of this method to identify previously unexplored variants of genes responsible for common biogeochemical functions. For example, by employing the transcriptomics approach in a coastal salt marsh in the southeastern United States, expression of community-specific variants of sulfur oxidation genes could be detected. At Mono Lake, an alkaline, hypersaline lake in California, a chitinase gene was detected that bore little similarity to known chitinases [7]. The utility of providing gene sequences of biogeochemical interest without constraints imposed by existing sequence data and with preference for those genes being actively expressed is evident. It is even conceivable for this method to provide full-length transcripts from the environment.

Environmental transcript libraries also have considerable potential for generating hypotheses about novel or unsuspected biogeochemical activities. Multiple variants of a polyamine deacetylase transcript found in the coastal marsh expression library have led to a hypothesis that polyamines may be an important, as yet unrecognized, source of carbon and nitrogen for this microbial community [7].

With the development of molecular tools and techniques that can be applied to natural microbial communities, it is now feasible to overcome many of the limitations associated with RNA work in the environment. Information gleaned from environmental transcriptomics is perhaps most powerful when used in conjunction with other methods of community analysis such as those involving genomics, qPCR, and microarrays.

ACKNOWLEDGEMENTS

This work was supported by NSF grants MCB-0084164 to the Sapelo Island Microbial Observatory and MCB-9977886 to the Mono Lake Microbial Observatory, and by the Gordon and Betty Moore Foundation. R.S.P. was funded by an NSF graduate research fellowship. The Annotation Engine service was provided by The Institute for Genomic Research as a result of funding from the DOE and the NSF.

REFERENCES

1. Béjà O, Suzuki MT, Koonin EV, Aravind L, Hadd A, Nguyen LP, Villacorta R, Amjadi M, Garrigues C, Jovanovich SB, Feldman RA, DeLong EF (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. Environ. Microbiol. 2:516-529

2. Belasco JG (1993) mRNA degradation in prokaryotic cells: an overview. In: Belasco JG, Brawerman G (eds) Control of messenger RNA stability pp. 3-11. Academic Press, San Diego

3. Binns MM, Boursnell MEG, Foulds IJ, Brown TDK (1985) The Use of a Random Priming Procedure to Generate cDNA Libraries of Infectious-Bronchitis Virus, a Large RNA Virus. Journal of Virological Methods 11:265-269

4. Bürgmann H, Widmer F, Sigler WV, Zeyer J (2003) mRNA extraction and reverse transcription-PCR protocol for detection of *nifH* gene expression by *Azotobacter vinelandii* in soil. Appl. Environ. Microbiol. 69:1928-1935

5. Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analytical Biochemistry 162:156-159

6. Liang P, Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257:967-971

7. Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M, Pate WM, Moran MA, Hollibaugh JT (2005) Analysis of Microbial Gene Transcripts in Environmental Samples. Appl. Environ. Microbiol. 71:4121-4126

8. Rodríguez-Valera F (2004) Environmental genomics, the big picture? FEMS Microbiol. Lett. 231:153-158

9. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR, Loiacono KA, Lynch BA, MacNeil IA, Minor C, Tiong CL, Gilman M, Osburne MS, Clardy J, Handelsman J, Goodman RM (2000) Cloning the soil metagenome: a strategy for

accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. 66:2541-2547

10. Sambrook J, Russell DW (2001) Preparation of cDNA libraries and gene identification. In Molecular cloning: a laboratory manual pp. 11.1-11.133. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

11. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43

12. Wawrik B, Paul JH, Tabita FR (2002) Real-time PCR quantification of rbcL (ribulose-1,5-bisphosphate carboxylase/oxygenase) mRNA in diatoms and pelagophytes. Applied and Environmental Microbiology 68:3771-3779

13. Zhou JH (2003) Microarrays for bacterial detection and microbial community analysis. Curr. Op. Microbiol. 6:288-294

14. Zhu BM, Xu F, Baba Y (2006) An evaluation of linear RNA amplification in cDNA microarray gene expression analysis. Molecular Genetics and Metabolism 87:71-79

Table 3.1: Role categories as determined by the TIGR Annotation Engine for transcripts from a coastal salt marsh transcript library.

Main Role	Number	%
Amino acid biosynthesis	**6**	2.5
Biosynthesis of cofactors	**1**	0.4
Cell envelope	**1**	0.4
Cellular processes	**13**	5.5
Central intermediary metabolism	**44**	18.5
DNA metabolism	**5**	2.1
Energy metabolism	**10**	4.2
Protein fates	**3**	1.3
Protein synthesis	**12**	5
Regulatory functions	**10**	4.2
Transport and binding proteins	**9**	3.8
Unknown functions	**2**	0.8
Unclassified	**28**	11.8
Conserved hypothetical proteins	**10**	4.2
Hypothetical proteins	**84**	35.3
rRNA	**44**	
Total number of clones	**282**	

Figure 3.1: Schematic representation of the environmental transcriptomics approach, including the two alternative methods for generating cDNA from environmental mRNA.

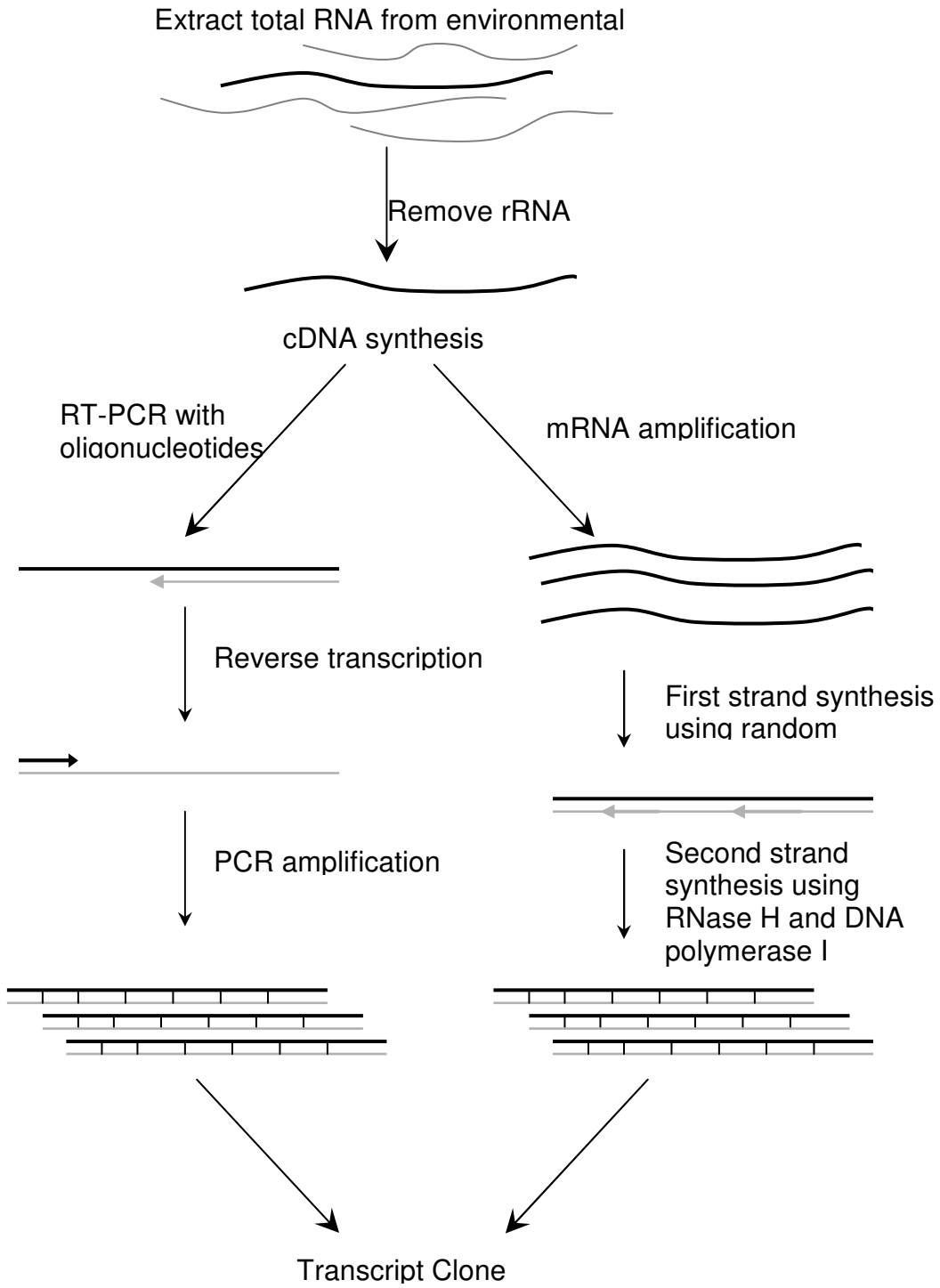
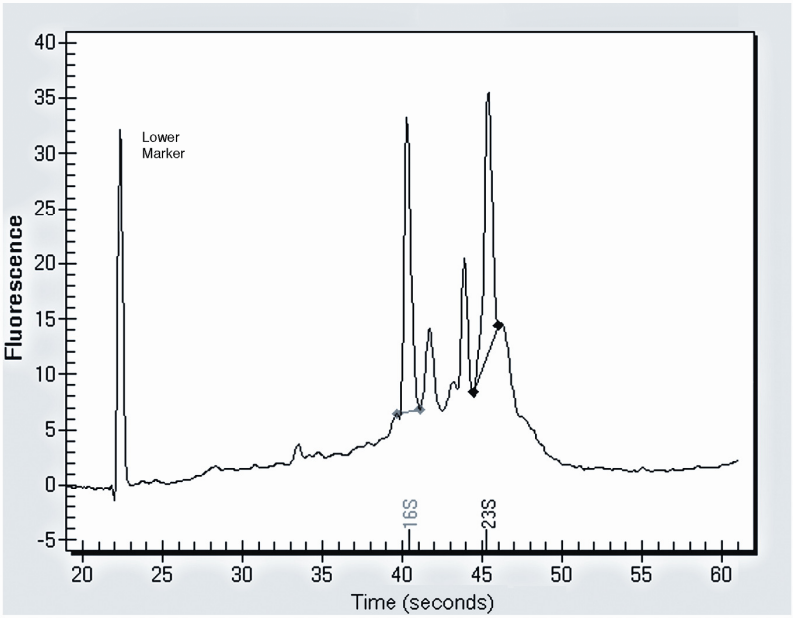

Extract total RNA from environmental
Remove rRNA
cDNA synthesis
RT-PCR with oligonucleotides
mRNA amplification
Reverse transcription
First strand synthesis using random
PCR amplification
Second strand synthesis using RNase H and DNA polymerase I
Transcript Clone

Figure 3.2: Experion traces of RNA extracted from seawater samples before (top) and after (bottom) removal of rRNA using the MICROB*Express* kit in combination with mRNA-ONLY Prokaryotic mRNA Isolation kit.


Lower
Marker
Fluorescence
Time (seconds)
16S
23S

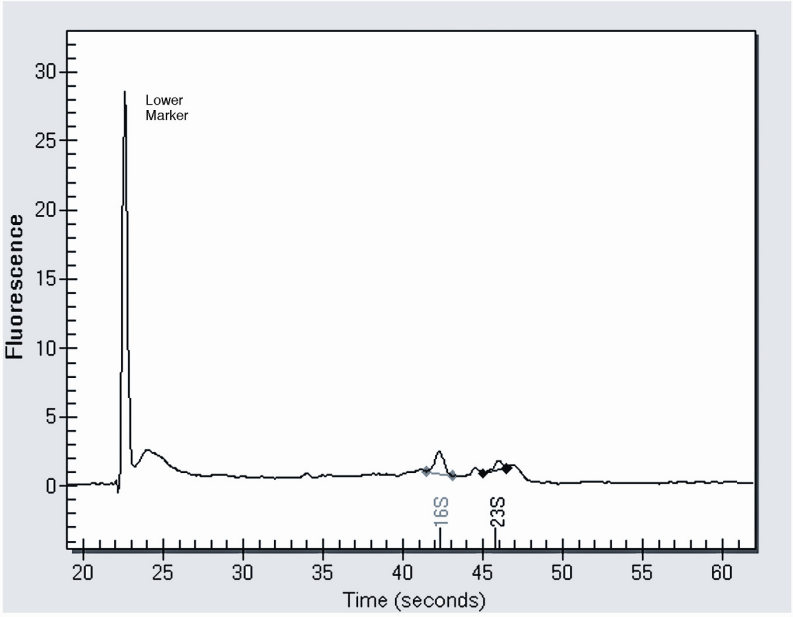
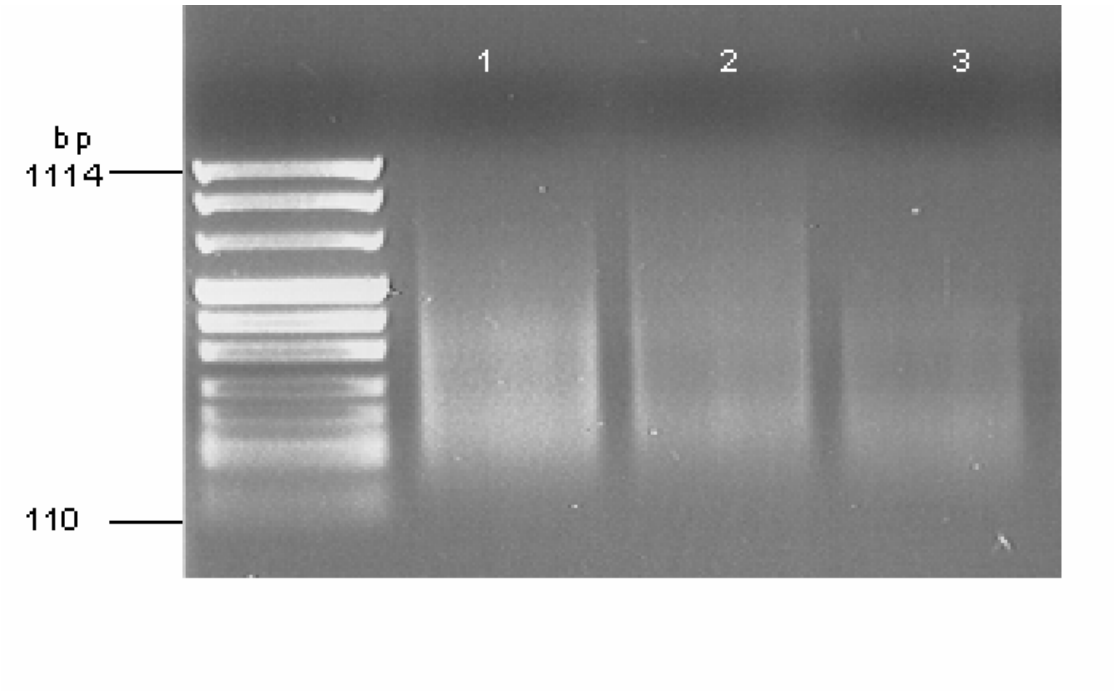

Lower
Marker
Fluorescence
Time (seconds)
16S
23S

Figure 3.3: 1% agarose gel depicting double-stranded cDNA generated from 5 µg aRNA (lane 1), 4 µg aRNA (lane 2) and 3 µg aRNA (lane 3). 100 ng of random hexamers were used for first-strand cDNA synthesis.


1
2
3
bp
1114
110

CHAPTER 4

DIEL METATRANSCRIPTOMIC ANALYSIS OF MICROBIAL COMMUNITIES IN THE NORTH PACIFIC SUBTROPICAL GYRE[1]

[1] Poretsky, R.S., I. Hewson, S. Sun, A. Allen, M.A. Moran, and J. Zehr. Submitted to *PLoS Biology*.

ABSTRACT

Metatranscriptomic analyses of microbial assemblages from surface water at the Hawaiian Ocean Time-Series (HOT) revealed community-wide metabolic activities and diel patterns of differential gene expression. Pyrosequencing produced 75,946 mRNA reads from a night transcriptome and 75,558 from a day transcriptome. Taxonomic binning of annotated mRNAs indicated that Cyanobacteria contributed a greater percentage to the transcript pools (54% of annotated sequences) than expected based on abundance (35% of cell counts and 21% 16S rRNA of libraries), and may represent the most actively transcribing cells in surface seawater. Major heterotrophic taxa contributing to the community transcriptome included α-proteobacteria (19% of annotated sequences, most of which were SAR11-related) and γ-proteobacteria (4%). The composition of the transcriptome was consistent with models of prokaryotic gene expression, including operon-based transcription patterns and an abundance of genes predicted to be highly expressed. Metabolic activities that are shared by many microbial taxa (e.g., glycolysis, citric acid cycle, amino acid biosynthesis, transcription, and translation machinery) were well represented among the community transcripts. Ongoing nitrogen transformations in this surface ocean community included ammonium and urea metabolism, denitrification, and N_2 fixation. Phosphorus was assimilated in the forms of phosphate and phosphonate. There was an overabundance of transcripts for photosynthesis, C1 metabolism, and oxidative phosphorylation in the day compared to night, and evidence that energy acquisition is coordinated with solar radiation levels for both autotrophic and heterotrophic microbes. In contrast, housekeeping activities such as amino acid biosynthesis, membrane synthesis and repair, and vitamin biosynthesis were overrepresented in the night transcriptome compared to the day. Direct

sequencing of these transcript pools has provided the most detailed information to date on metabolic and biogeochemical responses of a microbial community to solar forcing.

INTRODUCTION

Oceanic subtropical gyres make up 40% of the Earth's surface and play critical roles in carbon fixation and nutrient cycling. The Hawaii Ocean Time-Series (HOT) in the North Pacific subtropical gyre was established to provide a long-term perspective on oceanographic properties of such systems [1] and has served as the focus of substantial research into the role of marine microorganisms in ocean biogeochemistry [2,3,4]. Recent metagenomic sampling efforts at Station Aloha have provided information about the genes in the bacterioplankton community and how they are distributed with depth [5]. Understanding patterns of expression of these microbial genes and what factors induce their expression is the next critical step in probing the workings of this oceanic ecosystem.

Gene expression in natural environments can be assessed with functional gene probes or primer sets applied to messenger RNA (mRNA) extracted directly from natural communities [6,7]. Alternatively, approaches involving microarrays offer the opportunity to examine the presence and expression of multiple genes simultaneously in complex microbial assemblages [8]. While these methods provide valuable information on gene expression in natural communities, they are limited to genes whose importance is recognized and for which sequence information is already available. Analogous to metagenomics, environmental transcriptomics (metatranscriptomics) retrieves and sequences environmental mRNAs from a microbial assemblage without prior knowledge of what genes the community might be expressing [9]. Thus it provides the most unbiased perspective on community gene expression *in situ*.

Environmental transcriptomics protocols are technically difficult since prokaryotic mRNAs generally lack the poly(A) tails that make isolation of eukaryotic messages relatively straightforward [10] and because of the relatively short half lives of mRNAs [11]. In addition, mRNAs are much less abundant than rRNAs in total RNA extracts, thus an rRNA background often overwhelms mRNA signals. However, techniques for overcoming some of the difficulties of manipulating environmental prokaryotic mRNA have been recently developed.

A procedure for analyzing environmental transcriptomes by creating clone libraries using random primers to reverse-transcribe and amplify environmental mRNAs was recently described [12]. This method was successful in two different natural environments, but results were biased by selection of the random primers used to initiate cDNA synthesis. Recent advances in linear amplification of mRNA [9] now obviate the need for random primers in the amplification step. Further, amplification makes it possible to use less starting material [13], decreasing the collection and processing time of samples and thereby minimizing RNA degradation. *In vitro* transcription methods for amplifying mRNA involve polyadenylating the mRNA and incorporating a T7 promoter onto the 3' end of the transcript. Amplified RNA (aRNA) can then be converted to double stranded cDNA [14] using random hexamers [15] and directly sequenced by pyrosequencing. A first use of this method at Station ALOHA demonstrated its utility for characterizing microbial community gene expression [9].

Here we describe the application of a similar environmental transcriptomics approach to elucidate day/night differences in gene expression, also focusing on surface waters of the North Pacific subtropical gyre [1]. Together with the first gene expression study at Station Aloha [9], this community transcriptome analysis provides information on the dominant metabolic processes within the bacterioplankton assemblages, and demonstrates differential diel gene

expression that directly reflects *in situ* biogeochemical processes involving carbon fixation, carbon metabolism, and nitrogen acquisition.

RESULTS

cDNA sequence annotation

The cDNAs prepared from amplified RNA ranged in size from 100 bp to 1 kb, with the majority between 200- 500 bp. The average picoliter reactor pyrosequencing read length was 99 bp. After removing duplicate sequences, the 240,422 pyrosequences (106,907 night and 133,515 day; Table 1) were passed through an annotation pipeline (Figure 1) using empirically derived criteria for gene predictions from ~100 bp pyrosequences [16]. The first step in the pipeline was the removal of 88,916 (31,402 night and 57,514 day) predicted rRNA sequences based on sequence similarity to the nt database using BlastN. Relatively low rRNA sequence contamination (29% in the night library and 43% in the day) compared to the rRNA content of prokaryotic cells (>80% [17]), indicated that the steps for excluding rRNAs through selective degradation and subtractive hybridization were largely successful.

Following rRNA removal, 151,504 possible protein-encoding sequences remained (75,946 night and 75,558 day). BlastX against RefSeq indicated that about one-third of these possible protein-encoding sequences (24,515 night and 24,133 day) had hits in RefSeq that met the established annotation criteria. The sequences with significant RefSeq hits were further classified based on similarities to COG categories and the KEGG pathways, with 24,474 and 35,927 sequences, respectively, meeting the annotation criteria for these databases.

The sequences without RefSeq hits (51,431 night, 51,425 day) may represent genes that are present in marine microbial communities but not captured into RefSeq. These sequences

were queried by BlastX against the ORFs predicted from unassembled Global Ocean Sampling (GOS) data (CAMERA; http://camera.calit2.net/index.php). An additional 26,366 sequences (13,222 night and 13,144 day) had hits to the GOS database (accounting for 17% of the 151,504 possible mRNA sequences and 26% of the 102,856 sequences not identified in RefSeq; Figure 1). Finally, any remaining sequences were checked for similarity to the non-redundant (nr) database, which is not curated and includes more sequences than RefSeq. This rescued a few more sequences, although typically these were unannotated hits to sequences from "uncultured bacteria". At the end of the annotation pipeline, half of the possible protein-encoding sequences in each library (~38,000) had no significant hits to previously sequenced genes. These may be transcripts from poorly conserved regions of known genes or from novel genes.

Because the GOS sequences are not annotated, the translations of the 100 GOS ORFs that were most often hit by HOT transcripts without matches in RefSeq were queried by BlastP against nr to determine putative functions. The majority of these often-hit GOS proteins had highest similarities to hypothetical proteins and was attributed to a variety of organisms from all domains of life, including *Aspergillus* sp., *Methanosarcina mazei*, α- and γ- proteobacteria, (e.g., *Pelagibacter ubique* and *Pseudomonas mendocina*) and Cyanobacteria (e.g., *Nostoc punctiforme*). As expected, however, these best hits had poor statistical confidence (typically E > 1) (Supplemental Table 1).

Taxonomic origin of transcripts

The annotated HOT community transcriptome was dominated by Cyanobacteria-like transcripts (54%; Figure 2) most similar to sequences from *Prochlorococcus marinus* AS9601, *P. marinus* MIT 9301, and *P. marinus* MIT 9312. The second largest taxonomic bin was the α-

proteobacteria (19%), dominated by sequences with similarity to SAR11 group members *P. ubique* HTCC1002 and *P. ubique* HTCC1062 (~10% of prokaryotic transcripts; Table 2). γ-proteobacteria and Bacteroidetes/Chlorobi accounted for 4% and 3% of the total sequences, respectively. The γ-proteobacterial-like sequences had highest similarities to marine gamma proteobacterium HTCC2080 and *Hahella chejuensis* KCTC 2396. Sequences linked to the Bacteroidetes/Chlorobi phylum were mostly attributed to *Psychroflexus torquis* ATCC 700755, the genome sequence of which unfortunately includes contaminant metagenomic sequence [18]; thus the Bacteroidetes signal might be misleading. Approximately 2% of the total transcripts were of eukaryotic origin. Cyanobacteria contributed equally to the day and night transcriptome (Figure 3). Within the Proteobacteria, α- and γ-proteobacterial transcripts comprised a similar percentage of the day transcript pool relative to night (40% of the heterotrophic transcripts in the day, 45% at night for α-proteobacteria; 11% of the heterotrophic transcripts in the day, 8% in the night for *γ-proteobacteria*; Figure 3).

Predicted highly expressed genes in transcript pools

We asked whether the composition of the HOT community transcript pools fits biological models for prokaryotic gene expression. Genes most frequently transcribed by a cell have distinct patterns in codon usage [19]. We identified these predicted highly expressed (PHX) genes for genomes of six of the largest taxonomic bins: three *Prochlorococcus* genome bins (*P. marinus* MIT9313, AS9601, MIT9301), two SAR11 genome bins (*P. ubique* HTCC1062 and 1002) and five Roseobacter genomes combined together into a single Roseobacter bin. Environmental transcripts in the six corresponding taxonomic bins were likewise categorized with regard to PHX status based on orthology to the parent genome. For all taxa, and in

accordance with biological expectations, the environmental bins had significantly more PHX genes than a null distribution generated from 1000 random samples of the reference genome (Figure 4). This pattern was particularly evident for the Roseobacters (9% of the genes in the reference genomes are PHX vs. 30% of the transcripts; 3.1-fold enrichment) and for *Prochlorococcus* MIT9301 (4.6 vs. 12.9%; 2.8-fold enrichment). A larger proportion of PHX transcripts were found in the day for all *P. marinus* strains and the Roseobacter bin, suggesting that the most highly expressed genes are more often for daytime-biased processes. This trend was not evident, however, for the *P. ubique*-like transcripts.

Operon signature in environmental transcript pools

Genes that encode steps in the same metabolic pathway are frequently clustered into operons in prokaryotic genomes [20] to facilitate coordinated transcription. Thus a cell's transcript pool is anticipated to include more mRNAs from adjacent genes than what is expected from a random sampling of the genome. We tested this for the environmental transcript bins by counting the frequency with which transcripts from two adjacent genes on the reference strain genome (defined as $\leq$1 gene intervening) were both present in the bin, recognizing that the wild and reference organisms will not be fully syntenic. In all cases, the transcript bins had significantly more adjacent genes represented than a null distribution generated from 1000 random samples drawn from the reference genomes (Figure 5). Thus the environmental transcriptomic protocol captured operon-based expression patterns in natural marine bacterioplankton communities.

Comparisons with community composition

Prochlorococcus dominate the Cyanobacteria at Station ALOHA (>95% of cells; [21]) and in this study accounted for approximately 2 x 10^5 cell ml^{-1} (based on flow cytometric counting; http://hahana.soest.hawaii.edu/hot/hot-dogs/), or ~30% of the total prokaryotes. Heterotrophic bacteria accounted for ~5 x 10^5 cell ml^{-1}, and comprised 65% of the prokaryotes present during the time of sampling. Companion PCR-based 16S rRNA clone libraries from DNA collected in tandem with the RNA samples were also generated. A predictable relationship of one rRNA operon per ~1500 genes found when we surveyed all complete marine bacterial genomes (Supplemental Figure 1) suggests than an unbiased 16S rRNA gene inventory should provide a reasonable index of relative contributions to the community gene pool by taxon. Taxonomically binned mRNA sequences were therefore compared to community composition data to ask whether or not taxa contribute to the HOT community mRNA in direct proportion to their representation in the microbial assemblage (i.e., whether some taxa are more transcriptionally active on a per cell basis than others).

Cyanobacteria representation in the transcript libraries (54%) was about 2-fold higher than in the 16S rRNA gene libraries (21%; Figure 3) or by direct counts (~35%; Figure 3), suggesting that genes more genes are expressed in these autotrophic bacterioplankton or that expressed genes are more frequently transcribed than in co-occurring heterotrophs. When relative 16S rRNA abundance was calculated among just the heterotrophic groups (i.e., with Cyanobacterial sequences removed), many taxa had similar contributions to the transcript pool and gene pool, suggesting that they have relatively similar levels of transcriptional activity on a per-gene basis (Figure 3). The α-proteobacteria appeared to be the most transcriptionally active of the heterotrophic groups, although analyzed separately, the SAR11 subgroup was slightly less

active relative to abundance. The Actinobacteria, Firmicutes, β-proteobacteria also were represented by relatively more 16S rRNA genes than by transcripts in the HOT community transcriptome (β-proteobacteria, 11% of 16S rRNA amplicons vs. 1% of transcripts; Actinobacteria, 5% vs. 0.9%; and Firmicutes, 17% vs. 6%).

Gene function and metabolic pathways

We annotated the bacterioplankton transcripts to determine which metabolic pathways were most often represented. Transcript abundance is presented here as relative abundance within the collective community transcriptome rather than per-gene expression levels (see [9]).

By far, the majority of annotated transcripts (~45% according to COG and ~80% according to KEGG) were assigned to genes related to metabolism, and in particular to three categories: amino acid transport and metabolism, energy production and conversion (particularly oxidative phosphorylation, carbon fixation, and nitrogen metabolism), and carbohydrate transport (Figure 7). Also well represented were transcripts related to genetic information and processing (~26% according to COG and ~10% according to KEGG), including genes for transcription and translation as well as chaperones and enzymes involved in posttranslational modification. The environmental information and processing KEGG pathway category, which includes membrane transport and signal transduction pathways, was also common in the community transcriptome and was dominated by membrane transport genes, specifically those encoding ABC transporters. Of the ABC transporters for which a putative function could be assigned, those for uptake of amino acids, glycine betaine/L-proline, polyamines (spermidine and putrescine), iron, and nutrients in the form of nitrate, phosphate, and phosphonate were most abundant (Table 3).

We determined relative transcript levels for genes in five taxonomic bins (three *P. marinus* and two *P. ubique*; the Roseobacter bins had too few sequences for this analysis) by mapping the transcripts to the reference genome. All three of the *P. marinus*-like strains had photosynthesis genes among those most highly represented, including PsaA, PsaB, and the light-harvesting complex and RuBisCo. An ammonium transporter, and transcription-related genes were also abundant in the *P. marinus* bins (Figure 8a-c), and some of the most highly-represented transcripts were for hypothetical proteins with unknown function. Well-represented transcripts in the *P. ubique* bins included those for proteorhodopsin, a Na+/solute symporter, RNA polymerase, an ammonium transporter, and colicin V production, among others (Figure 8d-e).

Transcript mapping to all five of the reference genomes showed several gaps in which few transcripts were found, and these were mostly from hypothetical or phage-like genes. Such gaps have been identified previously for metagenomic datasets referenced against genomes of cultured bacteria [22,23], and are thought to be hypervariable regions originating in part from phage-mediated lateral gene transfer [24].The gaps in the *P. marinus* strains overlapped most of the genomic islands recently characterized in two high-light ecotypes, MIT9312 and MED4 [22]. Similarly, there were gaps in the SAR11 strain transcript maps, the largest of which contains ~50 kb and is consistent with the size and location of the largest hypervariable region identified in the *P. ubique* genomes [23].

