ECOLOGICAL INSIGHTS INTO MARINE MICROBIAL COMMUNITIES VIA EXPRESSION ANALYSES

by

SCOTT MICHAEL GIFFORD

(Under the Direction of Mary Ann Moran)

ABSTRACT

In the oceans, the transfer of energy and cycling of elements is predominantly controlled by bacterioplankton, such that any understanding of marine ecosystems requires knowledge about bacterial activities and functional capabilities. Metatranscriptomics, the direct retrieval and sequencing of environmental RNA, is a powerful tool that can identify active community members and their expressed functional capabilities. This dissertation is composed of three studies that used metatranscriptomics to gain fundamental insights into the ecology and biogeochemistry of coastal microbial communities. In the first study, an internal standard approach was developed to make absolute (per liter) estimates of transcript numbers, a significant advantage over proportional estimates. Expression levels of genes diagnostic for transformations in the marine nitrogen, phosphorus and sulfur cycles were determined, as well as the total size of the mRNA pool. By representing expression in absolute units, metatranscriptomics extends beyond relative comparisons, allowing for direct comparisons with other biogeochemical measurements. In the second study, a metatranscriptomic dataset revealed an unexpected abundance of transcripts to '*Candidatus* Nitrosopumilus maritimus', an ammonia oxidizing Archaea whose presence has significant implications in the carbon and nitrogen cycles.

Reads assigned to genes for ammonia uptake and oxidation accounted for 37% of *N. maritimus* transcripts. In contrast, transcripts from co-occurring ammonia oxidixing Bacteria were in much lower abundance, with no transcripts related to ammonia oxidation or carbon fixation. This study suggests that these two members of the ammonia oxidizing functional guild respond differently to the same environmental cues. The third study used metatranscriptomics to examine how differences in expression among taxa can be indicative of niche diversification. The sequencing of transcripts from four coastal bacterial communities revealed the expression and activity of thousands of different taxa. The genes carried by these taxa have extensive overlap, and the majority of highly expressed genes were for redundant functions. To identify unique ecological roles for these taxa, a method was developed to classify genes both by their expression level and their frequency in genomes. The results show clear functional delineations across broad phylogenetic groupings and provide insights into the diversity of lifestyle strategies that supports complex microbial assemblages.

INDEX WORDS: Marine, Ocean, Biogeochemistry, Expression, Metatranscriptomics, Ecology, Niche, Microbial communities, RNA, Bacteria, Plankton

# ECOLOGICAL INSIGHTS INTO MARINE MICROBIAL COMMUNITIES VIA EXPRESSION ANALYSES

by

SCOTT MICHAEL GIFFORD

B.S., The Ohio State University, 2003

M.S., San Francisco State University, 2006

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

2011

# ECOLOGICAL INSIGHTS INTO MARINE MICROBIAL COMMUNITIES VIA EXPRESSION ANALYSES

by

SCOTT MICHAEL GIFFORD

| | |
|---|---|
| Major Professor: | Mary Ann Moran |
| Committee: | James T. Hollibaugh |
| | William B. Whitman |
| | Melissa G. Booth |
| | Charles S. Hopkinson |

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2011

ACKNOWLEDGEMENTS

I'm especially indebted to all of the Marine Science students and staff that helped with the SIMO field work. This dissertation would not be possible without the time they generously volunteered. I have to thank my advisor, Dr. Mary Ann Moran, for her guidance and tireless patience. Her enthusiasm and work ethic are infectious, both of which have deeply shaped my interest in marine microbial ecology. My committee members, Drs. Tim Hollibaugh, Charles Hopkinson, Barny Whitman, and Melissa Booth, were instrumental in shaping this work. I would also like to thank all the past and present members of the Moran lab, whose advice, support, and friendship created an ideal environment for conducting research over the last five years. Finally, I would like to thank my wife Alecia, who in addition to providing emotional support and encouragement, sets a standard for scientific excellence that I am continuously striving to emulate.

TABLE OF CONTENTS

Page

# CHAPTER 1

## INTRODUCTION AND LITERATURE REVIEW

Of the major advances in ecology in the last century, the recognition that microbes are central to ecosystem function is arguably one of the most important. Nowhere is this more apparent than in the oceans (Pomeroy, 1974; Azam *et al.*, 1983) where the sheer abundance of marine bacteria (Whitman *et al.*, 1998) is strong evidence for their influence on energy transfer and elemental cycling. Indeed, bacteria are both the dominant primary producers (Chisholm *et al.*, 1988) and the dominant consumers of fixed carbon in the ocean (Azam, 1998) and have a major role in the sequestration and turnover of inorganic nutrients. Their genetic plasticity and physiological diversity allows them to occupy every imaginable niche and facilitates their interactions with members of the community ranging from fellow bacteria to large metazoans. Given their influence, understanding the genetic, ecological, and evolutionary forces that shape and control bacterial activities is critical to understanding ecosystem structure and function.

However, much of our current knowledge about bacterial activities is derived from bulk measurements of mixed species communities and cannot distinguish the identity or functional roles of individual taxa. This obscures details about the underlying processes that lead to emergent ecosystem properties; limiting our ability to extend insights gained in one system to another and making it difficult to predict how systems will respond to natural or anthropogenic perturbations.

An inherent difficulty in microbial taxonomy is a lack of morphological markers that allow individual cells to be grouped into species whose activities can be characterized. However, molecular markers have proved to be a powerful alternative and, since the late 1980's, have revolutionized the field of microbial ecology. Ribosomal RNA sequences have been used to catalog the taxonomic composition and phylogenetic relatedness of members of microbial communities (Giovannoni *et al.*, 1990; Amann *et al.*, 1995), taking the first steps in revealing *in situ* microbial populations. This allowed for the targeted isolation and cultivation of environmentally relevant organisms, whose phenotypic and genotypic properties could be examined directly in the laboratory (Gonzalez *et al.*, 2000; Rappe *et al.*, 2002; Moran *et al.*, 2004; Giovannoni *et al.*, 2005). Recent 16S rRNA surveys, particularly those using next generation sequencing technologies, have demonstrated that most microbial assemblages are composed of thousands of different taxa (Sogin *et al.*, 2006; Huber *et al.*, 2007 ), with only a small minority having relatives in culture. This high taxonomic richness raises questions as to how so many different types of bacteria can exist simultaneously (Hutchinson, 1961) and what biogeochemical roles each plays. Metagenomics can provide some insight into these questions by simultaneously examining the distribution of taxa and their potential functional mechanisms in a culture independent manner (Venter *et al.*, 2004; Rusch *et al.* 2007), but it can say nothing about which of these mechanisms are actively being expressed. Metatranscriptomics, the direct retrieval and sequencing of environmental RNA, looks at only those genes expressed in response to immediate environmental conditions and thus targets only active taxa and their realized functional capabilities.

The potential for metatranscriptomics to be a powerful tool for microbial community analysis has been demonstrated by several previous studies. Poretsky *et al.*, (2005) conducted the

first metatranscriptomic analysis using clone libraries and Sanger sequencing, identifying several hundred protein encoding transcripts and demonstrating the ability of the technique to capture environmentally representative transcripts, and identify expression of biogeochemically diagnostic genes. Since that time, advances in next generation sequencing technologies have increased sequencing capabilities by more than six orders of magnitude (Margulies *et al.*, 2005), providing an improvement in coverage of transcript pools (Stewart *et al.*, 2010). The first studies that used the new sequencing capabilities were largely descriptive, examining relative differences in community transcription segregated either temporally (Gilbert *et al.*, 2011; Poretsky *et al.*, 2009) or spatially (Frias-Lopez *et al.*, 2008; Hewson *et al.*, 2010; Shi *et al.*, 2011; Stewart *et al.*, 2011b) in a variety of open ocean and coastal habitats. However, metatranscriptomics can extend beyond purely descriptive studies to analyze differences in expression in experimental manipulations, such as dissolved organic matter (DOM) additions (Poretsky *et al.*, 2010; Mckarin *et al.*, 2010; Vila-Costa *et al.*, 2010) or artificially induced phytoplankton blooms (Gilbert *et al.*, 2008). The method has also illuminated the abundant expression of small, regulatory RNAs in microbial communities (Shi *et al.*, 2009) as well as provided insights into the molecular evolution of microbial genes (Stewart *et al.*, 2011a).

These studies have raised several methodological and ecological questions about expression in microbial communities, particularly in relation to quantification, coverage, and niche diversification. Because the total number of transcripts in any given sample is typically not known and only a small subsample is sequenced, analyses are limited to relative interpretations of read counts (i.e., as percent of transcriptome) rather than more environmentally meaningful units (i.e., per volume or mass). This has limited direct comparisons to other biogeochemical measurements, such as standing stocks or activity rates. Furthermore, with no knowledge of the

total transcript pool size, it is not possible to determine how deeply a sample is sequenced. An examination of collector's curves from several studies (Poretsky *et al.*, 2009; Stewart *et al.*, 2010) suggest that coverage of the community transcriptome is very low, and that the transcripts sequenced most often are those encoding basic cellular machinery that is shared among most taxa (Hewson *et al.*, 2009). Yet the simultaneous presence of so many active taxa suggests that the dominance of cellular housekeeping transcripts is likely obscuring less abundant transcripts for ecological or biogeochemical processes that may be indicative of niche diversification. Furthermore, for those taxa that do share a substantial amount of biogeochemical functional overlap, expression analysis has yet to resolve the heterogeneity in their responses to shared environmental conditions that may lead to niche separation between members of the same functional guild.

**Chapter Overview**

The aim of this dissertation is to examine the activities and functional organization of microbial communities via expression analysis, with a specific emphasis on using metatranscriptomics to make quantitative insights into biogeochemical activities and niche diversification. The focus of the research spans several levels of ecological organization, from characterizing the basic properties of mRNA pools, to the analysis of individual bacterial populations, and finally to a holistic examination of a community assemblage. The work will address the following broad questions:

**Question 1: Can an internal standard approach be used to make metatranscriptomic interpretations quantitative?**

**a) How deeply is the transcript pool sequenced?**

**b) What is the environmental concentration of transcripts for diagnostic biogeochemical genes?**

**c) Are current sequencing strategies sufficient to detect differences in gene transcription between samples?**

Previous metatranscriptomic studies have been limited to relative interpretations due to variations in sequencing depth between samples. In chapter 2, a quantitative approach to metatranscriptomics was developed based on the addition of an internal RNA standard. Two replicate samples from the Sapelo Island Microbial Observatory (SIMO) time series were processed for this analysis and sequenced with two 454 pyrosequencing runs each, producing over 2 million reads. Based on the sequence coverage of the internal standard added, we estimated the size of the transcript pool in each sample and how deeply it was covered. Sequence counts could then be transformed to concentrations in the original seawater sample, a more meaningful ecological unit that allowed for direct comparisons across different samples and with other biogeochemical measurements. The concentration of transcripts for 82 diagnostic genes of the marine nitrogen, phosphorus, and sulfur cycles was determined, as well as the ability to make statistical distinctions in their abundance between the two replicate samples.

Appendix A gives a detailed description of the quantitative metatranscriptomic protocol, including methods for sample collection, internal standard addition, custom rRNA reduction,

linear amplification, and cDNA synthesis. Considerations for designing internal standard sequences, as well as their construction are also given.

**Question 2: What information can metatranscriptomic analyses provide about the *in situ* distribution and ecology of marine ammonia oxidizing archaea?**

**a) How does the presence and activity of ammonia oxidizing populations vary with time?**

**b) What can transcripts tell us about the metabolism of these populations?**

**c) What external factors regulate competition between marine ammonia oxidizing archaea and other members of the same functional guild?**

One of the advantages of metatranscriptomics is that it requires no prior information on the community, offering the potential to make insights into the functional activities of taxa not previously known to be important to a system. Chapter 3 examines the unexpected presence and activity of ammonia oxidizing archaea in coastal waters of the southeastern U.S. The importance of archaea to the subeuphotic regions of the open ocean is well recognized, and though they have been examined in coastal sediments, they were not thought to be abundant or active in the water column of shallow coastal ecosystems. Metatranscriptomic analysis of a summer SIMO sample revealed an abundance of transcripts with high sequence homology to the Thaumarchaeota *Nitrosopumilus maritimus*, and qPCR analysis of samples collected throughout 2008 showed a distinct increase in their distribution in the summer sample. Transcript sequences covered almost half the genes in the *N. maritimus* genome and were particularly enriched for genes related to ammonia oxidation. In contrast to the archaea, transcript abundances for ammonia oxidizing

bacteria were consistently low. The analysis thus provided insight into the physiological and environmental factors that influence archaeal distributions in coastal waters, as well as the potential factors leading to niche segregation among members of the ammonia oxidizing functional guild.

**Question 3: Can expression analysis reveal the distinct ecological roles of taxonomic groups within the bacterioplankton?**

a) **Can expression data be used as a proxy for relative activity among groups?**
b) **Is there a difference in expression levels of genes shared among many taxa versus those shared among few?**
c) **Is niche diversification driven by differences in expression of shared functional capabilities or by unique functional capabilities?**

In contrast to macroorganism communities, where obvious delineations typically exist in functional roles among community members, the divisions of niche space within microbial communities has remained elusive. This is due in large part to the functional redundancy found in microbial genomes and metagenomes. In chapter 4, metatranscriptomics is used to look at expression patterns across phylogenetically distinct microbial groups sharing the same environment in order to identify their unique ecological niches. Illumina sequencing of four seasonal samples from the SIMO time series produced over 11 million protein-encoding reads, providing a robust view of the gene expression of hundreds of different taxa. The relative abundance of ribosomal proteins among the top 200 reference bins showed distinct differences in activity among taxa. A strong positive relationship between expression level and functional

redundancy provided a method of examining genes with atypically high expression within an ortholog group, which were indicative of genes for niche specialization. The results show clear delineations in gene expression patterns across broad taxonomic groupings and provide insights into the diversity of ecological strategies that characterize a complex microbial assemblage.

## References


Amann RI, Ludwig W, Schleifer KH (1995). Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiol Rev* **59:** 143-169.

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

Azam F (1998). Microbial control of oceanic carbon flux: The plot thickens. *Science* **280:** 694-696.

Chisholm SW, Olson RJ, Zettler ER, Goericke R, Waterbury JB, Welschmeyer NA (1988). A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334:** 340-343.

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

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

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B *et al.* (2011). The taxonomic and functional diversity of microbes at a temperate coastal site: A 'multi-omic' study of seasonal and diel temporal variation. *Plos One* **5**(11): e15545.

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

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:** 1242-1245.

Gonzalez JM, Simo R, Massana R, Covert JS, Casamayor EO, Pedros-Alio C *et al.* (2000). Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl Environ Micro* **66:** 4237-4246.

Hewson I, Poretsky RS, Dyhrman ST, Zielinski B, White AE, Tripp HJ *et al.* (2009). Microbial community gene expression within colonies of the diazotroph, Trichodesmium, from the Southwest Pacific Ocean. *ISME-J* **3:** 1286-1300.

Hewson I, Poretsky RS, Tripp HJ, Montoya JP, Zehr JP (2010).Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ Microbiol* **12:** 1940-1956.

Huber JA, Mark Welch D, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318:** 97-100.

Hutchinson GE (1961). The paradox of the plankton. *Am Nat* **95:** 137-145.

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

McCarren J, Becker JW, Repeta DJ, Shi YM, Young CR, Malmstrom RR *et al.* (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *P Natl Acad Sci USA* **107:** 16420-16427.

Moran MA, Buchan A, Gonzalez JM, Heidelberg JF, Whitman WB, Kiene RP *et al.* (2004). Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432:** 910-913.

Pomeroy LR (1974). The ocean's food web, a changing paradigm. *Bioscience* **24:** 499-504.

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M *et al.* (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Micro* **71:** 4121-4126.

Poretsky RS, Hewson I, Sun SL, Allen AE, Zehr JP, Moran MA (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11:** 1358-1375.

Poretsky RS, Sun S, Mou X, Moran MA (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* **12:** 616-627.

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

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *Plos Bio* **5:** 398-431.

Shi YM, Tyson GW, DeLong EF (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459:** 266-U154.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *P Natl Acad Sci USA* **103:** 12115-12120.

Stewart FJ, Ottesen EA, DeLong EF (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME-J*. **4**:896-907

Stewart F, Sharma A, Bryant J, Eppley J, DeLong EF (2011a). Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol* **12:** R26.

Stewart FJ, Ulloa O, DeLong EF (2011b). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol***:** 10.1111/j.1462-2920.2010.02400.x

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Vila-Costa M, Rinta-Kanto JM, Sun SL, Sharma S, Poretsky R, Moran MA (2010) Transcriptomic analysis of a marine bacterial community enriched with dimethylsulfoniopropionate. *ISME-J* **4:** 1410-1420.

Whitman WB, Coleman DC, Wiebe WJ (1998). Prokaryotes: The unseen majority. *P Natl Acad Sci USA* **95:** 6578-6583.

Shi Y, Tyson GW, Eppley JM, DeLong EF (2011) Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME-J* **5**: 999-1013

# CHAPTER 2

# QUANTITATIVE ANALYSIS OF A DEEPLY-SEQUENCED MARINE MICROBIAL METATRANSCRIPTOME[1]

[1]Gifford, S.M., Sharma, S., Rinta-Kanto, J.M, and Moran, M.A. 2011. *ISME-J* 5:461-472 doi:10.1038/ismej.2010.141; Reprinted here with permission of the publisher

**Abstract**

The potential of metatranscriptomic sequencing to provide insights into the environmental factors that regulate microbial activities depends on how fully the sequence libraries capture community expression (i.e., sample sequencing depth and coverage depth), and the sensitivity with which expression differences between communities can be detected (i.e. statistical power for hypothesis testing). Here we use an internal standard approach to make absolute (per liter) estimates of transcript numbers, a significant advantage over proportional estimates that can be biased by expression changes in unrelated genes. Coastal waters of the southeastern U.S. contain 1 x $10^{12}$ bacterioplankton mRNA molecules per liter of seawater (~200 mRNA molecules per bacterial cell). Even for the large bacterioplankton libraries obtained here (~500,000 possible protein-encoding sequences in each of two libraries after discarding rRNAs and small RNAs from >1 million 454 FLX pyrosequencing reads), sample sequencing depth was only 0.00001%. Expression levels of 82 genes diagnostic for transformations in the marine nitrogen, phosphorus, and sulfur cycles ranged from below detection (<1 x $10^{6}$ transcripts $L^{-1}$) for 36 genes (e.g., phosphonate metabolism gene *phnH*, dissimilatory nitrate reductase subunit *napA*) to >2.7 x $10^{9}$ transcripts $L^{-1}$ (ammonia transporter *amt* and ammonia monooxygenase subunit *amoC*). Half of the categories for which expression was detected, however, had too few copy numbers for robust statistical resolution, as would be required for comparative (experimental or time-series) expression studies. By representing whole community gene abundance and expression in absolute units (per volume or mass of environment),"omics" data can be better leveraged to improve understanding of microbially-mediated processes in the ocean.

## Introduction

Metatranscriptomics is a powerful tool for capturing gene expression patterns in natural microbial communities without prior assumptions as to the ongoing activities or dominant taxa (Frias-Lopez *et al.*, 2008; Poretsky *et al.*, 2005). In contrast to metagenomics, which provides an inventory of the community gene pool, metatranscriptomics identifies which of those genes are being transcribed in a given ecological context, including under experimentally manipulated conditions (Gilbert *et al.*, 2008; Poretsky *et al.*, 2010).

The advent of second generation sequencing has increased metatranscriptome library sizes by orders of magnitude (Frias-Lopez *et al.*, 2008; Hewson *et al.*, 2009a; Poretsky *et al.*, 2005; Poretsky *et al.*, 2009b; Urich *et al.*, 2008), yet how deeply a community transcriptome is "covered" by the sequence library remains a critical issue. If too shallow, libraries will be dominated by transcripts from metabolic pathways shared by most cells and poor in those representing specialized biogeochemical pathways (Hewson *et al.*, 2009b; Poretsky *et al.*, 2009b). As a consequence, unique expression patterns within a community may be missed, and comparative analyses between communities can be insensitive.

Variability in sample sequencing depth between community metatranscriptomes, regardless of coverage level, further limits the power of comparative analyses by restricting assessments to relative data (i.e., as a proportion of the transcriptome; Fig. 2.1). This is problematic because changes occurring in the abundance of some mRNAs in response to shifting conditions (Hannah *et al.*, 2008; Robinson and Oshlack, 2010; van de Peppel *et al.*, 2003) leads to changes in the percent representation of other mRNAs whose absolute abundance has not changed (Fig. 2.1). Comparative analyses based on ratios of mRNA copies to DNA copies [i.e., relative abundance in metatranscriptomes vs. relative abundance in metagenomes; (Frias-Lopez

*et al.*, 2008)] doesn't solve this problem, since both are similarly affected by unknown and possibly different sample sequencing depths. Thus as currently obtained, metatranscriptomic data provide information on enrichment or depletion of a transcript category in the community transcriptome, but not on the absolute abundance of these transcript categories per volume or mass of environment (Fig. 2.1), which is the most relevant comparison for biogeochemical studies and ecosystem modeling. Metagenomic-, metaproteomic-, and environmental microarray-based studies suffer these same proportional data constraints.

Here we report the deep sequencing of two replicate metatranscriptomes from southeastern U.S. coastal seawater to characterize microbial gene expression and address three critical questions about sequencing effort: 1) What was the sample sequencing depth for the bacterioplankton community transcriptome? We used an internal mRNA standard to estimate the number of transcripts in the natural sample compared to the number of transcripts sequenced. 2) What was the abundance of transcripts representing key bacterial transformations in the nitrogen, phosphorus, and sulfur cycles in coastal seawater? We used BLAST analysis normalized to internal standard recovery to estimate absolute transcript numbers for over 80 diagnostic steps in marine elemental cycles. 3) Was the sequencing strategy sufficient to detect differences in gene transcription between samples? We examined the effect of coverage depth on statistical comparisons of biogeochemically-diagnostic transcripts in the metatranscriptomes.

## Methods

**Sample Collection.** Two replicate seawater samples (FN56 and FN57) were collected at Marsh Landing, Sapelo Island, Georgia, U.S. (31° 25' 4.08 N, 81° 17' 43.26 W; www.simo.marsci.uga.edu) on 8 August 2008 at 2330 h local time, 1 h before high tide and 3 h

after sunset. These samples are part of a multi-year time series of the Sapelo Island Microbial Observatory (www.simo.marsci.uga.edu), in which collections are made every three months and each collection set consists of duplicate samples from 4 consecutive high tides (2 day and 2 night). Surface water (5.75 L from a depth of 0.5 m) was pumped directly through 3 μm and 0.22 μm filters. The 0.22 μm filter was placed in a Whirl-Pak® bag (Nasco, Fort Atkinson, WI) and immediately flash frozen in liquid $N_2$. Total time from the start of filtration to freezing was ten minutes. While the samples are considered biological replicates within the larger time series, we note that there was eight minutes between the end of the first collection (sample FN56) and the start of the second (sample FN57). Nutrient data are collected monthly at station GCE6 (~3 km from Marsh Landing) as part of the Georgia Coastal Ecosystems Long Term Ecological Research program (GCE-LTER; http://gce-lter.marsci.uga.edu).

**RNA Processing and Sequencing.** RNA processing in preparation for pyrosequencing was done as previously described (Poretsky *et al.*, 2009a; Poretsky *et al.*, 2009b), with the exception of the addition of an *in vitro* transcribed standard to the extraction tube before beginning the extraction. The standard was constructed by linearizing a pGem-3Z plasmid (Promega, Madison, WI) with ScaI restriction enzyme (Roche, Penzberg, Germany) and cleaned with a phenol:chloroform:isoamyl alcohol extraction. Complete digestion of the plasmid was confirmed with a 1% agarose gel. The DNA fragment was then *in vitro* transcribed using the Riboprobe in vitro Transcription System (Promega, Madison, WI) according to manufacturer's protocol; an SP6 RNA polymerase was used to create a 994 nt long RNA fragment. The pGem plasmid had another internal T7 promoter region, but it was present in the reverse complement sequence during *in vitro* transcription and aRNA amplification (see below) and did not interfere. Residual DNA was removed with RQ1 RNase-Free DNase, and the RNA was cleaned with a

phenol:chloroform:isoamyl alcohol extraction. The RNA standard was quantified with a Nanodrop spectrophotometer (Thermo Scientific, Wilmington, DE), and correct fragment size was confirmed with an Experion automated electrophoresis system (Bio-Rad, Hercules, CA, USA).

Twenty five nanograms (4.7 X $10^{10}$ copies) of the RNA standard was added to a 50 ml conical tube containing 8 ml RLT lysis buffer (Qiagen) and 3 g of RNA PowerSoil beads (Mo-Bio, Carlsbad, CA). The sample filters were removed from -80°C storage, shattered, and added to the extraction tubes. RNA was then extracted using an RNEasy kit (Qiagen, Valencia, CA, USA), and any residual DNA was removed using the Turbo DNA-free kit (Applied Biosystems, Austin, TX, USA). In order to reduce the number of rRNAs in the pyrosequencing reads, total RNA was treated in two ways to enrich for mRNA. Epicentre's mRNA-Only isolation kit (Epicentre, Madison, WI, USA) was first used to decrease rRNA contamination enzymatically. The samples were then treated with MICROB*Express* and MICROB*Enrich* kits (both from Applied Biosystems) which couple an oligonucleotide rRNA probe with magnetic separation to enrich for mRNA. Successful reduction of rRNA was confirmed by running both pre- and post-treated samples on an Experion automated electrophoresis system (Bio-Rad, Hercules, CA, USA). In order to obtain enough mRNA for pyrosequencing, the samples were linearly amplified using the MessageAmp II-Bacteria kit (Applied Biosystems). The amplified RNA was then converted to cDNA using the Universal RioboClone cDNA synthesis system with random primers (Promega, Madison, WI, USA), which produced cDNAs primarily in the size range of 200 to 600 bp. Residual reactants and nucleotides from cDNA synthesis were removed from the sample using the QIAquick PCR purification kit (Qiagen), and gel-based size selection was used to select fragments in the 250 bp to 500 bp range. cDNAs from each replicate sample were

loaded into ½ of each of four GS-FLX plates for 454 pyrosequencing. Sequences are deposited in the CAMERA Database (http://camera.calit2.net/about-camera/full-datasets) under accession name "CAM_PROJ_Sapelo2008".

**Read Annotation.** Duplicate clusters were identified using an online program (Gomez-Alvarez *et al.*, 2009). Ribosomal RNA sequences were identified with a BLASTn search against the small and large subunit SILVA database (http://www.arb-silva.de) with a bit score cutoff ≥ 50; sequences identified as rRNA were then removed from further consideration. To identify small, non protein encoding RNAs (Shi *et al.*, 2009), all non ribosomal reads were compared to the RFam database (http://rfam.janelia.org) using BLASTn with a bitscore ≥40, and hits were considered putative small RNAs (psRNAs) if the best hit in the RefSeq database was a hypothetical protein or if the RFam alignment was ≥ 95 nt (Fig. 2.S1).

Remaining reads were annotated using BLASTx searches against the NCBI RefSeq and Clusters of Orthologous Genes (COG) databases (Tatusov *et al.*, 2003) with a bit score cutoff ≥ 40. Taxonomic binning was based on RefSeq hit. Collector's curves were produced from a custom script in the R environment (R Core Development Team, 2009). Read coverage of proteorhodopsin PU1002_03206 gene bin and the internal standard was assessed by assembling reads against the reference sequence using Geneious version 4.8 (Biomatters Ltd., Auckland) with gaps and default scoring (word length = 18, max gap size = 1, max gaps per read = 20, max mismatches = 20, and max ambiguities = 4) and the consensus sequence representing the majority nucleotide at each position. Sequence variation at each nucleotide position was determined using a custom script in R with the BioStrings package (Pages *et al.*, 2009).

**Elemental Cycle Transcripts.** Reference diagnostic genes representing transformations in the N, P, and S cycles were selected from marine Alphaproteobacteria, Gammaproteobacteria,

and Bacteriodetes genomes (the three most common taxa in marine metagenomic libraries), or from other taxa if these three groups did not contain an ortholog to the gene of interest. These reference sequences were used as query sequences in BLASTx analysis against the metatranscriptomic data (bitscore $\geq 40$, E-value $\leq 10^{-3}$) and redundant hits were removed. Remaining hits were manually checked with BLASTx against the RefSeq database and discarded if the top 3 hits were not to a similar annotation as the original reference gene.

**Statistics.** Pairwise statistical comparisons were carried out with Xipe, a bootstrapped difference of means calculation developed by Rodriquez-Brito et al. (2006), using 20,000 bootstrap iterations and 95% confidence intervals, or with 2 x 2 contingency tables and the Fisher's Exact Test (White *et al.*, 2009) using $p < 0.05$. Subsampled libraries for Xipe analyses were created by sampling without replacement using R (R Core Development Team, 2009). The Benjamini-Hochberg correction was used to adjust the Fisher's Exact Test p-values as a control for the False Discovery Rate (FDR) using the R package "multtest" (Strimmer, 2008) and only those genes with an adjusted p-value $< 0.05$ were considered significant. A simulation analysis of Fisher's Exact Test significance threshold as a function of count number was carried out using an R script that ran 2 X 2 contingency tables at incrementing count values for library sizes of 125,000 reads.

## Results and Discussion

**Sequence Libraries.** cDNAs derived from two replicate coastal bacterioplankton samples [samples FN56 and FN57 in the Sapelo Island Microbial Observatory series (http://simo.marsci.uga.edu)] were sequenced in four GS-FLX 454 runs (Margulies *et al.*, 2005), with four technical sequencing replicates per biological replicate. Over a million reads averaging

210 nt in length were obtained per sample (Table 2.1). After removal of rRNAs and putative small RNAs [psRNAs; (Shi *et al.*, 2009)], there were ~500,000 possible protein encoding reads in each library (Table 1).

**Sample Sequencing Depth.** Sample sequencing depth is defined here as the percent of mRNA molecules present in a sample that is represented in the sequence library. The greater the sequencing depth of an mRNA pool, the more thorough the representation of microbial gene transcription. Further, if the volume or weight of the sample is also known, information on the sample sequencing depth allows absolute transcript abundance to be calculated for a given quantity of the environment, not just proportional abundance in the community transcriptome. To estimate sample sequencing depth, a known number of artificial RNA sequences serving as an internal standard was added immediately prior to cell lysis at the initiation of nucleic acid extraction. This approach may have some biases, for example if the internal standard is more susceptible to degradation than natural mRNA or if the efficiency of release of natural mRNA from cells is less than 100%, but it provides a consistent accounting across samples through extraction, processing, and sequencing steps. Similar approaches have been successfully applied to qPCR (Coyne *et al.*, 2005) and microarray studies (Hannah *et al.*, 2008).