Metatranscriptomic comparison of night and day samples

The diel samples provide a basis for comparison of dominant expression patterns between day and night in this oceanic bacterioplankton community. Among the 1,577 COGs and the 167

KEGG pathways represented, statistical comparisons identified 12 COGs that were better represented at night and 13 that were better represented in the day (Table 4), and similarly 4 KEGG pathways that were better represented at night and 6 that were better represented in the day (95% confidence level; Table 5). For the nighttime-biased COGs, amino acid and nucleotide transport and metabolism, lipid, membrane and nucleotide biosynthesis, and deoxyhypusine synthase, an enzyme involved in dehydrogenation of polyamines, were significantly overrepresented. The night-biased KEGG pathways included those for glycospingolipid biosynthesis and nucleotide sugars metabolism. For the daytime-biased COGs, genes involved in energy production and conversion, posttranslational modification of newly synthesized proteins, protein turnover, and inorganic ion transport and metabolism were identified. As expected, catalases for the degradation of peroxide were also significantly overrepresented during the day and were among a variety of transcripts involved in protection from light-induced damage (Table 4). The KEGG pathways significantly overrepresented in the day included photosynthesis and oxidative phosphorylation.

Statistically significant differences in the distribution of transcripts between the day and night samples was also assessed independently of COG and KEGG assignments in order to capture signals from genes not currently classified by these annotation systems. When all major categories of transcripts with synonymous annotations were collapsed and compared between day and night, there were 13 annotation categories that were better represented at night and 29 that were better represented in the day (95% confidence level; Supplemental Table 2). Among the significant categories for the night transcriptome were those for ABC-type spermidine/putrescine transport system permeases, RNA methyltransferases, glutathione reductases, enzymes involved in amino acid and membrane biosynthesis and signal transduction

histidine kinases (Supplemental Table 2). Proteorhodopsins, enzymes involved in photosynthesis and chlorophyll biosynthesis, cytochromes, and ammonium transporters were among those categories significantly overrepresented during the day.

Reconstruction of dominant KEGG pathways for three major taxa

The KEGG pathways exhibiting diel patterns in three major taxa, (*P. marinus*, *P. ubique*, and Roseobacters, Supplemental Table 3) were manually curated both to check that the automated annotation was reasonable and to determine whether any complete or nearly-complete metabolic pathways were represented. Pathway reconstructions are presented here only for the sample in which they were significantly overrepresented, although the other transcriptome may have had representatives of some steps.

Histidine biosynthesis was an overrepresented nighttime activity for the *P. marinus*- and *P. ubique*-like organisms. For both taxa, we found expression of nearly all genes in the pathway (10 of 10 for *P. marinus*, 9 of 10 for *P. ubique*; Figure 9) through to the synthesis tRNA(his).

A large number of sequences in the *P. ubique* and Roseobacter day transcriptomes mapped to a KEGG system for transfer of methyl groups, the one carbon (C1) pool by folate pathway. This pathway serves as a mediator of C1 interconversions, suggesting that methyl transfer leading to energy generation or biosynthesis is a particularly important heterotrophic process during the day in the surface ocean.

Several KEGG pathways showed differential day/night expression for *P. marinus* only. Metabolism of glutathione, a reductant with multiple detoxifying and cytoprotective capabilities, was overrepresented in the *P. marinus* night transcriptome. Transcripts for genes involved in the γ-glutamyl cycle were found in particularly high abundance in the night transcriptome and were

not captured at all in the day, indicating preferential transport of amino acids, including cysteine, by *P. marinus* at night. In contrast, *P. marinus*-like organisms invested heavily in the biosynthesis of steroids during the daytime, as all genes involved in the synthesis of phytoene, a precursor of carotenoids, were expressed (Figure 10). As expected, transcripts for the *P. marinus* photosynthesis pathway, including those related to the phycobilisome, photosystem I and II, cytochromes, and the ATP synthase, were preferentially captured in the day transcriptome.

KEGG pathways showing differential day/night expression only in the case of *P. ubique* included nucleotide sugars metabolism, glycosphingolipid biosynthesis, vitamin B6 metabolism, and carotenoid (β- carotene) biosynthesis, all in the night transcript pool. The Roseobacter bin provided evidence of two pathways that were significantly overrepresented in the night: the TCA cycle and protein export.

Eukaryotic sequences

The majority of eukaryotic transcripts were most closely affiliated with sequences from green-lineage organisms (Viridiplantae) such as the picoeukaryotic prasinophytes *Ostreococcus* spp. [25] and *Micromonas* spp. A relatively large number of transcripts also appear to be most closely related to genes known to be encoded in Chromalveoltae (Stramenopile or Alevolate) genomes. Chromist algae are also known to be major components of the picoeukaryotic phytoplankton (< 5 μm) [26]. In general gene transcripts that most closely matched photosynthetic eukaryotic reference sequences, particularly those known to be involved in photosynthesis, were more abundant in the day compared to night sample. Among the most highly expressed genes detected in eukaryotic organisms include those encoding for chlorophphyll binding proteins, light harvesting reactions, and photosynthetic machinery (Fig

11). Highly expressed genes related to photosynthesis include those that most closely match the photosystem II D1 reaction-center protein from the diatom *Thalassiosira psuedonana* as well as the plastid encoded photosystem I subunit protein, psaB, from the diatom *Odontella sinensis*. Evidence for stramenopile nitrogen metabolism via urea cycle activity was also detected based on several transcripts that most closely matched stramenopile carbamoyl phosphate synthetase (CPS) III. This represents the first evidence suggesting that the unique diatom urea cycle [27,28] is likely active in natural populations of stramenopile picophytoplankton.

qPCR quality control

The half-life of prokaryotic transcripts can be as short as 30 seconds based on studies of mRNAs of cultured bacteria [11], while processing times for environmental nucleic acid samples can take hours [29]. Linear amplification of RNA greatly reduces the time between initiation of sampling and capture of transcripts because sample volumes can be reduced, but it has potential to introduce bias into the sequenced mRNA pool. A previous test with mRNA from the cultured marine bacterium *Silicibacter pomeroyi* DSS-3 demonstrated minor bias and good repeatability during linear amplification [30]. Here, we assessed the full environmental transcriptomic sequencing protocol by comparing qPCR-based ratios of selected genes in day vs. night total RNA fractions to the pyrosequencing-based ratio of these same genes in the sequenced transcript pools.

Five genes common in the transcriptome were selected for qPCR analysis: *recA* and *psaA* from *P. marinus* AS9601, a proteorhodopsin and a Na+/solute symporter (Ssf family) gene from *P. ubique* HTCC1062, and a probable integral membrane proteinase attributed to *P. torquis* ATCC 700755. Results showed a strong positive correlation between night and day ratios in the

original RNA pool and the pyrosequence datasets (Figure 12, r =0.94) indicating that the sequenced metatranscriptome was representative of the original RNA pool.

DISCUSSION

The Hawaii Ocean Time-series program provides comprehensive, long-term oceanographic information for the oligotrophic North Pacific Ocean [1]. The dissolved organic constituents 25 m at Station ALOHA are typically 70-110 µM for carbon, 5-6 µM for nitrogen, and 0.2-0.3 µM for phosphorus (http://hahana.soest.hawaii.edu/hot/hot-dogs/). Ammonium concentrations in these waters are below the detection limit of standard nutrient analysis. Averaging surface water nutrient data over the past several decades for the month of November shows no discernable differences in organic and inorganic carbon, nitrogen and phosphorus concentrations at Station ALOHA on a diel basis.

Along with previous metagenomic analysis of the genetic potential of this system [5] and a recent environmental transcriptomic study [9], this diel environmental transcriptomics effort provides insight into the temporal patterns of bacterioplankton activity. While environmental transcript pools can provide unprecedented access to ongoing metabolic processes and ecological activities, two important caveats are that: (1) their composition may be shaped by responses to collection and filtration manipulations, and (2) mRNAs with intrinsically shorter half lives are more likely to degrade before they can be stabilized and sequenced. Nonetheless, the community transcriptome had properties consistent with expected attributes of the HOT ecosystem. The taxonomic affiliations of transcripts agreed with known bacterioplankton community composition; closely related *Prochlorococcus* strains that are members of high light clade eMIT9312 comprised the most populated transcript bin. This clade has been shown to dominate

in the upper euphotic zone (>50 m) at low and mid-latitudes (below 30°) [31], much like the HOT stations from which our samples were collected. SAR11 comprised the second largest taxonomic bin. This taxon is the most numerous heterotrophic marine bacterioplankton group, particularly in oligotrophic oceans where it makes up 30-40% percent of cells in the euphotic zone [32].

We address the issue of transcriptome coverage using three approaches (Supplemental Table 4). First, by using the 16S rRNA clone library data to establish a taxon-abundance model for the system at a similarity level of 99%, and based on the assumptions that each taxon produces 1000 transcripts at any given time [17] and that genome coverage follows a Lander-Waterman model [33], we estimate that the most abundant taxon in the day or night sample had over 90% transcriptome coverage, while the 15 most abundant taxa had more than half their transcriptome covered. Second, the *P. marinus* and *P. ubique* strains (Figure 8) were examined to determine how many genes in the reference genome were represented by a transcript. Sequences with homology to approximately half of the genes in these organisms' genomes were found in the transcriptome. Assuming that bacterioplankton cells produce 1000 transcripts at any given time, we estimate that the three *P. marinus* strains have over 95% coverage of their transcriptome, while the two *P. ubique* strains have over 50% coverage. Finally, we determined the singletons and doubletons in the COG categories and applied the Chao1 index of diversity to determine the theoretical abundance of COGs in the day and night. Based on this calculation, the sequencing effort captured about 80% of the COGs predicted to be present in the night transcriptome and 70% of the COGs predicted for the day transcriptome.

Estimates of taxonomic composition of assemblages in the ocean consistently show the numerical importance of α- and γ-proteobacteria and Cyanobacteria [5,32,34], but little is known

about how abundance specifically relates to activity levels. Based on comparisons of the relative abundance of taxa (flow cytometry counts and 16S rRNA amplicons) to their representation in the community transcriptome, by far the highest per-cell transcriptional activity level in the HOT ecosystem was seen for the Cyanobacteria. Assuming similar mRNA half-lives across the prokaryotic taxa, these dominant autotrophs produced more transcripts per gene than any co-occurring heterotrophic group in both the day and night (Figure 3). This may reflect an advantage of autotrophy over heterotrophy for maintaining cellular activity levels since the marine organic carbon fueling heterotrophic activity in the ocean is in low concentration and largely refractory [35]. Among heterotrophic bacteria, the SAR11 group had low apparent per-cell transcription levels relative to other α-proteobacteria; this is consistent with the small cell size and relatively few ribosomes found for SAR11 cells in seawater [36]. In general, comparisons of transcript representation with 16S rRNA gene amplicon pools suggest the only large discrepancy in transcript levels vs. abundance in this bacterioplankton community is between autotrophic and heterotrophic cells.

Bacterioplankton in the euphotic zone synthesized nearly four times as much mRNA for the same volume of seawater in the day compared to night. The finding is consistent with evidence in the transcriptome for preferential expression of RNA polymerase in the day for the *P. marinus-, P. ubique-*, and Roseobacter-like cells (Table 5). At night, bacterial community investment was skewed toward biosynthesis (specifically of membranes, amino acids, and vitamins), while during the day, energy acquisition and metabolism received greater investments. As expected, many transcripts involved in light-mediated processes, such as photosynthesis and proteorhodopsin activity as well as the synthesis of structures required for these activities, were among those contributing differentially to the community transcriptome in the day. Daytime C1

utilization by some heterotrophs suggests a source of C1 compounds or methyl groups. Dimethylsulfoniopropionate (DMSP), an organic sulfur compound produced in abundance by marine phytoplankton [37], is a rich source of methyl groups for surface ocean bacterioplankton, and tetrahydrofolate-mediated C1 transfer (i.e., the C1 pool by folate and methane metabolism pathways; Supplemental Table 4) has been shown to play a role in its metabolism [18]. C1 compounds such as methanol and formaldehyde [38,39,40], methane [41], and methylhalides [42,43] may also be available to heterotrophic bacterioplankton in surface seawater.

Synthesis of vitamin B6, essential for a variety of amino acid conversions including transaminations, decarboxylations, and dehydrations, in conjunction with evidence for the γ-glutamyl pathway for amino acid uptake, the overrepresentation of amino acid transport and metabolism COGs, and the histidine synthesis pathway (Table 4; Supplemental Table 3), indicate that amino acid acquisition in general may be a relatively more important metabolic activity in the nighttime. *P. marinus* has recently been shown to exhibit diel patterns of amino acid uptake, with acquisition occurring predominantly at dusk [44]. Our data agree with this and suggest that heterotrophic taxa likewise preferentially transport and synthesize amino acids at night. Nighttime accumulation of amino acids might be a mechanism for nitrogen storage by these organisms, particularly for *P. marinus* which undergoes cell division at night. The emphasis on histidine synthesis by both autotrophs and heterotrophs might indicate that histidine-rich proteins are synthesized preferentially at night. Alternatively, histidine is one of the most nitrogen-rich amino acids (only arginine has more amino groups) and thus may act as a nitrogen storage compound; histidine has also been shown to have antioxidant properties [45] that may be beneficial during the day.

Microbial processes expected to be differentially expressed over a diel cycle, such as photosynthesis, oxidative phosphorylation, and synthesis of light-driven cellular machinery such as proteorhodopsins and photosynthetic pigments, were captured in this metatranscriptomic analysis of the oligotrophic ocean. Other less anticipated processes that emerged from the comparative diel analyses included uptake and utilization of multiple nitrogen, carbon, sulfur and phosphorus compounds (Table 6), many of which are relevant to major biogeochemical cycles. The utility of metatranscriptomics for providing an inventory of ongoing ecologically-relevant processes as well as insights into their temporal patterns is clear. Elucidating such patterns will directly facilitate predictive modeling of environmental controls on oceanic processes.

METHODS

Sample collection

Samples were collected at the Hawaiian Ocean Time-series (HOT) stations ALOHA (22°45'N, 158°W) and WHOTS (22°46.1'N, 157°53.4'W) in November, 2005 (HOT-175). For RNA extraction, seawater was collected from a depth of 25 m using Niskin bottles on a conductivity-temperature-depth (CTD) rosette sampler. A night sample was collected at 03:00 on November 11, 2005, and a daylight sample was collected at 13:00 on November 13, 2005. During HOT-175, the peak PAR level was at 12:00, with sunrise occurring around 07:00 and sunset just before 18:00. 80L (night) and 40L (day) of seawater were pre-filtered through a 5 μm, 142 mm polycarbonate filter (GE Osmonics, Minnetonka, MN) followed by a 0.2 μm, 142 mm Durapore (Millipore) filter using positive air pressure. The 0.2 μm filters were placed in a 15 ml tube containing 2 ml Buffer RLT (containing β-mercaptoethanol) from the RNeasy kit (Qiagen, Valencia, CA) and flash-frozen in liquid nitrogen for RNA extraction. For DNA extraction, an

additional 20 L of seawater were simultaneously filtered using the protocol outlined above at both time points. The 0.2 μm filters were placed in Whirlpack bags and flash frozen. The total sampling time from initiation of collection until freezing in liquid nitrogen was approximately 1.5 hours. We obtained ~1 μg of total RNA from 40- 80 L of seawater. Following mRNA enrichment and amplification, 30-100 μg of mRNA was available for conversion to cDNA for sequencing. Typically, only 3-5 μg of DNA was required for pyrosequencing.

RNA and DNA preparation

DNA was extracted using a phenol:chloroform-based protocol [29]. Briefly, frozen filters inside Whirlpak bags were transferred to 50 ml Falcon centrifuge tubes. 10 ml extraction buffer [SDS (10% Sodium Doecyl Sulphate): STE (100 mM NaCl, 10mM Tris, 1 mM EDTA), 9:1] was added to the tubes and boiled in a water bath for 5 min. The extraction buffer was then removed from the tubes, placed into Oak Ridge round-bottom centrifuge tubes, to which 3 ml NaOAc and 28 ml 100% EtOH were added. Organic macromolecules were precipitated overnight at -20^{o}C, before the tubes were centrifuged for 1 h at 15,000 x *g*. The supernatant was decanted, and pellets dried for 30 min in the air. The pellets were resuspended in 600 μL deionized water, and sequentially extracted with 500 μl phenol, 500 μl phenol:chloroform:isoamyl alcohol (24:1:0.1), and 500 μl chloroform:isoamyl alcohol (9:1); after each extraction the organic phase was removed and discarded. The supernatant was removed into a fresh tube at the end of last extraction, amended with 150 μl NaOAc and 1.2 ml 100% EtOH, and precipitated overnight. The tube contents were then centrifuged at 15,000 x *g* for 1 h, the supernatant decanted, and pellets dried in a speed vacuum dryer for 10 minutes. The DNA pellets were resuspended in 100 μl DNAse and RNAse-free deionized water (Ambion).

RNA was extracted using a modified version of the RNeasy kit (Qiagen) that results in high RNA yields from material on polycarbonate filters [46]. Frozen samples were first thawed slightly for 2 min in a 40-50°C water bath and then vortexed for 10 min with RNase-free beads from the Mo-Bio RNA PowerSoil kit (Carlsbad, CA). Following centrifugation for 5 min at 3,000- 5,000 x *g*, the supernatant was transferred to a new tube. Beginning with the RNeasy Midi kit, one volume of 70% ethanol was added to the lysate and, in order to sheer large molecular weight nucleic acids, the lysate was drawn up through a 22-gauge needle several (~5) times. RNA extraction then continued with the RNeasy Mini kit according to the manufacturer's instructions.

Following extraction, RNA was treated with DNase using the TURBO DNA-*free* kit (Ambion, Austin, TX). Two methods were employed to rid the RNA samples of as much rRNA as possible. The RNA was first treated enzymatically with the mRNA-ONLY Prokaryotic mRNA Isolation Kit (Epicentre Biotechnologies, Madison, WI) that uses a 5'-phosphate-dependant exonuclease to degrade rRNAs. The MICROB*Express* kit (Ambion) subtractive hybridization with capture oligonucleotides hybridized to magnetic beads was subsequently used as an additional mRNA enrichment step.

In order to obtain µg quantities of mRNA, approximately 500 ng of RNA was linearly amplified using the MessageAmp II-Bacteria Kit (Ambion) according to the manufacturer's instructions. Finally, the amplified, antisense RNA (aRNA) was converted to double-stranded cDNA with random hexamers using the Universal RiboClone cDNA Synthesis System (Promega, Madison, WI). The cDNA was purified with the Wizard DNA Clean-up System (Promega). The quality and quantity of the total RNA, mRNA, aRNA, and cDNA was assessed

by measurement on the NanoDrop-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE) and the Experion Automated Electrophoresis System (Bio-Rad, Hercules, CA).

cDNA sequencing and quality control

6 μg of cDNA from each sample (night and day) were sequenced using pyrosequencing technology by 454 Life Sciences (Branford, CT) [47] resulting in 10,682,120 bp from 106,907 reads for the night sample and 13,255,704 bp from 133,515 reads for the day sample. The average sequence length was 99 bp. The sequences have been deposited with the accession numbers XXX-XXX.

rRNA identification and removal

Previously established criteria based on an *in silico* analysis were used for BLAST-based gene predictions of the ~100 bp pyrosequences [16]. The sequences were clustered at an identity threshold of 98% based on a local alignment (number of identical residues divided by length of alignment) using the program Cd-hit [48]. Ribosomal RNA sequences were identified by BlastN queries of the reference sequence of each cluster against the non-curated, GenBank nucleotide database (nt) [49] using cutoff criteria of E value $\leq 10^{-3}$, amino acid length ≥ 23 and percent identity $\geq 40\%$ as determined by the *in silico* tests. rRNA identification is complicated by misidentified sequences in the RefSeq protein database (that is, rRNA sequences that are incorrectly annotated as putative proteins) that sometimes give better hits to actual rRNA sequences. We therefore inferred that a sequence was rRNA-derived if any of the top three BlastN hits were to an rRNA regardless of results of the subsequent analysis against the protein database.

cDNA sequence annotation

The criteria for protein predictions generated using BlastX against the NCBI curated, non-redundant reference sequence database (RefSeq)[50] were also established with *in silico* tests and set as E-value < 0.01, similarity > 40%, and overlapping length > 65 bp to the corresponding best hit. Sequences with hits to RefSeq were assigned functional protein or pathway predictions based on the Clusters of Orthologous Groups (COG) database [51] or the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [52]. The cutoff criteria for functional protein prediction based on orthologous groups using BlastX analysis against the COG database were established as E value < 0.1, similarity > 40%, and overlapping length > 65 bp to the corresponding best hit. The COG cutoff criteria were also applied to the KEGG database for pathway prediction. Taxonomic binning of the sequences was carried out using MEGAN, a program that assigns likely taxonomic origin to a sequence based on the NCBI taxonomy of closest Blast hits, with the default settings for all parameters [53]. The taxonomic affiliations of the putative mRNA sequences predicted using MEGAN to the family level and the top Blast hit for lower level assignments. All non-rRNA sequences that had no RefSeq hits were BlastX-queried against the nr database as well as BlastX-queried against CAMERA un-assembled ORFs predicted from the Global Ocean Survey (GOS) reads [54].

Eukaryotic sequence annotation

Eukarytoic transcripts were initially binned by MEGAN. Subsequently the sequencing reads were used to search (BlastX) an in-house curated database proein sequences derived from all available complete eukaryotic organelle and nuclear genomes. As of this study, the database

contains 46 eukaryotic genomes. Transcript reads that matched a reference protein sequence with 60% identity and an E value < e-10 were retained and the reference protein was used for functional annotation. Functional annotation was performed using a java-based tool called Blast2go [55] that annotates genes based on similarity searches with statistical analysis and highlighted visualization on directed acyclic graphs.

Predicted highly expressed genes

Predicted highly expressed (PHX) genes were determined for cultured representatives of three prokaryotic taxa that were well represented in the transcript libraries (*Prochlorococcus*, Roseobacter, and SAR11) using an algorithm developed by Karlin and Mrázek [19]. The algorithm is based on comparisons to codon usage patterns in genes expected to be frequently transcribed in a prokaryotic genome (list here; ribosomal proteins, chaperone proteins, etc.). Environmental transcript sequences that had best BLAST hits to one of the PHX genes were similarly designated as PHX.

Statistical analysis

A statistical program designed for comparing gene frequency in metagenomic datasets [56] was used to compare the night and day mRNA sequences categorized based on COGs, KEGGs, and taxonomy. The program was run with 20,000 repeated samplings with a sample size of 10,000 for COGs, 9,000 for KEGGs, and 25,000 for taxonomic bins. The significance level (p) was set at < 0.05.

16S rRNA libraries

PCR amplification of ribosomal DNA was carried out using primers 27F and 1522R [57]. The PCR conditions were as follows: 3 min at 96°C, followed by 30 cycles of denaturation at 95°C for 50 s, annealing at 58°C for 50 s, primer extension at 72°C for 1 min, and a final extension at 72°C for 10 min. PCR products were cleaned using the QIAquick PCR Purification Kit (Qiagen) and cloned into pCR2.1 vector using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA). 192 clones from each sample were sequenced at the University of Georgia Sequencing Facility on an ABI 3100 (Applied Biosystems, Foster City, CA).

qPCR verifications

To confirm that the composition of the pyrosequence library was representative of the initial mRNAs, transcripts of five genes that were top hits to multiple sequences within the two transcript pools were quantified in the total RNA pool for each sample. Quantitative PCR (qPCR) primer sets were designed for the *Prochlorococcus marinus* str. AS9601 *RecA* and *PsaA*, a proteorhodopsin gene and a Na+/solute symporter (Ssf family) gene from *Pelagibacter ubique* HTCC1062, and a probable integral membrane proteinase attributed to *Psychroflexus torquis* ATCC 700755 (Sequences and annealing temps in Supplemental Table 5). Reverse transcription reactions were carried out on 200 ng of RNA using the Omniscript RT kit (Qiagen) in 20 µl volumes containing 1X RT buffer, 0.3 µg/µl of random hexamers (Invitrogen), 1 µl of 5 mM dNTPs, 2 U of reverse transcriptase, and 20 U of RNase inhibitor (Promega,) at 37°C for 1 h, followed by inactivation of the reverse transcriptase at 95°C for 2 min. The day: night ratio of each gene transcript in the RNA pools was determined by qPCR amplification of a serial dilution of cDNAs in triplicate, and calculation of the difference in cycle threshold values (ΔC_T) between

the two samples. Quantitative amplification was done using the iCycler iQ real-time PCR detection system (Bio-Rad) in a 20 µl reaction volume containing 10 µl of iQ SYBR Green Supermix (Bio-Rad), 0.4 µl each of 10 µM of the forward and reverse primers, and 1 µl of the cDNA template. PCR conditions included a preliminary denaturation at 95°C for 3 min followed by 45 cycles of 95°C for 15 s, annealing for 1.5 s, 95°C for 1 min, and 55°C for 1 min. A melt curve was generated following the PCR, beginning with 55°C and increasing 0.4°C every 10 s until 95°C. A PCR control without an initial RT step was included with every set of reactions.

ACKNOWLEDGEMENTS

We thank the Captain and crew of the R/V Kilo Moana and Dr. David Karl. Funding was provided by The Gordon and Betty Moore Foundation Marine Investigator grants (M.A.M. and J.P.Z.), the National Science Foundation MCB-0702125 (M.A.M.) and OCE-0425363 (J.P.Z.), and the NSF C-MORE Center for Microbial Oceanography.

REFERENCES

1. Karl DM, Lukas R (1996) The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. Deep Sea Research Part II: Topical Studies in Oceanography 43: 129-156.
2. Cavender-Bares KK, Karl DM, Chisholm SW (2001) Nutrient gradients in the western North Atlantic Ocean: Relationship to microbial community structure and comparison to patterns in the Pacific Ocean. Deep Sea Research Part I: Oceanographic Research Papers 48: 2373-2395.

3. Karl D, Letelier R, Tupas L, Dore J, Christian J, et al. (1997) The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. Nature 388: 533-538.

4. Zehr JP, Waterbury JB, Turner PJ, Montoya JP, Omoregie E, et al. (2001) Unicellular cyanobacteria fix N2 in the subtropical North Pacific Ocean. Nature 412: 635-638.

5. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science 311: 496-503.

6. Bürgmann H, Widmer F, Sigler WV, Zeyer J (2003) mRNA extraction and reverse transcription-PCR protocol for detection of *nifH* gene expression by *Azotobacter vinelandii* in soil. Applied and Environmental Microbiology 69: 1928-1935.

7. Wawrik B, Paul JH, Tabita FR (2002) Real-time PCR quantification of rbcL (ribulose-1,5-bisphosphate carboxylase/oxygenase) mRNA in diatoms and pelagophytes. Applied and Environmental Microbiology 68: 3771-3779.

8. Zhou JH (2003) Microarrays for bacterial detection and microbial community analysis. Current Opinion in Microbiology 6: 288-294.

9. Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, et al. (2008) Microbial community gene expression in ocean surface waters. Proceedings of the National Academy of Sciences of the United States of America 105: 3805-3810.

10. Liang P, Pardee AB (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 257: 967-971.

11. Belasco JG (1993) mRNA degradation in prokaryotic cells: an overview. In: Belasco JG, Brawerman G, editors. Control of messenger RNA stability. San Diego: Academic Press. pp. 3-11.

12. Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, et al. (2005) Analysis of microbial gene transcripts in environmental samples. Applied and Environmental Microbiology 71: 4121-4126.

13. Gelder RNV, von Zastrow ME, Yool A, Dement WC, Barchas JD, et al. (1990) Amplified RNA Synthesized from Limited Quantities of Heterogeneous cDNA. Proceedings of the National Academy of Sciences of the United States of America 87: 1663-1667.

14. Gubler U, Hoffman BJ (1983) A simple and very efficient method for generating cDNA libraries. Gene 25: 263-269.

15. Binns MM, Boursnell MEG, Foulds IJ, Brown TDK (1985) The use of a random priming procedure to generate cDNA libraries of infectious bronchitis virus, a large RNA virus. Journal of Virological Methods 11: 265-269.

16. Mou X, Sun S, Edwards RA, Hodson RE, Moran MA (2008) Bacterial carbon processing by generalist species in the coastal ocean. Nature 451: 708-711.

17. Ingraham JL, Maaløe O, Neidhardt FC (1983) Growth of the bacterial cell. Sunderland, Mass.: Sinauer Associates. 435 p.

18. Howard EC, Henriksen JR, Buchan A, Reisch CR, Burgmann H, et al. (2006) Bacterial Taxa That Limit Sulfur Flux from the Ocean. Science 314: 649-652.

19. Karlin S, Mrázek J (2000) Predicted Highly Expressed Genes of Diverse Prokaryotic Genomes. Journal of Bacteriology 182: 5238-5250.

20. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N (1999) The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America 96: 2896-2901.

21. Campbell L, Vaulot D (1993) Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawaii (Station ALOHA). Deep-Sea Research 40: 2043-2060.

22. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic Islands and the Ecology and Evolution of *Prochlorococcus*. Science 311: 1768-1770.

23. Wilhelm L, Tripp HJ, Givan S, Smith D, Giovannoni S (2007) Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. Biology Direct 2: 27.

24. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* Cyanophage Genomes: Signature Features and Ecological Interpretations. PLoS Biology 3: e144.

25. Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, et al. (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proceedings of the National Academy of Sciences of the United States of America 103: 11647-11652.

26. McDonald SM, Sarno D, Scanlan DJ, Zingone A (2007) Genetic diversity of eukaryotic ultraphytoplankton in the Gulf of Naples during an annual cycle. Aquatic Microbial Ecology 50: 75-89.

27. Allen AE, Vardi A, Bowler C (2006) An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. Current Opinion in Plant Biology 9: 264-273.

28. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, et al. (2004) The genome of the diatom Thalassiosira pseudonana: Ecology, evolution, and metabolism. Science 306: 79-86.

29. Fuhrman JA, Comeau DE, Hagstrom A, Chan AM (1988) Extraction from Natural Planktonic Microorganisms of DNA Suitable for Molecular Biological Studies. Applied and Environmental Microbiology 54: 1426-1429.

30. Bürgmann H, Howard EC, Ye W, Sun F, Sun S, et al. (2007) Transcriptional response of *Silicibacter pomeroyi* DSS-3 to dimethylsulfoniopropionate (DMSP). Environmental Microbiology 9: 2742-2755.

31. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, et al. (2006) Niche Partitioning Among *Prochlorococcus* Ecotypes Along Ocean-Scale Environmental Gradients. Science 311: 1737-1740.

32. Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, et al. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420: 806-810.

33. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics 2: 231-239.

34. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology 5: e77.

35. Bauer JE, Williams PM, Druffel ERM (1992) 14C activity of dissolved organic carbon fractions in the north-central Pacific and Sargasso Sea. Nature 357: 667-670.

36. Rappe MS, Connon SA, Vergin KL, Giovannoni SJ (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature 418: 630-633.

37. Kiene RP, Linn LJ, Bruton JA (2000) New and important roles for DMSP in marine microbial communities. Journal of Sea Research 43: 209-224.

38. Carpenter LJ, Lewis AC, Hopkins JR, Read KA, Longley ID, et al. (2004) Uptake of methanol to the North Atlantic Ocean surface. Global Biogeochemical Cycles 18: GB4027.

39. Giovannoni SJ, Hayakawa DH, Tripp HJ, Stingl U, Givan SA, et al. (2008) The small genome of an abundant coastal ocean methylotroph. Environmental Microbiology 10: 1771-1782.

40. Heikes BG, Chang WN, Pilson MEQ, Swift E, Singh HB, et al. (2002) Atmospheric methanol budget and ocean implication. Global Biogeochemical Cycles 16: 80.1-80.13

41. Ward BB, Kilpatrick KA, Novelli PC, Scranton MI (1987) Methane Oxidation and Methane Fluxes in the Ocean Surface-Layer and Deep Anoxic Waters. Nature 327: 226-229.

42. Woodall CA, Warner KL, Oremland RS, Murrell JC, McDonald IR (2001) Identification of methyl halide-utilizing genes in the methyl bromide-utilizing bacterial strain IMB-1 suggests a high degree of conservation of methyl halide-specific genes in gram-negative bacteria. Applied and Environmental Microbiology 67: 1959-1963.

43. Schaefer JK, Goodwin KD, McDonald IR, Murrell JC, Oremland RS (2002) *Leisingera methylohatidivorans* gen. nov., sp nov., a marine methylotroph that grows on methyl bromide. International Journal of Systematic and Evolutionary Microbiology 52: 851-859.

44. Mary I, Garczarek L, Tarran GA, Kolowrat C, Terry MJ, et al. (2008) Diel rhythmicity in amino acid uptake by Prochlorococcus. Environmental Microbiology: Online early.

45. Dahl TA, Midden WR, Hartman PE (1988) Some Prevalent Biomolecules as Defenses against Singlet Oxygen Damage. Photochemistry and Photobiology 47: 357-362.

46. Poretsky RS, N. Bano, A. Buchan, M. A. Moran, and J. T. Hollibaugh (2008) Environmental transcriptomics: a method to access expressed genes in complex microbial communities. In: Kowalchuk GA, de Bruijn FJ, Head IM, Akkermans ADL, van Elsas JD, editors. Molecular Microbial Ecology Manual. 3rd ed. Netherlands: Springer. pp. 1892-1904.

47. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

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

49. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2007) GenBank. Nucleic Acids Research 35: D21-25.

50. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33: D501-504.

51. Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28: 33-36.

52. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27-30.

53. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Research 17: 377-386.

54. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A community resource for metagenomics. PLoS Biology 5: 394-397.

55. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.

56. Rodriguez-Brito B, Rohwer F, Edwards R (2006) An application of statistics to comparative metagenomics. BMC Bioinformatics 7: 162.

57. Johnson JL (1994) Similarity analysis of rRNAs. In: Gerhardt P, Murray RGE, Wood WA, Krieg NR, editors. Methods for general and molecular bacteriology. Washington, D.C.: American Society for Microbiology. pp. 683–700.

Table 4.1. Annotation Pipeline results for night and day transcriptomes.

	NIGHT	% NIGHT	DAY	% DAY
Total Reads	106,907	100	133,515	100
Ribosomal RNA Reads	31,402	29	57,514	43
Possible protein-encoding sequences	75,946	71	75,558	57
RefSeq Identified	24,515	23	24,133	18
GOS (non-RefSeq) Identified	13,222	12	13,144	10
nr Identified	92	0	71	0
Unidentified in RefSeq, GOS, and nr	38,117	36	38,210	29
KEGG-assigned Reads	19,273	18	16,654	12
COG-assigned Reads	12,487	12	11,987	9

Table 4.2. . Putative taxonomy of the organisms contributing the most transcripts to the community transcriptome as determined by top BlastX hit to RefSeq.
∗ indicates genome sequences with known contamination.

	Night	Day
Prochlorococcus marinus str. MIT 9301	6309	6292
Prochlorococcus marinus str. AS9601	3214	2849
Pelagibacter ubique HTCC1002	2541	1851
Prochlorococcus marinus str. MIT 9312	1430	1264
Pelagibacter ubique HTCC1062	1308	944
∗ *Psychroflexus torquis* ATCC 700755	889	715
Prochlorococcus marinus str. MIT 9515	609	758
Nostoc punctiforme PCC 73102	308	499
Prochlorococcus marinus subsp. pastoris str. CCMP1986	283	267
Clostridium novyi NT	255	345
Synechococcus sp. RCC307	210	221
Aurantimonas sp. SI85-9A1	154	223
Emiliania huxleyi	93	129
Parvibaculum lavamentivorans DS-1	91	96
∗ Alpha proteobacterium HTCC2255	88	94
Hahella chejuensis KCTC 2396	88	139
Synechococcus sp. WH 7803	84	112
Cyanophage P-SSM2	78	94
Marine gamma proteobacterium HTCC2143	75	74
Listeria monocytogenes str. 4b H7858	74	136
Clostridium difficile QCD-32g58	71	111
Crocosphaera watsonii WH 8501	70	69
Rickettsia typhi str. Wilmington	69	126
Stappia aggregata IAM 12614	67	75
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	63	91
Rhodospirillum rubrum ATCC 11170	62	66
Marine gamma proteobacterium HTCC2207	61	44
Frankia sp. EAN1pec	59	90
Rhodobacterales bacterium HTCC2150	58	42
Marine gamma proteobacterium HTCC2080	54	67
Prochlorococcus marinus str. NATL1A	54	68
Gramella forsetii KT0803	53	48
Methylobacterium sp. 4-46	52	68
Unidentified eubacterium SCB49	51	36
Microscilla marina ATCC 23134	49	62
Dinoroseobacter shibae DFL 12	48	34
Flavobacteriales bacterium HTCC2170	47	42
Cyanophage P-SSM4	46	56
Roseobacter sp. SK209-2-6	46	39
Magnetospirillum magneticum AMB-1	45	47
Robiginitalea biformata HTCC2501	44	42
Vibrio cholerae V52	42	69
Jannaschia sp. CCS1	41	27
Roseobacter sp. CCS2	41	35
Sinorhizobium medicae WSM419	41	27
Bradyrhizobium sp. ORS278	40	34
Roseovarius sp. TM1035	40	42
Silicibacter pomeroyi DSS-3	39	30

Table 4.3. KEGG category distribution of genes in the night and day transcriptomes.

KEGG	% NIGHT	% DAY
Metabolism	84.78	83.25
Amino Acid Metabolism	22.80	20.81
Carbohydrate Metabolism	17.87	17.08
Energy Metabolism	13.66	17.25
Metabolism of Cofactors and Vitamins	8.61	8.66
Nucleotide Metabolism	6.33	6.31
Lipid Metabolism	4.97	4.53
Metabolism of Other Amino Acids	3.22	2.97
Xenobiotics Biodegradation and Metabolism	2.46	2.11
Glycan Biosynthesis and Metabolism	2.29	1.48
Biosynthesis of Secondary Metabolites	2.00	1.68
Biosynthesis of Polyketides and Nonribosomal Peptides	0.57	0.37
Genetic Information Processing	9.18	10.07
Translation	5.35	5.58
Transcription	1.90	2.74
Folding, Sorting and Degradation	1.11	1.00
Replication and Repair	0.81	0.75
Environmental Information Processing	3.15	2.85
Membrane Transport	2.10	1.92
Signal Transduction	1.05	0.92
Cellular Processes	1.40	1.48
Endocrine System	1.05	1.09
Cell Motility	0.17	0.17
Immune System	0.13	0.19
Cell Communication	0.04	0.01
Cell Growth and Death	0.01	0.02
Human Diseases	0.27	0.35
Metabolic Disorders	0.16	0.16
Cancers	0.08	0.13
Infectious Diseases	0.02	0.03
Neurodegenerative Disorders	0.02	0.04
Other	1.22	2.00

Table 4.4. COGs significantly overrepresented in the night (blue shading) and day (yellow shading) transcriptomes ($p < 0.05$).

COG ID	COG	Category
COG0404	Glycine cleavage system T protein (aminomethyltransferase)	Amino acid transport and metabolism
COG0498	Threonine synthase	Amino acid transport and metabolism
COG0445	NAD/FAD-utilizing enzyme apparently involved in cell division	Cell division and chromosome partitioning
COG1208	Nucleoside-diphosphate-sugar pyrophosphorylases involved in lipopolysaccharide biosynthesis/translation initiation factor eIF2B subunits	Cell envelope biogenesis, outer membrane
COG0513	Superfamily II DNA and RNA helicases	DNA replication, recombination and repair
COG0416	Fatty acid/phospholipid biosynthesis enzyme	Lipid metabolism
COG0104	Adenylosuccinate synthase	Nucleotide transport and metabolism
COG0151	Phosphoribosylamine-glycine ligase	Nucleotide transport and metabolism
COG0519	GMP synthase - PP-ATPase domain	Nucleotide transport and metabolism
COG1186	Protein chain release factor B	Translation, ribosomal structure and biogenesis
COG1899	Deoxyhypusine synthase	Translation, ribosomal structure and biogenesis
COG0644	Dehydrogenases (flavoproteins)	Energy production and conversion
COG0723	Rieske Fe-S protein	Energy production and conversion
COG0843	Heme/copper-type cytochrome/quinol oxidases, subunit 1	Energy production and conversion
COG1005	NADH:ubiquinone oxidoreductase subunit 1 (chain H)	Energy production and conversion
COG1290	Cytochrome b subunit of the bc complex	Energy production and conversion
COG1845	Heme/copper-type cytochrome/quinol oxidase, subunit 3	Energy production and conversion
COG0670	Integral membrane protein, interacts with FtsH	General function prediction only
COG4147	Predicted symporter	General function prediction only
COG0004	Ammonia permeases	Inorganic ion transport and metabolism
COG0376	Catalase (peroxidase I)	Inorganic ion transport and metabolism
COG3239	Fatty acid desaturase	Lipid metabolism
COG0443	Molecular chaperone	Posttranslational modification, protein turnover, chaperones
COG0459	Chaperonin GroEL (HSP60 family)	Posttranslational modification, protein turnover, chaperones
COG0542	ATPases with chaperone activity, ATP-binding subunit	Posttranslational modification, protein turnover, chaperones

Table 4.5. KEGG pathways significantly overrepresented in the night (blue shading) and day (yellow shading) transcriptomes ($p < 0.05$).

Pathway ID	Pathway	Category
path00520	Nucleotide sugars metabolism	Carbohydrate Metabolism
path00521	Streptomycin biosynthesis	Biosynthesis of Secondary Metabolites
path00602	Glycosphingolipid biosynthesis - neo-lactoseries	Glycan Biosynthesis and Metabolism
path00603	Glycosphingolipid biosynthesis - globoseries	Glycan Biosynthesis and Metabolism
path00190	Oxidative phosphorylation	Energy Metabolism
path00195	Photosynthesis	Energy Metabolism
path03010	Ribosome	Translation
path03020	RNA polymerase	Transcription
path04940	Chaperonin	#N/A
path05060	Chaperonin	#N/A

Table 4.6. Select biogeochemically-relevant genes and their occurances in the night or day (+). An asterisk indicates that the gene was significantly overrepresented.

			Night	Day
Nitrogen	Nitrogenase (N fixation)	nifH, nifU, nifS, nifB	+	+
	Ammonium transport	amt	+	+ *
	Ammonia monooxygenase	amoA		
	Assimilatory nitrate reductase	narB	+	
	Hydroxylamine oxidoreductase	hao		
	Nitrate permease	napA	+	
	Nitrite reductase	nirA	+	
	Dissimilatory nitrite reductase	nirK, nirS		
	Nitric oxide reductase	norQ	+	
	Nitrate transporter	narK	+	
	Urease	ureC, ureE, ureF	+	+
Methylotrophy	Serine-glyoxylate aminotransferase		+	+
	Formate dehydrogenase	fdh, fdsD	+	+
	Methylene tetrahydrofolate reductase	metF	+	+
	Methane monooxygenase	mmo		
	Methanol dehydrogenase	mxa		+
	Methenyltetrahydromethanopterin cyclohydrolase	mch	+	+
	Crotonyl-CoA reductase		+	+
	Formaldehyde-activating enzyme	fae		+
Polyamine degradation	Deoxyhypusine synthase	dys2	+ *	+
	Spermidine/putrescine transport system permease	potC	+ *	+
	Acetylpolyamine aminohydrolase	aphA		
Sulfur cycle	Sulfur oxidation	soxB, soxC, soxA, soxZ, soxF	+	+
	Dimethyl sulfoniopropionate demethylase	dmdA		
Glycine betaine	Dimethylglycine dehydrogenase	dmgdh	+	+
	Glycine cleavage system (amnomethyltransferase)	gcvT	+ *	+
Aromatic Compounds	Aromatic ring hydroxylase	chlP	+	+ *
	protocatechuate 3,4-dioxygenase	pcaH		
	Benzoyl-CoA oxygenase	boxA		+
carbon monoxide	Carbon monoxide dehydrogenase	cosS, coxM, coxL	+	+

Phototrophy and C fixation	Photosystem I	multiple	+	+ *
	Photosystem II	multiple	+	+ *
	Rubisco	rbcL, rbcS	+	+ *
	Photosynthetic reaction center, M subunit	pufM		+
	Proteorhodopsin		+	+ *
Phosphate assimilation	Phosphonate uptake	phnD, phnC	+	+
	Alkaline phosphatase	phoA	+	+
	Phosphate uptake	pstA, pstS	+	+
Amino acid metabolism	Glutamate synthase	gltB	+	+
	Glutathione reductase	gor	+ *	+
	Histidine kinase	baeS	+ *	+
	Threonine synthase	thrC	+ *	+
Trace metal uptake	Selenium		+ *	+
	Iron	tonB	+	+
	Arsenite		+	
	Arsenate reductase	arsC	+	+

Figure 4.1. The mRNA annotation pipeline developed for 454 transcript reads showing combined counts for the day and night transcriptomes. All percentages are relative to the total number of sequences.

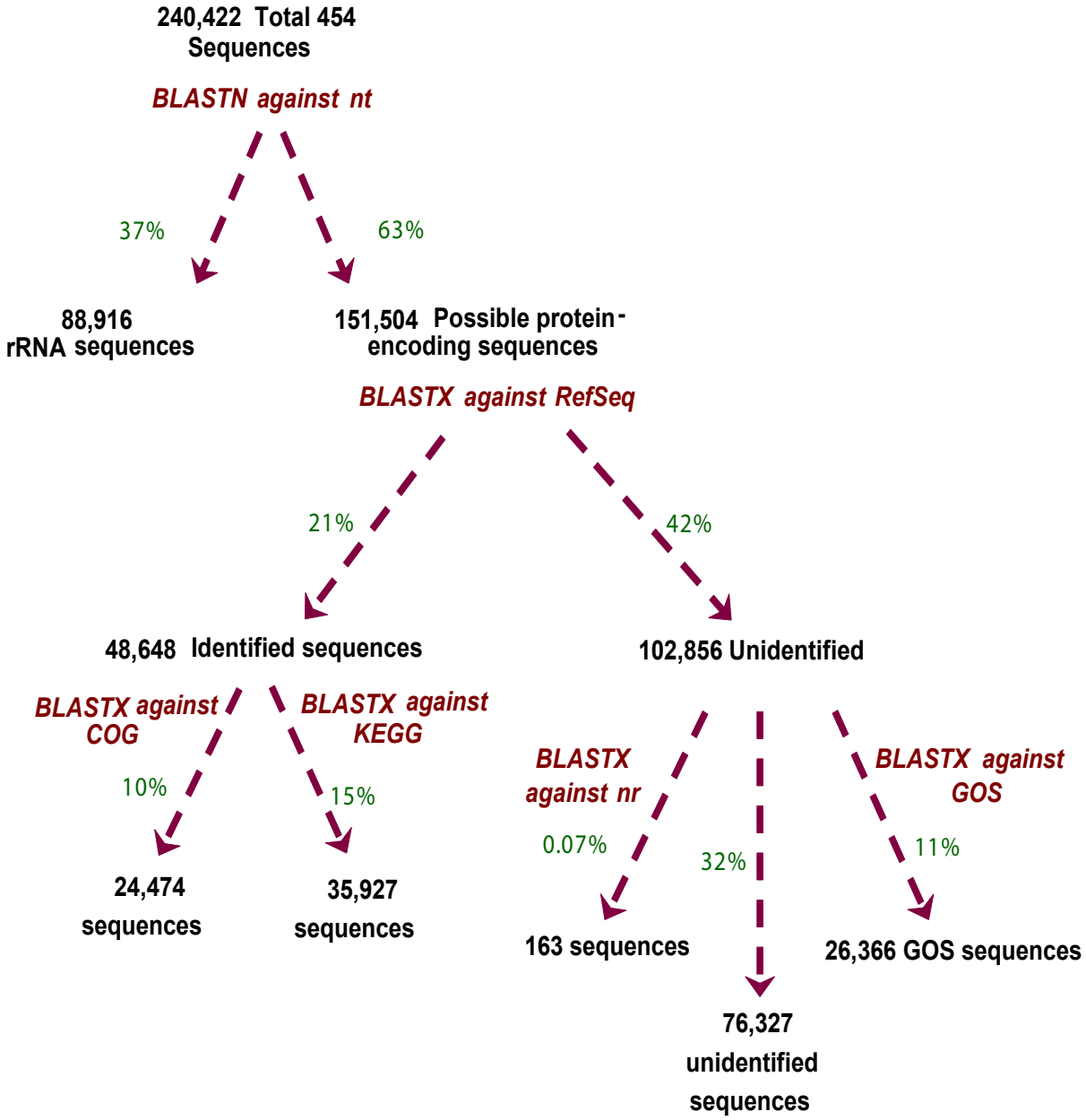

240,422 Total 454 Sequences
BLASTN against nt
37%
63%
88,916 rRNA sequences
151,504 Possible protein-encoding sequences
BLASTX against RefSeq
21%
42%
48,648 Identified sequences
102,856 Unidentified
BLASTX against COG
BLASTX against KEGG
10%
15%
24,474 sequences
35,927 sequences
BLASTX against nr
BLASTX against GOS
0.07%
32%
11%
163 sequences
26,366 GOS sequences
76,327 unidentified sequences

Figure 4.2. MEGAN-assigned taxonomic affiliations for day (light bars) and night (dark bars) transcripts at the phylum/domain and class/order levels, based on NCBI taxonomy, as a percentage of MEGAN-identified sequences in each library.

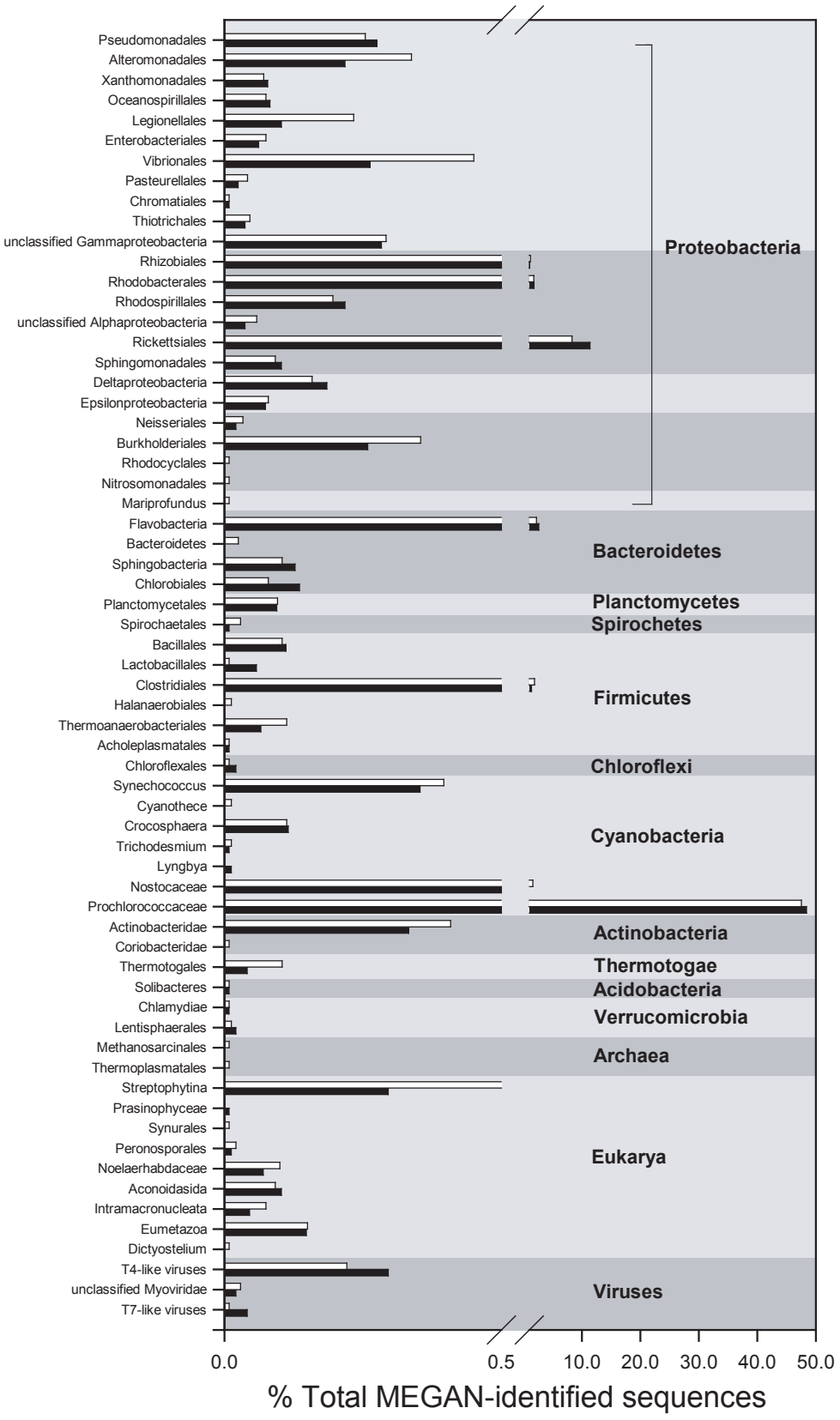

Pseudomonadales
Alteromonadales
Xanthomonadales
Oceanospirillales
Legionellales
Enterobacteriales
Vibrionales
Pasteurellales
Chromatiales
Thiotrichales
unclassified Gammaproteobacteria
Rhizobiales
Rhodobacterales
Rhodospirillales
unclassified Alphaproteobacteria
Rickettsiales
Sphingomonadales
Deltaproteobacteria
Epsilonproteobacteria
Neisseriales
Burkholderiales
Rhodocyclales
Nitrosomonadales
Mariprofundus
Flavobacteria
Bacteroidetes
Sphingobacteria
Chlorobiales
Planctomycetales
Spirochaetales
Bacillales
Lactobacillales
Clostridiales
Halanaerobiales
Thermoanaerobacteriales
Acholeplasmatales
Chloroflexales
Synechococcus
Cyanothece
Crocosphaera
Trichodesmium
Lyngbya
Nostocaceae
Prochlorococcaceae
Actinobacteridae
Coriobacteridae
Thermotogales
Solibacteres
Chlamydiae
Lentisphaerales
Methanosarcinales
Thermoplasmatales
Streptophytina
Prasinophyceae
Synurales
Peronosporales
Noelaerhabdaceae
Aconoidasida
Intramacronucleata
Eumetazoa
Dictyostelium
T4-like viruses
unclassified Myoviridae
T7-like viruses
Proteobacteria
Bacteroidetes
Planctomycetes
Spirochetes
Firmicutes
Chloroflexi
Cyanobacteria
Actinobacteria
Thermotogae
Acidobacteria
Verrucomicrobia
Archaea
Eukarya
Viruses
0.0
0.5
10.0
20.0
30.0
40.0
50.0
% Total MEGAN-identified sequences

Figure 4.3. Contribution of taxa to the 16S rRNA amplicon pool compared to the transcript pool in the day (open symbols) and night (filled symbols) communities. Cyanobacterial counts (triangles) are displayed as a percentage of total sequences, while the heterotrophic bacterial counts (circles) are displayed as a percentage of heterotrophic sequences only. Cyanobacterial transcript contribution abundance as determined by flow cytometry is indicated (star). α= *α*-proteobacteria; β= *β*-proteobacteria; γ= *γ*-proteobacteria; AP= All Proteobacteria; F= Firmicutes, Cyano= Cyanobacteria; A= Actinobacteria; C= Clostridia; S= SAR11; B= Bacteroidetes/Chlorobi. The line shows a 1:1 relationship.

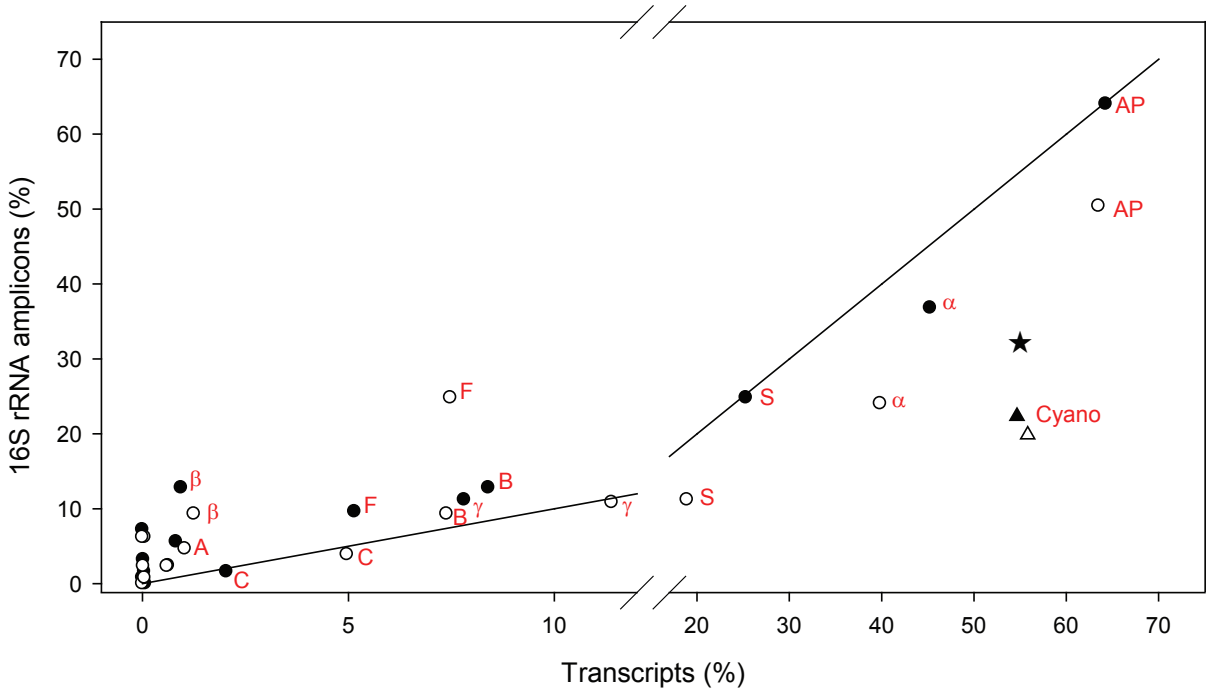

16S rRNA amplicons (%)
Transcripts (%)
0
5
10
20
30
40
50
60
70
AP
AP
α
α
Cyano
S
S
F
F
B
B
γ
γ
β
β
A
C
C

Figure 4.4. Comparison of the predicted highly expressed (PHX) transcripts in a taxonomic bin (dark bars), relative to the reference genome (transparent bars). All six had significantly more PHX in the transcript bins than expected by chance alone ($p < 0.05$).

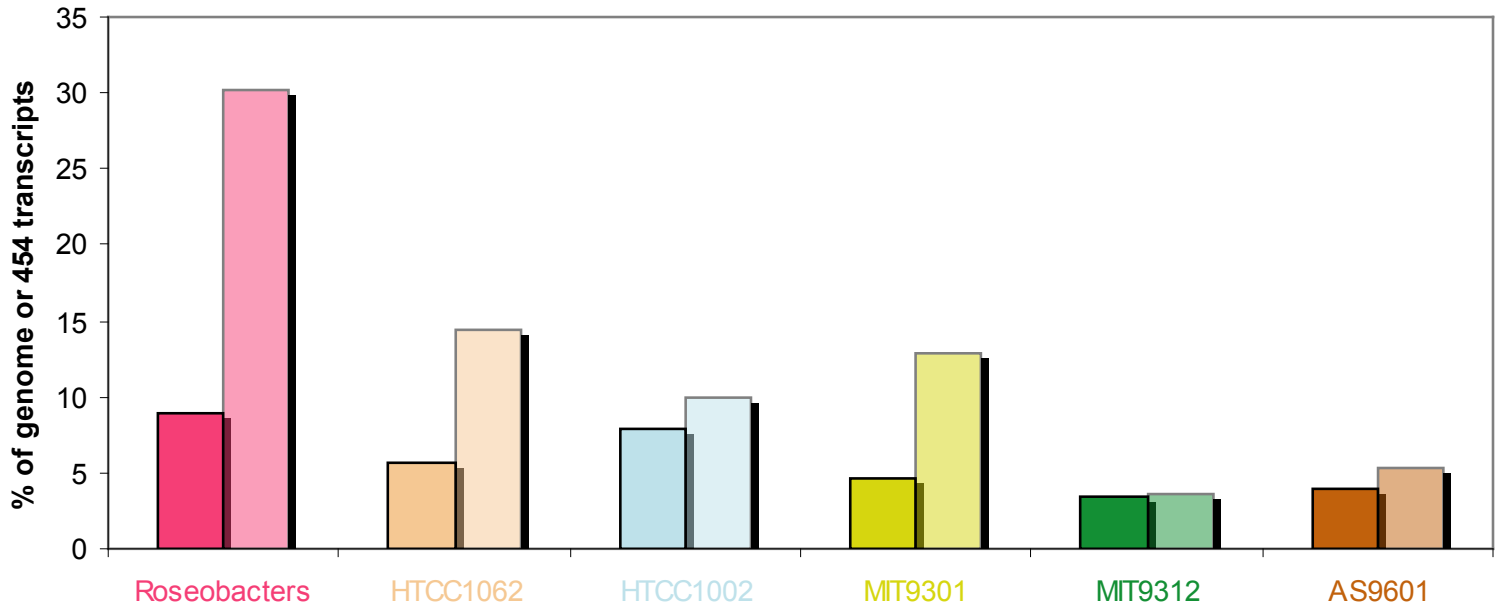

% of genome or 454 transcripts
35
30
25
20
15
10
5
0
Roseobacters
HTCC1062
HTCC1002
MIT9301
MIT9312
AS9601

Figure 4.5. The frequency of transcript sequences in each taxonomic bin that occurs with an adjacent gene on the reference genome as determined by *in silico* random sampling of reference genomes for the same number of genes found within the transcript sequences. All six had significantly more neighboring genes expressed in the transcript bins (triangles) than expected by chance alone ($p<0.05$).