A total of 4,014 internal standards were identified in the FN56 sequence library out of 4.7 x $10^{10}$ copies added prior to cell lysis, leading to an estimate of 1.0 x $10^{12}$ bacterioplankton mRNA molecules per L of coastal seawater (Table 2.2). Two other estimates of the size of the community transcriptome were derived using literature values for mRNA content of marine bacterioplankton (Table 2.2), and these were in reasonable agreement with the internal standard method. The sample sequencing depth was therefore ~0.00001%, or 1 in $10^{7}$ transcripts, with FN57 sequenced slightly deeper than FN56 (Table 2).

Direct cell counts indicated 4.2 x $10^9$ bacterioplankton cells $L^{-1}$ in the seawater samples, and therefore an average of 190 mRNA transcripts $cell^{-1}$ (Table 2.2). Laboratory cultures of *Escherichia coli* in exponential growth phase have ~1,400 transcripts $cell^{-1}$ (Neidhardt and Umbarger, 1996). The 7-fold lower estimate for coastal bacterioplankton was not unexpected, however, because the cells are considerably smaller in size (Azam and Hodson, 1977) and have much lower growth rates (Ducklow, 2000) than laboratory-grown *E. coli*. Based on this per cell abundance estimate, it can be deduced that transcript copy number was lower than gene copy number for most of the bacterial and archaeal genes present in this coastal ocean.

**Coverage Depth.** Coverage depth is defined here as the percent of the unique mRNAs present in a sample that is represented in the sequence library. Sample sequencing depth and coverage depth are not strictly coupled, since a low richness/high evenness community transcriptome will be well covered even with shallow sample sequencing.

We evaluated coverage depth for the coastal metatranscriptomes in terms of taxa, functional gene categories, and genes. Taxonomic coverage, as assessed by a collector's curve of NCBI taxonomy bins at the species or strain level, was approaching saturation for the library (Fig. 2.2 - inset); indeed, 75% of the total taxonomic richness emerging from this analysis would have been discovered with <15% of the sequencing effort. Saturating coverage was also found for functional gene assignments based on best hits to the Clusters of Orthologous Genes (COG) database (Table 2.1); 75% of the total richness would have been found with <10% of the sequencing effort (Fig. 2.2-Inset). However, these coverage assessments are constrained by the composition of the reference database, since apparent richness can be no higher than the number of reference bins available for transcript assignment. We found 1,909 taxon bins represented in

the metatranscriptomic libraries out of 8,054 entries in the NCBI taxonomy database, and 3,298 COGs out of 5,666 entries in the COG database.

When coverage was assessed based on gene assignments in the RefSeq database (>6 million accession numbers), transcripts binned to over 168,000 different genes (Table 2.1), and the collector's curve indicated that the metatranscriptome library was far from saturating (Fig. 2.2). Singletons made up 59% of the sequences (Fig. 2.S2), and abundant transcripts (>10 hits · accession number$^{-1}$) and highly abundant transcripts (>100 hits · accession number$^{-1}$) composed only 3% and 0.5% of the library, respectively. While RefSeq binning could overestimate transcript richness (because identical transcripts bin to different reference genes due to differences in the region sequenced or because of sequencing errors), it also underestimates richness (because a variety of sequence variants bin to the same reference gene; Fig. 2.3). In any event, despite efforts to sequence more deeply than typical, our libraries exhibited the same low coverage that has been reported in previous metatranscriptomic analyses of marine bacterioplankton communities (Frias-Lopez *et al.*, 2008; Poretsky *et al.*, 2010; Stewart *et al.*, 2010).

Metatranscriptomes might be expected to have lower richness compared to metagenomes if expression is limited to a small fraction of the bacterial genome at any one time. In this case, they would also have higher coverage than metagenomes for the same size sequence library (Gilbert *et al.*, 2008). Yet for this coastal metatranscriptome, the distribution of hits per gene (Fig. 2.S2) did not indicate dominance by a limited number of highly transcribed genes (Fig. 2.2). Similarly, a synchronized clonal population of *Bacillus anthracis* expressed 40-80% of genes under all growth conditions tested (Passalacqua *et al.*, 2009), suggesting that the population's transcriptome was only slightly less rich than its genome. Even allowing for

significant advances in sequencing technology, the extremely low sample sequencing depth found here suggests that most natural community transcriptomes will continue to be undersampled.

**Microdiversity.** Assembly of transcripts from the most highly expressed genes (> 1,000 reads for some) revealed significant variation within reference bins (see Chapter 3). For example, the 2,259 reads that binned to the *P. ubique* HTCC1002 proteorhodopsin gene (PU1002_03206; 1 of 28 proteorhodopsin bins in the libraries) had high sequence diversity (Fig. 2.3). That this observed diversity was in fact real biological variation was substantiated by an assembly of the internal standard reads (Fig. 2.3) which indicated a mean sequencing error rate in this study of 3.7 (±7.4) per 1000 bp compared to a mean sequence variation rate of 97.6 (± 28.0) per 1000 bp for transcripts binning to PU1002_03206. Although high diversity in proteorhodopsin genes has been found previously in the ocean (Campbell *et al.*, 2008; Rusch *et al.*, 2007), the metatranscriptomic data revealed simultaneous expression of scores of microdiverse sequence variants. Transcriptome coverage estimates based on gene binning to the RefSeq database are considerable underestimates of the true sequence richness.

**Detection of biogeochemically informative transcripts.** We determined the absolute abundance of transcripts for key genes representing the phosphorus cycle (25 diagnostic genes), nitrogen cycle (50 genes), and sulfur cycle (7 genes) (Fig. 2.4). Transcripts were found for 56% of genes surveyed. Most P and S cycle transformations were represented by at least one transcript. N cycle expression was dominated by ammonia transporter and ammonia monooxygenase transcripts, which had the highest copy numbers of any gene category ($2.7 \times 10^9$ transcripts $L^{-1}$; Fig. 2.4); many other N cycle genes were not detected at all.

To examine detection of biogeochemically-diagnostic mRNAs in theoretically smaller libraries, the full metagenomic libraries were randomly subsampled *in silico* to generate subsets. The majority of the elemental cycle transcripts detected in the full libraries were still evident in smaller libraries. For example, >80% of the P-cycle related genes would have had at least one hit in a library $1/4^{th}$ the size (Fig. 2.S3).

**Statistical resolution.** Comparative metatranscriptomics seeks to differentiate transcript abundance between samples, for example, across natural environmental gradients or in response to experimental manipulations. We examined the statistical power of comparative analyses as a function of library size, starting first with broad categories of gene function as represented by COG assignments. Subsets of each of the replicate libraries were generated *in silico* and the fold-difference criteria (high abundance count / low abundance count) needed for statistically significant differences were compared using a resampling method based on difference of medians (Xipe; Rodriguez-Brito *et al.*, 2006) . Even for libraries $1/4^{th}$ of the original size, there was little effect on the fold-difference threshold required for a COG category to be considered significantly different between samples (Fig. 2.5). This was true as well for an alternate statistical approach using contingency tables and Fisher's Exact Test (White *et al.*, 2009) (Fig. 2.5-inset), and also when analyzing libraries much smaller than the original (for example, the average fold-difference threshold for significance in a 10,000 read library was <1% greater than in a 500,000 read library).

Library size, however, had a direct impact on the number of counts in a transcript category, thereby affecting the power of statistical comparisons. Transcript categories with low copy numbers (defined here as $\leq$15 hits in the lower abundance sample) required from 2- to 8-fold difference between the two samples for statistical significance (Fig. 2.5). For smaller *in*

*silico* subsets of the libraries or for more specific transcript annotation categories (e.g., RefSeq gene bins), both of which result in lower counts per category, the power to detect statistical differences between two samples decreased. For example, 17 out of the 25 genes that mediate key steps in the marine phosphorus cycle fell into a low count category even with the full-size library (Fig. 2.4), and nearly all would do so if the library was $1/4^{th}$ of the original size. For metatranscriptomic libraries of the magnitude obtained here (>1,000,000 454 FLX reads), only those transcripts present at concentrations $>1 \times 10^6$ $L^{-1}$ had a good probability of being detected, and only those present at concentrations $>1.5 \times 10^7$ (which would exclude all singletons and other low-count transcript categories) could be compared across samples with good statistical power.

**Replication.** The need to improve sample sequencing depth competes with the need for replication in comparative metatranscriptomic analyses. Two important sources of variability that can be quantified through replication include technical variation during sample processing/sequencing, and natural biological variation within the environment sampled.

For the first type, 454 pyrosequencing is prone to artifacts in which single DNA fragments are sequenced more than once ("duplicate sequences"). While artifactual duplicates are recognized in metagenomes as sequences with identical 5' sequence and high identity throughout (Dinsdale *et al.*, 2008; Gomez-Alvarez *et al.*, 2009), true duplicate sequences can arise in metatranscriptomes from discrete mRNAs from highly expressed genes. In the present work, 24% of RefSeq reads were "duplicates" (same start site and $\geq$90% identity). Since each replicate biological sample was sequenced as four technical replicates (independent emulsion PCRs and sequencing runs), and assuming that artifactual duplicates arise during the emulsion PCR step (Gomez-Alvarez *et al.*, 2009; Stewart *et al.*, 2010), artifactual duplicates should have

uneven distributions across the four 454 runs while natural duplicates should be evenly distributed. We found that most duplicate clusters averaged ~25% per technical sequencing replicate (Fig. 2.S4), and a statistical comparison of COG assignments for all six within-sample pairwise combinations of the technical sequencing replicates indicated that only 0.2% fit the pattern for artifactual duplicates (significantly higher in one technical replicate compared to the other three). For the transcript with the highest copy number in the combined library (Rac prophage; ZP_03400590), removal of duplicate reads would have decreased the count by 98% (from 6,235 to 111 hits) despite evidence from technical replicates that many of these are natural (Fig. 2.S5). Duplicate removal from metatranscriptomic libraries based on sequence start position and percent identity (Gomez-Alvarez *et al.*, 2009, Stewart *et al.*, 2010) may therefore produce systematic underestimates of abundance for the most highly transcribed genes in the community, and statistical analysis of technical replicates is a recommended alternative.

For the second type of variation, within-treatment biological variability sets the false positive rate against which differences in gene expression patterns across treatments or environments can be evaluated (Poretsky *et al.*, 2010). In this study, patchiness in community gene transcription patterns was detectable in paired coastal seawater samples separated by ~300 m (based on tidal flushing rates past a fixed collection point). At the level of functional gene categories, pairwise comparisons indicated significant differences between the samples for 461 of 3,298 COGs (14%) (Xipe, $p < 0.05$). Only 9 significant COGs contained sequences from a putative artifactual duplicate cluster (see above), highlighting the benefit of technical replicate averaging for reducing spurious differences from sequencing artifacts. In accordance with other studies of environmental sequence libraries (Rodriguez-Brito *et al.*, 2006), as well as our

observations above, decreasing the library size had a major influence on the number of significant differences that were detectable (Table 2.S1).

Differences between replicate samples at the individual gene level (i.e. transcripts binned by RefSeq hits) was also examined, using Fisher's Exact Test coupled with a correction for the False Discovery Rate (FDR; Strimmer, 2008) to control for Type I errors arising when simultaneously conducting large numbers of statistical tests (in this case, for >186,000 different RefSeq bins). Eighty-three (0.05%) of the gene bins were statistically different between the two samples ($p < 0.05$ with Benjamini-Hochberg correction), including those representing phage genes, ammonia oxidation genes, and various genes for light driven energy acquisition (Fig. 2.S6). The replicate samples therefore established within-treatment variability (e.g., Fig. 2.4) for future between-treatment comparisons.

**Microbial gene expression in a coastal ocean.** Transcripts from the combined library binned to genes from 1,909 reference organisms. Thirty-three percent of the sequences had best hits to Alphaproteobacteria genes (with roseobacters accounting for 11% and *Pelagibacter ubique* for 7%) and 27% had best hits to Gammaproteobacteria genes (Fig. 2.6). Unexpectedly, 4% of the transcripts binned to the two archaeal genomes of *Nitrosopumilus maritimus* (3.3%) and *Cenarchaeum symbiosum* (0.1%). While Archaea are often abundant and active in deep ocean environments, they were not expected to contribute significantly to gene expression in this shallow coastal water system; the *N. maritimus* taxonomic bin was the second largest in the metatranscriptome (see Chapter 3). The Oligotrophic Marine Gammaproteobacteria (OMG) clades, which are usually in low abundance in oceanic 16S rRNA libraries (<3%; Cho and Giovannoni, 2004), were also unexpectedly well represented in the metatranscriptome (Fig. 2.6; 8% of transcripts), and transcripts binning to three genes from a Gammaproteobacteria prophage

(3% of transcripts) may indicate an ongoing infection of these OMG populations. Eukaryotic transcripts composed 6% of the total, with those binning to *Ostreococcus* spp. particularly well represented (20% of eukaryotic hits).

Copy numbers of transcripts representing 82 genes diagnostic for P, N, and S cycling were determined simultaneously from the metatranscriptomic data (Fig. 2.4). For P transformations, annotations suggest bacterioplankton were transporting phosphate by both high and low affinity transporters. The expression of low affinity transporters, along with polyphosphate storage genes, is consistent with elevated phosphate concentrations (1.2 μM) at the time of sampling, which is typical of late summer in this coastal ocean (Fig. 2.4). Expression patterns also indicated ongoing utilization of organic phosphorus, including phosphoesters (via *phoX, phoD,* and *phoA*) and phosphonates (although transcripts for the canonical C-P lyase pathway were near the limit of detection). For N transformations, sampling occurred during a local ammonia peak (2.6 μM; Fig. 2.4), and transcripts related to the uptake and oxidation of ammonia (*amt*, *amoA,B,C*) were orders of magnitude higher in abundance than genes mediating nitrate or nitrite processing (e.g., *nar, nap, and nir genes)* (Fig. 2.4). Transcripts for urea metabolism, the only representative of dissolved organic nitrogen utilization included in the analysis, made up the second most abundant group of N-related sequences (Fig. 2.4). Nitrogen is often the limiting nutrient (or co-limiting with carbon; Pomeroy *et al.*, 2000) to microbial activity in this ecosystem, and dissolved organic N is 2- to 200-fold higher in concentration than inorganic N. For S transformations, gene expression suggested substantial utilization of reduced sulfur compounds typically found in high concentrations in marsh-dominated coastal systems (Kiene and Capone, 1988; Pakulski and Kiene, 1992). Transcripts were found for metabolism of dimethylsulfoniopropionate (*dmdA*, *dddP* and *dddD*), as well as oxidation of sulfide/thiosulfate

(*sox* genes) (Fig. 2.4). This broad inventory of P, N, and S cycle transcripts represents an absolute benchmark against which time-series and experimentally manipulated transcriptomes in this ecosystem can be compared.

## Conclusions

Addition of an internal mRNA standard provides a significant advantage in metatranscriptomics protocols since it allows estimation of the fraction of the microbial transcriptome captured in the sequence library, as well as the absolute quantification of transcript copy number in the environment (Figs. 2.1, 2.4). While RT-qPCR approaches can also provide absolute transcript numbers, often with greater sensitivity (Church *et al.*, 2010), they are currently limited to a handful of functional genes at a time. Furthermore, the high microdiversity found in many natural gene populations (Fig. 2.3) makes primer design challenging (Varaljay *et al.*, 2010), and likely results in RT-qPCR only quantifying a subset of the total functional gene population. Multiple internal standards that vary in length and concentration (van de Peppel *et al.*, 2003) will allow for more robust calculations of sequencing depth in future studies, and better position "omics" data for integration with biogeochemical rate measurements.

Because low-count transcript categories are difficult to resolve statistically, library size had a critical effect on comparative metatranscriptomic analyses. Many of the biogeochemically diagnostic transcripts detected in our libraries would have been detected in ones that were $1/4^{th}$ or $1/10^{th}$ the size, but these theoretically smaller libraries resulted in a decreased ability to statistically differentiate between samples. Typical library sizes for metatranscriptomes ($10^5$ to $10^6$ sequence reads) are therefore sufficient for descriptive studies, but significant gains in comparative analyses of biogeochemically-informative gene expression patterns will require a

greater sequencing investment. Indeed, 54% of the 46 detected steps in the marine N, P, and S cycles would require at least a 2-fold difference in copy number between samples in order to meet statistical criteria for hypothesis testing (i.e., $p < 0.05$). While a 2-fold change in transcript abundance is an appropriate minimum criterion for expression studies of clonal bacterial cultures in synchronized growth (Bürgmann *et al.*, 2007), it may fail to catch smaller expression differences among complex microbial communities that are ecologically relevant.

Despite a sample sequencing depth of only 1 in $10^7$ transcripts, the libraries provided remarkable insights into gene expression in a marine microbial community, including evidence for active microbes not known previously to have a major role in the ecosystem and quantification of transcripts for scores of steps in marine elemental cycles. These data establish the foundation for comparative assessments of diel, seasonal, and annual changes in microbial gene expression that will provide insights into the regulation of biogeochemical processes in the coastal ocean.

**Acknowledgments**

We thank R. Newton for assistance with sample collection and bioinformatic analysis, L. Tomsho and S. Schuster at Penn State University for 454 sequencing expertise, J. T. Hollibaugh for comments and discussion on the manuscript, and S. Rathbun for helpful discussions on statistical methods. Nutrient data were provided by K. Hunter and S. Joye through the Georgia Coastal Ecosystems Long Term Ecological Research program (OCE-0620959).This project was supported by funding from the Gordon and Betty Moore Foundation and the National Science Foundation Microbial Observatories Program (MCB-0702125).

## References


Azam F, Hodson RE (1977). Size distribution and activity of marine microheterotrophs. *Limnol Oceanogr* 22: 492-501.

Bürgmann H, Howard EC, Ye WY, Sun F, Sun SL, Napierala S *et al.* (2007). Transcriptional response of *Silicibacter pomeroyi* DSS-3 to dimethylsulfoniopropionate (DMSP). *Environ Microbiol* 9: 2742-2755.

Campbell BJ, Waidner LA, Cottrell MT, Kirchman DL (2008). Abundant proteorhodopsin genes in the North Atlantic Ocean. *Environ Microbiol* 10: 99-109.

Cho J-C, Giovannoni SJ (2004). Cultivation and growth characteristics of a diverse group of oligotrophic marine gammaproteobacteria. *Appl Environ Microbiol.* 70: 432-440.

Church MJ, Wai B, Karl DM, DeLong EF (2010). Abundances of crenarchaeal *amoA* genes and transcripts in the Pacific Ocean. *Environ Microbiol* 12: 679-688.

Coyne KJ, Handy SM, Demir E, Whereat EB, Hutchins DA, Portune KJ *et al.* (2005). Improved quantitative real-time PCR assays for enumeration of harmful algal species in field samples using an exogenous DNA reference standard. *Limnol Oceanogr Meth* 3: 381-391.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al.* (2008). Functional metagenomic profiling of nine biomes. *Nature* 452: 629-632.

Ducklow HW (2000). Bacterial production and biomass in the oceans. In: Kirchman (ed). *Microbial Ecology of the Ocean*, 1st edn. Wiley-Liss: New York, NY.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al.* (2008). Microbial community gene expression in ocean surface waters. *PNAS* 105: 3805-3810.

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

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

Hannah MA, Redestig H, Leisse A, Willmitzer L (2008). Global mRNA changes in microarray experiments. *Nat Biotechnol* 26: 741-742.

Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR *et al.* (2009a). *In situ* transcriptomic analysis of the globally important keystone $N_2$-fixing taxon *Crocosphaera watsonii*. *ISME J* 3: 618-631.

Hewson I, Poretsky RS, Dyhrman ST, Zielinski B, White AE, Tripp HJ *et al.* (2009b). Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J* 3: 1286-1300.

Kiene RP, Capone DG (1988). Microbial transformations of methylated sulfur-compounds in anoxic salt-marsh sediments. *Microb Ecol* 15: 275-291.

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

Neidhardt FC, Umbarger HE (1996). Chemical composition of *Escherichia coli.* In: Böck A, Curtiss III R, Kaper JB, Karp PD, Neidhardt FC, Nyström T *et al* (eds). EcoSal—*Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, 2nd edn. ASM Press: Washington, DC.

Pages H, Aboyoun P, Gentleman R, DebRoy S. (2009). Biostrings: string objects representing biological sequences, and matching algorithms. R package version 2.14.8

Pakulski JD, Kiene RP (1992). Foliar release of dimethylsulfoniopropionate from *Spartina alterniflora*. *Mar Ecol-Prog Ser* 81: 277-287.

Passalacqua KD, Varadarajan A, Ondov BD, Okou DT, Zwick ME, Bergman NH (2009). Structure and complexity of a bacterial transcriptome. *J Bacteriol* 191: 3203-3211.

Pomeroy LR, Sheldon JE, Sheldon WM, Blanton JO, Amft J, Peters F (2000). Seasonal changes in microbial processes in estuarine and continental shelf waters of the south-eastern USA. *Estuar Coast Shelf S* 51: 415-428.

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M *et al.* (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* 71: 4121-4126.

Poretsky RS, Gifford SM, Rinta-Kanto J, Vila-Costa M, Moran MA (2009a). Analyzing gene expression from marine microbial communities using environmental transcriptomics. *JoVE* 24. (http://www.jove.com/index/details.stp?ID=1086)

Poretsky RS, Hewson I, Sun SL, Allen AE, Zehr JP, Moran MA (2009b). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* 11: 1358-1375.

Poretsky RS, Sun S, Mou X, Moran MA (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* 12: 616-627.

R Development Core Team. (2009). R: A language and environment for statistical computing v2.10.0. R Foundation for Statistical Computing: Vienna, Austria. (http://www.R-project.org)

Robinson M, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11: R25.

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

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* 5: 398-431.

Shi YM, Tyson GW, DeLong EF (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* 459: 266-U154.

Simon M, Azam F (1989) Protein-content and protein-synthesis rates of planktonic marine bacteria. Mar Ecol-Prog Ser 51:201-213.

Stewart FJ, Ottesen EA, DeLong EF (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J*.

Strimmer K (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics* 9:303.

Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.

Urich TA, Lanzen J, Qi DH, Huson DH, Schleper C, Schuster SC (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* 3.

van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FCP (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* 4: 387-393.

Varaljay VA, Howard EC, Sun SL, Moran MA (2010). Deep sequencing of a dimethylsulfoniopropionate-degrading gene (*dmdA*) by using PCR primer pairs designed on the basis of marine metagenomic data. *Appl Environ Microbiol* 76: 609-617.

White JR, Nagarajan N, Pop M (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 5 e1000352.

Xu L, Chen H, Hu XH, Zhang RM, Zhang Z, Luo ZW (2006). Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Mol Biol Evol 23:1107-1108.

**Table 2.1** Summary statistics for coastal ocean metatranscriptome datasets. Percentages are of total reads. RefSeq Hits = number of reads with significant homology to the RefSeq database; RefSeq Genes = number of unique accession numbers within those hits; Unassigned = number of possible proteins that did not have a significant hit to either the RefSeq or COG databases (47% of possible proteins).

| | FN56 | | FN57 | | Combined | |
|---|---|---|---|---|---|---|
| Total reads | 1,067,363 | | 1,114,536 | | 2,181,899 | |
| rRNA | 466,834 | (44%) | 623,804 | (56%) | 1,090,638 | (50%) |
| psRNA | 100,437 | (9%) | 25,213 | (2%) | 125,650 | (6%) |
| Possible proteins | 500,092 | (47%) | 465,519 | (42%) | 965,611 | (44%) |
| RefSeq Hits | 255,280 | | 260,739 | | 516,019 | |
| RefSeq Genes | 96,573 | | 109,395 | | 168,669 | |
| RefSeq Taxa | 1,707 | | 1,761 | | 1,909 | |
| COG hits | 162,925 | | 170,593 | | 333,518 | |
| Unassigned | 244,812 | | 204,780 | | 449,592 | |

**Table 2.2.** Estimation of the number of bacterioplankton mRNA molecules in coastal seawater and sequencing depth of the metatranscriptomic libraries.

| Calculation method | Sample | mRNA molecules per liter | mRNA molecules per cell[2] | Sequencing depth (%) |
|---|---|---|---|---|
| Internal standard[1] | FN56 | $1.0 \times 10^{12}$ | 238 | 0.000009 |
| | FN57 | $0.6 \times 10^{12}$ | 142 | 0.000015 |
| Extracted RNA mass[3] | FN56 | $0.2 \times 10^{12}$ | 48 | 0.000043 |
| | FN57 | $0.4 \times 10^{12}$ | 95 | 0.000020 |
| Per cell RNA content[4] | FN56 | $2.6 \times 10^{12}$ | 619 | 0.000003 |
| | FN57 | $2.6 \times 10^{12}$ | 619 | 0.000003 |

[1] The libraries contained 4,014 (FN56) and 6,865 (FN57) copies of the internal standard out of a total of 500,092 (FN56) and 465,519(FN57) potential protein encoding sequences. The standard was added at $4.7 \times 10^{10}$ copies/5.75 liters of seawater just prior to cell lysis for total RNA extraction (see Methods for details).
[2] Cell numbers in the 3 μM filtrate averaged $4.2 \times 10^{9}$ $L^{-1}$ based on epifluorescence microscopy.
[3] Extraction yields were 14.4 (FN56) and 32.9 (FN57) μg total RNA from 5.75 liters of seawater. Total RNA is assumed to contain 4% mRNA by mass (Neidhardt *et al.*, 1996) and bacterial mRNAs are assumed to average 924 nt (Xu *et al.*, 2006).
[4] Marine bacterial cells are assumed to contain 5.7 fg total RNA per cell [midpoint of 1.9-9.5 fg range reported by Simon and Azam (1989)]. See footnote 3 for estimate of percent mRNA by mass and footnote 2 for cell counts $L^{-1}$.

**Table 2.S1.** Number of significantly overrepresented COGs detected in comparisons of subsamples from either within or between the replicate metatranscriptomes using Xipe. For the 25% and 50% libraries, the bold numbers are the mean of 7 to 10 replicated subsamples with the standard deviation given in parentheses.

| | | Statistically Different COGs | | | | | |
|---|---|---|---|---|---|---|---|
| | | Sample A[1] | | Sample B[2] | | Total | |
| 25% | 56 vs 56 | **12** | (2.3) | **11** | (3.7) | **23** | (3.7) |
| | 57 vs 57 | **12** | (2.5) | **16** | (3.6) | **28** | (4.4) |
| | 56 vs 57 | **57** | (5.2) | **94** | (7.7) | **151** | (9.1) |
| 50% | 56 vs 56 | **12** | (2.9) | **14** | (2.7) | **26** | (3.4) |
| | 57 vs 57 | **16** | (3.5) | **18** | (3.5) | **34** | (4.0) |
| | 56 vs 57 | **97** | (6.2) | **166** | (6.0) | **264** | (9.3) |
| 75% | 56 vs 57 | **134** | (NA) | **241** | (NA) | **375** | (NA) |
| 100% | 56 vs 57 | **160** | (NA) | **301** | (NA) | **461** | (NA) |

[1,2] Number of statistically greater COGs in FN56 ([1]) or FN57 ([2]) when the comparison is between biological replicates (i.e. 56 vs 57).

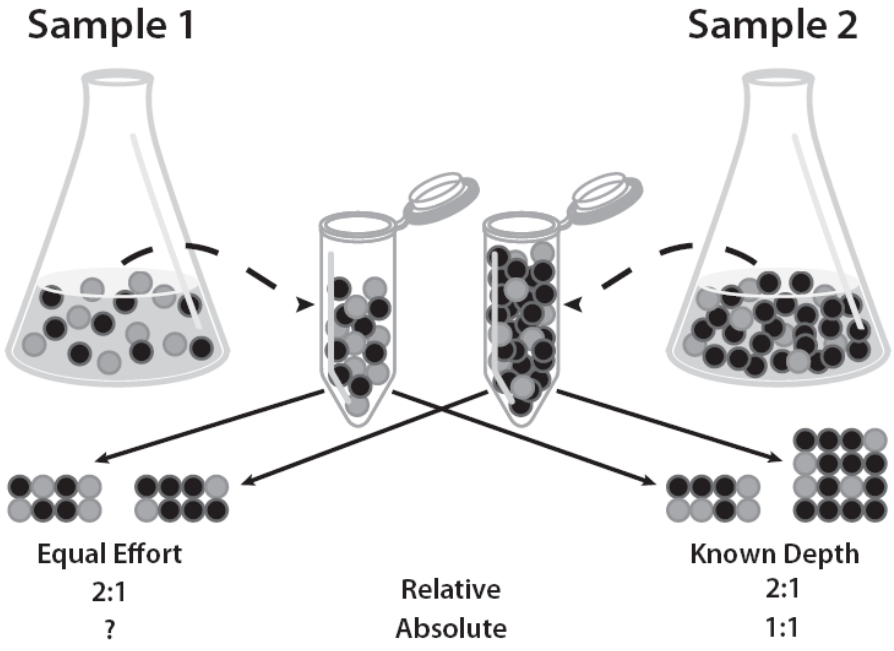



**Figure 2.1.** Effect of sample sequencing depth on quantification of transcripts (or genes) in environmental samples. 'Equal effort' sequences the same number of reads per sample volume regardless of the size of the mRNA pool, and therefore conveys only relative abundance. 'Known depth' sequences a known proportion of the transcript pool (50% for both, in this example) and therefore also conveys absolute copy numbers per sample volume. The latter is more relevant to biogeochemical rate measurements, since mRNAs of biogeochemical interest (gray dots) can make up different proportions in community transcriptomes yet have identical numbers in the environment.