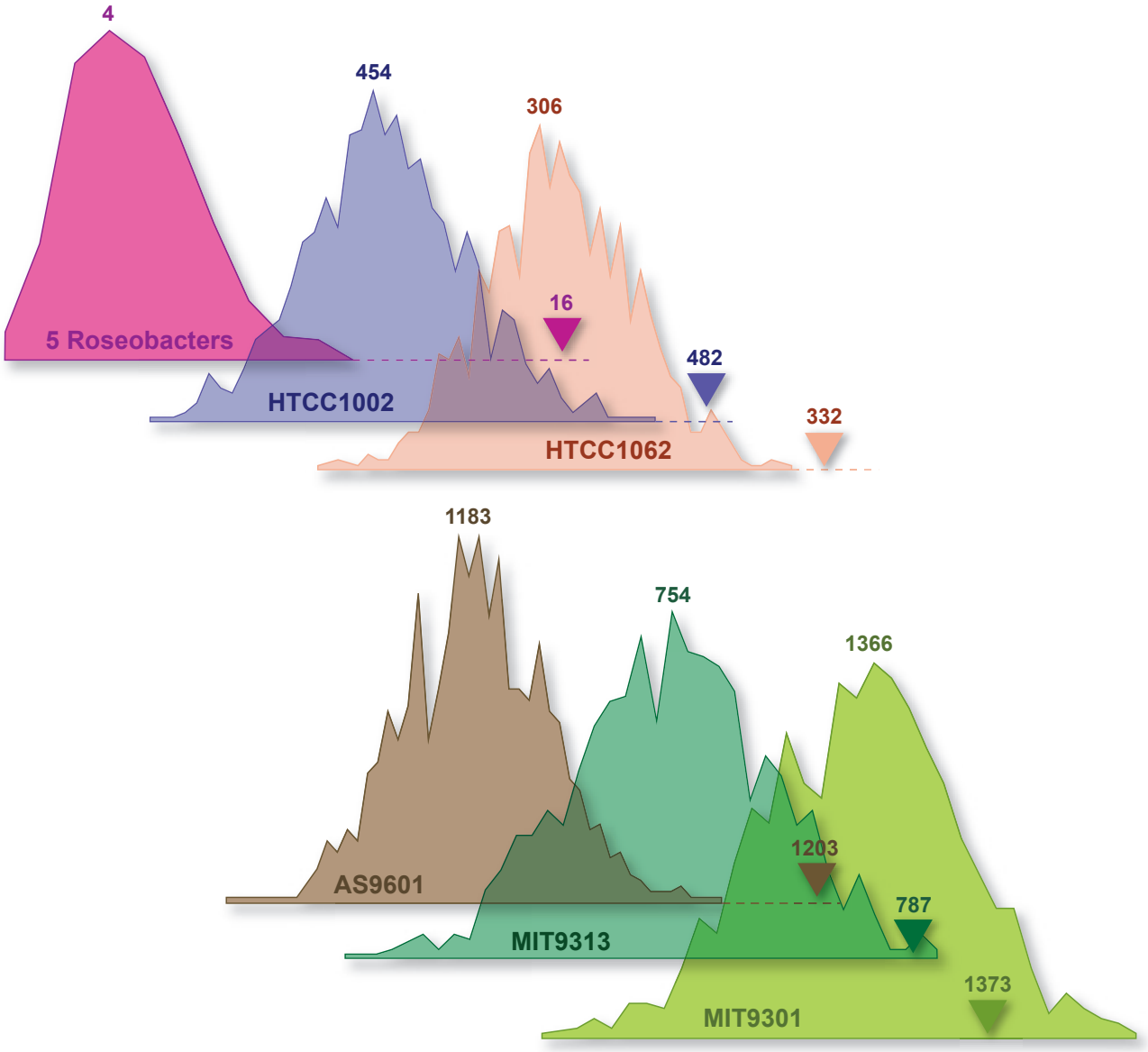

4
454
306
16
482
332
5 Roseobacters
HTCC1002
HTCC1062
1183
754
1366
1203
787
1373
AS9601
MIT9313
MIT9301

Figure 4.6. Depth profiles of *Prochlorococcus*-like, *Synechococcus*-like, and pigmented nanoeukaryotes as determined by flow cytometry. Heterotrophic bacteria counts are obtained using SYBR I and subtracting the Prochlorococcus concentration (obtained from the autofluorescing sample) from the SYBR I stained concentration. (http://hahana.soest.hawaii.edu/hot/hot-dogs/) at Station ALOHA. The horizontal line delineates the mixed layer depth.

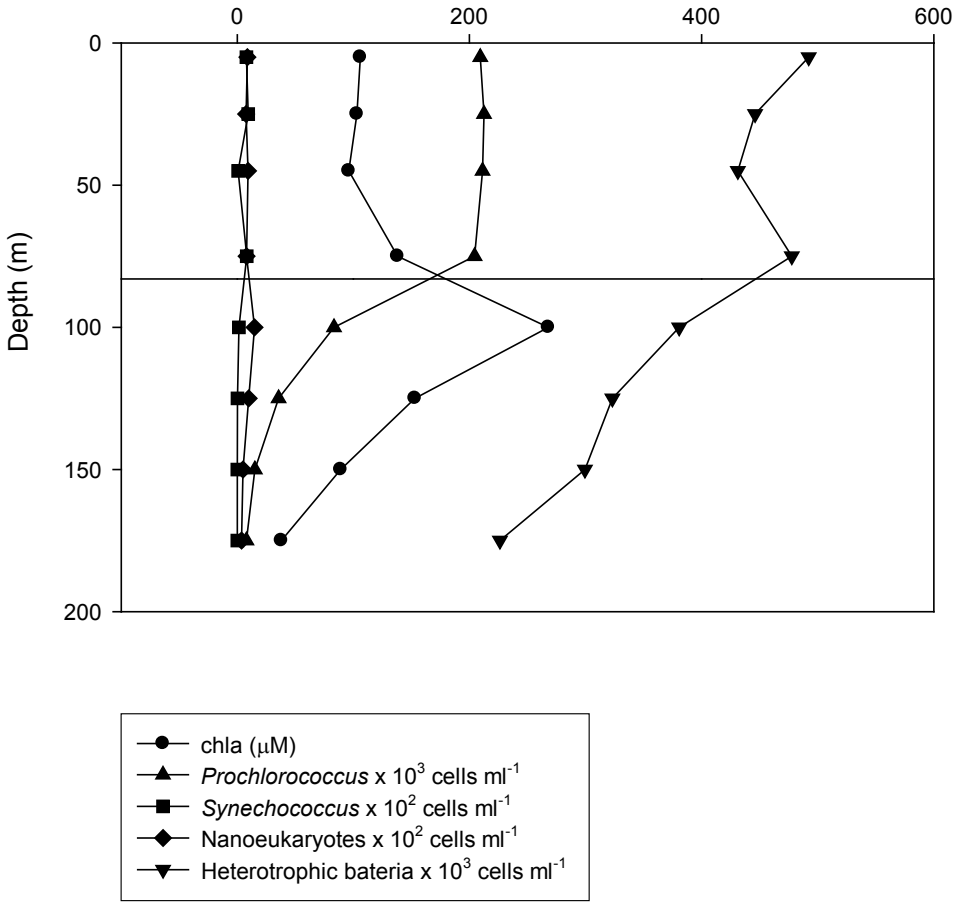

0
200
400
600
0
50
100
150
200
Depth (m)
chla (μM)
Prochlorococcus x 10^3 cells ml^-1
Synechococcus x 10^2 cells ml^-1
Nanoeukaryotes x 10^2 cells ml^-1
Heterotrophic bateria x 10^3 cells ml^-1

Figure 4.7. The 50 most abundant KEGG pathways in the day (open) and night (filled) transcriptomes. The pathways marked with a star were significantly overexpressed in one of the transcript pools.

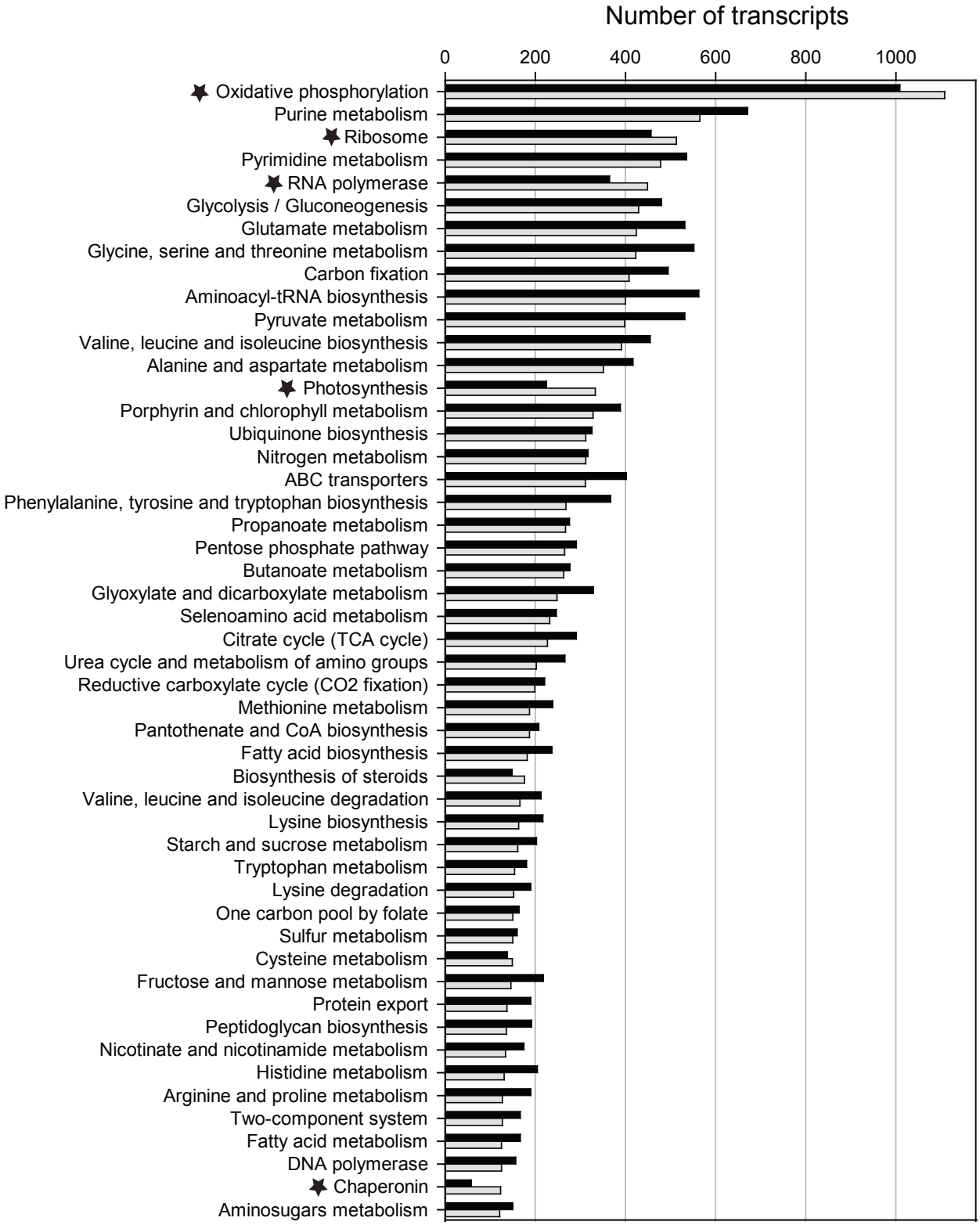
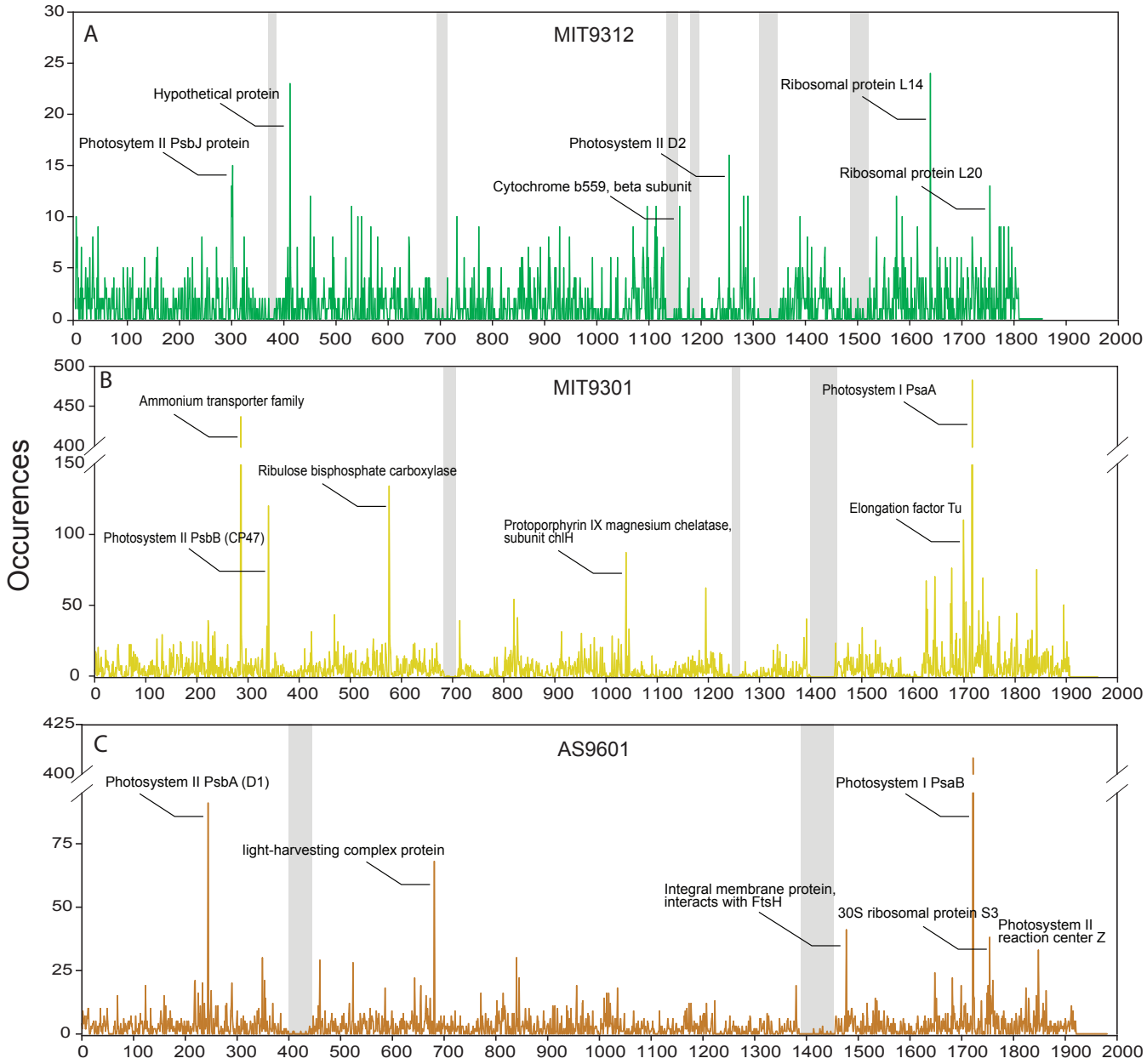

Number of transcripts
0
200
400
600
800
1000
Oxidative phosphorylation
Purine metabolism
Ribosome
Pyrimidine metabolism
RNA polymerase
Glycolysis / Gluconeogenesis
Glutamate metabolism
Glycine, serine and threonine metabolism
Carbon fixation
Aminoacyl-tRNA biosynthesis
Pyruvate metabolism
Valine, leucine and isoleucine biosynthesis
Alanine and aspartate metabolism
Photosynthesis
Porphyrin and chlorophyll metabolism
Ubiquinone biosynthesis
Nitrogen metabolism
ABC transporters
Phenylalanine, tyrosine and tryptophan biosynthesis
Propanoate metabolism
Pentose phosphate pathway
Butanoate metabolism
Glyoxylate and dicarboxylate metabolism
Selenoamino acid metabolism
Citrate cycle (TCA cycle)
Urea cycle and metabolism of amino groups
Reductive carboxylate cycle (CO2 fixation)
Methionine metabolism
Pantothenate and CoA biosynthesis
Fatty acid biosynthesis
Biosynthesis of steroids
Valine, leucine and isoleucine degradation
Lysine biosynthesis
Starch and sucrose metabolism
Tryptophan metabolism
Lysine degradation
One carbon pool by folate
Sulfur metabolism
Cysteine metabolism
Fructose and mannose metabolism
Protein export
Peptidoglycan biosynthesis
Nicotinate and nicotinamide metabolism
Histidine metabolism
Arginine and proline metabolism
Two-component system
Fatty acid metabolism
DNA polymerase
Chaperonin
Aminosugars metabolism

Figure 4.8. Mapping of transcripts to five reference genomes. A-C are *P. marinus* strains; D-E are *P. ubique* strains. Shaded areas represent regions of few mapped transcripts and may represent hypervariable regions. The reference genomes within each of these two species have within-taxon 16S rRNA sequence similarities of >99%.


A
MIT9312
Hypothetical protein
Photosytem II PsbJ protein
Photosystem II D2
Cytochrome b559, beta subunit
Ribosomal protein L14
Ribosomal protein L20
B
MIT9301
Ammonium transporter family
Ribulose bisphosphate carboxylase
Photosystem II PsbB (CP47)
Protoporphyrin IX magnesium chelatase, subunit chlH
Photosystem I PsaA
Elongation factor Tu
C
AS9601
Photosystem II PsbA (D1)
light-harvesting complex protein
Integral membrane protein, interacts with FtsH
30S ribosomal protein S3
Photosystem I PsaB
Photosystem II reaction center Z
Occurences

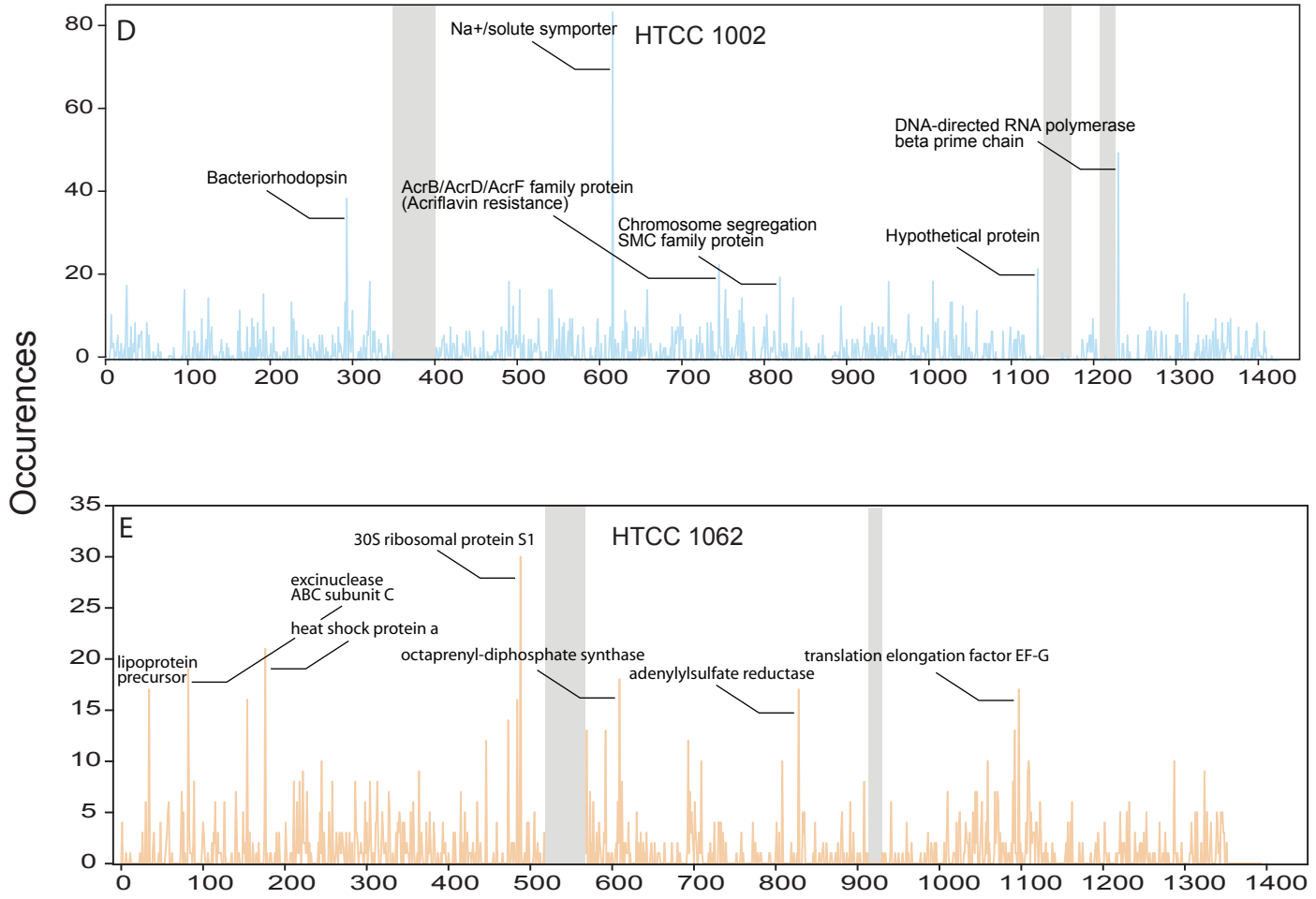
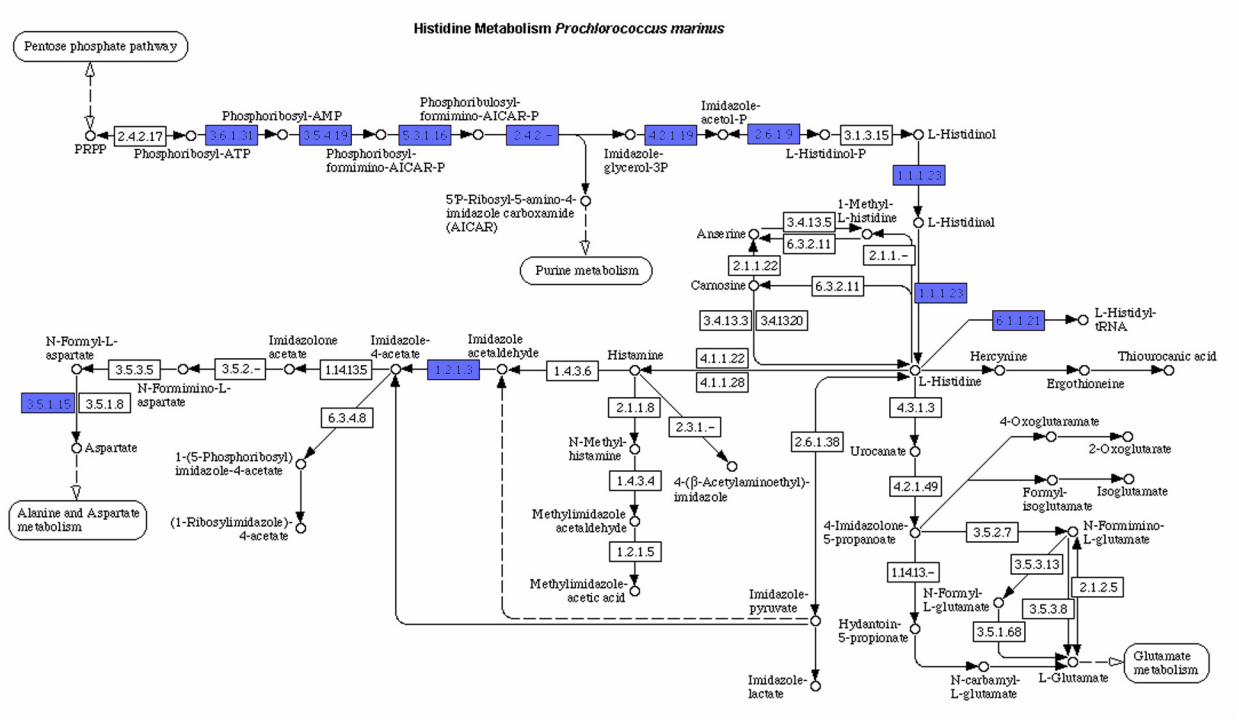

D
HTCC 1002
Na+/solute symporter
DNA-directed RNA polymerase
beta prime chain
Bacteriorhodopsin
AcrB/AcrD/AcrF family protein
(Acriflavin resistance)
Chromosome segregation
SMC family protein
Hypothetical protein
E
HTCC 1062
30S ribosomal protein S1
excinuclease
ABC subunit C
heat shock protein a
lipoprotein
precursor
octaprenyl-diphosphate synthase
adenylylsulfate reductase
translation elongation factor EF-G
Occurences

Figure 4.9. Histine metabolism pathways for *P. marinus* (top) and *P. ubique* (bottom). Blue shading indicates that transcripts were found in the night transcriptome; Grey shading indicates genes that are present in the genome, but no transcripts were found; White shading indicates genes that are not present in the reference genomes.


Histidine Metabolism Prochlorococcus marinus
Pentose phosphate pathway
PRPP
2.4.2.17
Phosphoribosyl-ATP
Phosphoribosyl-AMP
3.6.1.31
3.5.4.19
Phosphoribosyl-formimino-AICAR-P
5.3.1.16
Phosphoribulosyl-formimino-AICAR-P
2.4.2.-
Imidazole-glycerol-3P
4.2.1.19
Imidazole-acetol-P
2.6.1.9
3.1.3.15
L-Histidinol
L-Histidinol-P
1.1.1.23
L-Histidinal
1-Methyl-L-histidine
3.4.13.5
6.3.2.11
Anserine
2.1.1.22
Carnosine
2.1.1.-
1.1.1.23
5'P-Ribosyl-5-amino-4-imidazole carboxamide (AICAR)
Purine metabolism
3.4.13.3
3.4.13.20
6.1.1.21
L-Histidyl-tRNA
N-Formyl-L-aspartate
3.5.3.5
3.5.2.-
Imidazolone acetate
1.14.13.5
Imidazole-4-acetate
1.2.1.3
Imidazole acetaldehyde
1.4.3.6
Histamine
4.1.1.22
4.1.1.28
L-Histidine
Hercynine
Thiourocanic acid
Ergothioneine
3.5.1.15
3.5.1.8
N-Formimino-L-aspartate
Aspartate
Alanine and Aspartate metabolism
6.3.4.8
1-(5-Phosphoribosyl) imidazole-4-acetate
(1-Ribosylimidazole)-4-acetate
2.1.1.8
2.3.1.-
N-Methyl-histamine
1.4.3.4
4-(β-Acetylaminoethyl)-imidazole
Methylimidazole acetaldehyde
1.2.1.5
Methylimidazole-acetic acid
4.3.1.3
2.6.1.38
Urocanate
4-Oxoglutaramate
2-Oxoglutarate
4.2.1.49
Formyl-isoglutamate
Isoglutamate
4-Imidazolone-5-propanoate
3.5.2.7
N-Formimino-L-glutamate
3.5.3.13
1.14.13.-
2.1.2.5
N-Formyl-L-glutamate
3.5.3.8
Hydantoin-5-propionate
3.5.1.68
Imidazole-pyruvate
Imidazole-lactate
N-carbamyl-L-glutamate
L-Glutamate
Glutamate metabolism

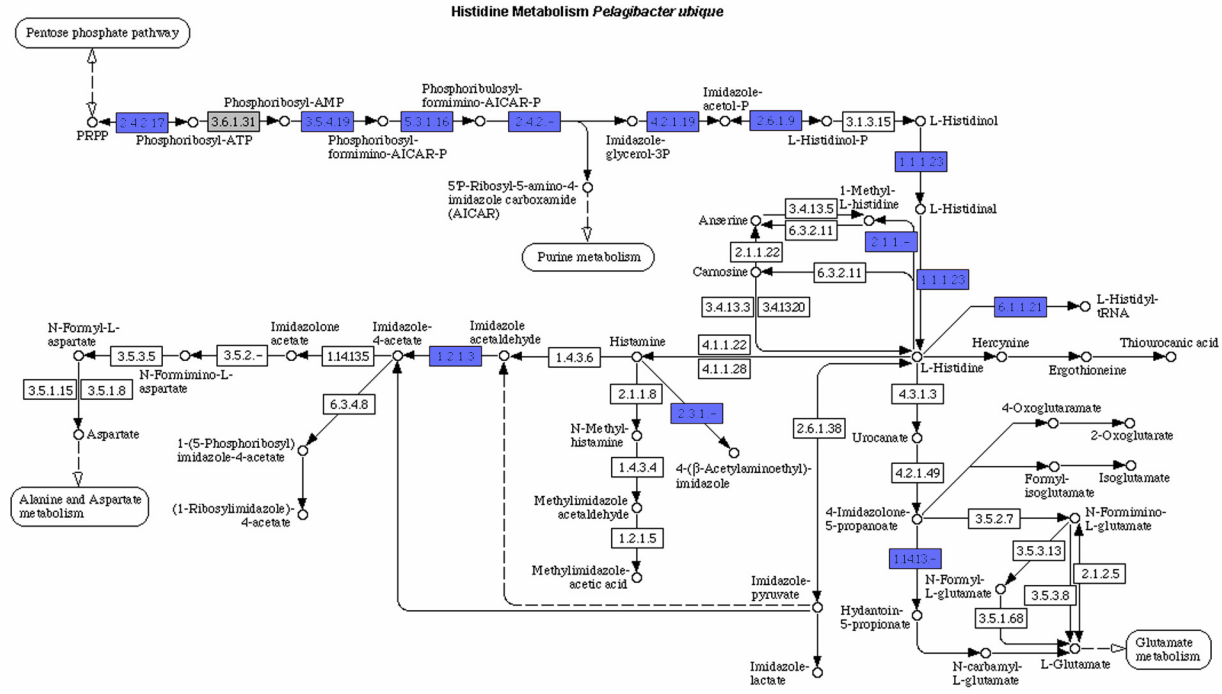

Histidine Metabolism Pelagibacter ubique
Pentose phosphate pathway
PRPP
2.4.2.17
Phosphoribosyl-ATP
Phosphoribosyl-AMP
3.6.1.31
3.5.4.19
Phosphoribosyl-formimino-AICAR-P
5.3.1.16
Phosphoribulosyl-formimino-AICAR-P
2.4.2.-
Imidazole-glycerol-3P
4.2.1.19
Imidazole-acetol-P
2.6.1.9
3.1.3.15
L-Histidinol
L-Histidinol-P
1.1.1.23
L-Histidinal
1-Methyl-L-histidine
3.4.13.5
6.3.2.11
Anserine
2.1.1.22
Carnosine
2.1.1.-
1.1.1.23
5'P-Ribosyl-5-amino-4-imidazole carboxamide (AICAR)
Purine metabolism
3.4.13.3
3.4.13.20
6.1.1.21
L-Histidyl-tRNA
N-Formyl-L-aspartate
3.5.3.5
3.5.2.-
Imidazolone acetate
1.14.13.5
Imidazole-4-acetate
1.2.1.3
Imidazole acetaldehyde
1.4.3.6
Histamine
4.1.1.22
4.1.1.28
L-Histidine
Hercynine
Thiourocanic acid
Ergothioneine
3.5.1.15
3.5.1.8
N-Formimino-L-aspartate
Aspartate
Alanine and Aspartate metabolism
6.3.4.8
1-(5-Phosphoribosyl) imidazole-4-acetate
(1-Ribosylimidazole)-4-acetate
2.1.1.8
2.3.1.-
N-Methyl-histamine
1.4.3.4
4-(β-Acetylaminoethyl)-imidazole
Methylimidazole acetaldehyde
1.2.1.5
Methylimidazole-acetic acid
4.3.1.3
2.6.1.38
Urocanate
4-Oxoglutaramate
2-Oxoglutarate
4.2.1.49
Formyl-isoglutamate
Isoglutamate
4-Imidazolone-5-propanoate
3.5.2.7
N-Formimino-L-glutamate
3.5.3.13
1.14.13.-
2.1.2.5
N-Formyl-L-glutamate
3.5.3.8
Hydantoin-5-propionate
3.5.1.68
Imidazole-pyruvate
Imidazole-lactate
N-carbamyl-L-glutamate
L-Glutamate
Glutamate metabolism

Figure 4.10. Biosynthesis of steroids and carotenoids pathway for *P. marinus*. Yellow shading indicates that transcripts were found in the day transcriptome; Grey shading indicates genes that are present in the genome, but no transcripts were found; White shading indicates genes that are not present in the reference genomes.