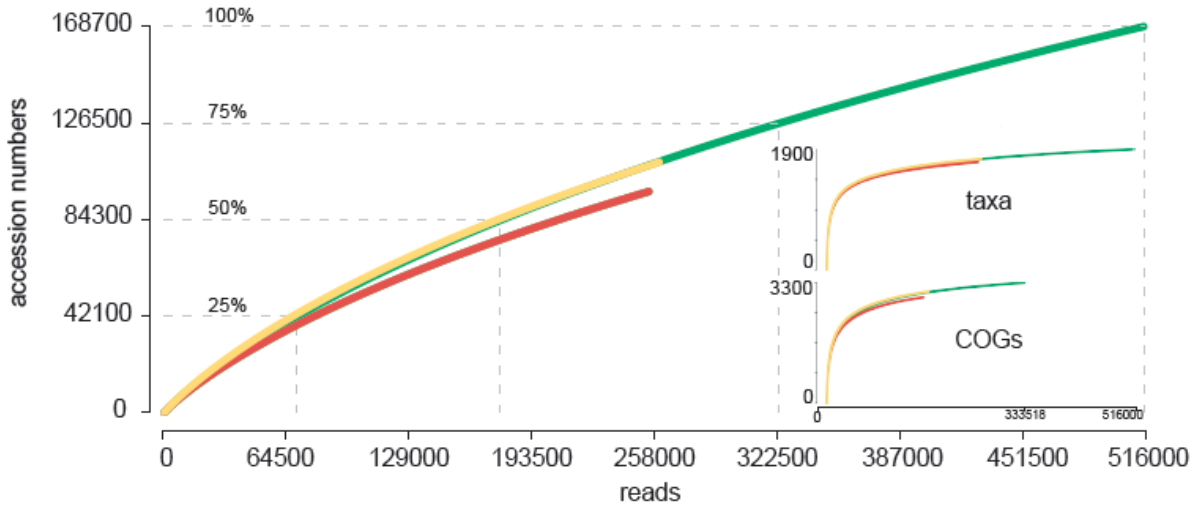



**Figure 2.2**. Collector's curve of gene richness as a function of reads analyzed. Red: FN56; Yellow: FN57; Green: combined libraries. Dashed lines indicate the number of reads needed to reach quarter percentiles of the total richness of the combined library. Inset: Collector's curves for taxonomic and functional gene category (COG) richness, with the y-axis corresponding to the number of unique reference organisms or COG numbers.

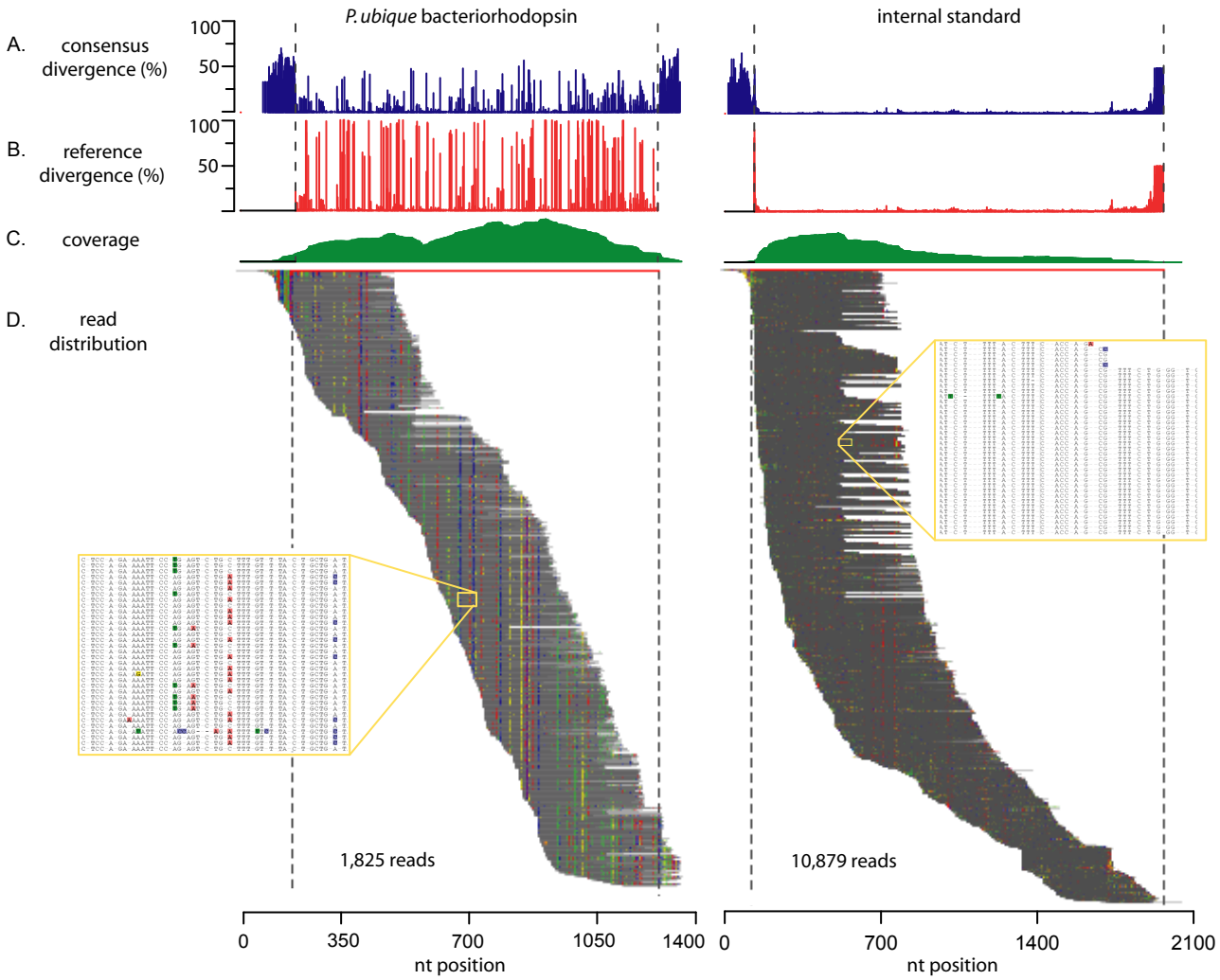
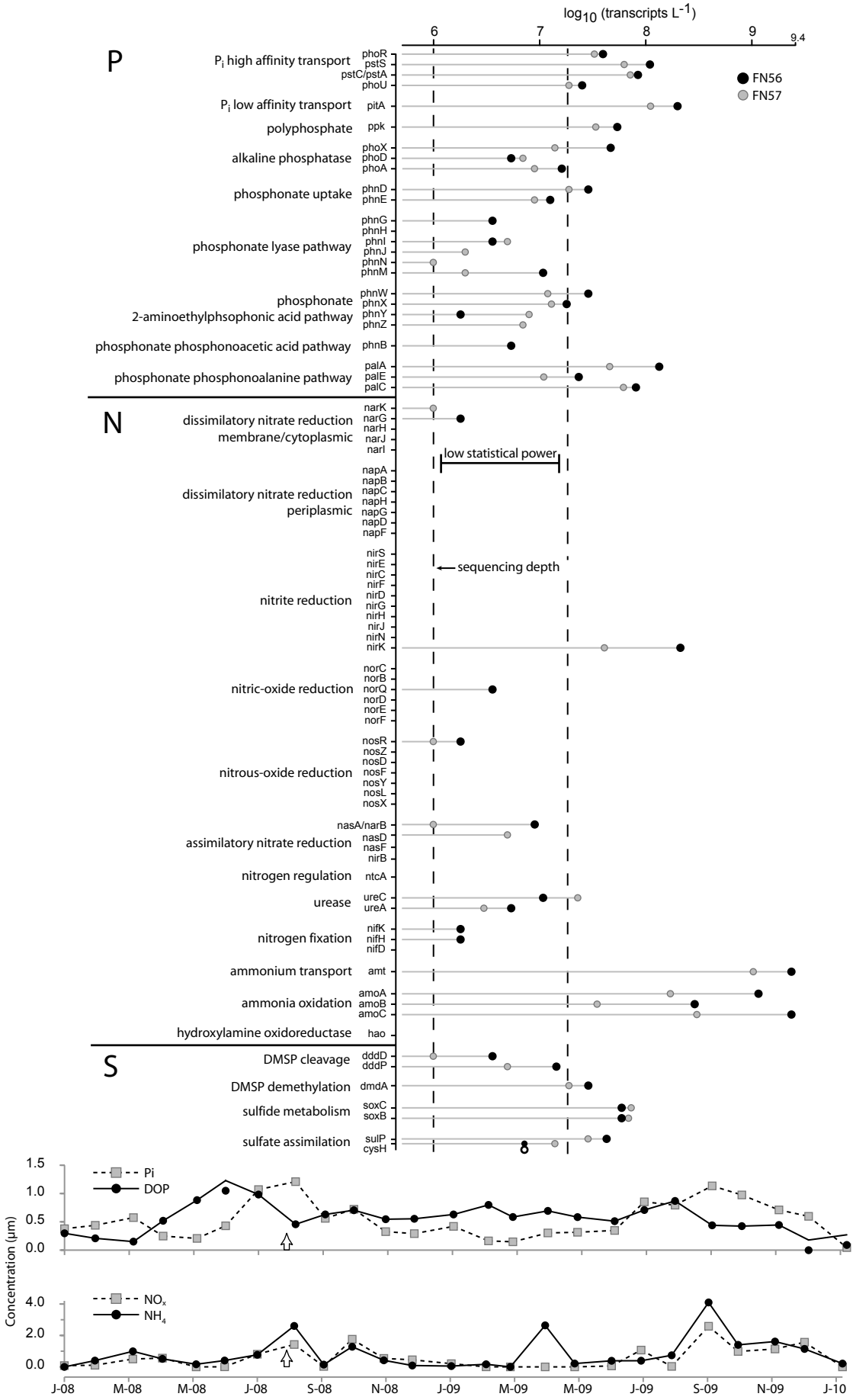



Figure 2.3. Assembly of 1,825 reads (out of 2,259 total) binning to the P. ubique HTCC1002 proteorhodopsin gene (PU1002_03206 (left), and of 10,879 reads (out of 10,879 total) binning to the internal transcript standard (right). A) Percent nucleotide divergence from the consensus sequence. B) Percent nucleotide divergence from the reference sequence. C) Coverage by nucleotide position. D) Read assembly to the reference gene (shown in red), with dashed lines indicating start and end positions of the reference. Note that the reference gene lengths are extended by assembly gaps. Divergence from the consensus sequence (i.e. the majority nucleotide at a given position) is indicated as follows: A = red, T = green, C = blue, G = yellow. Insets show close-up regions of assemblies.

**Figure 2.4** Copy numbers of phosphorus, nitrogen, and sulfur cycle transcripts in a coastal ocean microbial community. The left line represents the limit of detection for this study, and together with the right line defines the region where copy numbers are too low for robust statistical analysis (i.e., where the fold-difference requirement is >2). Symbols indicate copy numbers in biological duplicates. Bottom graphs show monthly nutrient concentrations for GCE LTER station 6. The arrows mark the date of sample collection.


log10 (transcripts L-1)
6
7
8
9
9.4
FN56
FN57
P
Pi high affinity transport
phoR
pstS
pstC/pstA
phoU
Pi low affinity transport
pitA
polyphosphate
ppk
alkaline phosphatase
phoX
phoD
phoA
phosphonate uptake
phnD
phnE
phosphonate lyase pathway
phnG
phnH
phnI
phnJ
phnN
phnM
phosphonate
2-aminoethylphsophonic acid pathway
phnW
phnX
phnY
phnZ
phosphonate phosphonoacetic acid pathway
phnB
phosphonate phosphonoalanine pathway
palA
palE
palC
N
dissimilatory nitrate reduction
membrane/cytoplasmic
narK
narG
narH
narJ
narI
low statistical power
dissimilatory nitrate reduction
periplasmic
napA
napB
napC
napH
napG
napD
napF
sequencing depth
nitrite reduction
nirS
nirE
nirC
nirF
nirD
nirG
nirH
nirJ
nirN
nirK
nitric-oxide reduction
norC
norB
norQ
norD
norE
norF
nitrous-oxide reduction
nosR
nosZ
nosD
nosF
nosY
nosL
nosX
assimilatory nitrate reduction
nasA/narB
nasD
nasF
nirB
nitrogen regulation
ntcA
urease
ureC
ureA
nitrogen fixation
nifK
nifH
nifD
ammonium transport
amt
ammonia oxidation
amoA
amoB
amoC
hydroxylamine oxidoreductase
hao
S
DMSP cleavage
dddD
dddP
DMSP demethylation
dmdA
sulfide metabolism
soxC
soxB
sulfate assimilation
sulP
cysH
Concentration (µm)
Pi
DOP
NOx
NH4
1.5
1.0
0.5
0.0
4.0
2.0
0.0
J-08
M-08
M-08
J-08
S-08
N-08
J-09
M-09
M-09
J-09
S-09
N-09
J-10

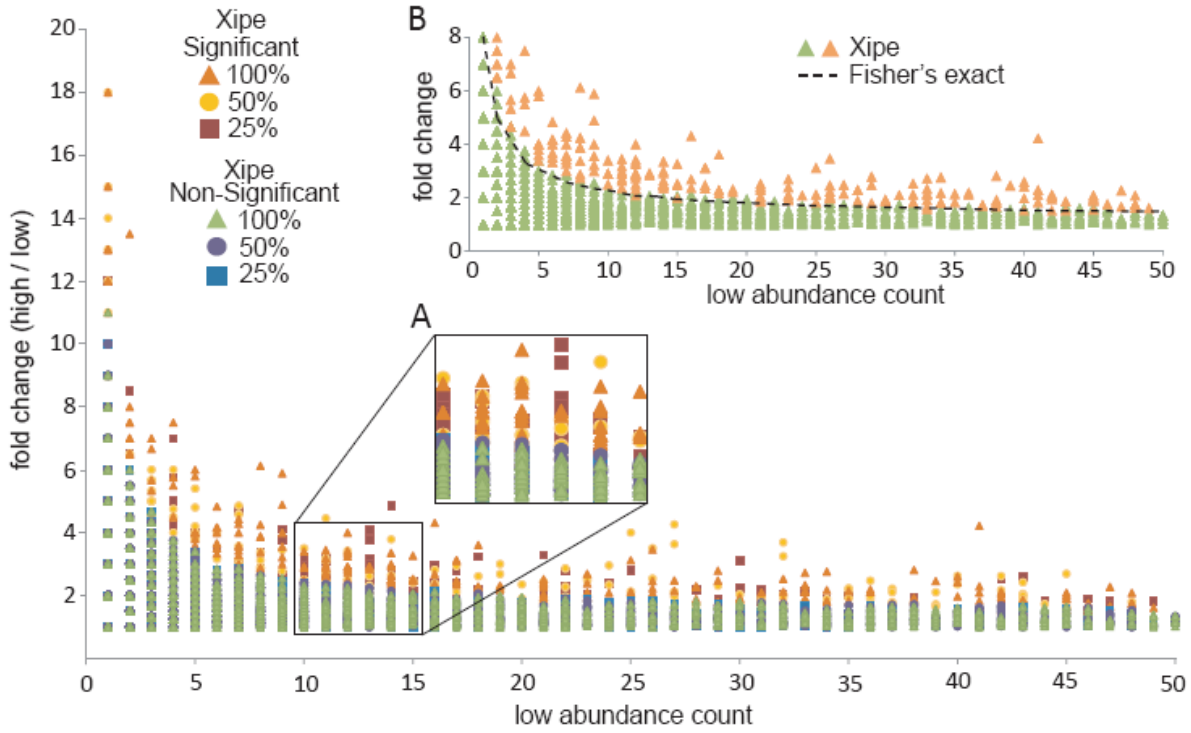



**Figure 2.5.** Minimum fold difference required for statistical significance (Xipe, $p < 0.05$) as a function of both the count in the lower abundance sample and the library size. Samples and subsamples were from the combined libraries (FN56 and FN57). Marker color is based on the statistical outcome (significant or non-significant) and library size (percent of full library). Inset A: Zoom of region in the main figure. Note that the minimum fold-difference for significance is independent of the three library sizes analyzed. Inset B: An alternative analysis of the significance threshold using contingency tables and Fisher's Exact Test. The minimum fold-difference threshold at which a low abundance count is significant by the Fisher's Exact method is plotted as a dotted black line. The results from the Xipe analysis (main figure) at the 100% library size are also shown in inset B for direct comparison with the Fisher's Exact method.

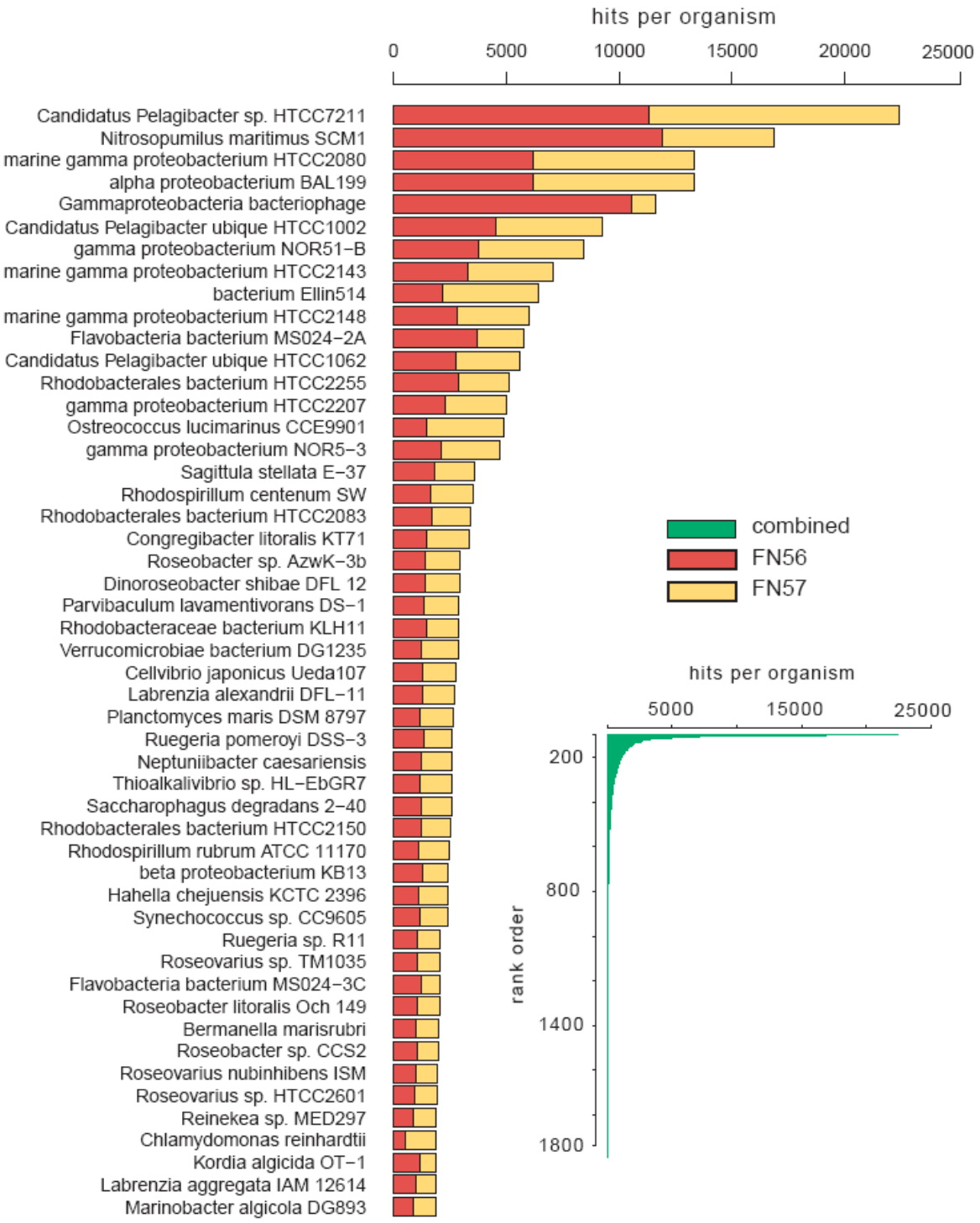



**Figure 2.6.** Rank order abundance of taxonomic bins (species or strain level). Main figure: top 50 taxonomic annotation bins; inset: all 1,909 taxonomic annotation bins.

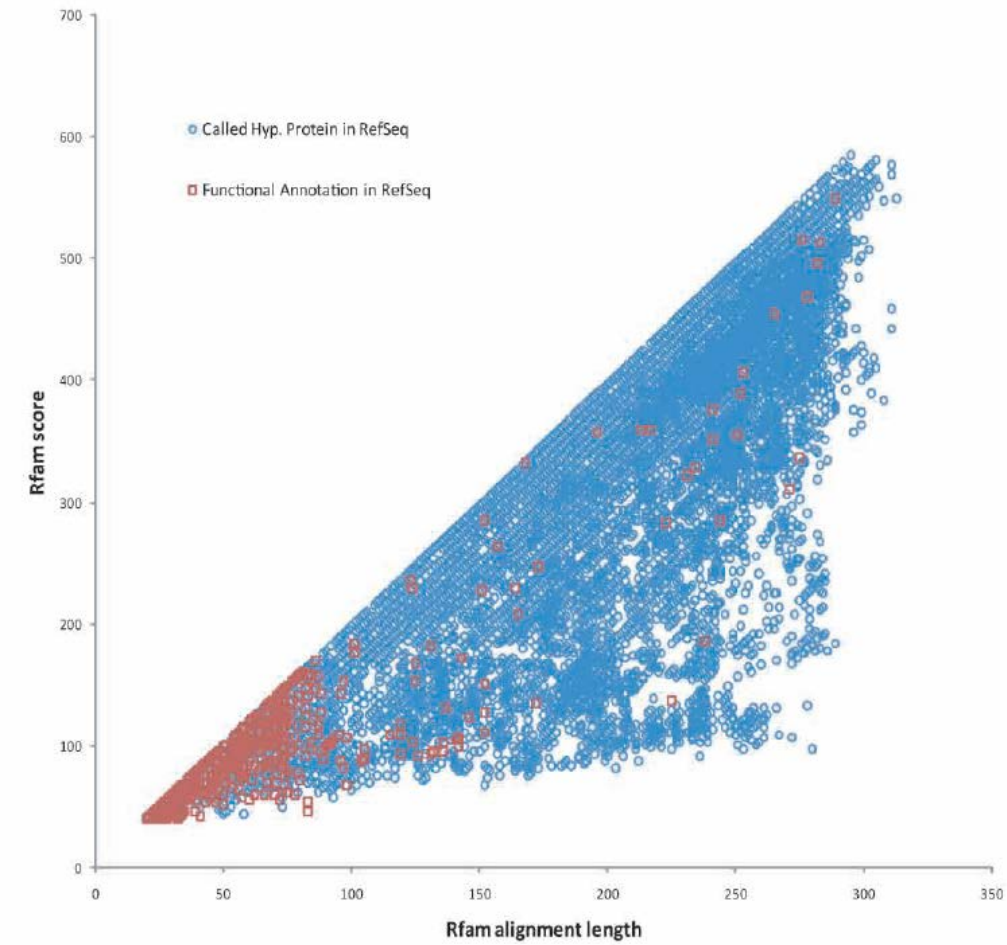



**Figure 2.S1.** Reads with sequence homology to both the Rfam and RefSeq databases. Points are colored by their categorization in RefSeq: blue symbols are annotated as hypothetical proteins; red symbols have been given a specific functional annotation. All sequences annotated as hypothetical proteins were considered psRNAs, along with those with a RefSeq functional annotation but an Rfam alignment length >95 nt.

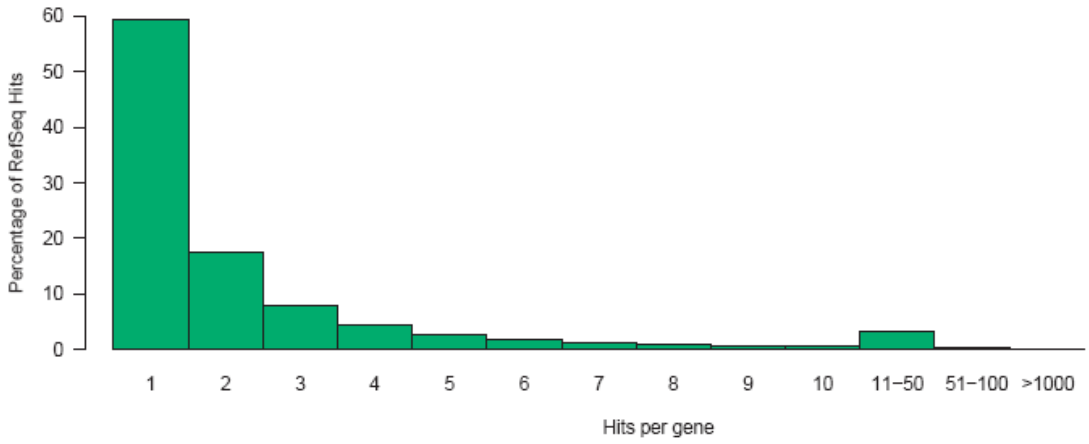



**Figure 2.S2.** Distribution of the number of reads binning to individual RefSeq genes as a percentage of the total reads in the combined library. One equals the percentage singletons, two the number of doubletons, etc.

| | | | | Combined | | | |
|---|---|---|---|---|---|---|---|
| | **$P_i$ High Affinity Transporters** | FN56 | FN57 | 100% | 75% | 50% | 25% |
| phoR | sensory kinase | 22 | 33 | 55 | 41 | 27 | 13 |
| pstS | phosphate periplasmic binding protein | 61 | 63 | 124 | 93 | 62 | 31 |
| pstC/pstA | phosphate ABC transporter / permease | 47 | 72 | 119 | 89 | 59 | 29 |
| phoU | phosphate uptake / response regulator | 14 | 19 | 33 | 24 | 16 | 8 |
| | **$P_i$ Low affinity transporters** | | | | | | |
| pitA | phosphate permease | 111 | 112 | 223 | 167 | 111 | 55 |
| | **Polyphosphate storage** | | | | | | |
| ppk | polyphosphate kinase | 30 | 34 | 64 | 48 | 32 | 16 |
| | **Alkaline phosphatases** | | | | | | |
| phoX | alkaline phosphatase | 26 | 14 | 40 | 30 | 20 | 10 |
| phoD | alkaline phosphatase | 3 | 7 | 10 | 7 | 5 | 2 |
| phoA | alkaline phosphatase | 9 | 9 | 18 | 13 | 9 | 4 |
| | **Phosphonate Uptake** | | | | | | |
| phnD | ABC transporter | 16 | 19 | 35 | 26 | 17 | 8 |
| phnE | ABC transporter | 7 | 9 | 16 | 12 | 8 | 4 |
| | **Phosphonate C-P lyase pathway** | | | | | | |
| phnG | carbon-phosphorus lyase complex subunit | 2 | 0 | 2 | 1 | 1 | 0 |
| phnH | carbon-phosphorus lyase complex subunit | 0 | 0 | 0 | 0 | 0 | 0 |
| phnI | carbon-phosphorus lyase complex subunit | 2 | 5 | 7 | 5 | 3 | 1 |
| phnJ | carbon-phosphorus lyase complex subunit | 0 | 2 | 2 | 1 | 1 | 0 |
| phnN | carbon-phosphorus lyase complex subunit | 0 | 1 | 1 | 0 | 0 | 0 |
| phnM | carbon-phosphorus lyase complex subunit | 6 | 2 | 8 | 6 | 4 | 2 |
| | **Phosphonate 2-aminoethylphosphonic acid pathway** | | | | | | |
| phnW | 2-aminoethylphosphonate-pyruvate transaminase | 16 | 12 | 28 | 21 | 14 | 7 |
| phnX | phosphonoacetaldehyde hydrolase | 10 | 13 | 23 | 17 | 11 | 5 |
| phnY | Phytanoyl-CoA dioxygenase | 1 | 8 | 9 | 6 | 4 | 2 |
| phnZ | HD domain protein | 4 | 7 | 11 | 8 | 5 | 2 |
| | **Phosphonate Phosphonoacetic acid pathway** | | | | | | |
| phnB | 2-phosphonopropionate transporter | 3 | 0 | 3 | 2 | 1 | 0 |
| | **Phosphonate Phosphonoalanine pathway** | | | | | | |
| palA | phosphonopyruvate hydrolase | 69 | 46 | 115 | 86 | 57 | 28 |
| palE | putative solute binding protein | 13 | 11 | 24 | 18 | 12 | 6 |
| palC | putative ABC transporter inner membrane component | 45 | 62 | 107 | 80 | 53 | 26 |

**Figure 2.S3.** Abundance of phosphorus related transcripts. Query gene symbols and descriptions are given in the two left columns. Counts indicate the number of hits to the query gene in the combined metatranscriptome library and in *in silico* subsets, with the intensity of shading increasing with hit abundance.


**Figure 2.S4.** Distribution of duplicate reads among technical replicates for the largest 100 duplicate clusters. The percentage of the cluster within a technical replicate is indicated by the bar color (replicate 1 = brown, replicate 2 = orange, replicate 3 = green, replicate 4 = red). Data are shown for both biological replicates: A = FN56 and B = FN57. The x-axis is ordered by cluster size.

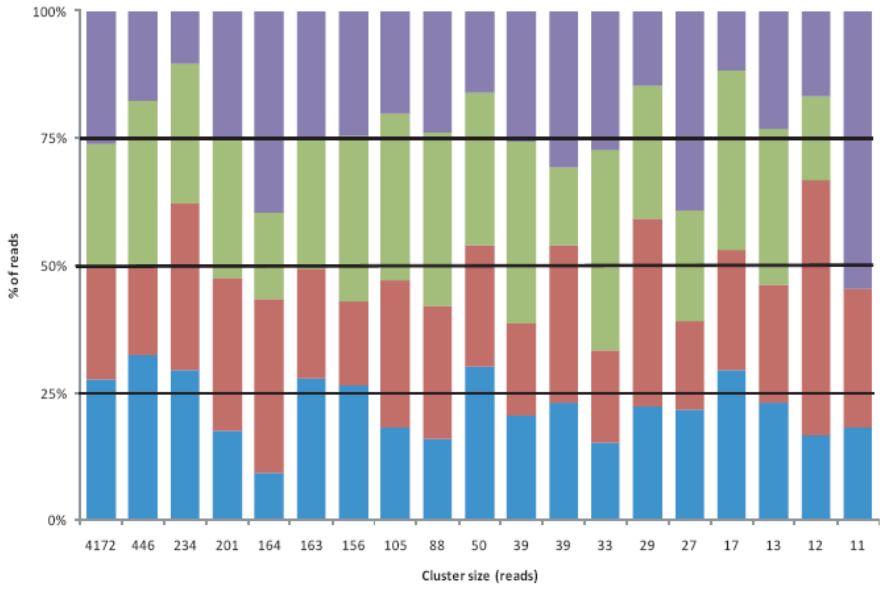



**Figure 2.S5.** Distribution of duplicate reads among technical replicates for reads binning to the Refseq protein ZP_03400590 (Rac prophage). The percentage of the cluster within a technical rep is indicated by the bar color (replicate 1 = blue, replicate 2 = red, replicate 3 = green, replicate 4 = purple). Data are shown for library FN56 only, which contained the majority of the prophage sequences.

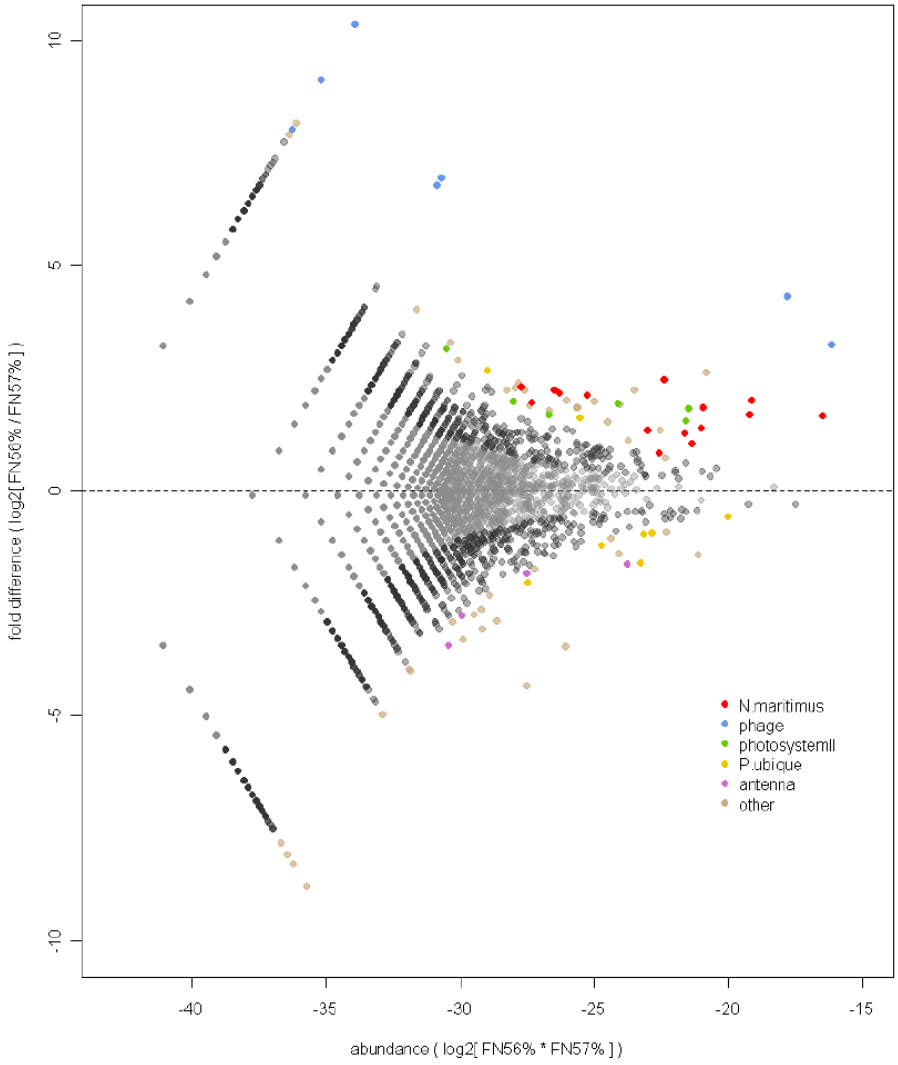



**Figure 2.S6.** M vs. A plot of relative gene abundance showing pairwise comparisons of samples FN56 and FN57. Sample percentage equalsf hits for a RefSeq gene over total possible protein encoding genes. Gray points = not statistically different (Fishers' Exact; $p < 0.05$). Black points = statistically different but did not meet the FDR criterion. All colored points were both statistically significant and met the FDR criterion. *N.maritimus* = *Nitrosopumulis maritimus* related, phage = phage related, photosystem = photosystem II protein D1 related, *P.ubique* = *Pelagibacter ubique* related, and other = all other significant/FDR points.

# CHAPTER 3

# METATRANSCRIPTOMIC ANALYSIS OF AMMONIA-OXIDIZING ORGANISMS IN AN ESTUARINE BACTERIOPLANKTON ASSEMBLAGE[1]

---

[1]Gifford, S.M.*, Hollibaugh, J.T.*, Sharma, J., Bano, B., and Moran, M.A. (2011). *ISME-J* 5:866-878. doi:10.1038/ismej.2010.172;

Reprinted here with permission of the publisher.

*Co-First authors who contributed equally to this work.

**Abstract**

Quantitative PCR analysis revealed elevated relative abundance (1.8% of prokaryotes) of marine group 1 Crenarchaeota (MG1C) in samples of southeastern U.S. coastal bacterioplankton collected in August 2008, compared with samples from the same site at other times (mean 0.026%). We analyzed the MG1C sequences in metatranscriptomes from these samples to gain insight into the metabolism of an MG1C population growing in the environment and for comparison with ammonia oxidizing bacteria (AOB) in the same samples. Assemblies revealed low diversity within sequences assigned to most individual MG1C ORFs and high homology with "*Candidatus* Nitrosopumilus maritimus" strain SCM1 genome sequences. Reads assigned to ORFs for ammonia uptake and oxidation accounted for 37% of all MG1C transcripts. We did not recover any reads for Nmar_1354-Nmar_1357, proposed to encode components of an alternative, nitroxyl-based ammonia oxidation pathway; however, reads from Nmar_1259 and Nmar_1667, annotated as encoding a multicopper oxidase with homology to *nirK*, were abundant. Reads assigned to 2 homologous ORFs (Nmar_1201 and Nmar_1547), annotated as hypothetical proteins, were also abundant, suggesting that their unknown function is important to MG1C physiology or ecology. Transcripts from other metabolic pathways (carbon fixation, TCA cycle,) were represented in the metatranscriptome, but at much lower levels than those for ammonia oxidation. Superoxide dismutase and peroxiredoxin-like transcripts were more abundant in the MG1C transcript pool than in the complete metatranscriptome, suggesting that the MG1C population was selectively exposed to oxidative stress. qPCR indicated low AOB abundance (0.0010 % of prokaryotes) and we found no transcripts related to ammonia oxidation and only one RuBisCO transcript among the transcripts assigned to AOB, suggesting they were not responding to the same environmental cues as the MG1C population.

## Introduction

Marine Group 1 Crenarchaeota (MG1C) are abundant and widespread in meso-pelagic, open ocean environments (Fuhrman and Hagström, 2008; Karner *et al.*, 2001). They have proved difficult to culture so that our knowledge of their metabolism is based primarily on culture-independent methods (Kirchman *et al.*, 2007; Ouverney and Fuhrman, 2000; Teira *et al.*, 2004). Metagenomic data (Treusch *et al.*, 2005; Venter *et al.*, 2004) suggested that MG1C might play a role in ammonia oxidation and more recent research (Beman *et al.*, 2008; Hallam *et al.*, 2006a; Hallam *et al.*, 2006b; Santoro *et al.*, 2010; Wuchter *et al.*, 2006) and the successful isolation of a representative MG1C (Konneke *et al.*, 2005) have confirmed this. As a consequence of these findings, the paradigm that members of the β- and γ-Proteobacteria are responsible for most of the ammonia oxidation in the ocean has come into question (reviewed in (Francis *et al.*, 2007; Prosser and Nicol, 2008).

Studies of the distributions of planktonic ammonia oxidizing organisms have shown that ammonia oxidizing Crenarchaeota (Ammonia Oxidizing Archaea, AOA) tend to be numerically dominant in the open ocean (Agogue *et al.*, 2008; Beman *et al.*, 2008; de Corte *et al.*, 2008; Kalanetra *et al.*, 2009; Mincer *et al.*, 2007; Santoro *et al.*, 2010; Wuchter *et al.*, 2006) and fjords (Urakawa *et al.*, 2010; Zaikova *et al.*, 2010). Most studies of AOA populations in estuaries (Beman and Francis, 2006; Bernhard *et al.*, 2010; Caffrey *et al.*, 2007; Francis *et al.*, 2005; Magalhaes *et al.*, 2009; Mosier and Francis, 2008; Santoro *et al.*, 2008) have focused on sediments. From these studies it appears that the relative abundance of ammonia-oxidizing Bacteria (AOB) increases in estuaries relative to coastal waters or the open ocean. The environmental factors responsible for the success of AOB versus AOA in estuarine and coastal waters are not known, but the shift correlates with salinity in some systems (Bernhard *et al.*,

2010; Caffrey *et al.*, 2007; Magalhaes *et al.*, 2009; Mosier and Francis, 2008; Santoro *et al.*, 2008). However, the success of one group over the other is not likely to be directly based on salinity as AOA can be dominant in the oligohaline reaches of some estuaries (Mosier and Francis, 2008) and in soils (Prosser and Nicol, 2008).

Estuaries are distinct from meso-pelagic open ocean environments in a number of important ways: salinity variation; trace metal availability; concentrations and types of organic carbon and other reduced substrates; and other factors known to influence microbes. As part of a program investigating the dynamics of microbial populations in estuarine waters and their response to fluctuating environmental variables (SIMO, http://simo.marsci.uga.edu/), samples of DNA and RNA from the plankton assemblage have been collected regularly at a station in Georgia coastal waters. Quantitative estimates of *amoA* gene abundance indicated elevated abundance of AOA in samples collected in August 2008. We analyzed the metatranscriptome of two samples collected at this time and studied the distribution of transcripts among MG1C ORFs to gain insight into the metabolism of a MG1C population growing in the environment. An additional goal was to understand the factors that regulate competition within the guild of ammonia oxidizing microorganisms. The data also allowed us to examine proposed pathways for ammonia oxidation in AOA.

## Methods

**Sample Collection.** Near-surface water samples were collected quarterly from a floating dock at Marsh Landing on the Duplin River, Sapelo Island, Georgia (31° 25' 4.08" N, 81° 17' 43.26" W; Supplemental Figure 1), ~6 km from the mouth of Doboy Sound, as described in chapter 2. Briefly, samples were collected twice per day at approximately noon and midnight, <1 hr before high tide, over a 2-day period during each sampling campaign. Samples used for

RNA extraction were collected in rapid succession in the middle of the first night of each sampling campaign. A sample (5.75 L) of surface (~0.5 m) sea water was pumped directly from the river through 3 μm pore size filters (Capsule Pleated 3 μm Versapor Membrane; Pall Life Sciences, Ann Arbor Michigan, USA) then through 0.22 μm pore size filters (Supor polyethersulfone; Pall Life Sciences, Ann Arbor Michigan, USA) using a peristaltic pump (Supplemental Figure 3.1). The 0.22 μm filter was placed in a Whirl-Pak® plastic bag and immediately flash-frozen in liquid nitrogen. Total time from the start of filtration to freezing was ~10 minutes. We began filtering the second sample (FN57) immediately (~5 minute delay) after the filter from the first sample (FN56) was placed in liquid nitrogen. We collected samples for DNA extraction concurrently by filling 20 L carboys with surface water while the RNA samples were filtering. Once the second RNA sample was frozen, we filtered 12 L of the DNA sample through 3 μm and 0.22 μm filters as above, and the 0.22 μm filters were flash frozen.

**mRNA isolation.** mRNA was isolated from the samples as described previously (chapter 2). Before beginning the extraction, 25 ng of a 994 nt RNA standard (derived from the pGEM cloning vector) was added to the sample in lysis buffer to serve as an internal standard (chapter 2). Total RNA was extracted from the filters using an RNAEasy kit (Qiagen, Valencia, CA, USA) and any residual DNA was removed by treating the sample twice with a Turbo DNA-Free Kit (Applied Biosystems, Austin, TX, USA).

The purified RNA preparations (containing 14 and 32 μg total RNA from samples FN56 and FN57, respectively, 2-5 mg of this RNA was taken through the rRNA removal steps) were treated in two ways to remove ribosomal RNA. Epicentre's mRNA-Only kit (Epicentre, Madison, WI, USA) was used first to decrease rRNA contamination enzymatically. The samples were then treated with MICROB*Express* and MICROB*Enrich* kits (both from Applied

Biosystems) that couple rRNA oligonucleotide hybridization probes with magnetic separation to enrich for mRNA. Initial and final RNA extracts were analyzed on an Experion automated electrophoresis system (Bio-Rad, Hercules, CA, USA) to verify successful removal of most of the rRNA. RNA remaining in the samples was amplified linearly using the MessageAmp II-Bacteria kit (Applied Biosystems). The amplified RNA was then converted to cDNA using the Universal RiboClone cDNA synthesis system (Promega, Madison, WI, USA) with random hexamer primers. Left over reactants and nucleotides from cDNA synthesis were removed from the sample using the QIAquick PCR purification kit (Qiagen).

**Sequencing and Annotation.** cDNA was sequenced in four GS-FLX runs. One half of each PicoTiter plate was loaded with cDNA from one replicate sample, resulting in each sample being sequenced to the equivalent of two full runs divided over four plates (Supplemental Figure 3.2; described in detail in chapter 2). Over 2 million sequence reads were produced. Ribosomal RNA sequences in these reads were identified by a BLASTn (Zhang *et al.*, 2000) search against the small and large subunit SILVA database (http://www.arb-silva.de) with a bit score cutoff of 50. Sequences identified as rRNA (~50% of the total) were excluded from further processing.

The remaining non-rRNA sequences were queried against NCBI's RefSeq database using BLASTx (Altschul *et al.*, 1997) with a bit score cutoff of 40. The top hit that exceeded this bit score was taken as the ORF assignment for that sequence. Approximately 50% of the non-rRNA sequences were assigned to annotated ORFs by this procedure. Replicate reads (defined according to (Gomez-Alvarez *et al.*, 2009) accounted for 24.4% of this total, but our sequencing protocol, which used technical replicates on different plates, allowed us to identify replicates arising from methodological artifacts as discussed in (Gomez-Alvarez *et al.*, 2009) versus biologically valid replicates (see chapter 2). Our analysis indicates that most of the replicates are

not artifacts, so we retained them in the data set. As discussed in chapter 2, for the purposes of the analyses that follow, we assume that the population of reads returned from the sequencing effort is an unbiased sampling of the transcripts present in the populations of Bacteria and Archaea *in situ*.

**Quantitative PCR and sequencing of 16S rRNA and *amoA* amplicons.** The abundance of MG1C and AOB *amoA* genes and of Crenarchaeota and Bacteria 16S rRNA genes was determined by quantitative, real-time PCR (qPCR) as described previously (Caffrey *et al.*, 2007). Primers are given in Supplemental Table 3.1. The abundance of individual genes (copies per ng of DNA extracted from the sample) was used to estimate the number of MG1C and AOB cells in the sample for the purposes of calculating the number transcripts per cell. Relative abundance of MG1C or AOB was calculated as follows:

RA = ([A]/(([Cren 16S]/GD) + ([Bact 16S]/GD)))

Where:

RA is the relative abundance of the organism of interest;

[A] is the concentration of the gene of interest, either Crenarchaeota 16S or AOB *amoA* measured by qPCR;

[Cren 16S] and [Bact 16S] are the concentrations of Crenarchaeota and Bacteria 16S rRNA genes; and

GD is the gene dosage of 16S rRNA genes per genome, taken to be 1 for Marine Group 1 Crenarchaeota (from genomes annotated in DOE's IMG database) and 1.8 as an average for marine bacteria (Biers *et al.*, 2009). We used the abundance of AOB *amoA* genes to estimate AOB abundance assuming a gene dosage of 2.5 amoA genes/AOB genome (an average from (Norton *et al.*, 2002).

The total number of MG1C or AOB cells in the sample was then calculated as:

RA * (total prokaryote abundance determined by epifluorescence microscopy [$4.2 \times 10^9$ cells/L]) * (sample volume [5.75 L]).

Crenarchaeota 16S rRNA and *amoA* genes were cloned and sequenced (primers listed in Supplemental Table 1) for phylogenetic analysis as described previously (Kalanetra *et al.*, 2009).

**Analysis of MG1C ribotypes**. We compared (using BLASTn) the rRNA reads removed from one of the pyrosequencing libraries (FN56) to the "*Candidatus* Nitrosopumilus maritimus" strain SCM1 16S (Konneke *et al.*, 2005) rRNA gene sequence (Nmar_R0029) to identify MG1C 16S rRNA sequences in our data set. We then queried the top 250 hits against the NCBI nr/nt database to obtain information from the annotations of the top hits on the distribution by habitat of ribotypes (ecotypes) related to the MG1C in our samples. We also assembled these reads using Nmar_R0029 as a scaffold to obtain a consensus sequence that was compared to sequences obtained by cloning and then sequencing PCR amplicons of 16S rRNA genes from the DNA sample.

**Assemblies.** The Geneious® (Drummond *et al.*, 2010) software package version 4.8 was used to assemble reads into contigs, for sequence manipulations (e.g. alignments) and for phylogenetic analyses. All assemblies were constructed using unedited cDNA sequences. Unless otherwise noted, the appropriate genomic reference sequence was used as a scaffold and assemblies required ≥25 bp of overlap and ≥75% identity between sequences in the overlapping portions. The gap/extend penalty was set at 18, mismatch score at -9, and match score at 5. With these assembly parameters, most of the reads assigned to MG1C ORFs assembled into one contig per ORF with good coverage over the entire length of the gene.

**Consensus sequences for *amoA* genes** were derived from reads assembled against the "*Ca.* N. maritimus" strain SCM1 *amoA* gene (Nmar_1500) sequence as a scaffold as described above. The consensus sequence for the majority genotype was determined by requiring >75% agreement at each position. Reads representing this majority consensus sequence were removed from the data set manually, then remaining reads representing less abundant, minority sequences were assembled as before and the consensus sequence was again recorded. Although inspection suggested additional diversity in the reduced data set, we did not attempt to recover additional consensus sequences as coverage was too low for reliable assembly and analysis.

**Environmental data.** The Georgia Coastal Ecosystems LTER (GCE-LTER) program collects data on a variety of environmental variables from the area surrounding our sampling site. These data and their accompanying metadata are available on the GCE-LTER website http://gce-lter.marsci.uga.edu/. The closest GCE-LTER water quality monitoring station, GCE6, is located in Doboy Sound, ~4.5 km from our sampling site (Supplemental Figure 3.1).

## Results and Discussion

**qPCR analysis**. Analysis of the abundance of Bacteria and MG1C 16S rRNA genes and of AOA and AOB *amoA* genes by quantitative PCR indicated elevated abundance of MG1C and of ammonia oxidizers, especially AOA, in water samples collected on 6-7 August, 2008 (Figure 3.1). MG1C *amoA* abundance was 35- to 781-fold greater in August than on other sampling dates, while MG1C 16S rRNA abundance was 43- to 1,658-fold greater (Figure 3.1A and B). MG1C relative abundance in the prokaryotic community averaged 1.8% (range 1.1-2.6%) for the August samples versus an average of 0.026 (range 0.0002-0.15%) on other dates (Figure 3.1D).

Both MG1C *amoA* and Crenarchaeota 16S rRNA abundance increased during the 2-day sampling campaign in August 2008 (Figure 3.1A, inset).

There was no correlation between the abundance of *amoA* genes from MG1C and AOB in these samples (linear regression, $r^2$=0.14, P>0.1). AOB *amoA* gene abundance was only 1.2- to 2.8-fold greater in August than on other sampling dates (Figure 3.1C), comparable to the increase in Bacteria 16S rRNA gene abundance (Figure 3.1D). AOB relative abundance was 0.0024% (range 0.0018-0.0045%) in August versus 0.0044% (range 0.0009-0.013%) on other dates and AOB *amoA* gene abundance did not increase during the 2 day sampling campaign (not shown). The ratio of MG1C to AOB *amoA* abundance (398:1) in August was more than 20-fold greater than on other dates (Figure 3.1C), suggesting selective growth of MG1C over AOB at the time of sampling. With the exception of the August samples, the ratios of AOA to AOB abundance are similar to our previous observations in sediment samples from nearby sites (Caffrey *et al.*, 2007).

**MG1C 16S rRNA ecotypes.** MG1C 16S rRNA reads retrieved from our libraries were most similar (250 sequences, all >96.9% identity with 235> 99% and 162 reads = 100% identity, significance values $1*e^{-159}$ to $1*e^{-113}$) to environmental sequences from coastal waters, coral symbionts or sediments (13 different studies) or to the "*Ca.* N. maritimus" SCM1 16S rRNA gene. These reads assembled into one contig (not shown). We compared the consensus sequence from the contig to nearly full-length sequences obtained by cloning and sequencing PCR amplicons from the original sample (Figure 3.2). We detected 2 sequence variants by inspection of the assembly, but only one of these, corresponding to the consensus and with >99% identity to the "*Ca.* N. maritimus" strain SCM1 16S rRNA gene, was captured in the clone library.

**Metatranscriptome properties.** We retrieved ~2 million cDNA pyrosequencing reads from the two samples (chapter 2). Analysis of this data set (Table 3.1) revealed that 17,386 sequences (median length 236 bp, range 47-360 bp) could be assigned to coding regions in the two MG1C genomes, "*Ca.* N. maritimus" strain SCM1 and Cenarchaeum symbiosum (Hallam *et al.*, 2006a; Hallam *et al.*, 2006b). For simplicity, we will refer to this subset of reads as the "MG1C metatranscriptome." MG1C thus accounted for 3.1 % of the reads identified as transcripts, which is comparable to their contribution to the population of prokaryotes (1.8%, Figure 3.1). In contrast, only 46 reads were assigned to ORFs from Euryarchaeota. The remaining reads were assigned primarily to ORFs from Bacteria or viruses.

Of the 17,386 reads that were assigned to MG1C ORFs, 16,914 were assigned to *"Ca.* N. maritimus" strain SCM1 while 472 were assigned to C. symbiosum (Table 3.1). These reads were assigned to 786 different *"Ca.* N. maritimus" strain SCM1 ORFs (Figure 3.3) representing 44% of the 1,797 coding regions annotated in this genome (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) and to 82 ORFs from the *C. symbiosum* genome (4% of 2,017 annotated coding regions). Since there are currently only 2 MG1C genomes represented in the RefSeq database and they share a great deal of homology (Walker *et al.*, 2010), reads assigned to one of them as the top hit were usually assigned (with similar significance values) to the other as the second hit. We noted gaps in the recruitment of reads against the "*Ca.* N. maritimus" strain SCM1 genome (Figure 3.3) that appear too long to be random consequences of low coverage, suggesting possible sites of indels (e.g., Nmar_124 to Nmar_154, Nmar_1147 to Nmar_1173, Nmar_1323 to Nmar_1357).

We estimate that each MG1C cell contained 168 transcripts (averaged for the two libraries, Supplemental Table 3.2). This is similar to the value seen in chapter 2 obtained for the

entire prokaryote metatranscriptome from these samples (190 transcripts/cell). The differences in the number of transcripts per MG1C cell between libraries (321 vs 79 for FN56 vs FN57) may reflect differences in the physiological state of MG1C community between samples. Ratios of the abundance of transcripts from specific ORFS between the two libraries were more variable than expected by random sampling error, suggesting differences in the physiological state of the bacterioplankton in the two samples (see chapter 2 for a detailed analysis). Our estimate of the average number of transcripts in AOB cells was 49,241 transcripts per cell (Supplemental Table 3.2). This value is unreasonably high and suggests either that non-AOB reads were incorrectly assigned to AOB ORFs or that our qPCR estimates of AOB abundance are too low, or both. The qPCR estimates of AOB *amoA* gene abundance upon which this calculation is based are typical of what we (Caffrey *et al.*, 2007) and others have reported for coastal waters. Comparison of the distributions of bit scores for hits to ORFs from these two populations (data not shown) suggests that read assignments are less reliable for AOB than for MG1C.

Most ORFs were represented by singletons or only a few reads (Figure 3.3); however, 34 MG1C ORFs were represented by 50 or more reads (Table 3.2), together accounting for 13,686 reads (78.8% of the reads in the MG1C metatranscriptome). The best-represented ORF (Nmar_1547, a hypothetical protein) accounted for 22% of the MG1C metatranscriptome.

**Nmar_1547 homologues.** Thirty-one percent of the reads in the MG1C metatranscriptome were assigned to a group of 6 homologous ORFs: Nmar_1547, Nmar_1201, CENSYa_0159, CENSYa_0161, CENSYa_2159 and CENSYa_2160; with most assigned to Nmar_1547 and Nmar_1201 (Table 3.2). The sequences of these ORFs are very similar (bit scores>1140, E-value=0) and the apparent duplication is noteworthy in genomes that appear to have undergone reduction as an adaptive strategy (Hallam *et al.*, 2006a; Walker *et al.*, 2010) and

that contain many important genes (e.g. *amoABC*) as single copies. Hallam *et al.* (2006a) first noted them in their analysis of the Cenarchaeum symbiosum metagenome and similar sequences (represented by ABZ07689) were reported to be abundant in cDNA libraries from 4000 m at Station ALOHA by (Shi *et al.*, 2009). Searches (BLASTn) of the "GOS All ORFS" dataset in the CAMERA database (http://camera.calit2.net/) identified 25 sequences, all from coastal samples, with significant homology (E-values $<e^{-53}$) to Nmar_1547.

Nmar_1547 and Nmar_1201 are large ORFs, >5000 nt in length. In contrast to most other MG1C ORFs (see Figure 3.4 for example), reads assigned to Nmar_1547 and Nmar_1201 did not assemble into one contig against the respective genome sequence as a scaffold (Supplementary Figure 3.6). Although the assemblies otherwise have very high coverage (up to 229-fold and up to 1,113-fold for Nmar_1201 and Nmar_1547, respectively), the regions where the assemblies break have increased sequence variability (Supplemental Figure 3.6), resulting in insufficient homology to the RefSeq sequence to support assembly. These "ORFs" may not be protein coding regions, though analysis of codon usage in a homologous sequence (ABZ07689, BLASTp score = 604, E-value=0) by Shi and colleagues (Shi *et al.*, 2009) suggest that they are. Annotation and BLASTx searches of the GenBank non-redundant protein database suggest functions distantly related to adhesins or hemaglutinin and hemolysin proteins and annotation indicates that they have leader sequences and transmembrane domains. These features suggest that they may be cell-surface proteins involved in some interaction with other cells, detrital particles, high molecular weight DOM, etc. By analogy with genes involved in ammonia processing (below), their abundance in the transcript pool suggests that their unknown function is important to MG1C physiology or ecology.

**Ammonia uptake and oxidation.** Transcripts from ORFs related to ammonia uptake and oxidation were among the most abundant in the MG1C transcript pool (Table 3.2). A total of 6,455 reads (37% of the reads in the MG1C metatranscriptome) were assigned to ORFs identified by (Walker *et al.*, 2010) as being related to the ammonia oxidation pathway. This includes 2,657 reads assigned to ammonia monooxygenase subunits: *amoA* (Nmar_1500, 836); *amoB* (Nmar_1503, 198); and *amoC* (Nmar_1502, 1,623), giving relative abundances in the transcript pool of 4.2:1.0:8.2 for *amoA*:*amoB*:*amoC*, which differs from the stoichiometry of the subunits in native ammonia monooxygenase (1:1:1). Inspection of the assemblies of these reads revealed 2 dominant genotypes in the population of *amoA* reads (Figure 3.4) and at least 2 in the *amoB* and *amoC* (data not shown) populations. Phylogenetic analysis of the consensus sequence for the dominant *amoA* genotype (Figure 3.5) placed it in a clade containing the "*Ca.* N. maritimus" strain SCM1 gene and a variety of shallow water column and sediment environmental sequences. It is also >99.5% identical to MG1C *amoA* gene nucleotide sequences retrieved previously from Georgia coastal waters (Hollibaugh, unpublished data) and to sequences obtained from the DNA samples collected in this study (Figure 3.5). The minor consensus sequence grouped separately from the dominant consensus sequence and was not recovered in the (small) clone library we sequenced. Half of the *amoA* gene sequences retrieved from the DNA sample were most similar to an environmental sequence from the sediments of an eutrophic Mexican estuary (Beman and Francis, 2006). Reads corresponding to this clade were not found in the metatranscriptome, suggesting the presence of an inactive sub-population of MG1C in these samples.

Reads assigned to ORFs annotated as ammonia transporters and permeases were also abundant with a total of 1,017 reads assigned to Nmar_1698 (757), Nmar_0588 (94) and

CENSYa_1453 (166). The relatively high abundance of ammonia permeases in the MG1C metatranscriptome (6% of the transcripts) seems at odds with the model of ammonia oxidation as a cell-surface process proposed in (Walker *et al.*, 2010). Nmar_1698 and CENSYa_1453 are very similar to each other (BLASTx bit score 705, E=0; BLASTn bit score 803, E=0, 71% identity) and reads assigned to them assembled into one contig against an Nmar_1698 scaffold (not shown). Nmar_0588 is divergent with no similarity to other MG1C genes and a best BLASTx hit to an ammonium transporter from the slime mold *Polysphondylium pallidum* PN500 (bit score of 365, $E=5*e^{-99}$). Our data thus indicate transcription by the MG1C population of two different ammonia transporters, possibly with different kinetic properties. Inspection of the Nmar_1698 and Nmar_0588 assemblies indicates additional diversity in the ammonia transporter genes transcribed by the MG1C population, with at least 3 variants of Nmar_1698 and possibly 2 variants of Nmar_0588.

(Walker *et al.*, 2010) propose 2 alternative pathways for ammonia oxidation in "*Ca.* N. maritimus" strain SMC1. One of the proposed pathways proceeds via hydroxylamine, but depends on a Cu-based alternative to the AOB heme-based hydroxylamine oxidoreductase. The second alternative mechanism proposes ammonia oxidation by an ammonia monooxygenase that produces a reactive nitroxyl intermediate instead of hydroxylamine. Both pathways transfer electrons to the quinone pool via a quinone reductase. Nmar_1226, proposed by Walker *et al.*, (2010) to serve this function as an analog of the AOB quinone reductase, was well- represented in the MG1C metatranscriptome (46 reads). However, we did not detect any transcripts from the genes (Nmar_1354 to Nmar_1357) proposed by Walker *et al.* (2010) to encode proteins involved in the nitroxyl-based alternative ammonia oxidation pathway. Instead, a total of 1,006 reads was assigned to two other ORFs, Nmar_1259 and Nmar_1667, that were similar to sequences

retrieved from the Sargasso Sea (EAH96098 and EAI84410; bit scores >498, E=0). These sequences have been identified as crenarchaeote homologues of Cu-containing nitrite reductases (*nirK*) by (Treusch *et al.*, 2005) and (Bartossek *et al.*, 2010). Their function *in vivo* is in question because, at least under the aerobic growth conditions they reported, "Ca. N. maritimus" strain SCM1 stoichiometrically converts ammonia to nitrite in culture (Konneke *et al.*, 2005; Martens-Habbena *et al.*, 2009). An experiment performed by (Bartossek *et al.*, 2010) to test the relationship between transcription of these genes and the expected activity (nitrous oxide production) failed to support a nitrite reductase function and (Bartossek *et al.*, 2010) speculated that the proteins encoded by these genes might exhibit "other or additional activities besides nitrite reduction." (Walker *et al.*, 2010) included them in the list of genes devoted to energy production from ammonia oxidation without specifying their function. These homologies and their elevated abundance in the transcript pool suggest that they play an important role in the primary ammonia oxidation pathway *in situ*.