Biosynthesis of Steroids and Carotenoid Biosynthesis *Prochlorococcus marinus*

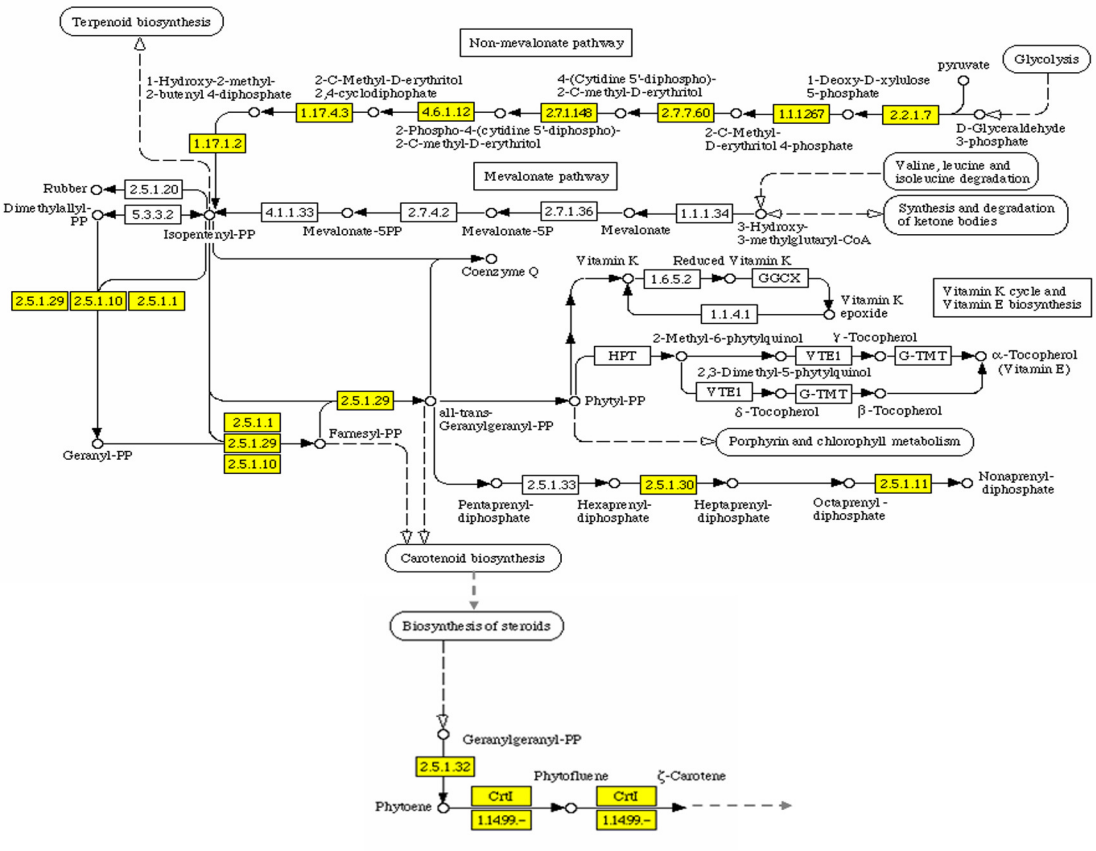


Figure 4.11. Number of eukaryotic transcripts in day (top bars) compared to night (bottom bars) samples. The relative contribution of Viridiplanteae (■), photosynthetic Chromist algae (■), and other Chromist (■) transcripts to each Gene Ontology (GO) annotation category are depicted.

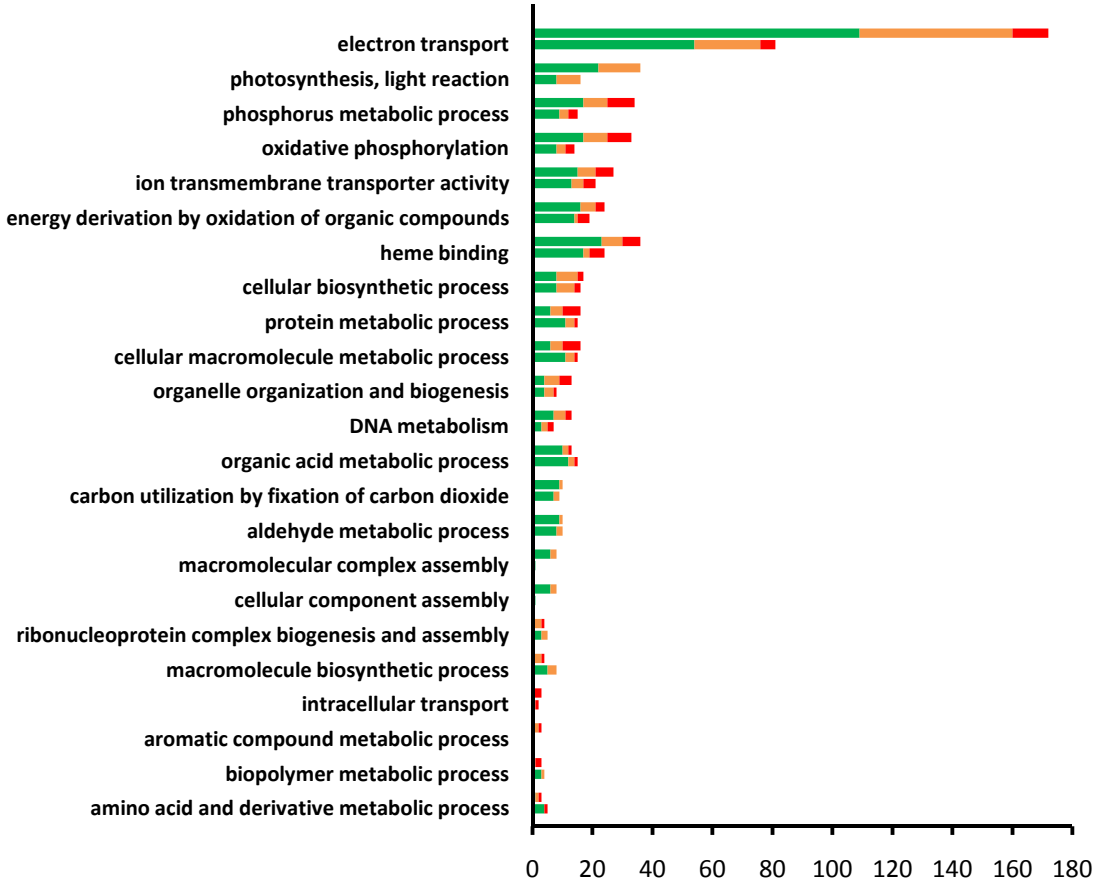
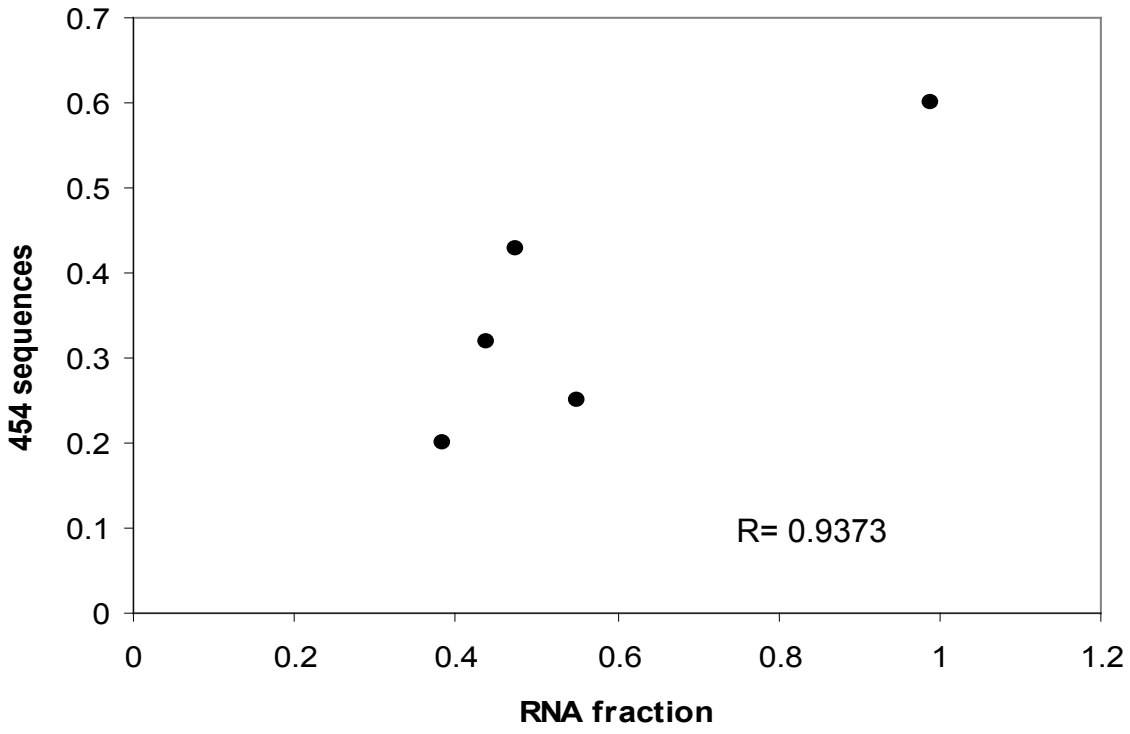

electron transport
photosynthesis, light reaction
phosphorus metabolic process
oxidative phosphorylation
ion transmembrane transporter activity
energy derivation by oxidation of organic compounds
heme binding
cellular biosynthetic process
protein metabolic process
cellular macromolecule metabolic process
organelle organization and biogenesis
DNA metabolism
organic acid metabolic process
carbon utilization by fixation of carbon dioxide
aldehyde metabolic process
macromolecular complex assembly
cellular component assembly
ribonucleoprotein complex biogenesis and assembly
macromolecule biosynthetic process
intracellular transport
aromatic compound metabolic process
biopolymer metabolic process
amino acid and derivative metabolic process
0
20
40
60
80
100
120
140
160
180

Figure 4.12. Quality control of the pyrosequences using qPCR verifications of transcript ratios for five genes: *recA* and *psaA* from *P. marinus* str. AS9601, a bacteriorhodopsin and a Na+/solute symporter (Ssf family) gene from *P. ubique* HTCC1062, and a probable integral membrane proteinase attributed to *P. torquis* ATCC 700755. The night:day ratio of transcripts in the pyrosequence libraries is plotted against the same ratio in the original total RNA fraction.


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
454 sequences
R= 0.9373
0
0.2
0.4
0.6
0.8
1
1.2
RNA fraction

Supplemental Table S4.1. Most abundant sequences with no BlastX hits in RefSeq or nr, but with blastP hits in the GOS database.

CAMERA id	454 hits NIGHT	454 hits DAY	Description	Organism	E-value	Percent id
JCVI_PEP_1105075490270	21	17	carbamoyl transferase, NodU family	Microscilla marina ATCC 23134	8.00E-07	39
JCVI_PEP_1105078753186	9	16	hypothetical protein STIAU_0963	Stigmatella aurantiaca DW4/3-1	0.49	27
JCVI_PEP_1105079515673	12	42	hypothetical protein AN5245.2	Aspergillus nidulans FGSC A4	1.00E-06	64
JCVI_PEP_1105080866231	8		hypothetical protein P3TCK_21760	Photobacterium profundum 3TCK	4.1	33
JCVI_PEP_1105081245293		12	importin 11	Xenopus tropicalis	5.6	30
JCVI_PEP_1105085592208	6		TPR Domain containing protein	Tetrahymena thermophila SB210	1.00E-13	34
JCVI_PEP_1105086313321	14	42	PREDICTED: hypothetical protein, partial	Danio rerio	3.2	40
JCVI_PEP_1105086541329	9		ENSANGP00000012099	Anopheles gambiae str. PEST	5.5	36
JCVI_PEP_1105088016445	18	26	conserved hypothetical protein	Aurantimonas sp. SI85-9A1	2.00E-13	73
JCVI_PEP_1105088081499	43	71	PREDICTED: hypothetical protein isoform 3	Pan troglodytes	9.5	38
JCVI_PEP_1105089605501		9	PREDICTED: similar to extra spindle poles like 1, partial	Ornithorhynchus anatinus	7.1	40
JCVI_PEP_1105091626633		12	fatty acid desaturase	Pseudomonas mendocina ymp	1.00E-39	34
JCVI_PEP_1105091912663	10	17	PREDICTED: similar to multidrug resistance protein 2; MRP2	Monodelphis domestica	4.1	32
JCVI_PEP_1105092263789	18	39	protein of unknown function DUF6, transmembrane	Pseudomonas mendocina ymp	0.64	32
JCVI_PEP_1105092773177		13	LOL3 (LSD ONE LIKE 3); caspase/ cysteine-type endopeptidase	Arabidopsis thaliana	0.037	33
JCVI_PEP_1105093885997	9	14	heparan sulfate D-glucosaminyl 3-O-sulfotransferase 4	Homo sapiens	0.076	35
JCVI_PEP_1105094620847	22	26	PREDICTED: hypothetical protein	Monodelphis domestica	9.4	50
JCVI_PEP_1105094768003	27	27	hypothetical protein BT_2907	Bacteroides thetaiotaomicron VPI-5482	7.1	29
JCVI_PEP_1105099326839	6		hypothetical protein NT01CX_0152	Clostridium novyi NT	4.00E-19	72
JCVI_PEP_1105101402207	42	83	hypothetical protein AN5245.2	Aspergillus nidulans FGSC A4	1.00E-06	68

JCVI_PEP_1105102681931	10	11	hypothetical protein Nwi_0408	Nitrobacter winogradskyi Nb-255	9.3	33
JCVI_PEP_1105103942839	9	12	ephA	Pseudomonas fluorescens Pf-5	0.58	26
JCVI_PEP_1105104181909		9	hypothetical protein PY00059	Plasmodium yoelii yoelii str. 17XNL	9.4	61
JCVI_PEP_1105105913673	10	18	hypothetical protein VSWAT3_19961	Vibrionales bacterium SWAT-3	7.00E-05	70
JCVI_PEP_1105107045545	9	26	hypothetical protein PFI0160w	Plasmodium falciparum 3D7	0.85	35
JCVI_PEP_1105108198893	6		hypothetical protein A9601_03531	Prochlorococcus marinus str. AS9601	1.00E-16	89
JCVI_PEP_1105110838419	9	21	hypothetical protein FG11306.1	Gibberella zeae PH-1	9.3	37
JCVI_PEP_1105111146131	21	35	conserved hypothetical protein	Aspergillus terreus NIH2624	2.00E-06	53
JCVI_PEP_1105111995181	6	19	PREDICTED: similar to RAS guanyl releasing protein 2 isoform 1	Macaca mulatta	1.9	33
JCVI_PEP_1105111995183	12	13	PREDICTED: hypothetical protein	Monodelphis domestica	9.4	50
JCVI_PEP_1105112567145	11	23	PREDICTED: similar to Nicotinic acid receptor 2 (G-protein coupled receptor 109B) (G-protein coupled receptor HM74) (G-protein coupled receptor HM74B)	Homo sapiens	2.5	29
JCVI_PEP_1105113205687	8	11	hypothetical protein RoseRS_0166	Roseiflexus sp. RS-1	7.1	44
JCVI_PEP_1105113249369		17	PREDICTED: hypothetical protein	Danio rerio	1.6	35
JCVI_PEP_1105113249371	14	23	PREDICTED: hypothetical protein	Mus musculus	0.49	30
JCVI_PEP_1105116114161	12	20	unnamed protein product	Kluyveromyces lactis	7.2	33
JCVI_PEP_1105119280025	12		hypothetical protein MAP0314	Mycobacterium avium subsp. paratuberculosis K-10	1.9	39
JCVI_PEP_1105120216769	6	13	Sec-independent protein translocase, protein	Methanosarcina mazei Go1	1.9	29
JCVI_PEP_1105120626697		278	NADH dehydrogenase subunit 4	Campanulotes bidentatus compar	3.2	38
JCVI_PEP_1105121521421	15	29	hypothetical protein LOC100037854	Xenopus tropicalis	9.3	33
JCVI_PEP_1105121637171	25	43	conserved hypothetical protein	Listeria monocytogenes str. 4b H7858 ref|ZP_00232117.1| conserved hypothetical	1.00E-29	78
JCVI_PEP_1105122403379	8		hypothetical protein PU1002_05581	Candidatus Pelagibacter ubique HTCC1002	4.00E-11	41
JCVI_PEP_1105122670937	11	13	prenyltransferase, UbiA family	Myxococcus xanthus DK 1622	2.5	41
JCVI_PEP_1105123442829	88	213	hypothetical protein	Neurospora crassa OR74A	5.4	48
JCVI_PEP_1105125074505	8		pyridoxal kinase	Entamoeba histolytica HM-	9.3	25

				1:IMSS		
JCVI_PEP_1105126127769	14	37	probable ABC transporter, permease protein	Sinorhizobium meliloti 1021	0.95	28
JCVI_PEP_1105126190105	8		hypothetical protein SAR11_0401	Candidatus Pelagibacter ubique HTCC1002	1.00E-56	55
JCVI_PEP_1105127214729	8		homocysteine S-methyltransferase family protein	Alteromonas macleodii 'Deep ecotype'	0.003	36
JCVI_PEP_1105128063073	11	20	Sec-independent protein translocase, protein	Methanosarcina mazei Go1	1.9	30
JCVI_PEP_1105128063077	20	23	67 kDa myosin-cross-reactive antigen family protein	Aspergillus clavatus NRRL 1	0.5	42
JCVI_PEP_1105130550191		11	hypothetical protein PD2058	Xylella fastidiosa Temecula1	5.4	39
JCVI_PEP_1105131294605	10	22	acetolactate synthase, large subunit, biosynthetic type	Caldicellulosiruptor saccharolyticus DSM 8903	9.3	25
JCVI_PEP_1105134702477		9	branched-chain amino acid transport system permease protein	Thermobifida fusca YX	5.00E-31	39
JCVI_PEP_1105135092565	18	17	hypothetical protein Npun02000733	Nostoc punctiforme PCC 73102	3.00E-05	91
JCVI_PEP_1105135202684	15	13	G protein-coupled receptor 55	Mus musculus	0.043	29
JCVI_PEP_1105135294490	6		hypothetical protein STIAU_0963	Stigmatella aurantiaca DW4/3-1	0.83	26
JCVI_PEP_1105135627263	15	39	Sec-independent protein translocase, protein	Methanosarcina mazei Go1	1.1	30
JCVI_PEP_1105136011803		9	ABC-type transport system, permease component.	Bdellovibrio bacteriovorus HD100	1.5	47
JCVI_PEP_1105137026599	14	20	TNF-receptor-associated factor 1 CG3048-PA, isoform A	Drosophila melanogaster	9.3	42
JCVI_PEP_1105139263541	7		Beta-lactamase class C related penicillin binding protein	Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	1.9	33
JCVI_PEP_1105139378317		10	hypothetical protein GLP_532_41019_43322	Giardia lamblia ATCC 50803	3.3	41
JCVI_PEP_1105141077537	6		hypothetical protein ST2383	Sulfolobus tokodaii str. 7	1.1	31
JCVI_PEP_1105144909953		14	erythrocyte binding protein, putative	Trichomonas vaginalis G3	0.29	29
JCVI_PEP_1105145635501	33	56	hypothetical protein SO_1887	Shewanella oneidensis MR-1	9.4	31
JCVI_PEP_1105146102115		10	hypothetical protein Npun02000733	Nostoc punctiforme PCC 73102	3.00E-05	91
JCVI_PEP_1105146135083	6		DNA polymerase I	Marinobacter algicola DG893	0.012	38
JCVI_PEP_1105146453441	11	62	hypothetical protein CHGG_07381	Chaetomium globosum CBS 148.51	2.4	33
JCVI_PEP_1105148401573		10	hypothetical protein	Paramecium tetraurelia	9.4	31

JCVI_PEP_1105149538093	8	16	Putative pilus assembly protein, CpaE-like	Burkholderia xenovorans LB400	3.2	34
JCVI_PEP_1105149538099		11	PREDICTED: similar to RAS guanyl releasing protein 2 isoform 1	Macaca mulatta	1.9	33
JCVI_PEP_1105149617711	8	11	hypothetical protein DDBDRAFT_0215968	Dictyostelium discoideum AX4	3.3	29
JCVI_PEP_1105151059157	6		conserved hypothetical protein	Aspergillus terreus NIH2624	0.85	37
JCVI_PEP_1105151090117	8		UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase	Orientia tsutsugamushi Boryong	3.00E-23	29
JCVI_PEP_1105153754375	6		hypothetical protein Shewmr4_1771	Shewanella sp. MR-4	7.1	39
JCVI_PEP_1105153799937	90	155	L-aspartate oxidase	Nitrococcus mobilis Nb-231	0.23	31
JCVI_PEP_1105154470825	12	35	conserved hypothetical protein	Trichomonas vaginalis G3	0.66	31
JCVI_PEP_1105155112969	6	10	hypothetical protein Teth39DRAFT_1059	Thermoanaerobacter ethanolicus ATCC 33223	0.14	28
JCVI_PEP_1105155761181		17	hypothetical protein	Plasmodium berghei strain ANKA	3.5	34
JCVI_PEP_1105155761187		10	M28.4	Caenorhabditis elegans	1.5	40
JCVI_PEP_1105156846539	45	57	related to iron-sulfur flavoprotein of Methanosarcina thermophila	Marinomonas sp. MED121	1.00E-15	38
JCVI_PEP_1105157830830	6	9	unnamed protein product	Saimiriine herpesvirus 2	9.2	44
JCVI_PEP_1105159069745	8	15	putative periplasmic protein	Campylobacter jejuni subsp. jejuni HB93-13	1.5	33
JCVI_PEP_1105163367345		10	hypothetical protein PCNPT3_12759	Psychromonas sp. CNPT3	3.2	36
JCVI_PEP_1105163914371	20	38	Putative protein of unknown function; overexpression confers resistance to the antimicrobial peptide MiAMP1	Saccharomyces cerevisiae	0.004	53
JCVI_PEP_1105164013575	17	19	hypothetical protein Tfu_1246	Thermobifida fusca YX	0.65	36
JCVI_PEP_1105164583019	6	9	PREDICTED: similar to hCG1812157	Ornithorhynchus anatinus	3.2	44
JCVI_PEP_1105164583021		19	ABC transporter	Bacillus licheniformis ATCC 14580	1.1	29
JCVI_PEP_1105164943519		10	Sec-independent protein translocase, protein	Methanosarcina mazei Go1	3.5	30

Supplemental Table S4.2. Genes significantly overrepresented in the night (blue shading) and day (yellow shading) transcriptomes ($p < 0.05$).

2-dehydro-3-deoxy-phosphoheptonate aldolase
aspartate Semialdehyde dehydrogenase
CDP-diacylglycerol
deoxyhypusine synthase-like protein
excinuclease ABC subunit C
Putative 6-phosphogluconolactonase (DevB, Pgl)
putative glutathione reductase (NADPH)
pyruvate phosphate dikinase
selenium-binding protein, putative
Signal transduction histidine kinase
spermidine/putrescine transport system permease protein potc
Sun protein (Fmu protein)
Xaa-Pro aminopeptidase
acetyl-coenzyme A synthetase
Ammonium transporter family
Aromatic-ring hydroxylase (flavoprotein monooxygenase)
Bacteriorhodopsin
cell wall-associated hydrolase
ClpC
conserved hypothetical protein
CrcB protein domain protein
cytochrome b6
Cytochrome c oxidase, subunit I
D2 reaction center protein of photosystem II
Na+/solute symporter (Ssf family)
NADH dehydrogenase I subunit M
PetD protein (subunit IV of the Cytochrome b6f complex)
Photosystem I PsaA protein
Photosystem I PsaB protein
photosystem II 44 kDa protein
photosystem II D2 protein (photosystem q(a) protein)
photosystem II protein D1
Photosystem II PsbA protein (D1)
Photosystem II PsbB protein (CP47)
Photosystem II PsbC protein (CP43)
protoporphyrin IX magnesium chelatase, subunit chlH
putative aminotransferase
putative neutral invertase-like protein
putative tricarboxylic transport TctA
ribonucleotide reductase (Class II)
RNA polymerase beta prime subunit

Supplemental Table S4.3. KEGG pathways for three taxonomic bins (*P. marinus*, *P. ubique*, and Roseobacters) significantly overrepresented in the night (blue shading) and day (yellow shading) transcriptomes ($p < 0.10$).

Organism	Pathway ID	Pathway	Category
P. marinus	path:ko00340	Histidine metabolism	Amino Acid Metabolism
P. marinus	path:ko00380	Tryptophan metabolism*	Amino Acid Metabolism
P. marinus	path:ko00480	Glutathione metabolism	Metabolism of Other Amino Acids
P. marinus	path:ko00643	Styrene degradation*	Xenobiotics Biodegradation and Metabolism
P. marinus	path:ko00650	Butanoate metabolism*	Carbohydrate Metabolism
P. marinus	path:ko00760	Nicotinate and nicotinamide metabolism	Metabolism of Cofactors and Vitamins
P. marinus	path:ko01053	Biosynthesis of siderophore group nonribosomal peptides*	Biosynthesis of Polyketides and Nonribosomal Peptides
P. marinus	path:ko00100	Biosynthesis of steroids	Lipid Metabolism
P. marinus	path:ko00190	Oxidative phosphorylation	Energy Metabolism
P. marinus	path:ko00195	Photosynthesis	Energy Metabolism
P. marinus	path:ko03020	RNA polymerase	Transcription
P. ubique	path:ko00340	Histidine metabolism	Amino Acid Metabolism
P. ubique	path:ko00520	Nucleotide sugars metabolism	Carbohydrate Metabolism
P. ubique	path:ko00602	Glycosphingolipid biosynthesis - neo-lactoseries	Glycan Biosynthesis and Metabolism
P. ubique	path:ko00603	Glycosphingolipid biosynthesis - globoseries	Glycan Biosynthesis and Metabolism
P. ubique	path:ko00750	Vitamin B6 metabolism	Metabolism of Cofactors and Vitamins
P. ubique	path:ko00906	Carotenoid biosynthesis	#N/A
P. ubique	path:ko00190	Oxidative phosphorylation	Energy Metabolism
P. ubique	path:ko00680	Methane metabolism	Energy Metabolism
P. ubique	path:ko00760	Nicotinate and nicotinamide metabolism	Metabolism of Cofactors and Vitamins
P. ubique	path:ko03020	RNA polymerase	Transcription
P. ubique	path:ko04940	Type I diabetes mellitus*	Metabolic Disorders
P. ubique	path:ko05060	Prion disease*	Neurodegenerative Disorders
Roseobacter	path:ko00020	Citrate cycle (TCA cycle)	Carbohydrate Metabolism
Roseobacter	path:ko00380	Tryptophan metabolism*	Amino Acid Metabolism
Roseobacter	path:ko03060	Protein export	Folding, Sorting and Degradation
Roseobacter	path:ko00670	One carbon pool by folate	Metabolism of Cofactors and Vitamins
Roseobacter	path:ko03020	RNA polymerase	Transcription

* The transcripts corresponding to the enzymes in these pathways are involved in many processes and are therefore may not be participating in the assigned pathway.