**Other metabolic functions.** Sixty-seven reads were assigned to an MG1C ORF annotated as superoxide dismutase (SOD, Nmar_0394, Table 3.2), which accounted for 26% of all hits to SOD in the complete (MG1C plus Bacteria) metatranscriptome. The proportion of reads assigned to SOD in the MG1C metatranscriptome (67 of 17,386 reads) is significantly ($X^2$=314, p=0) greater than the proportion (270 of 543,016) of SOD reads in the rest of the metatranscriptome. Assembly of the reads assigned to Nmar_0394 revealed a population with low diversity (2 of 67 reads that differ from the consensus and 98.5% overall average pairwise identity) that differs slightly from the "*Ca.* N. maritimus" strain SCM1 gene (93% identity at the nucleotide level). Superoxide dismutase catalyzes the decomposition of superoxide radicals to yield hydrogen peroxide, which is broken down by catalase in many Bacteria. The "*Ca.* N.

maritimus" strain SMC1 genome does not contain an ORF annotated as catalase; however, it contains 4 ORFs (Nmar_0275, Nmar_0560, Nmar_1438 and Nmar_1496) annotated as thiol-specific antioxidants (peroxiredoxins) that may serve the same function (Imlay, 2008). These ORFs were represented by a total of 45 hits in the MG1C metatranscriptome, with 34 of these hits assigned to one ORF, Nmar_0275. The complete metatranscriptome (Bacteria plus MG1C) contained 889 hits to ORFs annotated as "catalase," "peroxiredoxin," or "thiol-specific antioxidant." Thus, a statistically significantly ($X^2$=9.121, p=0.0025) greater portion of reads in the MG1C metatranscriptome was assigned to ORFs with functions related to catalase than in the Bacteria metatranscriptome. Finally, 14 MG1C reads were assigned the DNA repair gene *radA* (Nmar_1386), which was not different from the proportion of *recA* in the Bacteria metatranscriptome ($X^2$=0.004, p=0.95).

The overrepresentation of MG1C superoxide dismutase and hydrogen peroxidase-related transcripts suggests that MG1C may be subjected to greater exposure to superoxide or that they are more sensitive to it than the Bacteria in these samples. Increased exposure may be a consequence of reactions unique to their metabolism. In contrast, similar levels of transcripts for DNA repair enzymes (*radA* and *recA*) suggest that the two populations (MG1C and Bacteria) are responding similarly to agents that cause DNA damage, such as UV radiation.

MG1C are reported to fix carbon via the 3-hydroxypropionate/4-hydroxybutyrate pathway (Berg *et al.*, 2007; Hallam *et al.*, 2006a; Hallam *et al.*, 2006b; Kockelkorn and Fuchs, 2009; Konneke *et al.*, 2005; Walker *et al.*, 2010). The MG1C metatranscriptome contained 146 reads assigned to ORFs from this pathway. Fifty-nine reads were assigned to MG1C TCA-cycle genes. There is no evidence in the metatranscriptome that heterotrophy played a significant role in the nutrition of this MG1C population. MG1C reads were not assigned to COGS for

transporters of organic compounds, in contrast to the high proportion of reads in the Bacteria metatranscriptome that were assigned to transporters (~20% of the Bacteria reads assigned to the top 50 COGS, Gifford unpublished data) or found previously in a Bacteria-dominated metatranscriptome retrieved from a near-by site (Poretsky *et al.*, 2010). Also, the ratio of MG1C *amoA* genes to MG1C 16S rRNA genes in these samples was 0.51 (Figure 3.1B). While lower than values reported for ammonia oxidizing enrichments (Wuchter *et al.*, 2006) or cultures (Konneke *et al.*, 2005) (1:1 to 2.8:1) or the gene dosage in MG1C genomes (1:1) (Hallam *et al.*, 2006a; Walker *et al.*, 2010), this ratio is much higher than ratios used to infer heterotrophy in other populations (Agogue *et al.*, 2008; de Corte *et al.*, 2008; Kalanetra *et al.*, 2009).

**AOB transcripts.** The metatranscriptome also contained 2,651 reads (0.5% of reads assigned to ORFs) that were assigned to AOB ORFs (Table 3.1). Forty-nine percent of the AOB reads were assigned to *Nitrosococcus* ORFs, with the remainder assigned to *Nitrosomonas* (27%) and *Nitrosospira* (24%) ORFs. None of the reads attributed to AOB were assigned to ORFs known to be involved in ammonia uptake or oxidation and only one AOB read (from *Nitrosomonas*) was assigned to RubisCO, the enzyme responsible for carbon fixation in AOB (data not shown). Based on our PCR data, the abundance of AOB *amoA* transcripts in the metatranscriptome may have been below the limit of detection. The relative abundance of AOA versus AOB *amoA* genes in these samples as determined by qPCR averages 398:1 (Figure 3.1). Given the number of MG1C *amoA* transcripts in the metatranscriptome (836), if the relative abundance of *amoA* transcripts in the AOB transcript pool was comparable to that seen in the MG1C, we would expect to recover only 2.1 (836/398) AOB *amoA* transcripts from the complete metatranscriptome. In contrast, we would expect to encounter 131 AOB *amoA* reads if all of the transcripts assigned to AOB were actually from AOB and if the relative abundance of *amoA*

transcripts in the AOB transcript pool was comparable to that seen in the MG1C metatranscriptome (4.9%). As discussed above, these calculations and the number of transcripts per cell implied by our data (Supplemental Table 3.2) suggests that the majority of the reads attributed to AOB were misassigned, likely due to binning of phylogenetically related but non-AOB sequences to AOB genomes.

**Implications for competition between AOA and AOB.** Combined with the much greater abundance of MG1C 16S rRNA and *amoA* genes in the August sample relative to other sampling dates (Figure 3.1), the increasing abundance of Crenarchaeota 16S rRNA and AOA *amoA* genes over the 2 days we sampled (Figure 3.1), and the distribution of reads in the MG1C metatranscriptome, our data suggest that the MG1C population was actively growing - blooming - when sampled. Assuming that the MG1C population developed locally rather than being advected into the study area, we examined environmental data collected by the Georgia Coastal Ecosystem LTER (http://gce-lter.marsci.uga.edu/) for the weeks preceding this sampling for potential explanations for the elevated MG1C population. There are no obvious perturbations in the records for weather (wind, rainfall, runoff, tides) or environmental variables (temperature, nutrients, chlorophyll, etc.) that might indicate a resuspension event, a pulse of nutrient-rich water from runoff or upwelling, a phytoplankton bloom, etc. (data not shown).

Ammonia concentrations are variable in Georgia coastal waters (Supplemental Figure 3.7) and a late summer increase in ammonium concentration is a regular feature of this coastal environment (Verity, 2002). The abundance of Bacteria 16S rRNA genes was also greater in the August sample than on other sampling dates (Figure 3.1), consistent with an overall increase in heterotrophic metabolism, presumably leading to elevated ammonium regeneration at this time of the year. Although ammonium concentrations were elevated during the August 2008

sampling campaign (Supplemental Figure 3.7), elevated ammonium concentrations at other times of the year did not correspond to elevated MG1C abundance (compare Figure 3.1 with Supplemental Figure 3.7) suggesting that ammonium alone is not the driving variable. AOB abundance did not increase during the August sampling series and average August abundance was only slightly elevated compared to other sampling dates. Previous work (Caffrey *et al.*, 2007) documented a correlation between AOA (but not AOB) *amoA* gene abundance and potential nitrification rates in sediment samples from this site, suggesting that AOB are typically not very active at this site, even though they are present. One explanation for the difference in the response of AOA versus AOB is that the threshold ammonia concentration needed to stimulate growth of AOB may be higher than for AOA. Differences in ammonia uptake kinetics between MG1C and AOB (Martens-Habbena *et al.*, 2009) suggest that MG1C are better competitors at low ammonia concentrations (Martens-Habbena *et al.*, 2009). Alternatively, MG1C and AOB may be differentially limited or inhibited by environmental factors other than ammonia availability.

**Acknowledgements**

We thank R. Newton for assistance with sample collection. L. Tomsho and S. Schuster provided 454 sequencing expertise. S. Obrebski assisted with statistical analyses. The manuscript was greatly improved by the helpful comments of reviewers of earlier versions. We thank them for their efforts. This project was funded by grants from the Gordon and Betty Moore Foundation and the National Science Foundation (MCB0702125, OCE0620959 and OCE0352216).

## References


Agogue H, Brink M, Dinasquet J, Herndl GJ (2008). Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic. *Nature* **456:** 788-791.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25:** 3389-3402.

Bartossek R, Nicol GW, Lanzen A, Klenk H-P, Schleper C (2010). Homologues of nitrite reductases in ammonia-oxidizing archaea: diversity and genomic context. *Environmental Microbiology* **12:** 1075-1088.

Beman JM, Francis CA (2006). Diversity of ammonia-oxidizing Archaea and Bacteria in the sediments of a hypernutrified subtropical estuary: Bahia del Tobari, Mexico. *Applied and Environmental Microbiology* **72:** 7767-7777.

Beman JM, Popp BN, Francis CA (2008). Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. *ISME Journal* **2:** 429-441.

Berg IA, Kockelkorn D, Buckel W, Fuchs G (2007). A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. *Science* **318:** 1782-1786.

Bernhard AE, Landry ZC, Blevins A, de la Torre JR, Giblin AE, Stahl DA (2010). Abundance of ammonia-oxidizing Archaea and Bacteria along an estuarine salinity gradient in relation to potential nitrification rates. *Applied and Environmental Microbiology* **76:** 1285-1289.

Biers EJ, Sun S, Howard EC (2009). Prokaryotic genomes and diversity in the surface ocean: interrogating the Global Ocean Sampling metagenome. *Applied and Environmental Microbiology* **75:** 2221-2229.

Caffrey JM, Bano N, Kalanetra K, Hollibaugh JT (2007). Environmental factors controlling ammonia-oxidation by ammonia-oxidizing Bacteria and Archaea in Southeastern estuaries. . *ISME Journal* **1:** 660-662.

de Corte D, Yokokawa T, Varela MM, Agogue H, Herndl GJ (2008). Spatial distribution of Bacteria and Archaea and *amoA* gene copy numbers throughout the water column of the Eastern Mediterranean Sea. *ISME Journal* **3:** 147-158.

Drummond AJ, Ashton B, Cheung M, Heled J, M. K, Moir R *et al*. (2010). Biomatters Inc.

Francis CA, Beman JM, Kuypers MMM (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME Journal* **1:** 19-27.

Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB (2005). Ubiquity and diversity of ammonia-oxidizing Archaea in water columns and sediments of the ocean. *Proceedings of the National Academy of Sciences of the US* **102:** 14683-14688.

Fuhrman J, Hagström Å (2008). Bacterial and archaeal community structure and its patterns. In: Kirchman DL (ed). *Microbial Ecology of the Oceans*, Second edn. John Wiley & Sons, Inc.: Hoboken, N.J. pp 45-90.

Gomez-Alvarez V, Teal TK, Schmidt TM (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME Journal* **3:** 1–4.

Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y-i, Sugahara J *et al* (2006a). Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proceedings of the National Academy of Sciences* **103:** 18296-18301.

Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM *et al* (2006b). Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biology* **4:** e95.

Imlay JA (2008). Cellular defenses against superoxide and hydrogen peroxide. *Annual Review of Biochemistry* **77:** 755-776.

Kalanetra KM, Bano N, Hollibaugh JT (2009). Ammonia-oxidizing *Archaea* in the Arctic Ocean and Antarctic coastal waters. *Environmental Microbiology* **11:** 2434–2445.

Karner MB, DeLong EF, Karl DM (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409:** 507-510.

Kirchman DL, Elifantz HD, Ana I., Malmstrom RR, Cottrell MT (2007). Standing stocks and activity of Archaea and Bacteria in the western Arctic Ocean. *Limnology and Oceanography* **52:** 495-507.

Kockelkorn D, Fuchs G (2009). Malonic semialdehyde reductase, succinic semialdehyde reductase, and succinyl-coenzyme A reductase from Metallosphaera sedula: enzymes of the autotrophic 3-hydroxypropionate/4-hydroxybutyrate cycle in sulfolobales. *Journal of Bacteriology* **191:** 6352-6362.

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437:** 543-546.

Magalhaes CM, Machado A, Bordalo AA (2009). Temporal variability in the abundance of ammonia- oxidizing bacteria vs. archaea in sandy sediments of the Douro River estuary, Portugal. *Aquatic Microbial Ecology* **56:** 13-23.

Martens-Habbena W, Berube PM, Urakawa H, de la Torre JR, Stahl DA (2009). Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* **461:** 976-979.

Mincer TJ, Church MJ, Taylor LT, Preston C, Karl DM, DeLong EF (2007). Quantitative distribution of presumptive archaeal and bacterial nitrifiers in Monterey Bay and the North Pacific Subtropical Gyre. *Environmental Microbiology* **9:** 1162-1175.

Mosier AC, Francis CA (2008). Relative abundance and diversity of ammonia-oxidizing archaea and bacteria in the San Francisco Bay estuary. *Environmental Microbiology* **10:** 3002-3016.

Norton J, Alzerreca J, Suwa Y, Klotz M (2002). Diversity of ammonia monooxygenase operon in autotrophic ammonia-oxidizing bacteria. *Archives of Microbiology* **177:** 139-149.

Ouverney CC, Fuhrman JA (2000). Marine planktonic archaea take up amino acids. *Applied and Environmental Microbiology* **66:** 4829-4833.

Poretsky R, S., Sun S, Mou X, Moran MA (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environmental Microbiology* **12:** 616-627.

Poretsky RS, Bano N, Buchan A, Hollibaugh JT, Moran MA (2006). Environmental Transcriptomics: A method to access expressed genes in complex microbial communities. *Molecular Microbial Ecology Manual, 3rd Edition*, 3rd edn.

Poretsky RS, Gifford S, Rinta-Kanto J, Vila-Costa M, Moran MA (2009). Analyzing gene expression from marine microbial communities using environmental transcriptomics. *Journal of Visualized Experiments*.

Prosser JI, Nicol GW (2008). Relative contributions of archaea and bacteria to aerobic ammonia oxidation in the environment. *Environmental Microbiology* **10:** 2931-2941.

Santoro AE, Casciotti KL, Francis CA (2010). Activity, abundance and diversity of nitrifying archaea and bacteria in the central California Current. *Environmental Microbiology* **OnlineEarly:** 9999.

Santoro AE, Francis CA, de Sieyes NR, Boehm AB (2008). Shifts in the relative abundance of ammonia-oxidizing bacteria and archaea across physicochemical gradients in a subterranean estuary. *Environmental Microbiology* **10:** 1068-1079.

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

Teira E, Reinthaler T, Pernthaler A, Pernthaler J, Herndl GJ (2004). Combining catalyzed reporter deposition-fluorescence in situ hybridization and microautoradiography to detect

substrate utilization by Bacteria and Archaea in the deep ocean. *Applied and Environmental Microbiology* **70:** 4411-4414.

Treusch AH, Leininger S, Kletzin A, Schuster SC, Klenk H-P, Schleper C (2005). Novel genes for nitrite reductase and Amo-related proteins indicate a role of uncultivated mesophilic crenarchaeota in nitrogen cycling. *Environmental Microbiology* **7:** 1985-1995.

Urakawa H, Martens-Habbena W, Stahl DA (2010). High abundance of ammonia-oxidizing Archaea in coastal waters, determined using a modified DNA extraction method. *Applied and Environmental Microbiology* **76:** 2129-2135.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA *et al* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Verity PG (2002). A decade of change in the Skidaway River estuary. I. Hydrography and nutrients. *Estuaries* **25:** 944-960.

Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ *et al* (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proceedings of the National Academy of Sciences* **107:** 8818-8823.

Wuchter C, Abbas B, Coolen MJL, Herfort L, van Bleijswijk J, Timmers P *et al* (2006). Archaeal nitrification in the ocean. *Proceedings of the National Academy of Sciences of the USA* **103:** 12317-12322.

Zaikova E, Walsh DA, Stilwell CP, Mohn WW, Tortell PD, Hallam SJ (2010). Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environmental Microbiology* **12:** 172-191.

Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* **7:** 203-214.

**Table 3.1.** Distribution of cDNA reads among functional categories of the annotation pipeline and among marine group 1 Crenarchaeota and ammonia oxidizing Bacteria taxonomic groupings.

| **Category** | **Total**[1] |
|---|---|
| Total number of pyrosequencing reads | 2,181,899 |
| Total RefSeq hits | 560,389 |
| Number of reads assigned to Marine Group 1 Crenarchaeota | 17,386 |
| Reads assigned to "*Candidatus Nitrosopumilus maritimus*" strain SCM1 | 16,914 |
| Number of different N. maritimus ORFs | 786 |
| Number of N. maritimus ORFs hit >50 times | 32 |
| Reads assigned to C. symbiosum | 472 |
| Number of different C. symbiosum ORFs | 84 |
| Number of C. symbiosum ORFs hit >50 times | 2 |
| Number of reads assigned to Ammonia Oxidizing Bacteria (AOB) | 2,651 |
| Number of different AOB ORFs | 1,253 |
| Reads assigned to *Nitrosococcus* | 439 |
| Reads assigned to *Nitrosomonas* | 382 |
| Reads assigned to *Nitrosospira* | 272 |
| Number of AOB ORFs hit >50 times | 3 |

1) Sum of reads from both libraries.

**Table 3.2.** Crenarchaeota ORFs represented by 50 or more reads in metatranscriptomes retrieved from Georgia coastal waters. "Locus" refers to the Locus_Tag identifier assigned to the gene, "Sample Count" is the number of reads for which that locus was identified as the top hit by BLASTx. "Annotation" gives the identity of the gene product. The median and range of the bit scores for all hits to a particular gene are also given.

| Locus | Total Count | Annotation | Bit Scores | | |
|---|---|---|---|---|---|
| | | | Median | Low | High |
| Nmar_1547 | 3812 | hypothetical protein | 141 | 40 | 176 |
| Nmar_1502 | 1623 | AmoC | 149 | 40 | 191 |
| Nmar_1201 | 1492 | hypothetical protein | 110 | 40 | 178 |
| Nmar_1500 | 836 | AmoA | 149 | 41 | 183 |
| Nmar_1698 | 757 | ammonium transporter | 98 | 40 | 149 |
| Nmar_1667 | 633 | hypothetical protein (NirK?) | 107 | 41 | 169 |
| Nmar_1650 | 598 | hypothetical protein | 144 | 40 | 169 |
| Nmar_1501 | 588 | hypothetical protein | 114 | 41 | 153 |
| Nmar_0239 | 404 | 4Fe-4S ferredoxin iron-sulfur binding domain-containing protein | 154 | 40 | 175 |
| Nmar_1259 | 373 | hypothetical protein (NirK?) | 129 | 40 | 167 |
| Nmar_0188 | 198 | hypothetical protein | 161 | 40 | 185 |
| Nmar_1503 | 198 | AmoB | 127 | 40 | 176 |
| Nmar_0345 | 166 | hypothetical protein | 99 | 45 | 139 |
| CENSYa_1453 | 166 | ammonia permease | 111 | 42 | 135 |
| Nmar_0182 | 154 | hypothetical protein | 124 | 41 | 166 |
| Nmar_1507 | 142 | hypothetical protein | 86 | 42 | 164 |
| Nmar_1303 | 140 | iron-sulfur cluster assembly accessory protein | 143 | 48 | 157 |
| Nmar_1688 | 133 | H+transporting two-sector ATPase C subunit | 68 | 42 | 89 |
| Nmar_0343 | 119 | hypothetical protein | 120 | 42 | 164 |
| Nmar_0344 | 117 | hypothetical protein | 134 | 41 | 161 |
| Nmar_1537 | 97 | 4Fe-4S ferredoxin iron-sulfur binding domain-containing protein | 154 | 50 | 186 |

| | | | | | |
|---|---|---|---|---|---|
| Nmar_0588 | 95 | ammonium transporter | 109 | 42 | 171 |
| Nmar_1765 | 94 | 4Fe-4S ferredoxin iron-sulfur binding domain-containing protein | 165 | 42 | 188 |
| Nmar_0700 | 88 | hypothetical protein | 142 | 42 | 171 |
| Nmar_0561 | 82 | major intrinsic protein | 112 | 54 | 174 |
| Nmar_1102 | 76 | blue (type1) copper domain-containing protein | 110 | 40 | 164 |
| Nmar_0238 | 71 | 4Fe-4S ferredoxin iron-sulfur binding domain-containing protein | 141 | 40 | 173 |
| Nmar_1034 | 70 | elongation factor 1-alpha | 145 | 44 | 181 |
| Nmar_0627 | 69 | hypothetical protein | 116 | 42 | 161 |
| Nmar_0394 | 67 | superoxide dismutase | 159 | 43 | 197 |
| CENSYa_161 | 66 | hypothetical protein | 67 | 44 | 115 |
| Nmar_0183 | 59 | cytochrome c oxidase subunit II | 131 | 40 | 193 |
| Nmar_0959 | 54 | ketol-acid reductoisomerase | 145 | 52 | 181 |
| Nmar_0558 | 50 | hypothetical protein | 93 | 45 | 166 |

**Table 3.S1.** Primers used in this study. 16S rRNA primer names beginning with "Tm" are Taqman primers. Under "Use," Q = qPCR and S = sequencing. Cren. = Crenarchaeota.

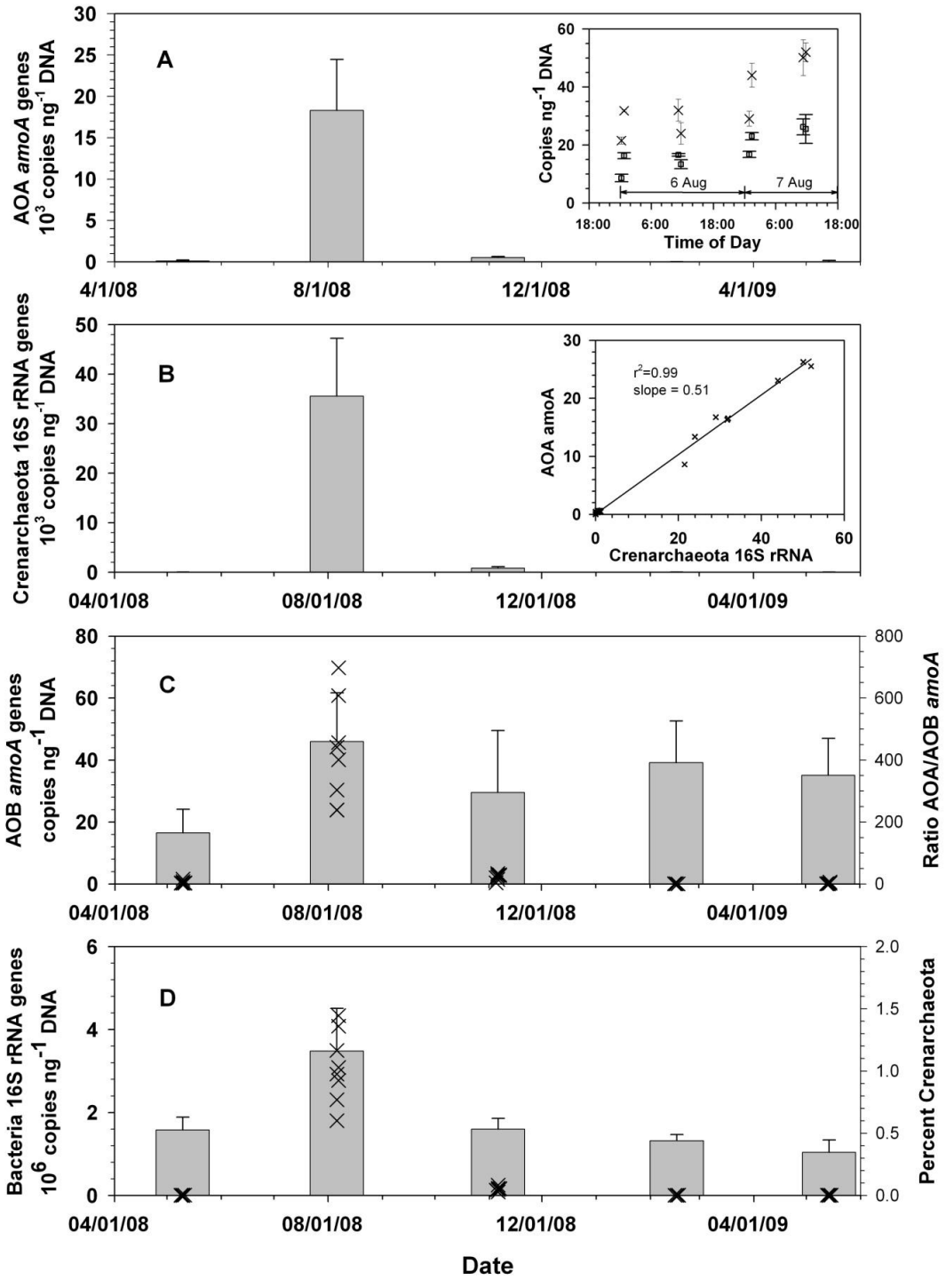
| **Target** | **Gene** | **Primer or Probe** | **Use** | **Sequence (5' to 3')** | ***Reference*** |
|---|---|---|---|---|---|
| Bacteria | 16S rRNA | BACT1369F | Q | CGGTGAATACGTTCYCGG | (Suzuki et al., 2000) |
| | 16S rRNA | PROK1492R | Q | GGWTACCTTGTTACGACTT | (Suzuki et al., 2000) |
| | 16S rRNA | Tm1389F | Q | CTTGTACACACCGCCCGTC | (Suzuki et al., 2000) |
| Cren. | 16S rRNA | ARCHGI334F | Q | AGATGGGTACTGAGACACGGAC | (Suzuki et al., 2000) |
| | 16S rRNA | ARCHGI554R | Q | CTGTAGGCCCAATAATCATCCT | (Suzuki et al., 2000) |
| | 16S rRNA | Tm519AR | Q | TTACCGCGGCGGCTGGCAC | (Suzuki et al., 2000) |
| Archaea | 16S rRNA | 21F | S | TTCCGGTTGATCCYGCCGGA | (DeLong, 1992) |
| | 16S rRNA | 958R | S | YCCGGCGTTGAMTCCAATT | (DeLong, 1992) |
| AOB | *amoA* | *amoA*-1F | Q | GGGGTTTCTACTGGTGGT | (Rotthauwe et al., 1997) |
| | *amoA* | *amoA*r NEW | Q | CCCCTCBGSAAAVCCTTCTTC | (Hornek et al., 2006) |
| AOA | *amoA* | Arch-*amoA*-for | Q | CTGAYTGGGCYTGGACATC | (Wuchter et al., 2006) |
| | *amoA* | Arch-*amoA*-rev | Q | TTCTTCTTTGTTGCCCAGTA | (Wuchter et al., 2006) |
| | *amoA* | Arch-*amoA*F | S | STAATGGTCTGGCTTAGACG | (Francis et al., 2005) |
| | *amoA* | Arch-*amoA*R | S | GCGGCCATCCATCTGTATGT | (Francis et al., 2005) |

**Table 3.S2.** Number of MG1C and AOB transcripts/cell in each of the samples calculated from metatranscriptomic data as explained in chapter 2. The total number of MG1C or AOB cells in the sample was calculated as described in the text from the volume filtered (5.75 L), total prokaryote abundance determined by epifluorescence microscopy (4.2 x $10^9$ cells/L) and relative abundance of MG1C and AOB determined by qPCR.

| **Category** | **Sample FN56** | **Sample FN57** | **Sum** |
|---|---|---|---|
| Total RefSeq hits | 287,137 | 273,252 | 560,389 |
| Total Marine Group 1 Crenarchaeota hits recovered in library | 12,246 | 5,140 | 17,386 |
| pGEM transcripts added | 4.70E+10 | 4.70E+10 | 9.4E+10 |
| pGEM hits recovered in library | 4,014 | 6,865 | 10,879 |
| Total number of MG1C hits expected in sample (from recovery of pGEM transcripts) | 1.4339E+11 | 3.52E+10 | 1.5E+11 |
| **Transcripts per MG1C cell** | **321** | **79** | **168** |
| | | | |
| Total *Nitrosococcus* Hits | 629 | 686 | 1315 |
| Total *Nitrosomonas* Hits | 369 | 341 | 710 |
| Total *Nitrosospira* Hits | 311 | 315 | 626 |
| All AOB Hits in Library | 1,309 | 1,342 | 2,651 |
| Total number of AOB hits expected in sample (from recovery of pGEM transcripts) | 1.5327E+10 | 9.19E+09 | 2.29E+10 |
| **Transcripts per AOB cell[1]** | **26,356** | **15,799** | **19,695** |

1) AOB abundance calculated from AOB *amoA* gene abundance (Figure 1) assuming a gene dosage of 2.5 copies of *amoA* per AOB genome (Norton et al. 2002).

**Figure 3.1.** Time series of quantitative, real-time PCR (qPCR) estimates of the abundance of *amoA* and 16S rRNA genes at the sampling site. Means (wide bars) and standard deviations (vertical lines) of 8 samples collected over 2 day periods are shown. In some cases, the bars are smaller than the abcissa. A. Archaeal *amoA* genes. Inset shows the time series of changes in amoA (□) and Crenarchaeota 16S rRNA (X) gene abundance on 6-7 August. Vertical bars are standard deviations of triplicate qPCR determinations for each sample. B. Marine Group 1 Crenarchaeota 16S rRNA gene abundance. Inset shows Archaeal *amoA* versus Crenarchaeota 16S rRNA gene abundance for each sample (regression line slope= 0.51, $r^2$=0.99). C. Bacterial *amoA* gene abundance. Crosses show the ratio of Archaeal *amoA* to Bacterial *amoA* for each sample. D. Bacteria 16S rRNA gene abundance. Crosses show the relative abundance of Crenarchaeota as a percentage of the prokaryotes (Bacteria + Crenarchaeota) in each sample assuming a gene dosage of 1 16S rRNA gene per genome for Marine Group 1 Crenarchaeota (from genomes annotated in DOE's IMG database) and 1.8 16S rRNA genes per genome as an average for marine bacteria (Biers *et al.*, 2009).


A
AOA amoA genes
10^3 copies ng^-1 DNA
Copies ng^-1 DNA
6 Aug
7 Aug
Time of Day
4/1/08
8/1/08
12/1/08
4/1/09
B
Crenarchaeota 16S rRNA genes
10^3 copies ng^-1 DNA
r^2=0.99
slope = 0.51
AOA amoA
Crenarchaeota 16S rRNA
04/01/08
08/01/08
12/01/08
04/01/09
C
AOB amoA genes
copies ng^-1 DNA
Ratio AOA/AOB amoA
D
Bacteria 16S rRNA genes
10^6 copies ng^-1 DNA
Percent Crenarchaeota
Date

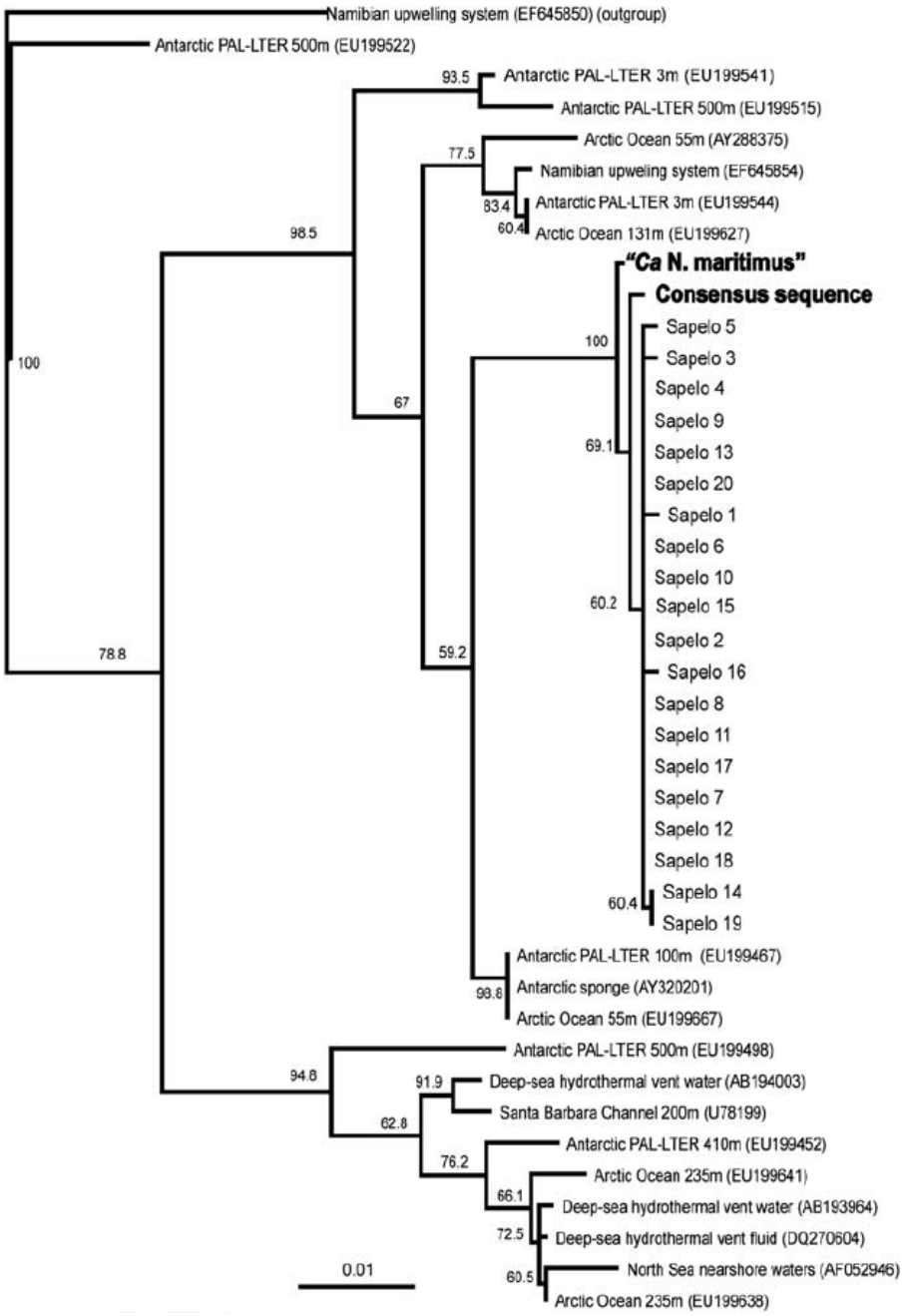



**Figure 3.2.** Phylogenetic analysis of the marine group 1 Crenarchaeota 16S rRNA sequences. The consensus sequence was obtained by assembling MG1C 16S rRNA reads contaminating the metatranscriptome. Sequences labeled "Sapelo" are from cloned PCR amplicons produced with DNA from the same sample. Reference sequences are shown in black except "*Candidatus* Nitrosopumilus maritimus" strain SCM1, which is shown in red. GenBank accession numbers are given in parentheses. This is a neighbor joining tree based on 876 bp sequences. Bootstrap analysis was used to estimate the reliability of phylogenetic reconstructions and support is shown if >50% (100 iterations).

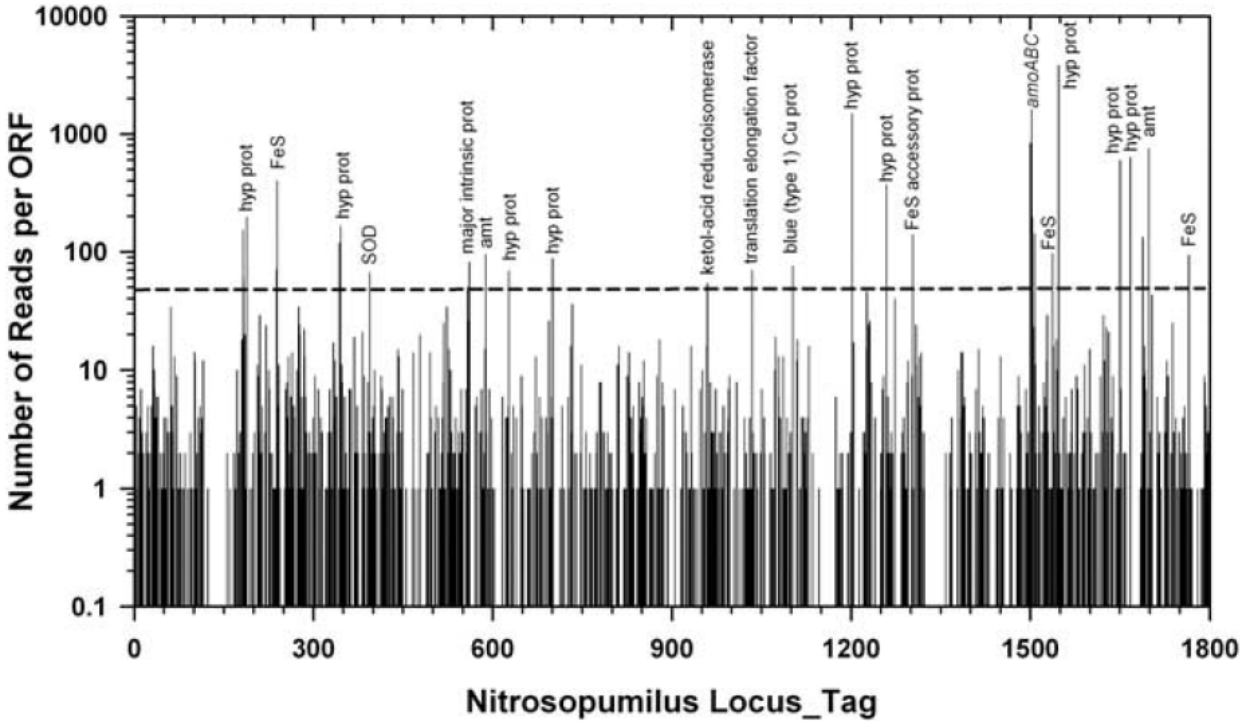



**Figure 3.3.** Distribution of pyrosequencing reads among *Nitrosopumilus* ORFs. The horizontal line is positioned at 50 hits per ORF and indicates the cutoff used to define highly expressed ORFS. Text over the longest bars identifies the annotation for that ORF (left to right): hyp prot – hypothetical protein; FeS – 4Fe-4S ferredoxin iron-sulfur binding domain-containing protein; amt – ammonium transporter; FeS accessory protein - iron-sulfur cluster assembly accessory protein; *amoABC* – ammonia monooxygenase subunits.

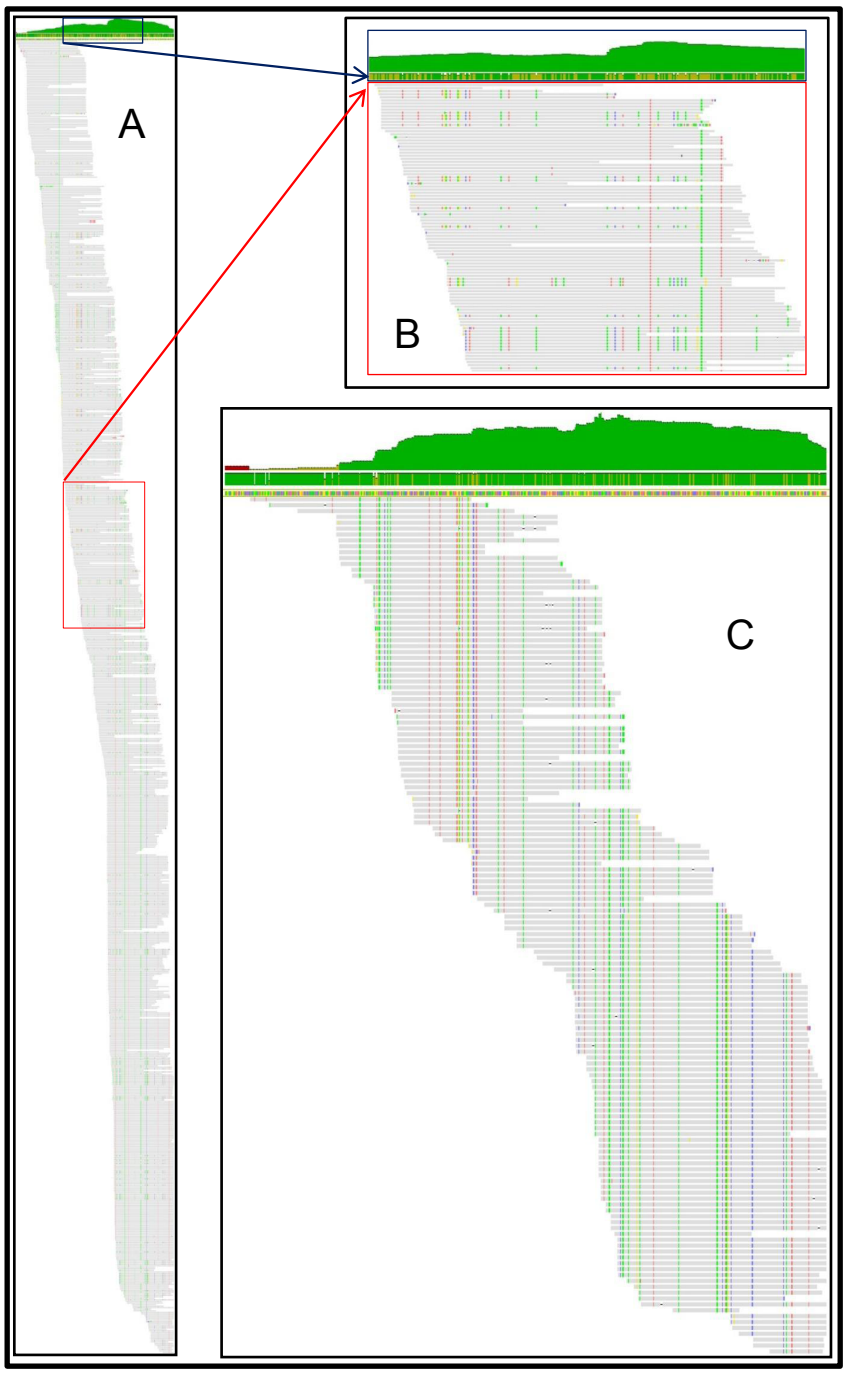



**Figure 3.4.** Assembly of 836 of 836 reads assigned to the "*Ca* N. maritimus" strain SCM1 *amoA* ORF (Nmar_1500) against the Nmar_1500 sequence as a scaffold. A. Coverage curve and assembly. B. Close-up of a portion of the assembly (origin indicated by boxes) showing primary and secondary sequence variants. C. Assembly of 147 minority sequence variants. Objects from top to bottom: coverage curve (green shape, 0-558); identity at each position (green and olive bar, 0-100%), reference sequence (not shown in Panel B), aligned reads. Highlighted positions in the aligned reads indicate disagreements with the reference sequence (code: red – A, blue – C, orange – G, green – T).

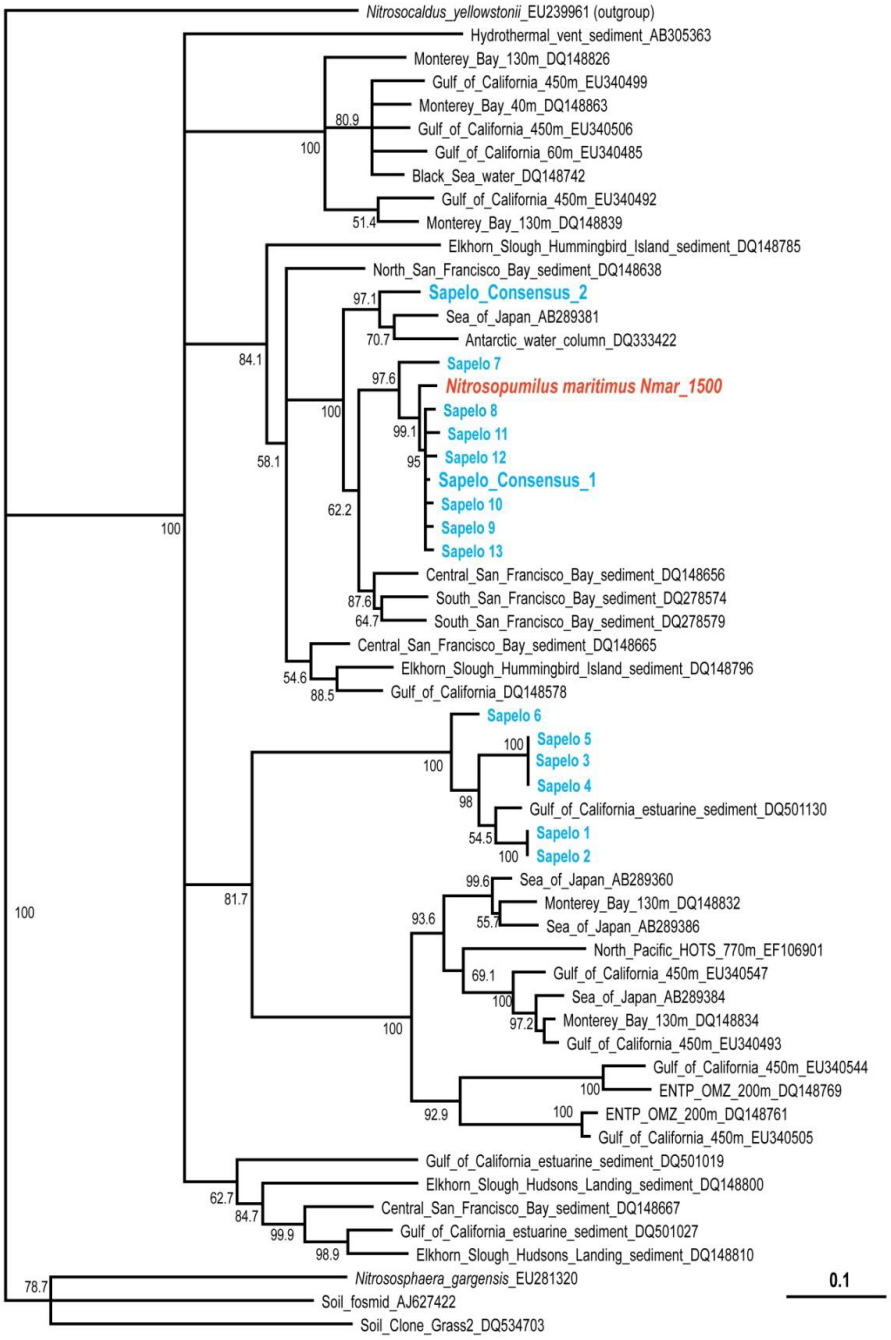



**Figure 3.5.** Phylogenetic analysis of marine group 1 Crenarchaeota *amoA* sequences in metatranscriptome and DNA samples. Consensus sequences (1 = dominant, 2 = minority) were obtained by assembling 16S *amoA* reads in the metatranscriptome. Sequences labeled ―Sapelo" are from cloned PCR amplicons produced with DNA from the same sample. Sequences were aligned with reference sequences using Clustal W. Minimum evolutionary distances were calculated using the Kimura two-parameter model. Reference sequences are shown in black except ―*Candidatus* Nitrosopumilus maritimus" strain SCM1, which is shown in red. GenBank accession numbers are given in parentheses. This is a neighbor joining tree based on 595 bp sequences. Bootstrap analysis was used to estimate the reliability of phylogenetic reconstructions and support is shown if >50% (100 iterations).

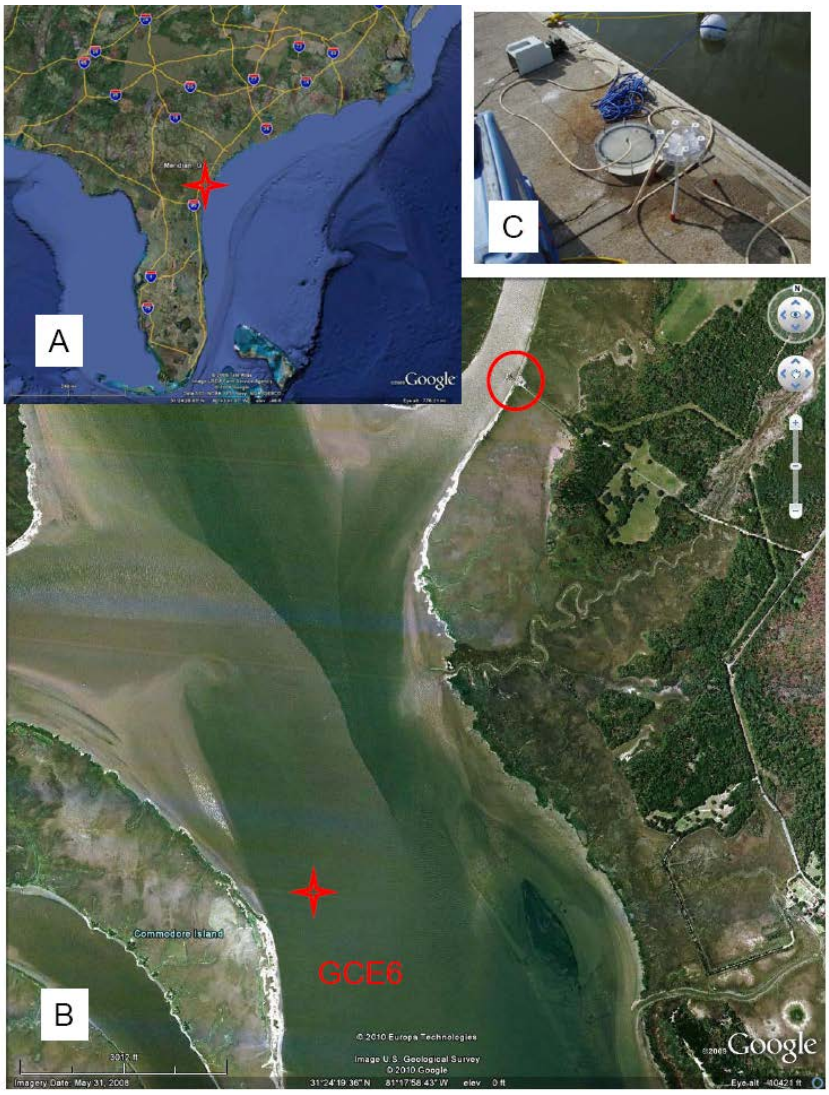



**Figure 3.S1.** A. Location of Sapelo Island, Georgia, USA (red star). B. Location of the sampling station on the Duplin River, Sapelo Island, Georgia (red circle) and of the nearest GCE-LTER station where nutrient data are collected (red star). C. Set-up of sampling gear on the dock shown in B.

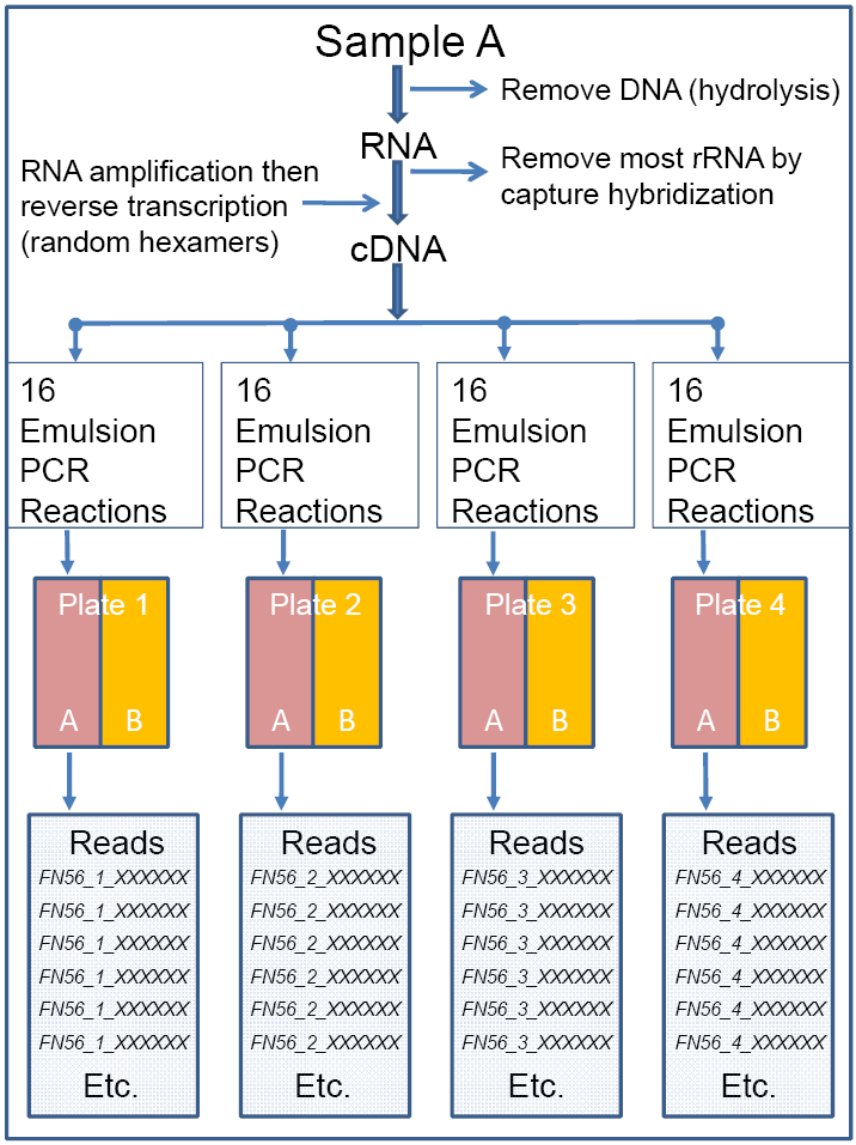



**Figure 3.S2.** Schematic of the mRNA preparation and pyrosequencing strategy used in this study.

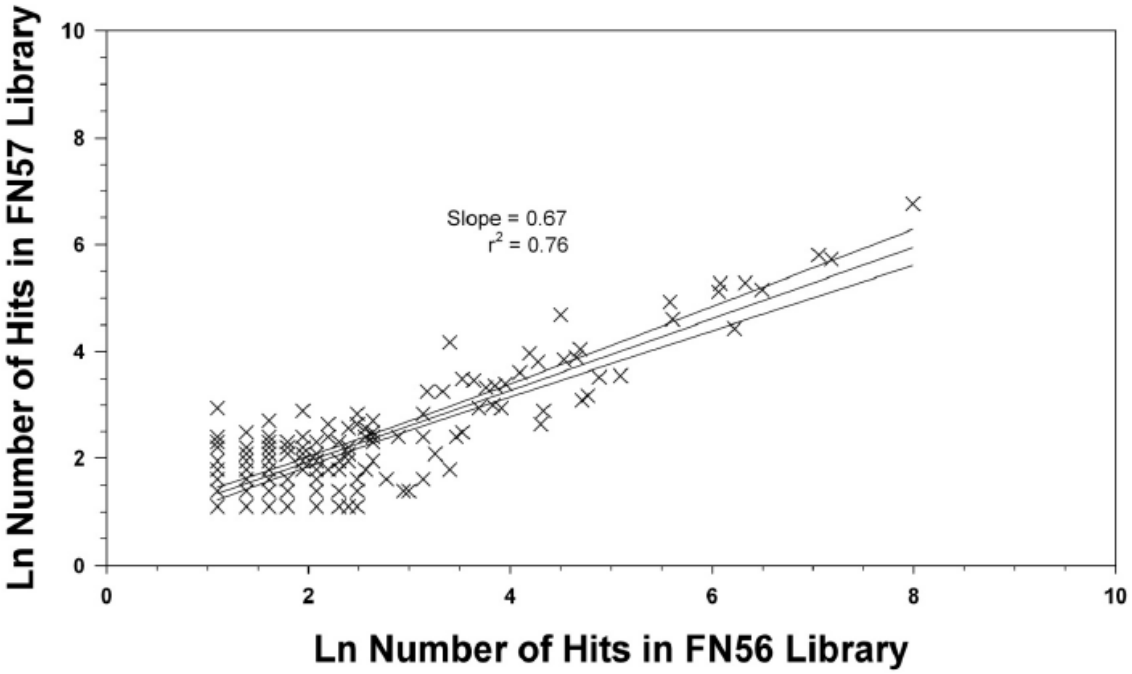



**Figure 3.S3.** Comparison of the number of transcripts assigned to the same ORF in metatranscriptomes from samples FN56 and FN57. The figure plots pairs for which at least 3 reads from each sample were assigned to that ORF (accounting for 89.8% of all reads assigned to MG1C, representing transcripts from 160 different ORFs). Numbers of transcripts per ORF are shown as natural logs. The Type 1 linear regression line and 95% confidence limits of data are shown. Deviations from this line indicate that a transcript is more (or less) abundant in one sample versus the other.

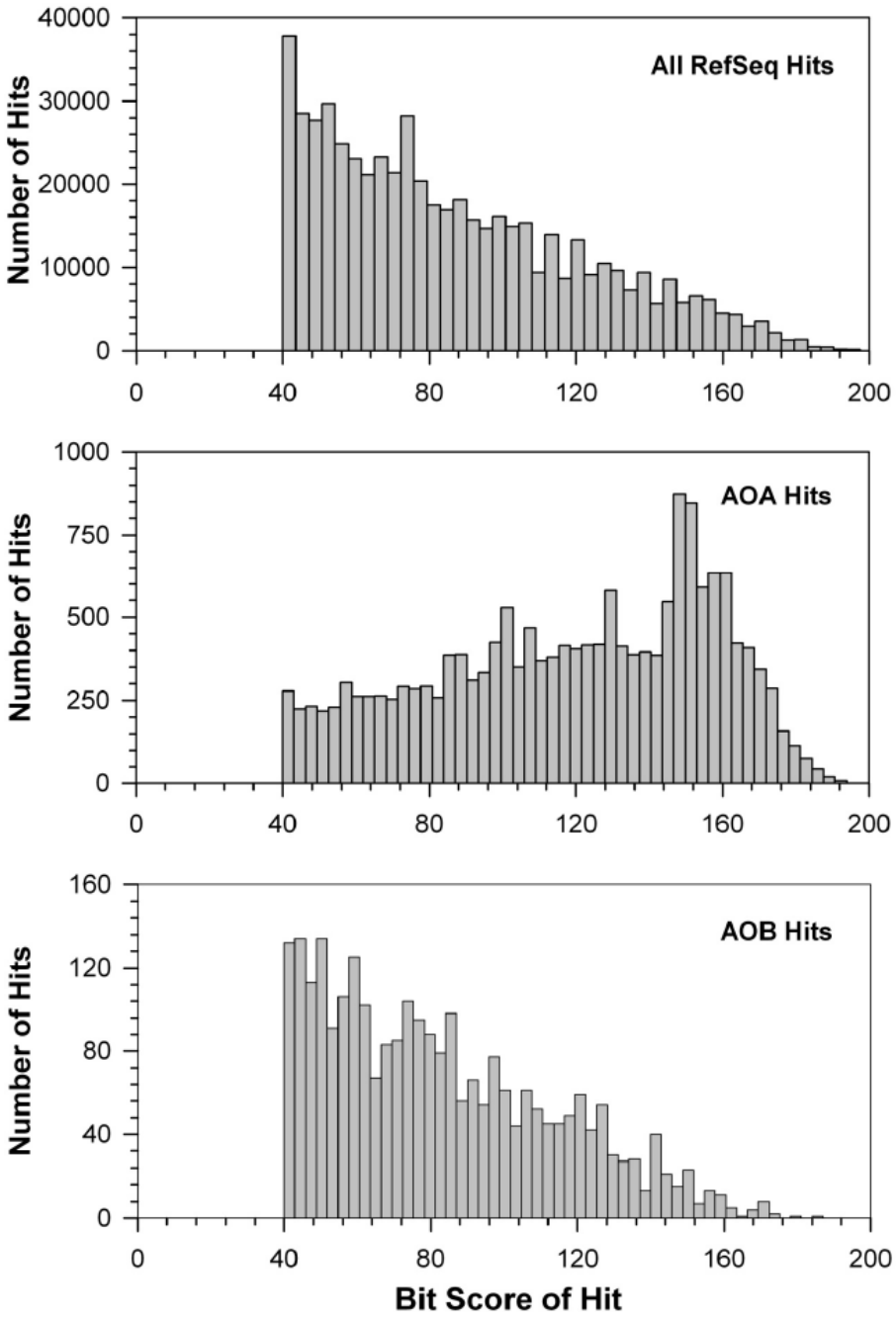



**Figure 3.S4.** Frequency distribution of bit scores for BLASTx hits of metatranscriptome sequences retrieved from our samples against the RefSeq database. A. All sequences in the metatranscriptome; B. All hits assigned to MG1C; C. All hits assigned to AOB.