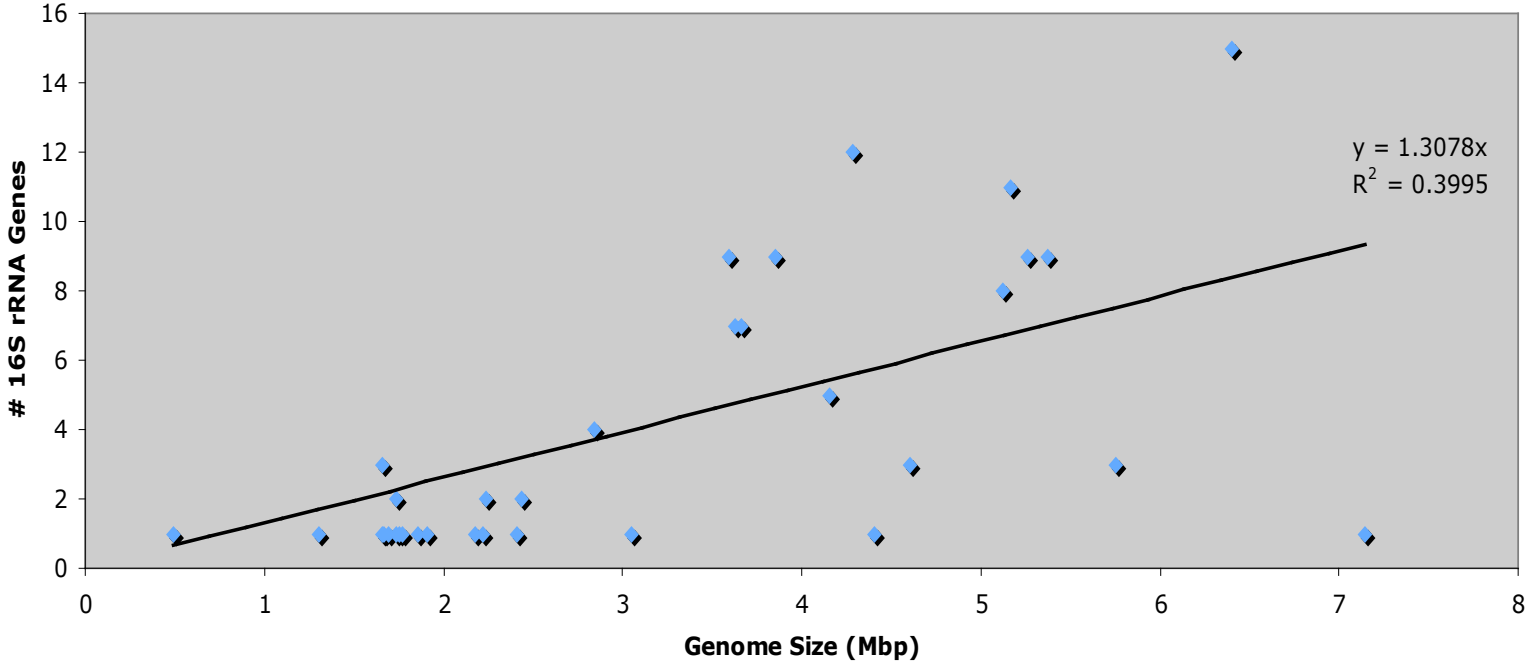
Supplemental Table S4.4. Estimates of coverage using the three different models. The Lander-Waterman model uses the 16S rRNA clone library data to establish a taxon-abundance model for the system at a similarity level of 99%, and is based on the assumptions that each taxon produces 1000 transcripts at any given time and all genes are expressed equally. The second model is based on *P. marinus* and *P. ubique* taxonomic bins and examines how many genes in the reference genome are represented by a transcript (the reference genomes have a within-taxon 16S rRNA sequence similarity of >99%). The third model uses the Chao1 richness estimators for COGs are computed using EstimateS (Version 8.0, R. K. Colwell, http://purl.oclc.org/estimates).

		% Coverage night		**% Coverage day**
Lander and Waterman model	**Taxon rank (night)**		Taxon rank (day)	
	1	89.5	1	94.1
	2	83.5	2- 4	89.6
	3- 4	74.1	5- 7	81.7
	5- 15	59.4	8- 14	67.8
	16- 62	36.3	15- 40	43.2
Genome coverage	*P. marinus* MIT9301	52.4		50.0
	P. marinus AS9601	43.8		39.8
	P. marinus MIT9312	33.4		29.8
	P. ubique HTCC1002	43.8		38.1
	P. ubique HTCC1062	36.9		29.0
Chao1 richness estimator for COGs		84.0		67.1

Supplemental Table S4.5. Primer sets used in qPCR.

Gene	Organism	F (5'- 3')	R (5'-3')	Annealing temperature	Product size
RecA	*P. marinus* AS9601	TTGTTGACTCGGTCGCAG	TGGCTTCCAATTACGTGATC	50	76
Proteorhodopsin	*P. ubique* HTCC1062	GGTGTTTCAGGTGTAGCAAACG	CGCCATTGACACAAGCCAG	50	87
Photosystem I PsaA protein	*P. marinus* AS9601	ACCTACTGCACGTCCCTGAG	GAATCAATTGTTTGGGCACAC	50	90
Na+/solute symporter (Ssf family)	*P. ubique* HTCC1062	ATTCGTTGCAATGGCAGGTG	CCACCAGTCCAACCGACAAG	50	75
probable integral membrane proteinase	*P. torquis* ATCC 00755	ACAGGCTGCTAGAGCAGATATG	CTCCTCGTGCTCTCGGTATC	50	83

Supplemental figure S4.1. The ratio of 16S gene number to genome size for all closed marine genomes as of January, 2008. The regression is significant ($p<0.02$), and indicates an average ratio of 1,500,000 bp per 16S rRNA gene. Assuming 1000 bp per gene, this is equivalent to one 16S rRNA gene per 1500 genes.


16S rRNA Genes
Genome Size (Mbp)
y = 1.3078x
R2 = 0.3995
0
2
4
6
8
10
12
14
16
0
1
2
3
4
5
6
7
8

CHAPTER 5

GENE EXPRESSION OF A MARINE ROSEOBACTER DURING EXPOSURE TO PHYTOPLANKTON EXUDATE[1]

[1] Poretsky, R.S., J. Oliver, P. Jasrotia, J. Cherrier and M.A. Moran. To be submitted to *Microbial Ecology.*

ABSTRACT

Extracellular release by marine phytoplankton is an important source of dissolved organic matter (DOM) in seawater. Compounds released by phytoplankton into the DOM pool are subsequently consumed by bacterioplankton for energy and biomass production. However, relatively little is known about the uptake and assimilation of individual compounds by marine bacterioplankton or the metabolic pathways by which they are catabolized. We are using *Silicibacter pomeroyi*, a cultured representative of the Roseobacter clade, as a model organism to assess the mechanisms and characteristics of bacterial utilization of phytoplankton-derived DOM. Marine roseobacters are known to physically associate with marine algae and consume various carbon and sulfur compounds produced by the algal cells. An axenic culture of a coastal marine phytoplankton, originally isolated from an environment similar to that of *S. pomeroyi,* was grown under conditions expected to release high molecular weight, labile DOM. *S. pomeroyi* was inoculated into the phytoplankton exudate and monitored over the next 12 h. Gene expression was assessed by collecting cells for RNA extraction and hybridizing resultant mRNA to a whole-genome microarray based on the *S. pomeroyi* genome sequence. Several *S. pomeroyi* genes appear upregulated in the presence of diatom DOM, including those involved in transport and utilization of amino acids, protocatechuate catabolism, and transcriptional regulation. These results provide a novel method for examining bacterial-phytoplankton associations on the level of gene expression and have implications for our understanding of carbon cycling between phytoplankton and bacteria in the marine microbial food web.

INTRODUCTION

Phytoplankton and bacteria have complex relationships in the marine environment. They can interact symbiotically, with each group taking advantage of products produced by the other, or competitively, e.g., for limiting nutrients (17). Specific physical interactions between groups of these organisms have been observed (4, 29, 55, 56). Additionally, the growth of phytoplankton has been shown to be influenced by the presence of bacteria (30, 58) or the compounds they produce (19). The classical role of bacteria in phytoplankton-bacterial interactions is the consumption and recycling of organic matter into the marine microbial food web (5). Phytoplankton production and bacterial uptake of organic matter is closely coupled (37). In an effort to better understand these relationships, the dynamics of such interactions have been receiving increasing attention (23, 30, 54, 56).

Phytoplankton and bacteria are dominant players in the cycling of dissolved organic matter (DOM). Marine DOM is one of the largest pools of reduced carbon on earth. Phytoplankton contribute to the DOM pool by release during grazing (35), viral lysis (25), or by exudation and excretion (6, 46). Approximately 10% of the carbon fixed by phytoplankton is released (6, 32) and is available for consumption by heterotrophic bacteria (3, 8, 14, 16, 37). Bacterial growth can be enhanced by the addition of phytoplankton-derived DOM (15). Recent work has demonstrated that high light stress stimulates the release of labile DOM by marine diatoms, more than 95% of which can be consumed by natural bacterial assemblages in less than 24 hours (Hamill and Cherrier, unpublished data). The composition of phytoplankton-released DOM is complex. Among the major components of this DOM are carbohydrates (44), proteins, nucleic acids, lipids, glycolate (24, 32), polyamines (39) and dimethylsulfoniopropionate (DMSP) (36). Although the utilization of many of these compounds by marine bacteria has been

investigated (16, 20, 22, 33, 42, 49, 51), the sheer complexity of the DOM pool makes it difficult to analyze bacterial decomposition of all of the DOM components. Furthermore, there remain additional gaps in our knowledge with respect to the metabolic activities of bacteria using phytoplankton-derived DOC.

Among the marine bacteria known to interact with phytoplankton and phytoplankton-derived exudates, members of the α-proteobacteria in general, and specifically the Roseobacter clade, are some of the most important. Roseobacters are abundant in the ocean (26, 27) and often found in the presence of phytoplankton blooms (28, 49, 52, 53). Members of this clade have been found to closely associate with dinoflagellates such as *Alexandrium* spp. (34), diatoms such as *Skeletonema costatum* (29, 55), and numerous other coastal phytoplankton (54). The Roseobacters are relatively amenable to culturing. Several have been isolated directly from the phytoplankton phycospheres (1, 2) and evidence exists for Roseobacter chemotaxis to phytoplankton products (40). Members of the Roseobacter clade have been shown to be consumers of phytoplankton-derived DOM compounds such as DMSP, amino acids, and glycolate (18, 38, 42).

The purpose of this study was to examine the response to and turnover of *S. costatum*-derived DOM by a model Roseobacter, *Silicibacter pomeroyi* DSS-3, via gene expression. Sufficient evidence exists to show that these organisms often co-exist in the coastal ocean, with some Roseobacters living attached to *S. costatum* and actively assimilating *S. costatum* DOM (7, 9, 29). The examination of *S. pomeroyi* gene expression using whole-genome microarrays has the extraordinary potential to provide insights into some of the processes involved in DOM processing by marine bacteria. By targeting gene expression rather than specific compounds, it is possible to identify important compounds within the DOM mixture and additional mechanisms

for DOM turnover that may not have been considered previously. We identified several genes that were significantly upregulated by *S. pomeroyi* in response to phytoplankton-derived DOM but not upregulated in the presence of acetate.

METHODS

Maintenance of phytoplankton culture

An axenic culture of the centric marine diatom *Skeletonema costatum* was obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (CCMP1332). Cultures were grown in sterile f/10 medium made with aged natural seawater (31) at 20°C in a circulating water tank. The cultures were maintained on a growth saturating 12:12-h light:dark cycle of approximately 110 µmol photons m^{-2} s^{-1} (Phillips 40 watt cool white fluorescent bulbs). Cultures were periodically tested for the presence of marine bacteria (General Marine Test Medium, CCMP) and were monitored for growth daily via *in vivo* fluorescence (Turner field fluorometer model 10-AU, Turner Designs). Cultures were considered acclimated when there was no significant difference between the slopes of 4 or more sequential growth curves (F-test, (62)). Twelve hour DOM release experiments were performed when cultures were in early-exponential growth to minimize the potential for nutrient limitation.

Isolation of phytoplankton-derived DOM

For the 12 h release experiment, steady state cultures in early-exponential growth were transferred into three 4 L flasks (6.7 ml of culture to 3 L f/10). Flasks were capped loosely with precombusted foil and returned to the tank at approximately 110 µmol photons $\cdot$ m^{-2} $\cdot$ s^{-1} and 20°C on a 12:12-h light:dark cycle. To enhance release of DOM by the cultures as a function of

varying light intensity, experimental incubation flasks were shifted to higher light (approximately 330 µmol photons m^{-2} s^{-1}; 110-330 µE) at the beginning of day 4, when cultures were again in early exponential growth. In order to monitor the net increase in DOC following light shock, samples for DOC analysis were collected just prior to artificial "sunrise" (T_0), after 6 h (T_1) and after 12 h (T_2). After 12 h, diatoms were removed via gentle vacuum filtration through precombusted (525° C for 4 hours) 0.7 µm nominal pore size GF/F filters (Whatman Inc.) and approximately 3 L of filtrate was kept at 4°C until bacterial inoculation.

Experimental design

Cultures of *S. pomeroyi* strain DSS-3 were grown in ½ YTSS medium (2 g yeast extract, 1.25 g tryptone peptone, 20 g sea salts in 1 L of DI water) at 30°C to an OD_{600} of 0.3 (mid-exponential growth phase). The cultures were then centrifuged for 10 min at 8000 x g and washed twice with 1X PBS. After the final wash, the cells were resuspended to a final density of ~ 5 x 10^7 cells ml^{-1} in triplicate flasks of 750 ml of either high light induced *S. costatum* DOM at a concentration of 105 µM C or in sodium acetate, an easily degradable source of organic carbon, dissolved in f/10 medium (to better mimic the background of the phytoplankton-derived DOM) at a concentration of 100 µM (200 µM C). All six flasks were supplemented with 100 µM NH_4Cl and 10 µM NaH_2PO_4 to prevent N or P limitation. Cultures were incubated at 25°C and shaking at 200 rpm in the dark.

The cultures were sampled immediately after inoculation (T_0), and after incubation of 20 min (T_1), 40 min (T_2), 6 h (T_3) and 12 h (T_4). The short (12 h) incubation time was chosen in order to capture significant transcriptional responses. At each time point, subsamples of 25 ml were removed from each flask into a 50 ml tube, mixed with 2 ml RNA stop solution (95%

ethanol, 5% phenol), and centrifuged for 10 min at 5,000 x g at 4°C. Cell pellets were then stored at -80°C for subsequent RNA extraction.

Additional aliquots of 100 ml were taken from each triplicate incubation flask at T_0, T_3, and T_4. This volume was filtered through a precombusted (525° C for 4 hours) 0.7 µm GF/F filter (Whatman Inc.) under gentle vacuum pressure. The filtrate was then distributed into a series of precombusted EPA VOA vials with 10% HCl- washed Teflon-lined caps and stored at -20°C for analysis of dissolved organic carbon (DOC), dissolved organic nitrogen (DON), nitrate plus nitrite ($NO_3^- + NO_2^-$), ammonium (NH_4^+), and monosaccharides (MCHO). Between time points, vacuum flasks, funnels and filter bases were rinsed with a 10% HCl solution and bio-grade DI water.

At each time point, 1 ml was removed from each flask for cell counts. Bacterial abundances were determined using 4,6-diamidino-2-phenylindole (DAPI) staining and counting with epifluorescence microscopy. For each sample, 10 fields were counted.

Analysis of Dissolved Organic Carbon and Nitrogen

DOC concentrations were measured using a Shimadzu TOC-V_{CPH} analyzer, employing a modification of the high-temperature catalytic oxidation process outlined by Suzuki et al. (59). Briefly, samples were acidified with 2 M ultrapure HCl (Sigma-Aldrich, St. Louis, MO) and sparged with CO_2-free air for 2 min to remove inorganic carbon and then combusted to measure organic carbon content. Standards were obtained from Shimadzu and prepared with Milli-pure water. The blank was determined to be less than 2 µM C. Analytical precision was 0.5 µM C. For each of 12 sample vials per time point, the TOC analyzer performed 3-5 injections. DOC results were referenced against certified reference materials (CRM) obtained from the University

of Miami, Rosensteil School of Marine and Atmospheric Sciences (RSMAS). Additional DOC measurements were obtained at the Laboratory for Environmental Analysis at the University of Georgia.

Total dissolved nitrogen (TDN) was measured concurrently with the TOC using an in-line Shimadzu TNM-1 analyzer. For each of 12 sample vials per time point, the TNM analyzer performed 3-5 injections. TDN results were also referenced against the CRM obtained from the RSMAS. Dissolved organic nitrogen (DON) was determined by subtracting the sum of inorganic nitrogen constituents from TDN.

Analysis of Inorganic Nitrogen Constituents

Analysis of nitrate and nitrite was performed via vanadium III reduction and chemilumunescence (10) with an Antek 7000 elemental analyzer and an Antek 745 nitrate/nitrite attachment. Ammonium concentrations were determined via colorimetric techniques outlined by Solorzano (57) using a Milton-Roy Spectronic 601 spectrophotometer.

Analysis of Monosaccharides (MCHO)

MCHO analysis was performed via the spectrophotometric method outlined by Myklestad et al. (45). Absorbance was determined using a Milton-Roy Spectronic 601 spectrophotometer. Standards were made using (D+) glucose (Sigma) in milli-pure water. The analytical blank was < 2 µM C.

RNA extraction and amplification

Frozen cell pellets were resuspended in 1.5 ml of RNAwiz (Ambion, Austin, TX) and transferred into 2 ml tubes containing ~ 0.5 ml zirconia beads. Bead-beating by vortexing at high speed for 10 min was followed by centrifugation at top speed for 5 min at 4°C. The supernatant was then transferred to a new 2 ml tube, and the RNA was extracted with 0.2 volumes of chloroform (~200 µl). Following 10 min incubation at room temperature, the phases were separated by centrifugation at >10,000 rpm for 5 min. The RNA was then precipitated by adding 0.5 volumes of RNase-free water (~350 µl) and 1 volume (~1 ml) of isopropanol to the aqueous phase and incubating at -20°C for 30 min, followed by centrifugation at top speed at 4°C for 20 min. The RNA pellet was washed twice with cold 70% ethanol and resuspended in 90 µl RNase-free water. Residual DNA was removed using the TURBO DNA-*free* kit (Ambion) and ribosomal RNA was removed enzymatically with the mRNA-ONLY Prokaryotic mRNA Isolation Kit (Epicentre Biotechnologies, Madison, WI). In order to obtain µg quantities of mRNA, approximately 500 ng of RNA was amplified linearly using the MessageAmp II-Bacteria Kit (Ambion), but substituting the UTP in the T7 reaction with amino-allyl labeled UTP (Ambion) for subsequent indirect labeling. Total RNA and mRNA yield and quality were assessed at each step by measurement on the NanoDrop-1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE) and the Experion Automated Electrophoresis System (Bio-Rad, Hercules, CA).

Microarray design

Custom microarrays with 12,000 probes (CombiMatrix, Mukilteo, WA) were designed previously based the *S. pomeroyi* complete genome (13). Briefly, 4161 out of the 4348 identified

genes in the *S. pomeroyi* genome were represented on the arrays with two probes per gene for most genes. Genes without probes were either duplicates of other genes in the genome or did not match probe design criteria. 1928 probes were replicated in triplicate, while 6215 probes had one spot each on the arrays. Fifteen probes with 1, 2, and 3 mismatches were used to assess probe specificity. The array also had 545 built-in quality control spots from the manufacturer and 149 empty spots for background correction.

RNA labeling and microarray hybridization

Microarrays were hybridized competitively. The pooled, triplicate T_0 DOM samples and the pooled, triplicate T_0 acetate samples were hybridized against each additional time point in the DOM and acetate time series, respectively. Each pair of RNAs was labeled immediately prior to microarray hybridization. The dyes used for the T_0 sample and the other samples were alternated with each array to minimize possible labeling biases. To label the amplified, anti-sense RNA (aRNA), 20 μg were dried using a vacuum concentrator and then resuspended in 9 μl coupling buffer (25 mg ml^{-1} sodium bicarbonate). The dyes, AlexaFluor 555 or 647 (Invitrogen, Carlsbad, CA), were dissolved in 11 μl DMSO, combined with the aRNA, and incubated in the dark for 30 min, occasionally vortexing to mix during the incubation. The RNA was then purified using the MEGAclear kit (Ambion) and ethanol precipitation to concentrate to 10 μl. Label incorporation and RNA concentration was assessed (NanoDrop) and the RNA was fragmented using RNA fragmentation reagents (Ambion). This fragmented RNA was hybridized for 16 h at 45°C to the array as previously described (13) according to the manufacturer's instructions. Following analysis, each array was stripped using the CustomArray Stripping Kit (CombiMatrix) and reused up to three times.

Microarray analysis

Arrays were scanned with an Axon GenePix 4000B microarray scanner (Molecular Devices Corporation, Sunnyvale, CA) at 5 µm and analyzed with GenePix Pro 6.0 software. Background correction was achieved by subtracting the mean of the 5 closest empty spots on the array. Bad spots were identified visually and by mean vs. median plots for each color channel, and flagged accordingly. The detection limit was defined per array as average Sum of Medians (SM) of the 149 empty spots + 2 x their standard deviation. Result files for all arrays using the AlexaFluor 555 dye for the T_0 sample were color channel-swapped. The raw data were then imported into Acuity 4.0 for processing and analysis and the resulting datasets were normalized for signal intensity using Lowess (locally weighted scatter plot smoothing) normalization, a non-linear normalization method that corrects for imbalances in the data. Low intensity spots that fell below the detection limit, features flagged as bad, empty features, and quality assurance features were excluded from the analysis. All analyses were performed on the Lowess normalized log ratio data, M $[\log_2 (F_{635}\,\text{Median}- B_{635})/(F_{532}\,\text{Median} - B_{532})]$. Changes in transcription over time within a treatment were analyzed based on at least a two-fold change in expression level as well as clustering to discern patterns and structure of expression over time. Lists of genes fulfilling fold change criteria were created based on $|M|>1$ in at least two of the triplicate arrays, indicating a two-fold change relative to T_0. Although clustering is used for organizing the data, Gap Statistic (based on within-cluster dispersion compared to a reference null distribution (60)) was used to determine the optimal number of clusters in each dataset. This cluster number was subsequently used for Self-Organizing Maps (SOM) on mean-combined log ratio data. Principal Component Analysis (PCA) was used as an additional data reduction method to classify the sample. For each treatment, the mean Lowess normalized log ratio data was used for PCA to

determine the variance in expression over time. The patterns contributing to variance over time were compared to the patterns determined by clustering. Statistical analysis testing the null hypothesis that a substance is not differentially expressed at one time point relative to another was carried out using Student's t-test assuming equal variances. Significant genes were those that met the both the fold change criteria at a *P* value of <0.05 and the appropriate SOM/PCA pattern. Genes with significant change in expression over time for the DOM treatment but not with a significant change over time for the acetate treatment were considered to be unique responses to DOM.

RESULTS

DOM production and utilization

The starting DOC concentration in each treatment was increased by ~100 μM due to carry-over from the bacterial inoculum. The multiple washes of the pelleted inoculum with PBS reduced the DOC carry-over by more than 500 μM from initial concentrations. Numerous efforts to further minimize this carry-over were not successful. Growth of *S. pomeroyi* in minimal medium reduced the initial [DOC], but multiple washes resulted in a final [DOC] similar to obtained from *S. pomeroyi* washed after growth in rich medium. Both the DOM and acetate treatments, however, received the same carry-over from the bacterial inoculum. Both treatments also had an f/10 medium background, either from the medium in which the *S. costatum* was grown or the fresh f/10 to which acetate was added.

Under moderate light conditions, *S. costatum* released 36 μM DOC. The DOC released by this diatom following 12 h high light shock increased to 105 μM, a net increase of 69 μM. This high light released DOM was used for the experimental bacterial incubations and was

compared to the control incubations with 100 μM sodium acetate. Following the 12 h incubation with *S. pomeroyi*, the DOC in the acetate samples was reduced by more than 60% (Figure 5.1). Assuming all of this could be attributed to acetate consumption and given a background DOC concentration of ~100 μM in the medium before the acetate was added, more than 90% of the acetate was consumed in 12 h. In contrast, there was no discernable decrease in the DOC concentration of the phytoplankton DOM amended samples. In fact, [DOC] increased slightly (+ 5 ± 2 μM) over 12 h (Figure 5.1).

TN changed little over the time in incubations with both phytoplankton-derived DOM and acetate (+ 12 ± 0.3 μM, DOM; - 2 ± 1 μM, acetate). [DON] increased slightly during the DOM incubation (+ 6 ± 2 μM) while [CHO] did not change (Figure 5.2). DIN concentrations changed minimally over the course of the incubation with DOM. Although [NO_3] did not change appreciably, [NH_4] increased slightly (+ 9 ± 0.2 μM; Figure 5.2). [DIN] and [DON] were not measured for the acetate treatment.

Growth of S. pomeroyi

S. pomeroyi abundance increased during the incubations in both the acetate and DOM samples, but cells grew faster in the acetate samples than in the presence of phytoplankton-derived DOM (Figure 5.3). In both cases, the greatest cell number increase was in the last 6 h of the incubation at which time cells in both treatments appeared to begin entering exponential growth. Both sets of incubations began with 5.18 x 10^7 cells ml^{-1}. At the conclusion of the 12 h incubation, the cells in the DOM samples reached a density of 7.38 x 10^7 cells ml^{-1} while the cells in the acetate samples reached a density of 1.18 x 10^8 cells ml^{-1}. The specific growth rates

of *S. pomeroyi* in the phytoplankton-derived DOM and acetate control samples were 0.42 d^{-1} and 0.74 d^{-1}, respectively.

Quality of microarray hybridizations

Prior quality control tests showed high probe specificity for these microarrays and good reproducibility of the hybridization process (13). High signal intensity was also seen in this study, with fewer than 5% of the probes on each array falling below the detection level. Hybridizations to mismatch probes agreed with previous findings that the probes are highly specific and three mismatches decrease the signal intensity by up to 70%. The Lowess M log ratio data for each probe of the color-swapped arrays correlated well with their technical replicates (acetate $r = 0.81 \pm 0.11$, DOM $r = 0.87 \pm 0.06$) at each time point. Similarly, the Lowess M log ratios for the biological replicates were well correlated (acetate $r = 0.80 \pm 0.04$, DOM $r = 0.85 \pm 0.10$). Replicate arrays were therefore averaged for analyses when possible. In comparison, the correlation of the mean Lowess M log ratio data between each time point of the DOM and acetate arrays was lower ($r = 0.57 \pm 0.13$).

Overview of transcriptome response

Hierarchical clustering of the microarray data was used to evaluate the global structure of the expression datasets. It was evident from this clustering that the acetate and DOM samples formed two distinct groups (Figure 5.4). Among the DOM samples, the 40 min sample was most different from the other three time points. The acetate samples form a separate cluster from the DOM samples, with the 20 and 360 min samples clustering together within it (Figure 5.4). This expression pattern was confirmed with non-hierarchical clustering in the form of Self-Organizing

Maps (SOM) (Figure 5.5) showing either initial upregulation of some genes at 40 min (Figure 5.5, top) or two-stage upregulation i.e., intial upregulation followed by further upregulation of other genes at the end of the incubation (Figure 5.5, bottom) for the DOM treatment. Most acetate probes did not change much, particularly after the early time points (20 or 40 min). According to the SOM, most probes on the arrays (3941 out of 7956 Figure 5.5, center) showed little or no change in expression level over the course of the experiment.

Identification of upregulated genes

The Lowess M log ratio values were used for all analyses. The composite (triplicate arrays averaged for each time point) gene expression patterns for the acetate and DOM samples were analyzed using PCA and SOM. The pattern described by the first principal component of the DOM-amended samples corresponded closely to the SOM cluster showing a profile pattern with two-stage upregulation relative to T_0. 1329 probes out of the 1795 identified DOM treatment probes with this pattern were common between the PCA and the SOM clustering of DOM-amended samples. 1217 probes that were upregulated at 40 min and subsequently downregulated, were identified using the same combination of PCA and SOM clustering. For the DOM-amended samples only, probes that were upregulated after 12h that were significantly different from those upregulated at the earlier time points were designated “late” response. Likewise, probes that were upregulated at 20 or 40 min and that were significantly different from those upregulated at the later time points were designated “early” response.

The probes with the above expression patterns that were at least 2-fold upregulated relative to T_0 in at least two of three microarrays were identified. 505 early response probes and 869 late response probes were found. 150 early response probes and 277 late response probes

were significantly different from the upregulated probes at the other time points at a *p* value of <0.05 (Figure 5.6). Of these, 72 early probes covering 46 genes (Table 5.1) and 171 late probes covering 78 genes (Table 5.2) were statistically different from the upregulated genes in acetate. Ten genes were upregulated after 20 min and remained upregulated throughout the incubation, but the upregulation was significantly higher after 12 h (Table 5.2). Twenty-five genes were upregulated with the acetate addition only (Table 5.3). All of the acetate-upregulated genes responded within the early part (20- 40 min) of the incubation.

Functions of upregulated DOM genes

The upregulated genes in the DOM arrays that were not upregulated in acetate were placed into TIGR role categories based on their assigned functions. The most abundant early and late genes assigned to categories with identified functions were related to transport and binding, energy metabolism, and regulatory functions. Approximately 23% of both the early and late genes were unknown genes and more than 10% were conserved hypothetical, for which no function has been assigned (Table 5.4); although the functions of some genes in the unknown function category can often be identified, they are designated as such because the functions are not covered by the other TIGR categories.

Transport and binding protein genes comprised 10% of the early upregulated genes and ~20% of the late upregulated genes in response to DOM (Table 5.4). Among the early transport and binding genes, five of the six were affiliated with amino acids, peptides, and amines. The transport and binding gene *tolQ* encodes a proton transporter thought to be involved with uptake of cations and iron-carrying compounds. Two branched chain amino acid transporters located next to each other on the *S. pomeroyi* chromosome, SPO2530 and SPO2531, were among the

upregulated amino acid transport genes. *potF* (SPO3469), encoding a putrescine ABC transporter in the amino acid transporter category, was also upregulated. Four of the transport proteins among the late response genes were affiliated with polyamine uptake: a spermidine/putrescine ABC transporter (SPO1849), and three opine/polyamine ABC transporters (SPO2699, SPO2700 and SPO2702). In addition, three components of TRAP transporters specific for dicarboxylates encoded by SPO2628, SPO0591 (DctP) and SPO1463 (DctM) were upregulated.

The early response genes assigned to energy metabolism were for cytochromes and an oxidoreductase and were affiliated primarily with electron transport. The gene *kbl* (SPO3360) encodes a 2-amino-3-ketobutyrate coenzyme A ligase predicted to be involved in metabolism of amino acids. A cytochrome gene was upregulated at the end of the DOM incubation (*ccoN*, SPOA0190). Three acetoin catabolism genes, *acoR*, *acoC*, and *acoX*, (SPO3788, SPO3790, and SPO3793) were also late response genes. Acetoin is a metabolic product excreted by many prokaryotic and eukaryotic microorganisms. Several bacteria utilize acetoin, particularly in the absence of readily usable carbon sources (61). Other late response genes involved in energy metabolism encoded N-dimethylarginine dimethylaminohydrolase (SPOA0064), phosphogluconate dehydratase (SPO3032), aromatic 1,2-dioxygenase (SPO1451), and sensor histidine kinase RegB. Two of the nitrate reductase genes in the partial denitrification pathway of *S. pomeroyi*, *nirG* (SPOA0225) and *nirN* (SPOA0228) were among the late response genes.

Most of the early response genes assigned to a regulatory function were putatively involved in transcriptional regulation, belonging to the AraC (SPO1584), TetR (SPOA0023), and LuxR (SPOA0102) families. The AraC (SPO1002) and LuxR (SPO2575, SPOA0102) family transcriptional regulators were also upregulated towards the end of the incubation by the late

response genes in addition to members of the LysR (SPO0832 and SPO3530), AsnC (SPO0233), IclR (SPOA0143) and ArsR (SPOA0427) families.