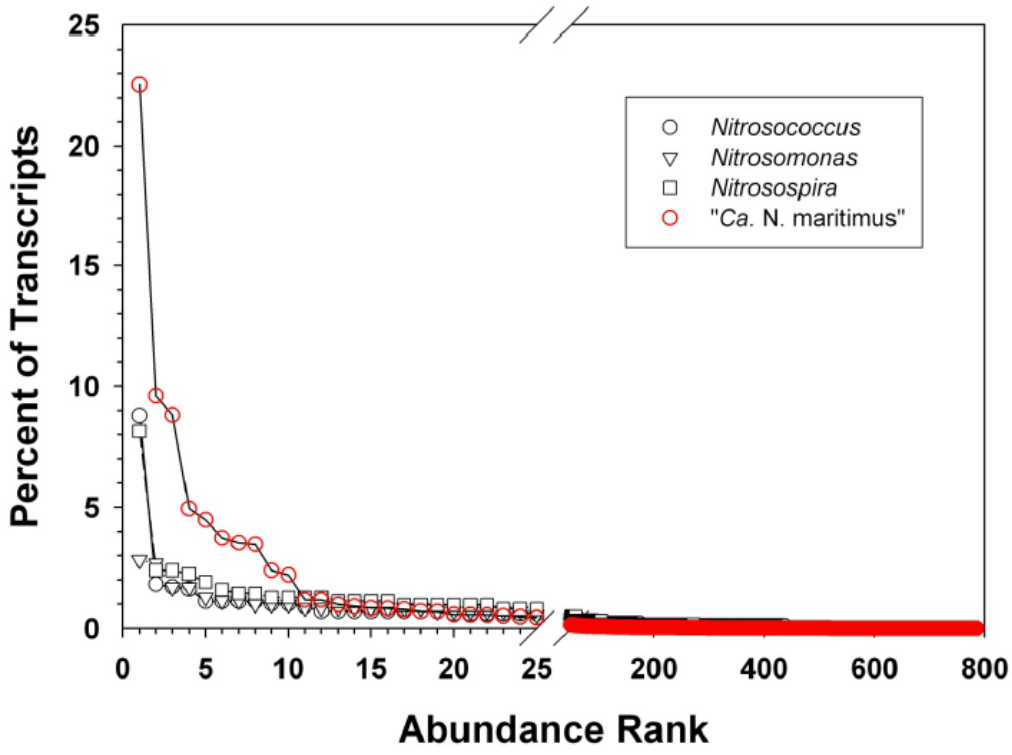



**Figure 3.S5.** Frequency distributions of reads among *Nitrosococcus, Nitrosomonas, Nitrosospira* or "*Ca*. N. maritimus" ORFs. The number of reads assigned to each ORF is normalized as a percentage of all of the reads assigned to each taxon. The distributions for hits to MG1C and AOB ORFs are statistically significantly different ($p<0.05$, jack-knife estimate of 95% CL for Pielou's Evenness; $p<0.0001$ for Mann-Whitney U-test) with AOB reads being more evenly distributed across ORFS than MG1C reads.

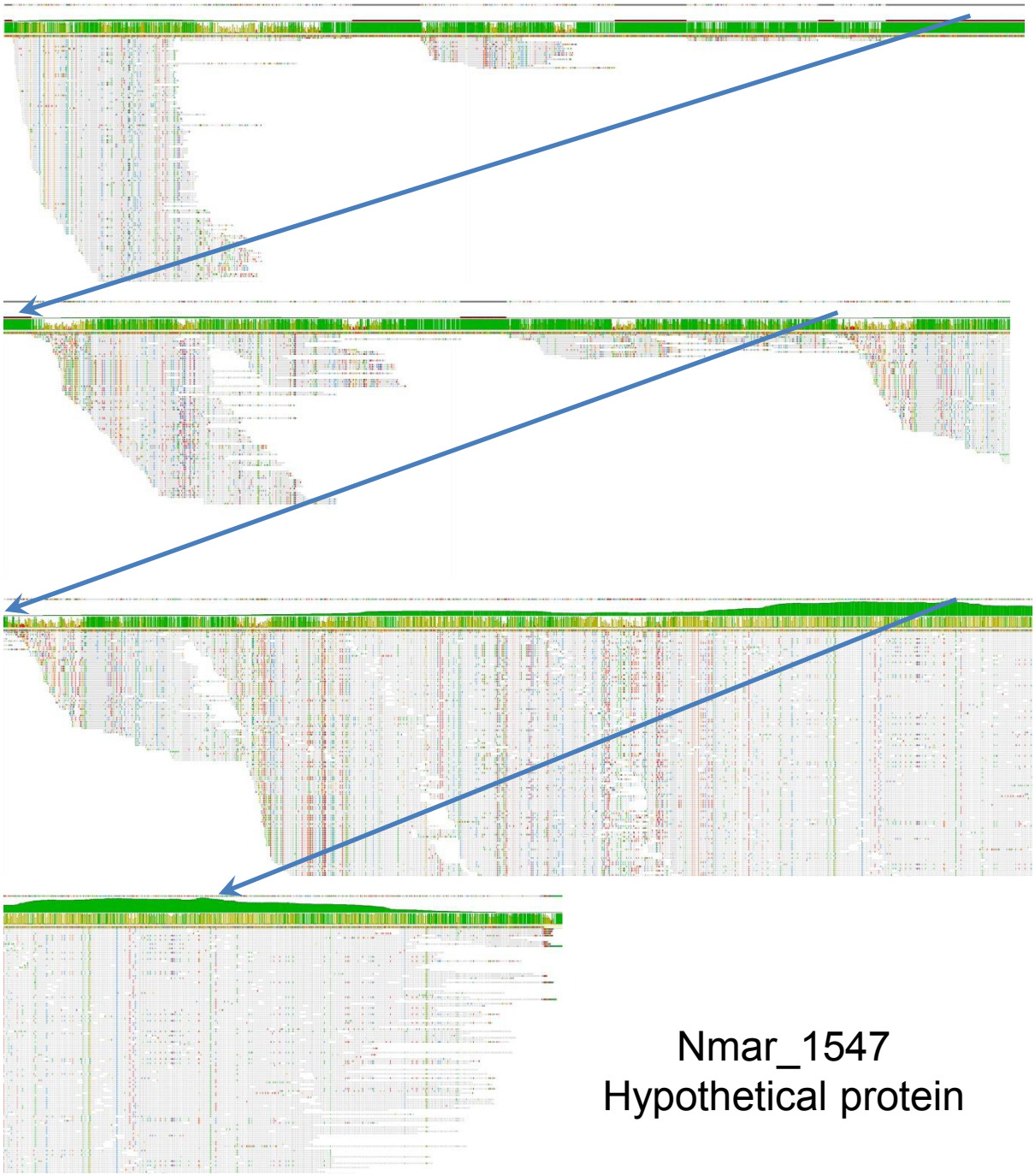



**Figure 3.S6.** Assembly of 2,803 of 3,812 reads assigned to Nmar_1547. Shown (top of figure to bottom) are the consensus sequence (≥50% of sequences in agreement for each position, bases disagreeing with the reference sequence are highlighted); the coverage curve (green/olive, range 0-1113), identity at each position (green/olive, 0-100%), the reference sequence, and the aligned reads with disagreements to the reference sequence highlighted (code: red – A, blue – C, orange – G, green – T). Not all of the reads in the alignment are shown in this vertically compressed display. Blue lines connect the same position in successive panels.

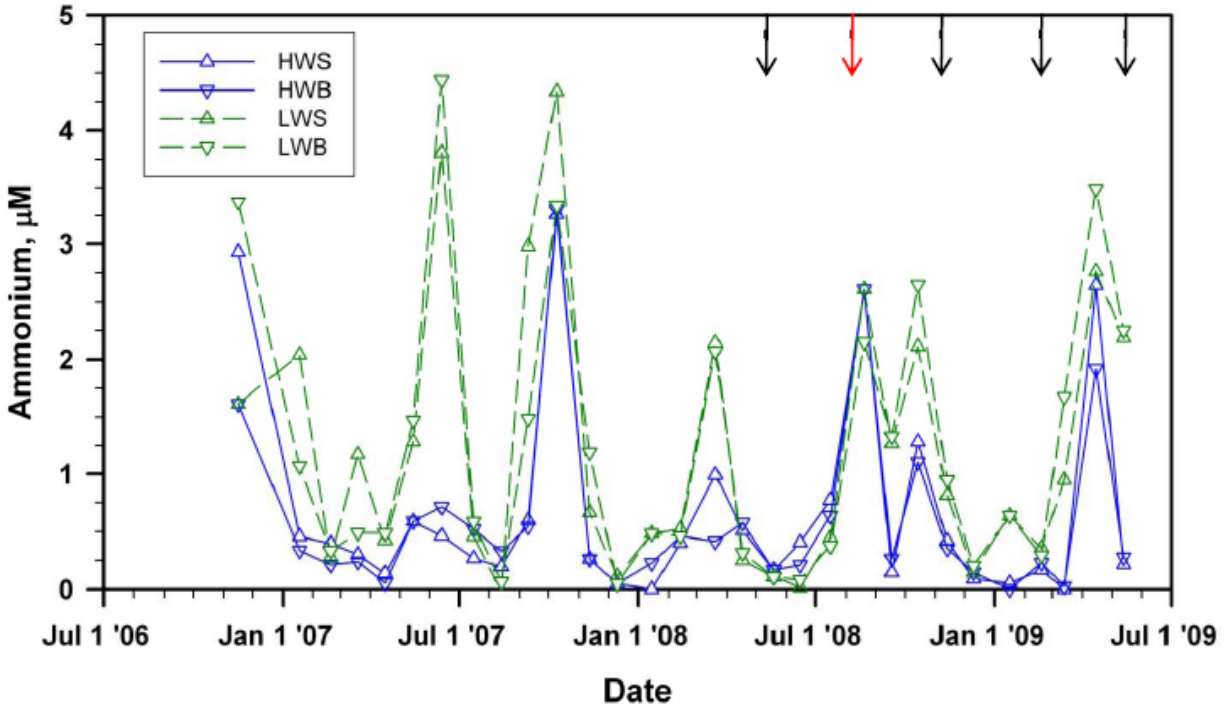



**Figure 3.S7.** Time series of ammonium concentration at station GCE6, 3.0 km from the sampling location. HWS – high tide, collected at 0.2 m depth; HWB – high tide, collected 0.5 m above the bottom (~6 m depth); LWS low tide, collected at 0.2 m depth; LWB low tide, collected 0.5 m above the bottom (~4 m depth). Arrows along the top of the panel indicate when samples for qPCR analysis were collected, red arrow indicates metatranscriptome sample.

# CHAPTER 4

# EXPRESSION PATTERNS REVEAL NICHE DIVERSIFICATION IN A MARINE MICROBIAL ASSEMBLAGE[1]

[1]Gifford, S.M., Sharma, S., Booth, M., and Moran, M.A. To be submitted to *International Society of Microbial Ecology Journal.*

**Abstract**

Resolving the ecological niches of coexisting marine microbial taxa is challenging due to the high species richness found in microbial communities and the extensive functional redundancy evident in marine bacterial genomes and metagenomes. Metatranscriptomics provides information on dynamic gene expression and can be useful for distinguishing biogeochemical activities of individual taxa that share the same environment. Here, we examined bacterioplankton transcription patterns in a well-mixed coastal ocean to characterize taxon-specific gene expression. Sequencing with the Illumina platform (producing >11 million protein encoding reads) allowed for the simultaneous examination of the activities of thousands of microbial groups. The >200,000 ribosomal protein reads found in the metatranscriptome libraries showed distinct patterns in abundance among these taxonomic bins, indicative of differences in *in situ* growth rate. For 16 genome bins chosen for closer inspection, gene expression levels related to functional overlap; that is, counts of transcripts increased in proportion to how common the gene was in the other genomes (referred to as 'ortholog number'). Genes showing atypically high expression within an ortholog number category were used to obtain insight into that taxon's unique functional role in the community. Genes for the transport and metabolism of a wide variety of substrates and for physical interactions with the biotic and abiotic environment were typical indicator genes for the various groups. Expression analyses identified distinct roles of individual microbial taxa within this highly complex community, providing insight into how the assemblage is maintained

## Introduction

Relationships between taxonomic composition and ecological function have been difficult to establish for marine bacterioplankton communities. Recent genomic and metagenomic inventories have unquestionably improved understanding of the potential functional roles of marine taxa as reflected in their gene repertoires (Moran *et al.*, 2004; Giovannoni *et al.*, 2005; Rusch *et al.*, 2007; Delong *et al.*, 2006). Nonetheless, genome comparisons of major marine bacterial groups reveal that many genes of known biogeochemical or ecological relevance are broadly distributed taxonomically (Moran, 2008).

The competitive exclusion principle, originally developed to conceptualize the organization of macroorganism communities (Hardin, 1960), posits that species richness is maintained by niche differentiation. While this idea has also been applied to microbial communities (Fuhrman *et al.*, 2006; Mou *et al.*, 2008), microbes have presented difficulties for a competitive exclusion framework from its beginning (Hutchinson, 1961). The apparent broad overlap in potential ecological roles, which has now been reinforced by genomic and metagenomic data, suggests that gene inventories will not be sufficient to assess functional niches of microbes. Thus, it remains a challenge to establish clear and unique ecological roles for individual bacterial taxa that will lead to better understanding and prediction of marine ecosystem processes.

One potentially important component of ecological function that has not typically been measured for bacteria is heterogeneity in how they sense and respond to the environment. Thus it is possible that bacteria from two different taxa that share the same functional gene have different regulation strategies for expression. Metatranscriptomics provides information on dynamic gene expression of individual microbial taxa sharing the same environment (Poretsky *et*

*al.*, 2005; Frias-lopez *et al.*,2008) and therefore has the power to address this additional aspect of functional niche.

Here, we examined bacterioplankton transcription patterns in a well-mixed coastal ocean to characterize taxon-specific gene expression. The Illumina GAIIx platform provided deep coverage of the community transcriptome, allowing assessment of differential gene transcription across representative genome bins. We used the relationship between transcription level and ortholog number (a proxy for functional redundancy) to identify genes with atypically high transcription levels encoding unique functional capabilities. This analysis reveals new details of the functional niches occupied by members of a marine bacterioplankton community and provides insights into the diversity of strategies that support this complex microbial assemblage.

## Methods

**Sample collection.** Samples were collected as part of the Sapelo Island Microbial Observatory (SIMO, http://www.simo.marsci.uga.edu), a multiyear time series examining expression in microbial communities of the coastal Southeastern U.S. Quarterly sampling expeditions, representing the winter, spring, summer, and fall seasons, are conducted at Marsh Landing (31°25'4.08N, 81°17'43.26W), Sapelo Island, Georgia, USA. Four samples representative of each season were chosen for this analysis: FN96 (7 November 2008), FN116 (17 February 2009), FN125 (14 May 2009), and FN158 (14 August 2009). All samples were collected at night, 4 to 6 h after sunset and 1 h before high tide. Cell collection for RNA extraction was conducted as described previously (Poretsky *et al.*, 2009; chapter 2). Briefly, water was drawn from approximately 1 m depth using a peristaltic pump and passed through a 3 μm prefilter (Capsule Pleated 3 mm Versapor Membrane; Pall Life Sciences, Ann Arbor, MI,

USA) and 0.22 μm collection filter (Supor polyethersulfone; Pall Life Sciences). After filtering 6 to 8 L, the 0.22 μm filter was placed into a WhirlPak© bag and flash frozen in liquid nitrogen. Total time from the start of filtration to flash freezing was 11 to 14 min.

**RNA processing and sequencing.** RNA processing in preparation for sequencing was done as described by Poretsky *et al.* (2009) and in chapter 2. The 0.22 μm collection filters were shattered, placed into 50 ml falcon tubes with 8 ml of RLT buffer (Qiagen, Valencia, CA, USA) and 2 ml of PowerSoil beads (MO BIO, Carlsbad, CA, USA) and vortexed for 10 min on a MO BIO vortex adapter. RNA was extracted from the 50 ml tubes using an RNeasy kit (Qiagen), and any contaminating DNA was digested using TurboDNAse (Applied Biosystems, Austn, TX, USA). Ribosomal RNA (rRNA) was reduced using a two step approach. The samples were first treated enzymatically with the mRNA only isolation kit (Epicentre, Madison, WI, USA) and then by subtractive hybridization using MicrobeExpress and MicrobeEnrich kits (Applied Biosystems). The enriched mRNA sample was then linearly amplified using the Message Amp II-Bacteria kit (Applied Biosystems), reverse transcribed to cDNA with the Universal Riboclone cDNA synthesis system (Promega,Madison, WI, USA), and purified with the QIAQuick PCR purificaton kit (Qiagen). The four cDNA samples were sheared to ~300 bp, barcoded, and sequenced in one lane of an Illumina GAIIx run.

**Bioinformatics pipeline.** A 25,000 read subsample of each library was searched against the SILVA database of large and small ribosomal genes sequences (www.arb-silva.de) with BLASTn (bitsocre ≥ 50). The 25K subsample BLAST was then repeated, except the search was against a small set of select rRNA sequences found in the first analysis to be most similar to the samples. The first and second BLAST runs were compared to identify sequences missed using the custom rRNA subject database. Sequences for the missing taxa were then added to the

custom rRNA sequence database, and the processes repeated until all rRNA hits identified in the full SILVA database BLAST were found with the custom rRNA database BLAST. The complete metatranscriptome libraries were then searched against the custom database to identify and remove rRNA sequences.

All remaining, non-rRNA sequences were compared to NCBI's RefSeq database (version 43) using BLASTx with a bitscore cutoff $\geq$40 to identify protein encoding sequences. A read's taxonomic affiliation was assigned based on the top RefSeq hit. Ribosomal protein encoding sequences were identified by a text based query of the read annotations.

**Ortholog identification.** For a set of 16 genome bins selected based on their representation in the transcriptome and their taxonomic breadth, orthologous genes were identified in a two-step process. Each gene in a subject genome was reciprocally blasted against the other 15 genomes. Genes with a reciprocal hit with an E value < $10^{-4}$ were considered orthologs. The process was repeated using each of the 15 genomes as the subject to build a complete list of reciprocal best-hit ortholog pairs among the 15 genomes. The results of the ortholog search were compiled into a single table. Initially the table consisted of 46,339 rows (one for each of the genes in the 15 genomes), with each row containing the subject gene and any orthologous genes in the other genomes. The rows of the table were randomly shuffled to reduce the potential for biasing ortholog selection towards any one genome. Beginning at the top, rows with common orthologs were retrieved and the most inclusive row (that which contained the most genes) was retained as the final ortholog group. Any other occurrences of the genes in this row were removed from the master table. This procedure was repeated until every row had been processed.

**Statistical analysis.** Statistically significant differences between gene sets grouped by ortholog count within a genome were identified using the non-parametric Wilcoxon rank-sum test ($P < 0.05$). Genes that had read counts >1.5 times the interquartile range of their ortholog count group were labeled as outliers. Rows in the ortholog master table (described above) that contained IDs of outlier genes were then retrieved to make a table of outlier orthologous relationships. The percent transcriptome for each outlier gene was calculated as the gene hits over the total number of hits to the reference genome. The resulting matrix was then transformed by multiplying by a factor of 1000 to separate genome bins that had detectable expression of a gene from those that had either no expression or no ortholog. A non-metric multidimensional scaling plot was created from the transformed data matrix using the metaMDS function (Oksanen *et al.*, 2011) in R with Wisconsin double standardization and a Bray Curtis dissimilarity matrix. MDS plots were generated with up to 10 dimensions, and a scree plot revealed that beyond 6 dimensions the stress did not decrease appreciatively. The first two dimensions explained 61% of the total variance and were plotted. To visualize how individual ortholog groups map with the genome positions, ortholog group scores are calculated in the metaMDS function based on a weighted average of genome NMDS scores multiplied by a gene's proportion of the transcriptome, and then expanded so that the variance within the ortholog groups scores matched that of the genome scores.

**Indicator gene analysis.** The indicator species analysis approach of Dufrene and Legendre (1997) was used for indicator gene identification with the genomes considered the samples and the genes considered species. A gene’s hit count as a proportion of the total reference genome hits was used as the abundance metric. The 16 genomes were divided into groups based on their phylogenetic relationships. The indicator value (IV) is as a product of the

proportion of expression a group contributed to the total expression of a gene times the proportion of genomes in the group expressing the gene, and is calculated as follows:

The mean percent transcriptome within a group is calculated as:

$a_{jk}$ = percent transcriptome of gene j in the genome i of group k

$n_k$ = number of genomes in group k

$$X_{kj} = \frac{\sum_{i=1}^{n_k} a_{ijk}}{n_k}$$

The specificity of expression towards one genome group over another is calculated as:

g = number of groups

$$A_{jk} = \frac{X_{kj}}{\sum_{k=1}^{g} X_{kj}}$$

The relative frequency, or the degree to which expression of a gene occurs in all of a group's genomes, is calculated as:

$b_{ijk}$ = presence or absence (1/0) of expression of gene j in genome i of group k.

$$F_{jk} = \frac{\sum_{i=1}^{n_k} b_{ijk}}{n_k}$$

The indicator value is the product of expression specificity and relative frequency, expressed as a percentage:

$$IV_{jk} = \left(A_{jk} \times F_{jk}\right) \times 100$$

Note that when there was only one genome in a group, the second term (relative frequency) is equal to 1, and the IV reduced to a measure of the proportion a genome contributed to the total global expression of a gene. Indicator genes were identified as those genes having an IV > 50,

expect for the 'All' group, in which case only genes with an IV > 93 (expressed in at least 15 out of 16 genomes) were considered.

## Results and Discussion

**Samples and sequencing.** Metatranscriptome samples were collected from Marsh Landing, Sapelo Island, Georgia, USA as part of the Sapelo Island Microbial Observatory program (SIMO; http://www.simo.marsci.uga.edu). The site is characteristic of nearshore coastal habitats of the Southeastern United States, with marsh, freshwater, and coastal influences. Samples were collected at night an hour before high tide in four months representative of the summer, fall, winter, and spring. After RNA processing and conversion to cDNA, Illumina Genome Analyzer IIx sequencing yielded 31 million reads. A BLASTn search against an in-house rRNA database revealed that 62% of the reads were rRNAs. The remaining 11 million potential protein encoding reads were compared to NCBI's RefSeq database via BLASTx. Over 4.1 million reads had significant hits (bit score > 40).

**Active community members.** Based on the highest scoring hit from the RefSeq BLAST, the reads binned to ~4,000 taxa (Table 4.1). The distribution of hits within these bins was log normal, with the top 200 bins accounting for 75% of all the hits, followed by a long tail of bins with very few hits. The Alphaproteobacteria, Gammaproteobacteria, Betaproteobacteria, and Bacteroidetes were the dominant transcript-producing groups (Table 4.1). The recently sequenced genome of "*Candidatus* Puniceispirillum marinum", the only SAR116 representative, recruited the most reads of any taxon. Small, streamlined genomes, such as those from "*Candidatus* Pelagibacter ubique", Betaproteobacterium KB13, Flavobacteria bacterium MS024-2A and MS024-3C, and *Nitrosopumilus maritimus*, were highly represented. Reference bins of

medium to large genomes, likely representative of ecological generalists, were also abundant, particularly those from the roseobacter clade and the 'oligotrophic marine gamma' (OMG) group. Hits to the Betaproteobacteria were predominantly to genomes of methylotrophic taxa. Archaea had few transcript hits in general, with the exception of *N. maritimus*, which was the 5$^{th}$ highest transcript-recruiting bin. Members of the Verrucomicrobiales, a recently described phylum identified in both terrestrial and aquatic habitats, were surprisingly well represented. Overall, the reference bins are indicative of a complex and highly diverse active coastal bacterioplankton community.

The average percent amino acid identity, which served as an index of how well transcripts matched the reference genomes to which they were assigned, ranged from 65 to 93% for the top 5 taxa in each taxonomic group (Table 4.1). The metatranscriptomic reads with the greatest similarity to their reference genomes were photosynthetic taxa belonging to both Bacteria and Eukaryota. The Archaea and Verrucomicrobia had the lowest transcript identities, though Verrucomicrobia identities were uniformly low throughout while the Archaea showed a greater variance.

**Ribosomal protein expression as a growth rate indicator.** Ribosomal proteins (RPs) are an essential translation component in all cells and typically account for 50-60 genes in a bacterial genome. Their well-conserved sequences make them good phylogenetic markers. Over 218,000 reads were annotated as RPs in the metatranscriptomes (Table 4.1) and these fell into 1,903 different taxonomic bins. The proportion of RP genes with at least one hit in a genome bin was correlated with overall transcript abundance (Fig. 4.S1), and the majority of RP genes were represented at least once in bins with >1,000 total transcripts.

We confirmed that the reference genomes were good taxonomic matches for the microbial groups present in the coastal samples by comparing the number of RP hits in a reference genome bin to the total hits to that bin. The distribution of RPs among the 1,563 bins with ≥ 100 reads had a significant positive relationship with bin size (linear regression of log transformed data; $R^2 = 53$, $P < 10^{-16}$; Fig. 4.1). One possible source of the residual variability seen in this relationship is sampling error for reference bins with low transcript coverage. Indeed, many of the bins with atypical percentages of RP hits had less than 1,000 total reads, suggesting this error source was confined to the low abundance bins. Only one high-recruiting genome bin (SAR11 HTCC1002) was an obvious outlier, with over 88,000 total hits but only 300 RP hits (Fig. 4.1). This may be related to cross-hits between HTCC1002 and HTCC1062, two very closely related SAR11 strains. Regardless, most well-covered reference genome bins had a % RP transcript value within one standard deviation of the mean value (5.3% ±3.9), suggesting they represent taxonomically coherent groups.

The relative abundance of ribosomal protein transcripts (% RP) was next used to estimate the activity level of populations represented by a reference bin, since levels of ribosomal protein transcripts are well correlated with growth rate in yeast (Eisen *et al.*,1998) and bacteria (Wei *et al.*, 2001). The % RP hits among the top 200 reference bins ranged from 0.05 (*Candidatus Pelagibacter ubique* HTCC1002) to 20.5% (*Chryseobacterium gleum*) and showed distinct phylogenetic patterns (Fig 4.2A). Gammaproteobacteria were clearly enriched in ribosomal proteins (mean 8.8%) and represented many of the highest % RP genome bins including *Teredinibacter turnerae* T7901 (16.2%), *Saccharophagus degradans* 2-40 (14.3%), *Marinomonas sp. MWYL1* (13.9%), and *Cellvibrio japonicus* Ueda107 (12.9%). Bacteroidetes had a more dispersed distribution, with % RP transcripts ranging from 2.6 to 20.5%, and they

included the two bins with the highest % RPs, *Chryseobacterium gleum* ATCC 35910 (20.5%) and *Capnocytophaga gingivalis* ATCC 33624 (18.8%). The reference bins for Flavobacteria bacterium MS024-2A and MS024-3C, two streamlined Bacteriodetes genomes with relatively high total transcripts, fell in the lower distribution of % RP for this phylum (4.6%). The SAR116 Candidatus *Puniceispirillum marinum* IMCC1322, which recruited the most transcripts, was in the mid % RP range (6.9%). Reference genomes for the Roseobacters were dispersed throughout, ranging from 1.8 (*Roseobacter litoralis* Och 149) to 9.2% (*Citreicella sp*. SE45), though their mean (4.2%) was in the lower range. Finally, despite their dominance in total transcript abundance, members of the SAR11 clade had the lowest % RP of all the groups examined, ranging from 0.05 to 1.7 %RP.

Supporting evidence that these variations in % RP transcripts reflect different *in situ* growth rates was seen in the seasonal shifts in % RP within taxa (Fig 4.2B and C), with most having maximum % RP in the summer (56% of taxa), followed by spring, winter, and fall (20, 18, and 7% of taxa, respectively). Bacterial secondary production rates estimated from $^{3}$H-leucine uptake rates made concurrently with RNA sample collection likewise mirrored the seasonal trends in % RP transcripts (summer: 2.8 X $10^{-6}$ g C $L^{-1}$ $hr^{-1}$; spring: 1.8 X $10^{-6}$ g C $L^{-1}$ $hr^{-1}$; fall: 0.4 X $10^{-6}$ g C $L^{-1}$ $hr^{-1}$; and winter: 0.3 X $10^{-6}$ $10^{-6}$ g C $L^{-1}$ $hr^{-1}$).

**Transcriptome characteristics.** To more fully characterize differences in expression patterns among phylogenentically distinct taxa that shared the same coastal habitat, we focused on 16 reference genome bins that had ample coverage in the metatranscriptomes (four seasonal samples combined) and spanned the range of % RP (Table 4.2).

More than 84% of the genes in the *Candidatus* Puniceispirillum marinum IMCC1322 reference genome had at least one homolog in the metatranscriptome (Fig 4.3). The most highly

represented genes were for a Na+/solute symporter, sugar ABC transporter, TRAP dicarboxylate transporter, and a V-type H(+)-translocating pyrophosphatase (Rinta-Kanto *et al.* 2011), along with energy transduction (Cytochrome c oxidase) and transcription/translation machinery (elongation factors G and Tu, and RNA polymerase). Several regions of the genome had low expression and low functional redundancy (i.e., with no or very few orthologs in the other 15 genomes) and were flanked by phage integrases (Fig 4.3). These areas are indicative of genome islands, regions highly specific to the reference genome and likely missing from the sampled populations. *Nisea* sp. BAL199, a marine Rhodospirallales and the closest relative to IMCC1322, has a much larger genome and many more unique genes (i.e., those with no orthologs in the other 15 reference genomes) (Table 4.2). However, several of the most highly expressed genes were similar to those for IMC1322, including a Na+/solute symporter, ABC-type branched-chain amino acid transporter, and a TRAP dicarboxylate transporter.

The two representatives of the Roseobacter clade, *Roseobacter* sp. AzwK-3b and *Citreicella* sp. SE45, have large genomes and many likely genomic islands (Fig. 4.4). The most highly expressed genes for Azwk-3B were hypothetical proteins or those involved in Aerobic Anoxygenic Photosynthesis (AAnP)-related processes, including light harvesting proteins, antenna complexes, and photosynthetic reaction centers. *Citreicella* sp. SE45, which does not contain AAnP genes, had high transcript recruitment to two subunits of formate dehydrogenase, as well as to genes for transcription and translation (RNA polymerase, chaperonin GroL, translation elongation factor G and Tu), and energy transduction (ATP synthase).