Two of the early response genes, *hisC* (SPO3177) and a gene for the ankrin repeat protein (SPO0621), were also upregulated in the late part of the DOM incubation. *hisC* encodes a histidinol-phosphate aminotransferase, an enzyme involved in histidine biosynthesis while ankrin repeat proteins are involved in protein-protein interactions.

Early response genes in the unknown function category encoded several methyltransferases (SPO2650 and SPO3491) and aminotransferases (SPOSPO3400 and SPO3417). Late response genes in this category were for an aminomethyltransferase (SPO0635), two hydrolases (SPO0025, putatively involved in repairing oxidative damage and SPOA0316, an aminohydrolase), two decarboxylases (SPO1589 and SPO3342), a lactoylglutathione glyoxyalase (SPO1620), and a monooxygenase (SPO3153). SPO1589, a member of the carboxymuconolactone decarboxylase family, is potentially involved in catabolism of protocatechuate. SPO3342 is a possible lysine decarboxylase.

DISCUSSION

This study aimed to investigate the transcriptional response of the marine Roseobacter *S. pomeroyi* to DOM derived from the diatom *S. costatum*. Based on previous studies showing light-mediated release of labile diatom DOM (Hamill and Cherrier, unpublished), phytoplankton-derived DOM as a high quality substrate for bacteria (7, 16, 17, 42), and the prevalence of members of the Roseobacter clade in association with phytoplankton or blooms (28, 48, 56), we expected *S. pomeroyi* to rapidly consume *S. costatum* DOM released under high light. The consumption of this complex DOM over several hours was expected to be captured by the *S.*

pomeroyi transcriptional response, particularly in comparison to *S. pomeroyi* consumption of acetate, a single, simple compound and an easily biodegradable carbon source.

Despite the increased concentration of *S. costatum*-derived DOM following exposure of *S. costatum* to high light levels, there was no detected consumption of the DOC, DON, DIN, or CHO components of this exudate during the 12 h incubation with *S. pomeroyi*. The absence of detectable DOC utilization in this study could potentially be attributable to the failure of the light shock to induce the release of labile DOM by *S. costatum*. Evidence for this lies in the initial concentration of CHO in the DOM (1.24 μM ± 0.24; Figure 5.2), which is ~10% of the initial *S. costatum* DOM, far less than the 74% of *S. costatum* DOC reported to be CHO (9) and the ~ 30 μM CHO released by *Thalassiosira weisflogii* under similar high light treatment (Hamill and Cherrier, unpublished). Alternatively, even if the initial DOM was bioavailable, the time prior to inoculation with *S. pomeroyi* as well as during the course of the incubation could have allowed diagenetic changes to render the phytoplankton exudates recalcitrant. The detection limit of DOC is another factor that could have prevented measurable changes in DOM constituents in this study. Small changes in [DOC] would be difficult to detect against the high background DOC, especially given the carry-over from the inoculum mentioned above.

Despite the lack of detectable changes in phytoplankton-derived DOM during the course of incubation, microarray characterizations of *S. pomeroyi* gene expression showed a marked difference in transcriptional response to DOM compared to similar incubations with acetate, suggesting that there was a specific response to the different substrates. The lack of measurable consumption could be related to *S. pomeroyi* substrate preference, with the differential gene expression simply reflecting exposure to different material. Recent work has shown that coastal bacterioplankton may be comprised of generalists that can use a wide range of carbon sources

(43), but the genome sequence of *S. pomeroyi* indicates that it might have some, albeit broad, substrate preferences. *S. pomeroyi* shows the genetic capability for processing a variety of compounds that are potentially produced by phytoplankton including peptides, polyamines, amino acids, and DMSP, but the absence of glycosidases indicates that it may not be capable of hydrolyzing polysaccharides (41). The overall responses to the two substrates used in the present study indicate that acetate, an easily degradable compound, can be consumed rapidly with few genes experiencing significant changes in expression level over time (0.6% of the genome). In contrast, *S. costatum*-derived DOM, a more complex substrate that is more difficult to degrade, stimulates substantial transcriptional changes involving more genes over time (2% of the genome) (Figure 5.5).

The two categories of upregulated genes in response to *S. costatum*-derived DOM, those early response genes that were upregulated at the start of the incubation and those late response genes that were upregulated at the end of the incubation, demonstrate a differential response over time. Early response genes suggest that *S. pomeroyi* initially devoted a portion of its transcriptome to sensing and responding to compounds in the phytoplankton-derived DOM. A gene encoding a flagellar protein was also among the early response genes. Although it has full capabilities for motility, *S. pomeroyi* lacks any of the known genes for chemotaxis and the mechanisms (if present) for directing swimming are unknown (41). The upregulation of a number of signaling genes (*luxR*, *araC*, *tetR*, and *lysR*) in response to DOM suggests a mechanism for responding to compounds in the environment and appropriately modulating gene expression. Several of the late response genes, including the many transcriptional response genes and the acetoin utilization genes, are likely a result of the absence of easily metabolizable compounds.

Among both the early and late response genes, many encoding proteins affiliated with amino acid uptake and utilization were upregulated in response to *S. costatum* DOM, indicating that amino acids could be an important, available component of the DOM for *S. pomeroyi*. Toward the end of the 12 h incubation, several additional genes were upregulated, including a number involved in polyamine transport as well as additional aminohydrolases. Polyamines are found in phytoplankton release (47) and expression of genes affiliated with these compounds has been documented in coastal seawater (50). Other late response genes encoded TRAP transporters, implicated in the transport of carboxylic acids and other labile components of DOM. *S. pomeroyi* has a high number of such transporters in its genome (41). Two genes involved in the degradation of aromatic compounds, one for an aromatic 1,2-dioxygenase and one for carboxymuconolactone decarboxylase (protocatechuate catabolism), were upregulated in response to DOM. Degradation of aromatic substrates by *S. pomeroyi* has previously been demonstrated (11) and it has been suggested that phytoplankton-derived DOM can be a source of such compounds (21) which can subsequently serve as a carbon and energy source for Rosoebacters (12).

Taken together, the results of this study indicate that *S. pomeroyi* has a distinct transcriptional response to diatom-derived organic matter. Although growth was faster and presumably easier with acetate as a carbon source, there was evidence that *S. pomeroyi* could respond to and possibly take advantage of some components of the DOM such as amino acids, aromatic compounds, and polyamines. The functions of several of the *S. pomeroyi* genes that were upregulated could not be readily identified either because the functions are general (e.g., methyltransferases) or because the functions are unknown (e.g., conserved hypothetical proteins). Some of the phytoplankton-DOM upregulated genes in this study were similar to genes

upregulated previously in response to DMSP (13). In addition, there was some overlap between the upregulated genes in response to acetate and phytoplankton-derived DOM. Thus, upregulation of some *S. pomeroyi* genes may be somewhat general in response to a variety of DOM constituents, including those that are readily bioavailable and those that are not. If given more bioavailable DOM than the *S. costatum* material used in this study, *S. pomeroyi* might be able to respond more quickly and grow faster by using some of the same genes. However, it was shown here that *S. pomeroyi* could tolerate and even grow in conditions that include the absence of readily available compounds. The use of whole-genome microarrays was shown to be a powerful tool for assessing molecular responses, but further work is necessary to better understand the dynamics of the processing of phytoplankton-derived DOM by coastal bacterioplankton. To achieve this, studies using other coastal bacteria and diverse DOM substrates obtained from a variety of phytoplankton can be conducted in conjunction with traditional biogeochemical approaches.

REFERENCES

1. Alavi, M., T. Miller, K. Erlandson, R. Schneider, and R. Belas. 2001. Bacterial community associated with *Pfiesteria*-like dinoflagellate cultures. Environ. Microbiol. 3:380-396.
2. Allgaier, M., H. Uphoff, A. Felske, and I. Wagner-Dobler. 2003. Aerobic anoxygenic photosynthesis in Roseobacter clade bacteria from diverse marine habitats. Appl. Environ. Microbiol. 69:5051-5059.
3. Amon, R. M. W., and R. Benner. 1996. Bacterial utilization of different size classes of dissolved organic matter. Limnol. Oceanogr. 41:41-51.
4. Ashen, J. B., and L. J. Goff. 2000. Molecular and ecological evidence for species specificity and coevolution in a group of marine algal-bacterial symbioses. Appl. Environ. Microbiol. 66:3024-3030.
5. Azam, F. 1998. Microbial control of oceanic carbon flux: The plot thickens. Science 280:694-696.
6. Baines, S. B., and M. L. Pace. 1991. The Production of Dissolved Organic-Matter by Phytoplankton and Its Importance to Bacteria - Patterns across Marine and Fresh-Water Systems. Limnol. Oceanogr. 36:1078-1090.
7. Bell, W. H. 1983. Bacterial Utilization of Algal Extracellular Products .3. the Specificity of Algal-Bacterial Interaction. Limnol. Oceanogr. 28:1131-1143.
8. Bell, W. H., and E. Sakshaug. 1980. Bacterial Utilization of Algal Extracellular Products .2. a Kinetic-Study of Natural-Populations. Limnol. Oceanogr. 25:1021-1033.

9. Biersmith, A., and R. Benner. 1998. Carbohydrates in phytoplankton and freshly produced dissolved organic matter. Mar. Chem. 63:131-144.

10. Braman, R. S., and S. A. Hendrix. 1989. Nanogram Nitrite and Nitrate Determination in Environmental and Biological-Materials by Vanadium(Iii) Reduction with Chemi-Luminescence Detection. Anal. Chem. 61:2715-2718.

11. Buchan, A., L. S. Collier, E. L. Neidle, and M. A. Moran. 2000. Key Aromatic-Ring-Cleaving Enzyme, Protocatechuate 3,4-Dioxygenase, in the Ecologically Important Marine Roseobacter Lineage. Appl. Environ. Microbiol. 66:4662-4672.

12. Buchan, A., J. M. Gonzalez, and M. A. Moran. 2005. Overview of the marine Roseobacter lineage. Appl. Environ. Microbiol. 71:5665-5677.

13. Burgmann, H., E. C. Howard, W. Y. Ye, F. Sun, S. L. Sun, S. Napierala, and M. A. Moran. 2007. Transcriptional response of *Silicibacter pomeroyi* DSS-3 to dimethylsulfoniopropionate (DMSP). Environ. Microbiol. 9:2742-2755.

14. Carlson, C. A., S. J. Giovannoni, D. A. Hansell, S. J. Goldberg, R. Parsons, M. P. Otero, K. Vergin, and B. R. Wheeler. 2002. Effect of nutrient amendments on bacterioplankton production, community structure, and DOC utilization in the northwestern Sargasso Sea. Aquat. Microb. Ecol. 30:19-36.

15. Cherrier, J., and J. E. Bauer. 2004. Bacterial utilization of transient plankton-derived dissolved organic carbon and nitrogen inputs in surface ocean waters. Aquat. Microb. Ecol. 35:229-241.

16. Cherrier, J., J. E. Bauer, and E. R. M. Druffel. 1996. Utilization and turnover of labile dissolved organic matter by bacterial heterotrophs in eastern north Pacific surface waters. Mar. Ecol. Prog. Ser. 139:267-279.

17. Cole, J. J. 1982. Interactions between Bacteria and Algae in Aquatic Ecosystems. Annu. Rev. Ecol. Syst. 13:291-314.

18. Cottrell, M. T., and D. L. Kirchman. 2000. Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. Appl. Environ. Microbiol. 66:1692-1697.

19. Croft, M. T., A. D. Lawrence, E. Raux-Deery, M. J. Warren, and A. G. Smith. 2005. Algae acquire vitamin B-12 through a symbiotic relationship with bacteria. Nature 438:90-93.

20. Dafner, E. V., and P. J. Wangersky. 2002. A brief overview of modern directions in marine DOC studies - Part I. Methodological aspects. J. Environ. Monit. 4:48-54.

21. Duval, B., K. Shetty, and W. Thomas. 1999. Phenolic compounds and antioxidant properties in the snow alga *Chlamydomonas nivalis* after exposure to UV light. J. Appl. Phycol. 11:559-566.

22. Edenborn, H. M., and C. D. Litchfield. 1987. Glycolate Turnover in the Water Column of the New-York Bight Apex. Mar. Biol. 95:459-467.

23. Ferrier, M., J. L. Martin, and J. N. Rooney-Varga. 2002. Stimulation of *Alexandrium fundyense* growth by bacterial assemblages from the Bay of Fundy. J. Appl. Microbiol. 92:706-716.

24. Fogg, G. E. 1983. The Ecological Significance of Extracellular Products of Phytoplankton Photosynthesis. Bot. Mar. 26:3-14.

25. Fuhrman, J. A. 1999. Marine viruses and their biogeochemical and ecological effects. Nature 399:541-548.

26. Giovannoni, S. J., and M. S. Rappé. 2000. The uncultured microbial majority, p. 47-84. *In* D. L. Kirchman (ed.), Microbial ecology of the oceans. Wiley-Liss, New York.

27. Gonzalez, J. M., and M. A. Moran. 1997. Numerical dominance of a group of marine bacteria in the alpha-subclass of the class Proteobacteria in coastal seawater. Appl. Environ. Microbiol. 63:4237-4242.

28. Gonzalez, J. M., R. Simo, R. Massana, J. S. Covert, E. O. Casamayor, C. Pedros-Alio, and M. A. Moran. 2000. Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. Appl. Environ. Microbiol. 66:4237-4246.

29. Grossart, H. P., F. Levold, M. Allgaier, M. Simon, and T. Brinkhoff. 2005. Marine diatom species harbour distinct bacterial communities. Environ. Microbiol. 7:860-873.

30. Grossart, H. P., and M. Simon. 2007. Interactions of planktonic algae and bacteria: effects on algal growth and organic matter dynamics. Aquat. Microb. Ecol. 47:163-176.

31. Guillard, R. R. L. 1975. Culture of phytoplankton for feeding marine invertebrates, p. 26-60. *In* W. L. Smith and M. H. Chanley (ed.), Culture of Marine Invertebrate Animals. Plenum Press, New York, USA.

32. Hellebust, J. A. 1965. Excretion of Some Organic Compounds by Marine Phytoplankton. Limnol. Oceanogr. 10:192-206.

33. Hodson, R. E., F. Azam, A. F. Carlucci, J. A. Fuhrman, D. M. Karl, and O. Holmhansen. 1981. Microbial Uptake of Dissolved Organic-Matter in Mcmurdo-Sound, Antarctica. Mar. Biol. 61:89-94.

34. Jasti, S., M. E. Sieracki, N. J. Poulton, M. W. Giewat, and J. N. Rooney-Varga. 2005. Phylogenetic diversity and specificity of bacteria closely associated with *Alexandrium* spp. and other phytoplankton. Appl. Environ. Microbiol. 71:3483-3494.

35. Jumars, P. A., D. L. Penry, J. A. Baross, M. J. Perry, and B. W. Frost. 1989. Closing the Microbial Loop - Dissolved Carbon Pathway to Heterotrophic Bacteria from Incomplete Ingestion, Digestion and Absorption in Animals. Deep-Sea Research Part a-Oceanographic Research Papers 36:483-495.

36. Kiene, R. P., L. J. Linn, and J. A. Bruton. 2000. New and important roles for DMSP in marine microbial communities. J. Sea Res. 43:209-224.

37. Larsson, U., and A. Hagstrom. 1979. Phytoplankton Exudate Release as an Energy-Source for the Growth of Pelagic Bacteria. Mar. Biol. 52:199-206.

38. Lau, W. W. Y., and E. V. Armbrust. 2006. Detection of glycolate oxidase gene glcD diversity among cultured and environmental marine bacteria. Environ. Microbiol. 8:1688-1702.

39. Lee, C., and N. O. G. Jorgensen. 1995. Seasonal Cycling of Putrescine and Amino-Acids in Relation to Biological Production in a Stratified Coastal Salt Pond. Biogeochemistry 29:131-157.

40. Miller, T. R., K. Hnilicka, A. Dziedzic, P. Desplats, and R. Belas. 2004. Chemotaxis of *Silicibacter* sp. Strain TM1040 toward Dinoflagellate Products, p. 4692-4701, vol. 70.

41. Moran, M. A., A. Buchan, J. M. Gonzalez, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B. Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Y. Ye, T. R. Miller, J. Carlton, D. A. Rasko, I. T. Paulsen, Q. H. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S.

Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut, and N. Ward. 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 432:910-913.

42. Mou, X. Z., R. E. Hodson, and M. A. Moran. 2007. Bacterioplankton assemblages transforming dissolved organic compounds in coastal seawater. Environ. Microbiol. 9:2025-2037.

43. Mou, X. Z., S. L. Sun, R. A. Edwards, R. E. Hodson, and M. A. Moran. 2008. Bacterial carbon processing by generalist species in the coastal ocean. Nature 451:708-U4.

44. Myklestad, S. M. 1995. Release of Extracellular Products by Phytoplankton with Special Emphasis on Polysaccharides. Sci. Total Environ. 165:155-164.

45. Myklestad, S. M., E. Skanoy, and S. Hestmann. 1997. A sensitive and rapid method for analysis of dissolved mono- and polysaccharides in seawater. Mar. Chem. 56:279-286.

46. Nagata, T. 2000. Production mechanisms of dissolved organic matter, p. 121-152. *In* D. L. Kirchman (ed.), Microbial ecology of the oceans. Wiley-Liss, New York.

47. Nishibori, N., A. Yuasa, M. Sakai, S. Fujihara, and S. Nishio. 2001. Free polyamine concentrations in coastal seawater during phytoplankton bloom. Fish. Sci. 67:79-83.

48. Pinhassi, J., M. M. Sala, H. Havskum, F. Peters, O. Guadayol, A. Malits, and C. L. Marrase. 2004. Changes in bacterioplankton composition under different phytoplankton regimens. Appl. Environ. Microbiol. 70:6753-6766.

49. Pinhassi, J., R. Simo, J. M. Gonzalez, M. Vila, L. Alonso-Saez, R. P. Kiene, M. A. Moran, and C. Pedros-Alio. 2005. Dimethylsulfoniopropionate turnover is linked to the composition and dynamics of the bacterioplankton assemblage during a microcosm phytoplankton bloom. Appl. Environ. Microbiol. 71:7650-7660.

50. 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.

51. Rich, J. H., H. W. Ducklow, and D. L. Kirchman. 1996. Concentrations and uptake of neutral monosaccharides along 140 degrees W in the equatorial Pacific: Contribution of glucose to heterotrophic bacterial activity and the DOM flux. Limnol. Oceanogr. 41:595-604.

52. Riemann, L., G. F. Steward, and F. Azam. 2000. Dynamics of bacterial community composition and activity during a mesocosm diatom bloom. Appl. Environ. Microbiol. 66:578-587.

53. Rink, B., S. Seeberger, T. Martens, C. D. Duerselen, M. Simon, and T. Brinkhoff. 2007. Effects of phytoplankton bloom in a coastal ecosystem on the composition of bacterial communities. Aquat. Microb. Ecol. 48:47-60.

54. Rooney-Varga, J. N., M. W. Giewat, M. C. Savin, S. Sood, M. LeGresley, and J. L. Martin. 2005. Links between Phytoplankton and bacterial community dynamics in a coastal marine environment. Microb. Ecol. 49:163-175.

55. Sapp, M., A. S. Schwaderer, K. H. Wiltshire, H. G. Hoppe, G. Gerdts, and A. Wichels. 2007. Species-specific bacterial communities in the phycosphere of microalgae? Microb. Ecol. 53:683-699.

56. Schafer, H., B. Abbas, H. Witte, and G. Muyzer. 2002. Genetic diversity of 'satellite' bacteria present in cultures of marine diatoms. FEMS Microbiol. Ecol. 42:25-35.

57. Solorzan.L. 1969. Determination of Ammonia in Natural Waters by Phenolhypochlorite Method. Limnol. Oceanogr. 14:799-&.

58. Stoderegger, K. E., and G. J. Herndl. 2005. Dynamics in bacterial surface properties of a natural bacterial community in the coastal North Sea during a spring phytoplankton bloom. FEMS Microbiol. Ecol. 53:285-294.

59. Suzuki, Y., E. Tanoue, and H. Ito. 1992. A High-Temperature Catalytic-Oxidation Method for the Determination of Dissolved Organic-Carbon in Seawater - Analysis and Improvement. Deep-Sea Research Part a-Oceanographic Research Papers 39:185-198.

60. Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. J. Roy. Stat. Soc. Ser. B. (Stat. Method.) 63:411-423.

61. Xiao, Z. J., and P. Xu. 2007. Acetoin metabolism in bacteria. Crit. Rev. Microbiol. 33:127-140.

62. Zar, J. H. 1996. Biostatistical analysis, 3rd ed. Prentice Hall, Upper Saddle River, N.J.

Table 5.1. Genes significantly upregulated in the early part of the incubation with phytoplankton-derived DOM.

Locus Tag	COG	Annotation	Function
SPO0002	COG0357	gidB glucose-inhibited division protein B	Unknown function
SPO0049	COG0476	thiamine biosynthesis protein ThiF	Biosynthesis of cofactors, prosthetic groups, and carriers
SPO0192	COG1749	flagellar hook protein FlgE putative	Cellular processes
SPO0568	COG4231	pyruvate ferredoxin/flavodoxin oxidoreductase family protein	Energy metabolism
SPO0616	COG0604	oxidoreductase zinc-binding dehydrogenase family	Unknown function
SPO0621	COG0666	ankrin repeat protein	Unknown function
SPO0834	COG3383	formate dehydrogenase alpha subunit	Energy metabolism
SPO0845		hypothetical protein	No Data
SPO0867	COG2911	conserved hypothetical protein	Hypothetical proteins
SPO0880		conserved hypothetical protein	Hypothetical proteins
SPO1105	COG1884	methylmalonyl-CoA mutase	Energy metabolism
SPO1142	COG1171	threonine dehydratase putative	Amino acid biosynthesis
SPO1221		hypothetical protein	No Data
SPO1341		alkane-1 monooxygenase putative	Energy metabolism
SPO1470	COG2513	conserved hypothetical protein	Unknown function
SPO1537		twin-arginine translocation pathway signal sequence domain protein putative	Unknown function
SPO1584	COG4977	transcriptional regulator AraC family	Regulatory functions
SPO1669		hypothetical protein	No Data
SPO1688	COG2050	thioesterase family protein	Unknown function
SPO2163	COG0625	conserved hypothetical protein	Hypothetical proteins
SPO2255		conserved hypothetical protein	Hypothetical proteins
SPO2306		hypothetical protein	No Data
SPO2389	COG0243	oxidoreductase molybdopterin-binding	Unknown function
SPO2530	COG4177	branched-chain amino acid ABC transporter permease protein	Transport and binding proteins

SPO2531	COG0559	branched-chain amino acid ABC transporter permease protein	Transport and binding proteins
SPO2632	COG0007	cobA-1 uroporphyrin-III C-methyltransferase	Biosynthesis of cofactors, prosthetic groups, and carriers
SPO2650		methyltransferase FkbM family	Unknown function
SPO2664	COG1126	polar amino acid uptake family ABC transporter ATP-binding protein	Transport and binding proteins
SPO2791	COG0365	acsA acetyl-coenzyme A synthetase	Energy metabolism
SPO2886	COG1804	CAIB/BAIF family protein	Unknown function
SPO3112	COG0811	tolQ proton transporter TolQ	Transport and binding proteins
SPO3177	COG0079	hisC histidinol-phosphate aminotransferase	Amino acid biosynthesis
SPO3360	COG0156	kbl 2-amino-3-ketobutyrate coenzyme A ligase	Energy metabolism
SPO3368	COG1012	aldehyde dehydrogenase family protein	Energy metabolism
SPO3400	COG0404	aminomethyl transferase family protein	Unknown function
SPO3410	COG2172	anti-sigma B factor putative	Regulatory functions
SPO3417	COG0436	aminotransferase classes I and II	Unknown function
SPO3469	COG0687	potF putrescine ABC transporter periplasmic putrescine-binding protein	Transport and binding proteins
SPO3491	COG2226	methyltransferase UbiE/COQ5 family	Unknown function
SPO3713		conserved hypothetical protein	Hypothetical proteins
SPOA0023	COG1309	transcriptional regulator TetR family	Regulatory functions
SPOA0102	COG2197	transcriptional regulator LuxR family	Regulatory functions
SPOA0117	COG3473	Asp/Glu/hydantoin racemase family protein	Unknown function
SPOA0193	COG2332	cycJ cytochrome c-type biogenesis protein CycJ	Energy metabolism
SPOA0203		hypothetical protein	No Data
SPOA0359		cytochrome c family protein	Energy metabolism

Table 5.2. Genes significantly upregulated in the late part of the incubation with phytoplankton-derived DOM.

Locus Tag	COG	Annotation	Function
SPO0025	COG0494	hydrolase NUDIX family	Unknown function
SPO0122		YGGT family protein	Unknown function
SPO0140	COG3816	conserved hypothetical protein	Hypothetical proteins
SPO0160	COG1690	conserved hypothetical protein	Hypothetical proteins
SPO0175[a]	COG2063	flgH flagellar L-ring protein FlgH	Cellular processes
SPO0201		conserved domain protein	Hypothetical proteins
SPO0233	COG1522	transcriptional regulator AsnC family	Regulatory functions
SPO0463[a]	COG0589	universal stress family protein	Cellular processes
SPO0488	COG0197	rplP ribosomal protein L16	Protein synthesis
SPO0591	COG1638	TRAP dicarboxylate transporter DctPsubunit	Transport and binding proteins
SPO0621	COG0666	ankrin repeat protein	Unknown function
SPO0635	COG0404	aminomethyl transferase family protein	Unknown function
SPO0832	COG0583	transcriptional regulator LysR family	Regulatory functions
SPO0858	COG1858	methylamine utilization protein MauG putative	Energy metabolism
SPO1002	COG4977	transcriptional regulator AraC family	Regulatory functions
SPO1017[a]	COG0410	branched-chain amino acid ABC transporter ATP-binding protein	Transport and binding proteins
SPO1070	COG3573	conserved hypothetical protein	Hypothetical proteins
SPO1128		hypothetical protein	No Data
SPO1132	COG4175	glycine betaine/proline ABC transporter ATP-binding protein	Transport and binding proteins
SPO1135		conserved hypothetical protein	Hypothetical proteins
SPO1155	COG0140	phosphoribosyl-ATP pyrophosphohydrolase	Amino acid biosynthesis
SPO1165		conserved domain protein	Hypothetical proteins
SPO1235	COG5293	conserved hypothetical protein	Hypothetical proteins
SPO1236		hypothetical protein	No Data
SPO1241		conserved domain protein	Hypothetical proteins

SPO1446	COG1878	cyclase putative	Unknown function
SPO1451	COG4638	aromatic 1 2-dioxygenase alpha subunit	Energy metabolism
SPO1463	COG4664	TRAP dicarboxylate transporter DctM subunit	Transport and binding proteins
SPO1563	COG2346	protozoan/cyanobacterial globin family protein	Transport and binding proteins
SPO1589	COG0599	carboxymuconolactone decarboxylase family protein	Unknown function
SPO1620	COG0346	glyoxalase family protein	Unknown function
SPO1849	COG0411	branched-chain amino acid ABC transporter ATP-binding protein	Transport and binding proteins
SPO1934[a]	COG0433	conserved hypothetical protein	Hypothetical proteins
SPO2009	COG1176	spermidine/putrescine ABC transporter permease protein	Transport and binding proteins
SPO2163	COG0625	conserved hypothetical protein	Hypothetical proteins
SPO2193[a]	COG2987	hutU urocanate hydratase	Energy metabolism
SPO2344	COG4583	sarcosine oxidase gamma subunit family	Amino acid biosynthesis
SPO2465	COG1733	conserved hypothetical protein	Hypothetical proteins
SPO2491	COG1399	conserved hypothetical protein	Hypothetical proteins
SPO2507		membrane protein putative	Cell envelope
SPO2535	COG0123	histone deacetylase/AcuC/AphA family protein	Unknown function
SPO2575[a]	COG2197	DNA-binding response regulator LuxR family	Regulatory functions
SPO2590	COG4783	peptidase M48 family	Protein fate
SPO2628	COG1638	TRAP transporter solute receptor DctP family	Transport and binding proteins
SPO2699	COG1177	opine/polyamine ABC transporter permease protein	Transport and binding proteins
SPO2700[a]	COG1176	opine/polyamine ABC transporter permease protein	Transport and binding proteins
SPO2702	COG3842	opine/polyamine ABC transporter ATP-binding protein	Transport and binding proteins
SPO2817		YeeE/YedE family protein	Unknown function
SPO2867	COG1010	cobJ precorrin-3B C17-methyltransferase	Biosynthesis of cofactors, prosthetic groups, and carriers
SPO3032	COG0129	edd phosphogluconate dehydratase	Energy metabolism
SPO3110	COG5373	tonB domain protein putative	Transport and binding proteins
SPO3130[a]	COG4974	xerC tyrosine recombinase XerC	DNA metabolism
SPO3142	COG3106	conserved hypothetical protein	Hypothetical proteins
SPO3152	COG2030	MaoC domain protein	Unknown function

SPO3153	COG0654	monooxygenase putative	Unknown function
SPO3177	COG0079	hisC histidinol-phosphate aminotransferase	Amino acid biosynthesis
SPO3342	COG1611	decarboxylase family protein	Unknown function
SPO3516	COG0250	nusG transcription termination/antitermination factor NusG	Transcription
SPO3530	COG0583	transcriptional regulator LysR family	Regulatory functions
SPO3564	COG0701	permease putative	Transport and binding proteins
SPO3736	COG0678	antioxidant AhpC/Tsa family	Cellular processes
SPO3788	COG3284	acoR acetoin catabolism regulatory protein	Energy metabolism
SPO3790	COG0508	acoC acetoin dehydrogenase complex E2 component dihydrolipoamide acetyltransferase	Energy metabolism
SPO3793	COG3199	acoX acetoin catabolism protein X	Energy metabolism
SPO3867	COG0642	regB sensor histidine kinase RegB	Energy metabolism
SPO3868		conserved hypothetical protein	Hypothetical proteins
SPOA0029	COG1805	nqrB NADH:ubiquinone oxidoreductase Na(+)-translocating B subunit	Transport and binding proteins
SPOA0064	COG1834	NG NG-dimethylarginine dimethylaminohydrolase putative	Energy metabolism
SPOA0102[a]	COG2197	transcriptional regulator LuxR family	Regulatory functions
SPOA0143[a]	COG1414	transcriptional regulator IclR family	Regulatory functions
SPOA0153		hypothetical protein	No Data
SPOA0190	COG3278	ccoN-2 cytochrome c oxidase cbb3-type subunit I	Energy metabolism
SPOA0225	COG1522	nirG nitrite reductase heme biosynthesis G protein	Biosynthesis of cofactors, prosthetic groups, and carriers
SPOA0228	COG2010	nirN nitrite reductase protein N	Central intermediary metabolism
SPOA0316	COG1574	amidohydrolase domain protein	Unknown function
SPOA0369	COG3667	copper resistance protein B putative	Cellular processes
SPOA0422	COG3558	conserved hypothetical protein	Hypothetical proteins
SPOA0427	COG0640	transcriptional regulator ArsR family	Regulatory functions

[a] These genes were upregulated throughout the incubation but the upregulation was significantly higher after 12 h ($p < 0.05$)

Table 5.3. Genes significantly upregulated during incubation with acetate.

Locus Tag	COG	Annotation	Function
SPO0019		membrane protein putative	Cell envelope
SPO0229		rpsU ribosomal protein S21	Protein synthesis
SPO0322	COG5570	conserved hypothetical protein	Hypothetical proteins
SPO0369	COG2937	acyltransferase family protein	Unknown function
SPO0405	COG2087	cobinamide kinase/cobinamide phosphate guanyltransferase	Biosynthesis of cofactors, prosthetic groups, and carriers
SPO0508	COG4188	conserved hypothetical protein	Hypothetical proteins
SPO0584	COG0436	aspC-1 aspartate aminotransferase	Amino acid biosynthesis
SPO0723		conserved hypothetical protein	Hypothetical proteins
SPO0813	COG1304	L-lactate dehydrogenase putative	Energy metabolism
SPO0829		conserved hypothetical protein	Hypothetical proteins
SPO1075	COG3333	membrane protein putative	Cell envelope
SPO1106		conserved hypothetical protein	Hypothetical proteins
SPO1349		lipoprotein putative	Cell envelope
SPO1443	COG0513	rhlE ATP-dependent RNA helicase RhlE	Transcription
SPO1536	COG0438	glycosyltransferase group 1	Cell envelope
SPO1757	COG3562	kpsS capsular polysaccharideexport protein KpsS	Cell envelope
SPO2649		conserved domain protein	Hypothetical proteins
SPO2745	COG2138	conserved domain protein	Hypothetical proteins
SPO2853	COG0714	cobalt chelatase CobS subunit	Biosynthesis of cofactors, prosthetic groups, and carriers
SPO2855	COG4547	cobalt chelatase, pCobT subunit	Biosynthesis of cofactors, prosthetic groups, and carriers
SPO3088	COG2885	OmpA family protein	Cell envelope
SPO3308	COG4095	conserved hypothetical protein	Hypothetical proteins
SPO3322		conserved hypothetical protein	Hypothetical proteins
SPO3424	COG4886	leucine rich repeat protein	Unknown function
SPO3706	COG0411	branched-chain amino acid ABC transporter ATP-binding protein	Transport and binding proteins

Table 5.4. Distribution of significantly upregulated genes among functional role categories for phytoplankton-derived DOM and acetate treatments.

Functional role category	DOM early	% DOM early	DOM late	% DOM late	Acetate	% Acetate
Amino acid biosynthesis	2	4.35	3	3.85	1	4.00
Biosynthesis of cofactors, prosthetic groups, and carriers	2	4.35	2	2.56	3	12.00
Cell envelope			1	1.28	6	24.00
Cellular processes	1	2.17	4	5.13		
Central intermediary metabolism			1	1.28		
DNA metabolism			1	1.28		
Energy metabolism	9	19.57	10	12.82	1	4.00
Conserved hypothetical proteins	5	10.87	15	19.23	9	36.00
No Data	5	10.87	3	3.85		
Protein fate			1	1.28		
Protein synthesis			1	1.28	1	4.00
Regulatory functions	4	8.70	8	10.26		
Transcription			1	1.28	1	4.00
Transport and binding proteins	5	10.87	14	17.95	1	4.00
Unknown function	13	28.26	13	16.67	2	8.00
Total	**46**		**78**		**25**	

Figure 5.1. Change in DOC concentrations during incubations with either phytoplankton-derived DOM (filled circles) or acetate (open circles). Data points represent an average triplicate incubations. Error bars represent one standard deviation.

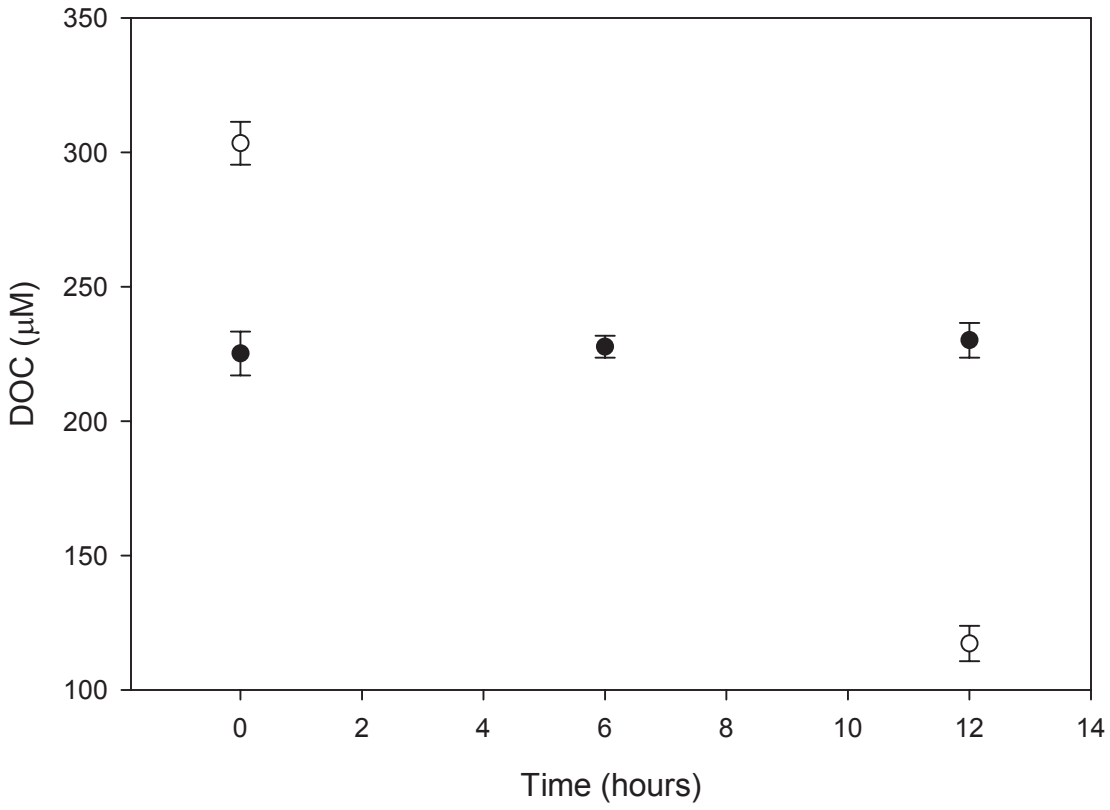

350
300
250
200
150
100
DOC (µM)
0
2
4
6
8
10
12
14
Time (hours)

Figure 5.2. Change in DON, NO_3, NH_4, CHO, and TDN during the course of the incubation with phytoplankton-derived DOM. Data points represent an average triplicate incubations. Error bars represent one standard deviation.

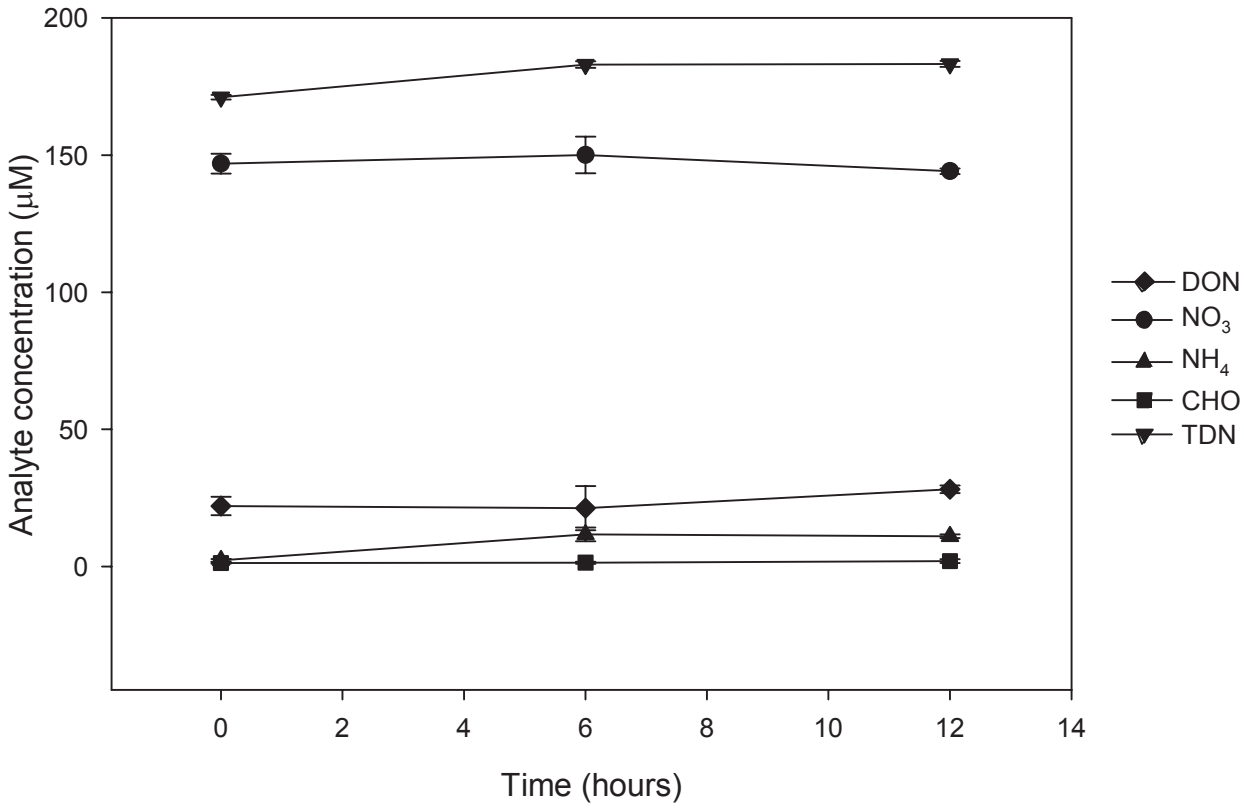

200
150
100
50
0
Analyte concentration (μM)
0
2
4
6
8
10
12
14
Time (hours)
DON
NO3
NH4
CHO
TDN

Figure 5.3. Growth of *S. pomeroyi* during the 12 h incubations with either phytoplankton-derived DOM (filled circles) or acetate (open circles) as determined by DAPI counts. Data points represent an average triplicate incubations. Error bars represent one standard deviation.

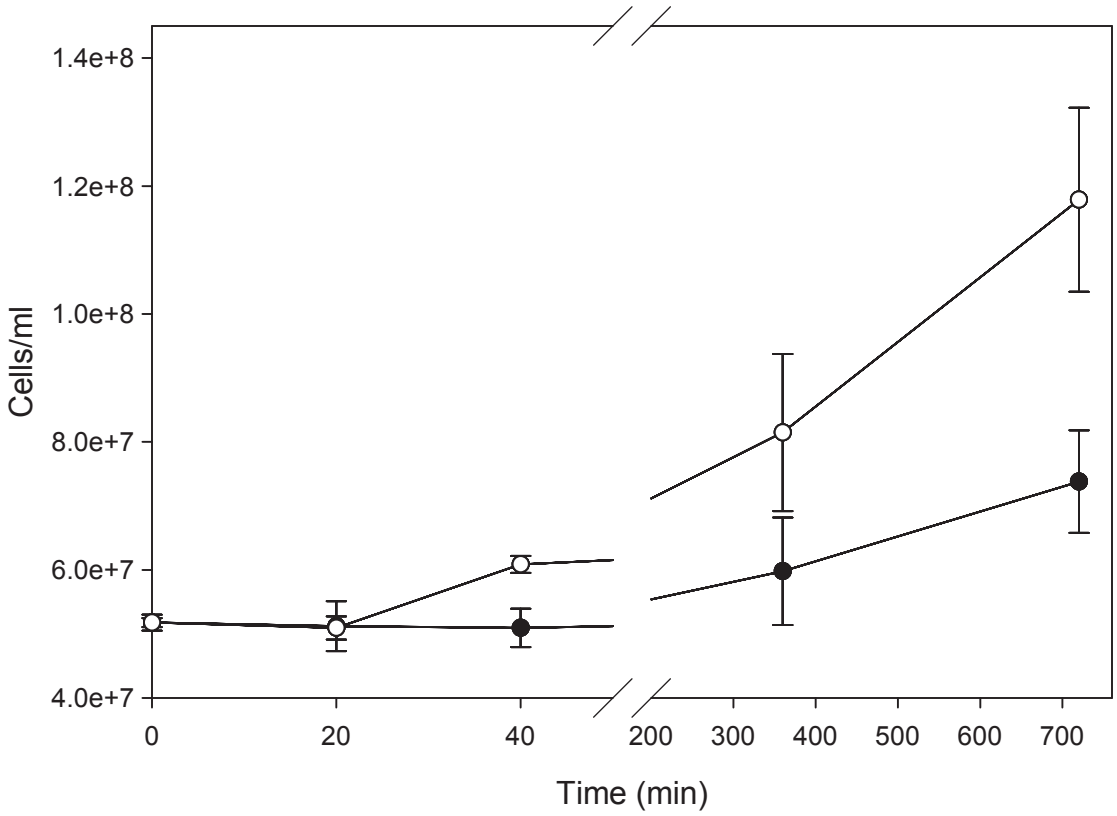

1.4e+8
1.2e+8
1.0e+8
8.0e+7
6.0e+7
4.0e+7
Cells/ml
0
20
40
200
300
400
500
600
700
Time (min)

Figure 5.4. Hierarchical clustering of expression patterns of 8143 *S. pomeroyi* probes using Pearson Correlation similarity metric with average linkage. Each row represents a single gene expression, while each column represents an identical treatment (green, upregulated; red, downregulated).

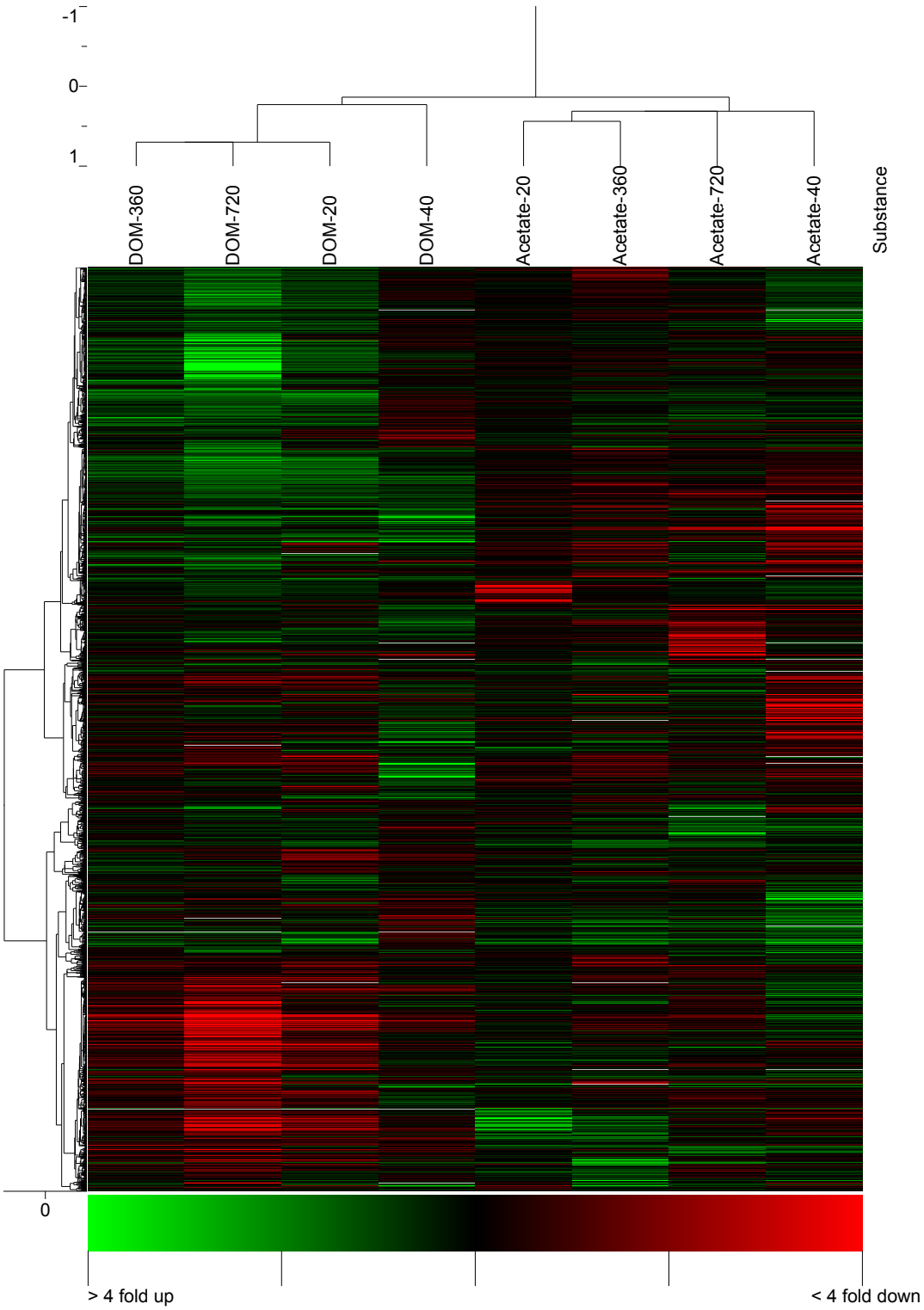

-1
0
1
DOM-360
DOM-720
DOM-20
DOM-40
Acetate-20
Acetate-360
Acetate-720
Acetate-40
Substance
0
> 4 fold up
< 4 fold down

Figure 5.5. Self-organizing maps (SOM) analysis of averaged Lowess normalized log ratio data (*M*). Green denotes probes upregulated relative to the T_0 control, red indicates downregulated. Lines show average behavior for each cluster. The treatments are indicated at the top of the figure along with the timepoints (20, 40, 360, and 720 min). The data were averaged for the three experimental replicates.

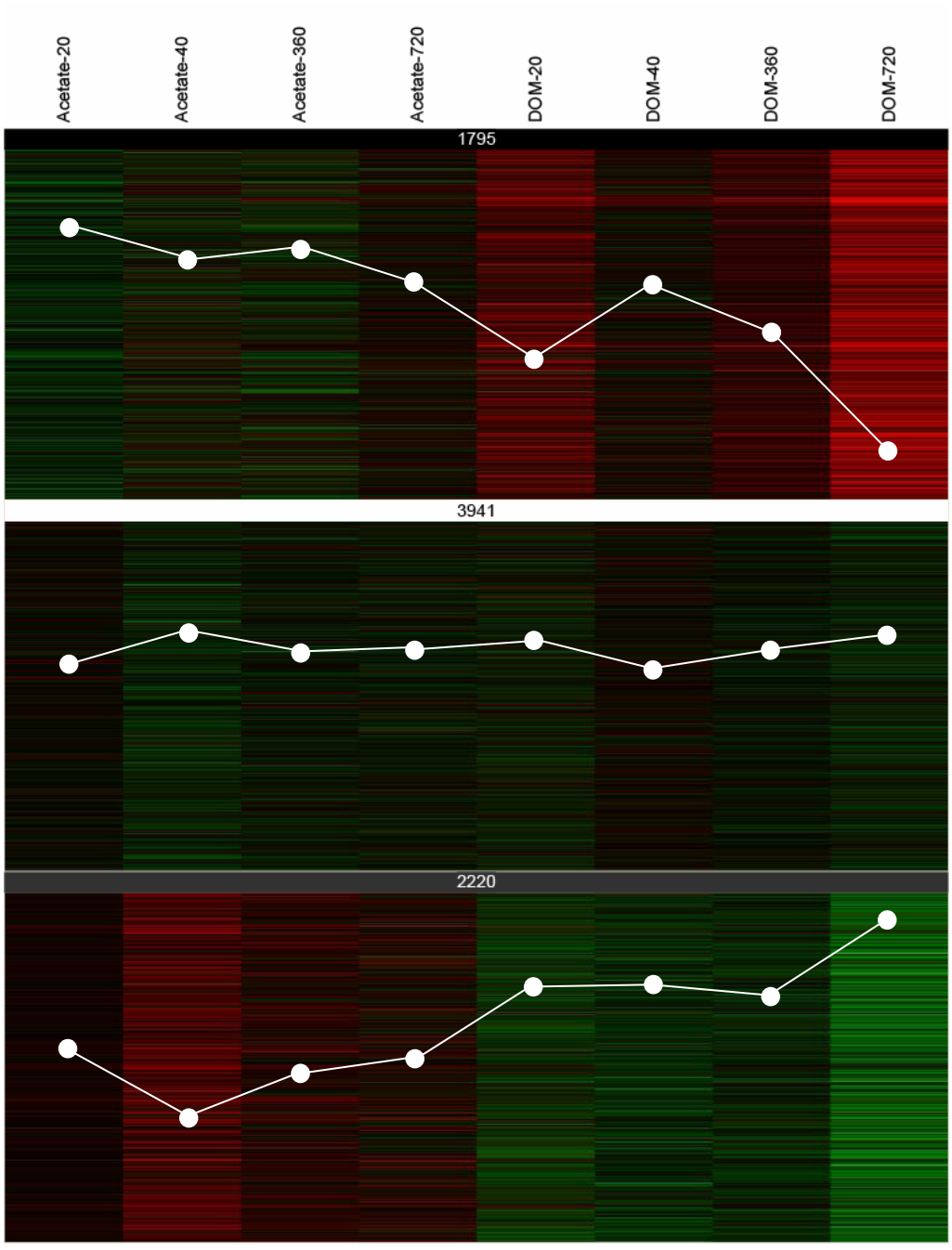
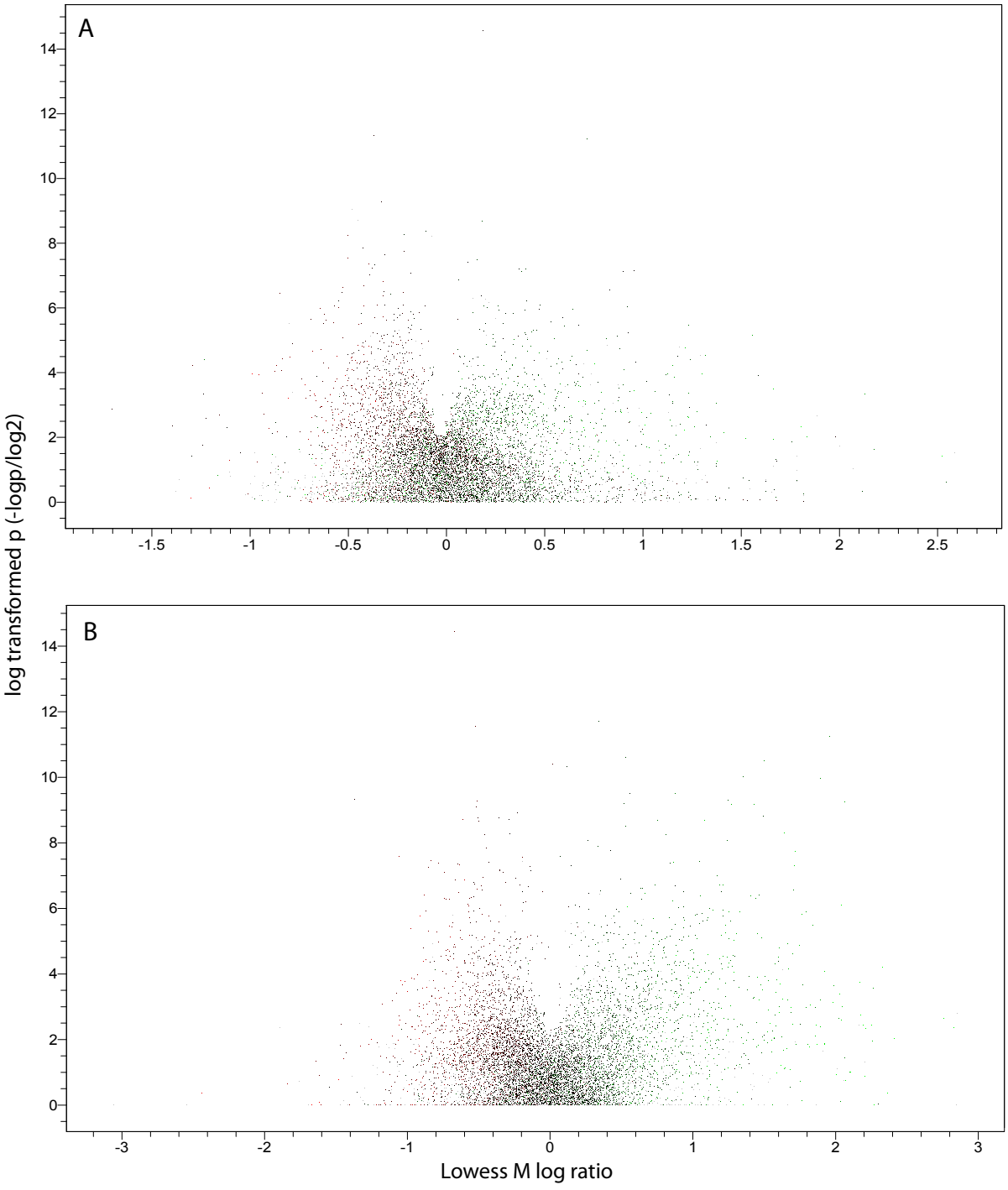

Acetate-20
Acetate-40
Acetate-360
Acetate-720
DOM-20
DOM-40
DOM-360
DOM-720
1795
3941
2220

Figure 5.6. The negative log transformed *p*-values of the Student's t-test plotted against the Lowess normalized log ratio data (*M*) of the DOM-treated samples at 40 min (A) and 720 min (B) averaged for the triplicate arrays. Significantly upregulated probes were those with a fold change >2 (|M|>1) and a *p*-value <0.05.


A
B
log transformed p (-logp/log2)
Lowess M log ratio
14
12
10
8
6
4
2
0
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
-3
-2
-1
0
1
2
3

CHAPTER 6

SUMMARY

The purpose of the studies presented here was to access marine microbial transcriptomes in order to better understand the roles of bacteria in marine systems. The relevance of this work is demonstrated by the known abundance (7), diversity (5), and ecosystem importance of marine bacteria (1), but the deficit of information on patterns of and controls on microbial activity. Two approaches were used to capture bacterial gene expression: 1) environmental transcriptomics involving the direct isolation and analysis of mRNA from the environment and 2) whole genome microarrays, which offer the opportunity to examine the presence and expression of multiple genes of a specific organism simultaneously.

Recent metagenomic analyses of bacterioplankton in the open ocean have revealed a wide suite of genomic potential activities (2, 6, 8), yet there have been few *in situ* studies involving bacterioplankton gene transcripts. Environmental transcriptomics, or metatranscriptomics, provides a way to survey an intact community for gene expression without the constraints of targeting a specific organism, phylogenetic group, or metabolic pathway. Initial efforts were successful in developing a technique for analyzing environmental transcriptomes by creating clone libraries using random primers to reverse-transcribe and amplify environmental mRNAs. This approach was applied to two different systems, Sapelo Island and Mono Lake, and provided evidence that these communities were actively incorporating and metabolizing labile components of the dissolved organic matter pool (amino acids, peptides, carbohydrates, and alcohols) (Table 2.4), metabolizing inorganic and C1 compounds (Table 2.3), and protecting themselves against antimicrobial agents (Table 2.3). The

putative transcripts with no known function (12% with matches to conserved genes, 35% with no matches) were of particular interest. This collection of environmental transcripts provided novel material for environmental microarrays and quantitative PCR primer sets to investigate patterns of activity in natural microbial assemblages.

Recognizing the biases involved with the selection of the random primers used to initiate cDNA synthesis in the first approach, an improved technique was developed that involves linear amplification of mRNA (by polyadenylating the mRNA and carrying out *in vitro* transcription) followed by synthesis of double stranded cDNA with random hexamers. This environmental transcriptomics approach was used to elucidate day/night differences in gene expression in surface waters of the North Pacific subtropical gyre (3). Analysis of transcripts from this system provided verification of processes known to be differentially expressed over a diel cycle, such as photosynthesis, oxidative phosphorylation, and synthesis of light-driven cellular machinery such as proteorhodopsins and photosynthetic pigments. Other processes emerged from the comparative analyses to exhibit a diel signal, both for global analyses of the community transcriptome and for individual autotrophic and heterotrophic taxonomic bins. The snapshot of bacterioplankton gene expression that emerged from the transcriptomes showed activity relevant to major biogeochemical processes, including uptake and utilization of nitrogen, carbon, sulfur and phosphorus compounds (Table 4.6).

The results of both of these studies demonstrate the potential for metatranscriptomics to elucidate dominant active metabolic processes within bacterioplankton assemblages. By examining patterns of microbial gene expression, we can further our understanding and improve predictive modeling of environmental controls on ecologically relevant processes in the ocean.

Finally, a whole genome microarray approach was used to monitor the transcriptome of *Silicibacter pomeroyi*, a member of the abundant Roseobacter clade (4), in response to complex DOM of known origin, i.e. exudate from an axenic culture of the marine diatom *Skeletonema costatum*, with the expectation that this model system could provide insights into similar interactions *in situ*. Several *S. pomeroyi* genes were upregulated in the presence of diatom DOM, including those involved in transport and utilization of amino acids, protocatechuate catabolism, and transcriptional regulation. These results provide a novel method for examining bacterial-phytoplankton associations on the level of gene expression and have implications for our understanding of carbon cycling between phytoplankton and bacteria in the marine microbial food web.

REFERENCES

1. Azam, F. 1998. Microbial Control of Oceanic Carbon Flux: The Plot Thickens. Science 280:694-696.
2. DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N.-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. 2006. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science 311:496-503.
3. Karl, D. M., and R. Lukas. 1996. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. Deep Sea Research Part II: Topical Studies in Oceanography 43:129-156.
4. Moran, M. A., A. Buchan, J. M. Gonzalez, J. F. Heidelberg, W. B. Whitman, R. P. Kiene, J. R. Henriksen, G. M. King, R. Belas, C. Fuqua, L. Brinkac, M. Lewis, S. Johri, B.

Weaver, G. Pai, J. A. Eisen, E. Rahe, W. M. Sheldon, W. Ye, T. R. Miller, J. Carlton, D. A. Rasko, I. T. Paulsen, Q. Ren, S. C. Daugherty, R. T. Deboy, R. J. Dodson, A. S. Durkin, R. Madupu, W. C. Nelson, S. A. Sullivan, M. J. Rosovitz, D. H. Haft, J. Selengut, and N. Ward. 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature 432:910-913.

5. Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. USA 103:12115-12120.

6. Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. W. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.

7. Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. Prokaryotes: The unseen majority. Proc. Natl. Acad. Sci. USA 95:6578-6583.

8. Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J.-M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter. 2007. The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. PLoS Biol. 5:e16.