The small and streamlined SAR11 genomes (bins HTCC1002, HTCC1062, HTCC7211, and HIMB114) had high transcript coverage (>90% in HTCC7211, and >60% in the others; Fig. 4.4, Table 4.2). The number of potential genome islands was relatively small, though two

possible islands were identified in HTCC7211 (Fig. 4.4). Given their streamlined genome and dense coverage, it is surprising that these genomes had the lowest transcriptome evenness, with 37 to 64% of all hits to just ten genes. Hits to the Na+/solute symporter alone accounted for 13 to 25% of all transcripts binning to SAR11 genomes. Proteorhodopsin, TRAP and ABC transporters, and ammonia transporter genes were also highly expressed by the SAR11 populations. Betaproteobacterium KB13 is similar to the SAR11s in genome size, but this transcript bin reflected the highly specialized metabolism of a methylotroph. It also had low transcriptome evenness, due largely to methanol dehydrogenase (39% of all transcripts) and other methylotrophy-related processes. Xanthorhodopsin, a glucose/sorbosone dehydrogenase, citrate lyase, and V-type H(+)-translocating pyrophosphatases were also highly expressed in the KB13-like population.

Gammaproteobacteria representatives HTCC2080 (Fig. 4.4), NOR-51, and HTCC2207 have mid-size genomes (Table 4.2), with potential genomic islands seen throughout. The Gammaproteobacteria genome bins had more even transcript distribution, with the top 10 genes making up only 10 to 26% of hits. All three bins were highly enriched in genes for TonB dependant transport and for phototrophy (AAnP for HTCC2080 and NOR-51; proteorhodopsin for HTCC2207). Similar to patterns noticed in other genomes (Roseobacter *Citreicella* sp. SE45, in particular), the transcriptome of NOR51 populations was enriched in genes for transcription and translation (RNA polymerase, ribosomal proteins, translation elongation factor Tu), and cytochromes. This was not the case for the other two Gammaproteobacteria, in line with the % RP data suggesting NOR51 populations are growing faster than those binning to the other two Gammaproteobacteria genomes.

Bacteroidetes representative Flavobacteria MS024-2A has a small genome (Table 4.2; Fig. 4.4) and few detectable genome islands, and is phylogenetically distant from the other 15 reference genomes. The most highly expressed genes were bacteriorhodopsin, TonB-dependent receptors, and a V-type H(+)-translocating pyrophosphatase. Verrucomicrobia member *Pedosphaera parvula* Ellin514 has the largest genome (Table 4.2), with the lowest percentage of orthologs and transcript coverage. The most highly expressed Ellin514 genes were for transcription and translation (RNA polymerase, elongation factors G and Tu, chaperonin GroEL), as well as a pyruvate phosphate dikinase, methionine aminopeptidase, and a type II/III secretion system protein (*pulD*). For the Archaea *Nitrosopumilus maritimus* SCM1, the top 10 genes (45% of the transcriptome) were for ammonia uptake and oxidation, including two ammonia transporters and all three subunits of ammonia monooxygenease. For the cyanobacterium *Synechococcus sp.* WH8109 bin, genes for carbon fixation (ribulose bisphosphate carboxylase; RuBisCO), photosynthesis core proteins, and transcription (RNA polymerase) dominated.

**Relationship between expression level and ortholog number.** We found a statistically significant positive relationship between gene expression level (number of reads recruited to a gene in a reference bin) and functional redundancy (number of genomes with orthologs to that gene) (Wilcoxon rank-sum test, $P < 0.05$; Fig 4.5, 4.S2). Thus most highly expressed genes were shared by multiple taxa, although SAR11s HTCC1002, HTCC1062, and HIMB114, as well as MS024-2A and *N. maritimus* SCM1 did not have as strong a pattern as the other genomes (Fig. 4.S2). Previous observations from marine metatranscriptomic data (Hewson 2009, Stewart 2011) support this conclusion, and lead to the conclusion that an analysis restricted to only the more highly expressed genes is likely to miss unique functional capabilities that distinguish taxa. In

order to address niche-defining features, we instead focused on those genes whose expression was higher than expected based on their ortholog count group; these expression outliers had a transcript abundance 1.5 times greater than the interquartile range of the ortholog count group to which they belonged (Fig. 4.5). Thus this approach identified informative genes by considering both expression level and functional redundancy.

**Expression patterns distinguish genome bins.** The ~46,000 genes represented in the 16 reference genomes were reduced to 22,000 ortholog groups based on a modified reciprocal best-hit approach (see methods). Ortholog groups that contained genes identified as expression outliers were used in a Non-metric Multidimensional Scaling (NMDS) analysis, with a gene's transcript count as a proportion of the total reference bin counts used as the abundance metric and excluding the *N. maritimus* SCM1 bin to get better separation among the 15 bacterial bins (Fig. 4.6A). There was clustering of taxonomically-related strains, suggesting a phylogenetic signal to the expression outlier genes. Genome bins representing the SAR11 group formed a cluster that was distinct from a Gammaproteobacteria bin cluster and the rest of the Alphaproteobacteria bins. Genome bins with no close relatives tended to have a distinct location on the NMDS plot, although Verrucomicrobium Ellin514 and Cyanobacterium *Synechococcus* sp. WH8109 were closely oriented to one another. An NMDS analysis based on all ortholog groups using just presence/absence of a gene rather than expression level (Fig. 4.S3) showed different groupings of the 15 bacterial genomes, indicating that information on gene expression in response to shared environmental conditions, not just presence/absence of the genes, provides an additional ecological dimension for analysis of bacterioplankton niches. To visualize how individual outlier ortholog groups influenced genome placement on the NMDS plot, the weighted scores for each ortholog group was calculated (see Methods) (Fig. 4.6B). Outlier

groups with few orthologs dominated the periphery while those with many orthologs clustered in the center.

**Indicator genes.** We adopted the Indicator Species Analysis approach of Dufrene and Legendre (1997) to identify indicator genes, defined here as those whose expression best distinguished the activities of a bacterioplankton taxon. An indicator value (IV) for each gene in the outlier ortholog groups was calculated based on the contribution of a reference bin to the total expression (see Methods).

The indicator genes for SAR116 populations suggested a motile taxon that was using a variety of substrates (Fig. 4.6C). Expression of a gene for methanesulfonate (MS) oxidation (methanesulfonate monooxygenase subunit; SAR116_2109), a potentially abundant compound in the marine environment generated from the oxidation of DMSO (Kelly, 1999), distinguished this genome bin. A second indicator gene in the SAR116 bin was a nitrate/sulfonate/bicarbonate permease (SAR116_2101), which neighbors the SAR116 MSO genes and is homologous to the MS transporter found in *Methylosulfonomonas methylovora* (Jamshed *et al.*, 2006). Indeed, the entire SAR116 genome region (SAR116_2098-2109) shares high homology and synteny with the *M. methylovora* MS operon. Indicator gene ketopantoate hydroxymethyltransferase (SAR116_2112) was located adjacent to a formate tetrahydrofolate ligase in the neighborhood of the MSO genes, and may process the methyl groups derived from MS.

Populations of both SAR116 and its relative *Nisaea* sp. BAL199 expressed genes for the degradation of aromatic compounds. Three SAR116 indicator genes fell in an operon encoding enzymes of the protocatechuate pathway (SAR116_0936, 0937, 0940), an intermediate important in the degradation of lignin derivatives from coastal marshes (Buchan *et al.*, 2000) and aromatic compounds synthesized by phytoplankton (Vernet and Whitehead, 1996). BAL199 had indicator

genes for chlorobenzene degradation (carboxymethylenebutenolidase, BAL199_05144) and a TRAP transporter for chloroaromatic compounds, but their annotations carried lower confidence. Other indicator genes in the BAL199 population bin included 3 subunits of an Fe(III) transporter and genes for taurine uptake and metabolism.

Indicator genes for C1 carbon metabolism were found for Bal199 as well as the two Roseobacter genomes, *Citreicella* sp. SE45 and *Roseobacter* sp. AzwK-3b, dominated by strong formate dehydrogenase expression (8, 11, and 3% of the Bal199, SE45, and Azwk-3B transcriptomes, respectively). The Bal199 bin also had four indicator genes for glyoxylate degradation and cycling through formate (BAL199_21019,21024,26437,27586), suggesting this may be the source of the C1 compounds being processed by Bal199-like populations. Expression for all three of these genome bins was highly enriched in TRAP dicarboxylate transporters, consistent with the transport of glyoxylate or other small organic acids. The two roseobacters also shared indicator genes for ureases, branched chain amino acid transporters, taurine metabolism genes, and sulfur oxidation through *sox* pathway genes.

Indicator transcripts in the SAR11 coastal isolate HTCC1062 genome bin pointed to the importance of sugar metabolism. Five genes from two putative sugar transporters were among the indicator genes; transporter SAR11_0769-0772 was hypothesized previously to target glucose (Schwalbach *et al.*, 2010). For SAR11_0269-0271, a neighboring indicator gene in the metatranscriptome with homology to sorbitol dehydrogenase (SAR11_0272), the close proximity of a highly expressed TRAP mannitol transporter genes (although they did not meet the indicator gene cut-off), and an indicator gene for fructose-bisphosphate aldolase (SAR11_0584) together suggest the uptake and metabolism of C6 sugars. This appears to be a distinguishing ecological function for HTCC1062-like populations.

The SAR11 open ocean isolate HTC7211 bin was characterized by the uptake of compatible solutes. HTCC7211 had two indicator genes for uptake of ectoine or hydroxyectoine (PB7211_776,1327), which serve as compatible solutes during osmotic stress (Mulligan *et al.*, 2011) and as bacterial carbon and nitrogen sources (Lecher *et al.*, 2009). The HTCC7211 bin also contained indicator genes for glycine betaine metabolism, including two ABC transporter genes (PB7211_147,194), a putative sarcosine oxidase (PB7211_683), and a glycine cleavage system T protein (PB7211_1405). Tripp *et al.* (2008, 2009) reported that SAR11 growth was significantly improved by addition of glycine betaine to the medium, stressing the importance of glycine to SAR11 metabolism.

SAR11 clade members HTCC1002 and HIMB114 had indicator genes for taurine transport (PU1002_02371) and metabolism (HIMB114_0332), respectively. HIMB114's indicator genes also included four separate genes for putative tricarboxylic transport membrane proteins from a recently described system (HIMB114_0326,0339,0341, and1737) (Antoine 2005). It is possible that the tricarboxylic acid tartrate is the substrate of one of these transporters, based on the presence of another HIMB114 indicator gene for tartrate dehydrogenase (HIMB114_0953). While little is currently known about tartrate sources in the marine environment, it can be secreted by marine algae (Marsh *et al.*, 1992). Finally, a high-affinity Fe permease was found among the HIMB114 indicator genes (HIMB114_0552).

The identification of two subunits of adenylyl-sulfate reductase (APS reductase) among the HTCC7211 indicator genes (7211_563, 1116) raised the possibility that sulfur oxidation is a shared characteristic of the SAR11 bins. While these genes are typically associated with dissimilatory sulfate reduction in anaerobic bacteria, they have been proposed to operate in the opposite direction to oxidize reduced sulfur (Meyer 2008). The APS reductase genes in

HTCC7211 have close homologs in HTCC1002 and HTCC1062 (which were also well expressed), as well as in known sulfur- and iron-oxidixing bacteria (Meyer 2008). SAR11 members are unable to reduce inorganic sulfur from the surrounding environment (Tripp et al., 2008), suggesting the source of reduced sulfur for APS reductase is likely originating from intracellular metabolic pools such as methionine and dimethylsulfoniopropionate (DMSP) degradation products.

The three Gammaproteobacteria indicator genes were highly enriched in TonB dependant transporters, which have traditionally been associated with iron and vitamin uptake, though recent studies indicate they likely have a much wider substrate range, including carbohydrates (Schauer et al., 2008). Transcripts for fatty acid metabolism were also enriched in all three Gammaproteobacteria bins. Both HTCC2080 and NOR51-B were expressing lipase 4, which releases fatty-acids from triglycerides. HTCC2080 had nine additional indicator genes for fatty acid metabolism, including four acyl-CoA dehydrogenases and a 3-ketoacyl-CoA thiolase involved in fatty acid β-oxidation. These findings are in line with McCarren *et al.* (2010), who found that Gammaproteobacteria related to *Idiomarina* and *Alteromonas* spp. responded to marine dissolved organic matter with a proportional increase in fatty acid metabolism gene transcription.

Gammaproteobacterium HTCC2207 indicator genes revealed a motile population binding to and degrading complex carbohydrates (Fig. 4.6C). There were 24 indicator genes for flagellar assembly and chemotaxis. There were four indicator genes (GB2207_00005, 00010, 06108, 06128) with cadherin domains that are potentially involved in the metabolism of complex carbohydrates by increasing cell aggregation and allowing direct binding to cellulose, xylan, and related compounds (Fraiberg *et al.* 2010, 2011). There were six indicator genes for breaking

glycosydic bonds, including four annotated as general glycosyl hydrolases, possibly targeting the β1-4 linkages found in cellulose, and two genes annotated as β1-3 glucanase (GB2207_09841) and laminarninase (GB2207_10126), possibly targeting the β1-3 linkages of laminarin (a storage glucan found in brown algae) or chrysolaminarin (a storage glucan of diatoms). *Ostreococcus*, a small photosynthetic picoeukaryote active in these coastal waters (Table 4.1), synthesizes B1-3 glucans such as a callose (Monnier *et al.*, 2010) and may be another source of complex carbohydrates.

Indicator genes for populations binning to Flavobacteria MS024-2A were similar to HTCC2207, included genes for attachment (cadherins) and the breakdown (2 glycosyl hydrolases) and synthesis (a glycogen synthase; Flav2ADRAFT_0634) of complex carbohydrates. MS024-2A was also motile, with indicator genes for gliding motility proteins *gldJMO*, likely involved in translocation across a solid surface. Three subunits of Na+-transporting NADH:ubiquinone oxidoreductase were also indicator genes (Flav2ADRAFT_1288, 1290, 1291), depicting cells invested in maintaining a sodium membrane potential that may be coupled with Na+/solute symporters.

Betaproteobacterium KB13 and Thaumarchaeote *Nitrosopumilus maritimus* SCM1 indicator genes reflected high degrees of specialization (Fig. 4.6C). For KB13, indicator genes for methanol dehydrogenase indicated ongoing methylotrophy. This genome bin also contained bacterioferritan as an indicator gene (KB13_1091), suggesting the ability to store Fe. For *N. maritimus*, dominant indicator genes included the ammonia monoxygenase genes, and two *nirK*-like genes (though the exact function of the later is uncertain, see Chapter 3). Four blue (type 1) copper domain proteins were identified as indicator genes; while the function of these proteins is

not yet known, the use of copper in metalloenzymes may decrease competition between *N. maritimus* and bacteria for iron (Urakawa *et al.*, 2011).

The Verrucomicrobium *Pedosphaera parvula* Ellin514 indicator genes had a strong signal for biofilm formation (Fig. 4.6C), including the Type IV pili and the Type II secretion systems, twitching motility genes, polysaccharide synthesis, and capsular exopolysaccharide synthesis. There was also a set of indicator genes for sugar metabolism, particularly those related to xylose, a potential exopolysaccharide component (Gilbert *et al.*, 2007). Three genes for ABC-Type II transporters (Cflav_PD3611, 3988, 3989), which transport polysaccharides to the outside of the cell, were also indicator genes. While there have been few phenotypic studies of Verrucomicrobia, particularly in aquatic environments, characterization of *Lentisphaera araneosa* from the sister phylum *Lentisphaerae* showed it too was an abundant producer of exopolysaccharides (Cho *et al.*, 2004, Thrash *et al.*, 2010). Together, the *P. parvula* indicator genes are suggestive of biofilm formation, possibly for a pathogenic or symbiotic lifestyle. This is supported by indicator genes for degrading cell wall polymers and homologs to myrosinases, a group of genes that cleave glucose from glucosinolates, which are secondary metabolites of plants.

The only photosynthetic organism in the 15 genomes was *Synechococcus sp.* WH 8109, and as expected, the indicator genes had a strong signal for photosynthesis-related processes (Fig. 4.6C). An indicator gene for peroxiredoxin may provide protection against the abundant reactive oxygen species produced during light harvesting (SH8109_1863). Three separate indicator genes were annotated as the cell division protein *ftsH* (SH8109_0218, 1424, 1960), which acts to maintain membrane protein quality (Ito and Akiyama 2005) and may have a role in maintaining membranes, including cell membranes during division and thylakoid membranes.

Several *Synechococcus* indicator genes are involved in sulfur assimilation, including a sulfate permease (SH8109_1514) and sulfite reductase (SH8109_1751). Furthermore, *Synechococcus* populations expressed the indicator gene UDP-sulfoquinovose synthase (*sqdB;* SH8109_0458), the diagnostic gene for sulfolipid synthesis. Open ocean Cyanobacteria have sulfolipid-enriched membranes hypothesized to decrease their need for phospholipids in a low phosphorus environment (<0.010 μM) (Van Mooy *et al.*, 2006, 2009). However phosphorus concentrations were high in the coastal site sampled here (>1.0 μM at all four sample dates), suggesting that sulfolipid synthesis may be a universal strategy for niche diversification in *Synechococcus.*

**Conclusions**

Hutchinson's paradox of the plankton (Hutchinson, 1961) presupposes three axioms: 1) the presence of many functionally overlapping taxa within a community, 2) an environment with limited niches, and 3) competitive exclusion as the mechanism that constrains the taxa:niche ratio to one. While Hutchinson was originally interested in how this paradox applied to the coexistence of tens of different phytoplankton species, contemporary 16S rRNA gene studies suggest bacterioplankton assemblages contain orders of magnitude more coexisting taxa. Indeed, the mRNA sequences examined here indicates the coexistence of literally thousands of active taxa with hundreds of overlapping functions in this coastal microbial community

One resolution to the apparent paradox is that more niches exist than we have the technical ability to observe (Hutchinson, 1961). The availability of 11 million microbial transcript sequences in this study allowed the evaluation of this possibility, since they provided information on co-occurring microbial processes at an unprecedented level of resolution. The majority of transcriptional effort by the microbial community was devoted to core metabolic processes that are shared among many taxa (e.g., elongation factors, ribosomal proteins,

ATPases), consistent with the positive correlation between expression level and functional redundancy of a gene (Fig. 4.5). This has an important effect of concentrating comparative analyses on the most highly conserved bacterial genes (Hewson et al.,2009), making it difficult to identify unique taxon-specific functions. Thus we focused instead on identifying taxonomic indicator genes based on expression outliers: genes for which expression level was atypically high based on their conservation status across genomes (i.e., ortholog number). The resulting indicator genes included a number that mediated utilization of substrates not previously considered important to marine bacteria, such as tartrate metabolism by SAR11 HIMB114, taurine metabolism by several Alphaproteobacteria, methanesulfonate by SAR116, and ectoine transport and metabolism by SAR11 HTCC7211. Gammaproteobacteria HTCC2207 and Bacteroidetes MS024-2A were distinguished by utilization of complex carbohydrates, a signal largely missing from the Alphaproteobacteria. Another major category of indicator genes involved mechanisms for energy acquisition, including sulfur oxidation for both Roseobacter bins using the *sox* system, sulfur oxidation for SAR11 HTCC7211 using the APS reductase system, and bacteriochlorophyll-based proton pumping for two Gammaproteobacteria and one Roseobacter genome. Finally, a number of indicator genes were related to physical interactions with cells or non-living surfaces in the environment, including gliding motility in Bacteroidetes MS024-2A, twitching motility in Verrucomicrobium *Pedosphaera parvula* Ellin514, a strong chemotaxis signal in Gammaproteobacterium HTCC2207, and surface adhesion genes in HTCC2207 and MS024-2A. The metatranscriptomic data therefore suggests that the diverse assemblage of coastal bacterioplankton is maintained, at least in part, by the presence of previously unrecognized niche dimensions.

Another resolution to the apparent paradox would be if functional activities of bacterioplankton taxa can indeed stably overlap (Hutchinson 1961), characterized by situations in which competitive exclusion is not the major ecological force driving community composition. For example, competitive exclusion may be less important if the time scale of substrate availability is short compared to the growth rate of the competing cells; or if top-down controls by viral or protist predators increase mortality rates among the best competitors (the "kill the winner hypothesis"; Suttle, 2007). We suspected that release from competitive exclusion would most likely occur in the case of fast-growing 'opportunitrophs' if they grow primarily at the expense of short-term fluxes of labile substrates (Moran et al., 2004). Using the % RP proxy for *in situ* growth rate to identify oligotrophic (slow-growing) versus opportunitrophic (fast-growing) taxa, we observed a higher degree of transcriptome specialization in slow-growing taxa, consistent with a narrower niche. In contrast, rapidly growing taxa exhibited higher transcriptome evenness and more diverse substrate uptake and metabolism, consistent with a generalist strategy in which many taxa are able to exist on the same resources given that the time frame for complete competitive exclusion is greater than that for substrate availability.

While genome sequences of cultured bacteria and metagenomes of bacterioplankton communities can address niche differentiation based on gene content (Rocap *et al.*, 2003), metatranscriptomics adds the additional ecological dimension of gene expression patterns under shared environmental conditions. For example, a gene upregulated in one genome under existing environmental conditions but not in another may represent a key niche dimension that would not be evident from gene inventories alone. Our examination of *in situ* gene expression patterns is based on four combined seasonal samples from southeastern coastal waters, providing a robust, time-averaged view of microbial activities that is not biased by any particular environmental

condition. Future work will extend the insights gained here to examine temporal variability in gene expression and seasonal dynamics of niche defining biogeochemical activities.

## References


Antoine R, Huvent I, Chemlal K, Deray I, Raze D, Locht C *et al.* (2005). The periplasmic binding protein of a tripartite tricarboxylate transporter is involved in signal transduction. *J Mol Bio* **351:** 799-809.

Buchan A, Collier LS, Neidle EL, Moran MA (2000). Key aromatic-ring-cleaving enzyme, protocatechuate 3,4-dioxygenase, in the ecologically important marine Roseobacter lineage. *App Environ Microb* **66:** 4662-4672.

Chang WH, Tolbert NE (1970). Excretion of glycolate, mesotartrate and isocitrate lactone by synchronized cultures of ankistrodesmus-braunii. *Plant Physiol* **46:** 377-&.

Cho JC, Vergin KL, Morris RM, Giovannoni SJ (2004). *Lentisphaera araneosa* gen. nov., sp nov, a transparent exopolymer producing marine bacterium, and the description of a novel bacterial phylum, Lentisphaerae. *Environ Microbiol* **6:** 611-621.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU *et al.* (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311:** 496-503.

Dufrene M, Legendre P (1997). Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecol Monogr* 67**:** 345-366.

Eisen MB, Spellman PT, Brown PO, and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *P Natl Acad Sci USA*. Vol. 95:14863–14868

Fraiberg M, Borovok I, Weiner RM, Lamed R (2010). Discovery and Characterization of Cadherin Domains in *Saccharophagus degradans* 2-40. *J Bacteriol* **192:** 1066-1074.

Fraiberg M, Borovok I, Bayer EA, Weiner RM, Lamed R (2011). Cadherin Domains in the Polysaccharide-Degrading Marine Bacterium *Saccharophagus degradans* 2-40 Are Carbohydrate-Binding Modules. *J Bacteriol* **193:** 283-285.

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

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *P Natl Acad Sci USA* **103:** 13104-13109.

Gilbert M, Mandrell RE, Parker CT, Li JJ, Vinogradov E (2007). Structural analysis of the capsular polysaccharide from *Campylobacter jejuni* RM1221. *Chembiochem* **8:** 625-631.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:** 1242-1245.
Hardin G (1960). Competitive exclusion principle. *Science* **131:** 1292-1297.

Hewson I, Poretsky RS, Dyhrman ST, Zielinski B, White AE, Tripp HJ *et al.* (2009). Microbial community gene expression within colonies of the diazotroph, Trichodesmium, from the Southwest Pacific Ocean. *ISME-J* **3:** 1286-1300.

Hutchinson GE (1961). The paradox of the plankton. *Am Nat* **95:** 137-145.

Ito K, Akiyama Y (2005). Cellular functions, mechanism of action, and regulation of FtsH protease. *Annu Rev Microbiol* **59:** 211-231.

Jamshad M, De Marco P, Pacheco CC, Hanczar T, Murrell JC (2006). Identification, mutagenesis, and transcriptional analysis of the methanesulfonate transport operon of Methylosulfonomonas methylovora. *Appl Environ Microb* **72:** 276-283.

Kelly DP, Murrell JC (1999). Microbial metabolism of methanesulfonic acid. *Arch Microbiol* **172:** 341-348.

Lecher J, Pittelkow M, Zobel S, Bursy J, Bonig T, Smits SHJ *et al.* (2009). The Crystal Structure of UehA in Complex with Ectoine-A Comparison with Other TRAP-T Binding Proteins. *J Mol Bio* **389:** 58-73.

Legendre P, Oksanen J, ter Braak CJF (1997). Testing the significance of canonical axes in redundancy analysis. *Method Ecol Evol* **2:** 269-277.

Malik R, Viola RE (2010). Structural characterization of tartrate dehydrogenase: a versatile enzyme catalyzing multiple reactions. *Acta Crystallogr D* **66:** 673-684.

Marsh ME, Chang DK, King GC (1992). Isolation and characterization of a novel acidic polysaccharide containing tartrate and glyoxylate residues from the mineralized scales of a unicellular coccolithophorid alga *pleurochrysis-carterae*. *J Biol Chem* **267:** 20507-20512.

McCarren J, Becker JW, Repeta DJ, Shi YM, Young CR, Malmstrom RR *et al.* (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *P Natl Acad Sci* **107:** 16420-16427.

Meyer B, Kuever J (2008). Homology Modeling of Dissimilatory APS Reductases (AprBA) of Sulfur-Oxidizing and Sulfate-Reducing Prokaryotes. *Plos One* **3** (1): e1514. doi:10.1371/journal.pone.0001514

Monnier A, Liverani S, Bouvet R, Jesson B, Smith JQ, Mosser J *et al.* (2010). Orchestrated transcription of biological processes in the marine picoeukaryote Ostreococcus exposed to light/dark cycles. *BMC Genomics* **11**.

Moran MA, Buchan A, Gonzalez JM, Heidelberg JF, Whitman WB, Kiene RP et al. (2004). Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* 432: 910-913.

Moran MA (2008). Genomics and Metagenomics of Marine Prokaryotes. In: Kirchman (ed). *Microbial Ecology of the Oceans*, 2nd edn.

Mou XZ, Sun SL, Edwards RA, Hodson RE, Moran MA (2008). Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451:** 708-U4.

Mulligan C, Fischer M, Thomas GH (2011). Tripartite ATP-independent periplasmic (TRAP) transporters in bacteria and archaea. *FEMS Microbiol Rev* **35:** 68-86.

Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE *et al.* (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME-J* **4:** 784-798.

Oksanen J, Blanchet FG., Kindt R, Legendre P,O'Hara RB, Simpson GL, *et al.* (2011). vegan: Community Ecology Package. R package version 1.17-10. http://CRAN.R-project.org/package=vegan

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M *et al.* (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microb* **71:** 4121-4126.

Poretsky RS, Gifford SM, Rinta-Kanto J, Vila-Costa M, Moran MA (2009). Analyzing gene expression from marine microbial communities using environmental transcriptomics. http://www.jove.com/details.php?id=1086 doi: 10.3791/1086. *J Vis Exp*. 24

Rinta-Kanto JM, Burgmann H, Gifford SM, Sun SL, Sharma S, del Valle DA *et al.* (2010). Analysis of sulfur-related transcription by Roseobacter communities using a taxon-specific functional gene microarray. *Environ Microb* **13:** 453-467.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA *et al.* (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424:** 1042-1047.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al* (2007). The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *Plos Biology* **5:** 398-431.

Stewart FJ, Sharma AK, Bryant JA, Eppley JM, DeLong EF (2011). Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol* **12**.

Schauer K, Rodionov DA, de Reuse H (2008). New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci* **33** (7): 330-338

Schwalbach MS, Tripp HJ, Steindler L, Smith DP, Giovannoni SJ (2010). The presence of the glycolysis operon in SAR11 genomes is positively correlated with ocean productivity. *Environ Microb* **12:** 490-500.

Suttle CA (2007). Marine viruses - major players in the global ecosystem. *Nat Rev Microbiol* **5:** 801-812.

Thrash JC, Cho JC, Vergin KL, Morris RM, Giovannoni SJ (2010). Genome Sequence of *Lentisphaera araneosa* HTCC2155(T), the Type Species of the Order Lentisphaerales in the Phylum Lentisphaerae. *J Bacteriol* **192:** 2938-2939.

Tripp HJ, Kitner JB, Schwalbach MS, Dacey JWH, Wilhelm LJ, Giovannoni SJ (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* **452:** 741-744.

Tripp HJ, Schwalbach MS, Meyer MM, Kitner JB, Breaker RR, Giovannoni SJ (2009). Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. *Environ Microb* **11:** 230-238.

Urakawa H, Martens-Habbena W, Stahl DA (2011). Physiology and genomics of ammonia-oxidizing archaea. In: Ward BB, Arp DJ and Klotz MG (eds). *Nitrification* ASM Press: Washington, D.C. pp 117-155.

Van Mooy BAS, Rocap G, Fredricks HF, Evans CT, Devol AH (2006). Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *P Natl Acad Sci USA* **103:** 8607-8612.

Van Mooy BAS, Fredricks HF, Pedler BE, Dyhrman ST, Karl DM, Koblizek M *et al* (2009). Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. *Nature* **458:** 69-72.

Vernet M, Whitehead K (1996) Release of ultraviolet-absorbing compounds by the red-tide dinoflagellate *Lingulodinium polyedra*. *Mar Biol* **127**(1): 35-44

Wei Y, Lee JM, Richmond C, Blattner FR, Rafalski JA, LaRossa RA (2001). High-Density Microarray-Mediated Gene Expression Profiling of *Escherichia coli*. *J Bacteriol* 183(2):545-556.

Table 4.1. Taxonomic binning of coastal metatranscriptomic reads based on the highest scoring pair from the BLASTx search against RefSeq. Results are for all four seasonal datasets combined. Taxa bins= number of different taxa the hits binned to for a given group. Rank= Rank abundance of a bin based on the total number of reads recruited. AA%ID = Mean percent amino acid identity of reads to genes in the reference genome. Ribosomal proteins = Total number of reads annotated as ribosomal proteins in a given bin.

| | taxa bins | hits | rank | mean AA%ID | ribosomal proteins |
|---|---|---|---|---|---|
| **Total** | **3,902** | **4,151,833** | | | **218,198** |
| **Alphaproteobacteria** | **249** | **1,745,196** | | | **70,882** |
| Roseobacter | 38 | 536,207 | | | 24,429 |
| *Roseobacter sp. AzwK-3b* | | 37,682 | 13 | 77 | 870 |
| *Rhodobacterales bacterium HTCC2083* | | 30,145 | 19 | 78 | 1,046 |
| *Silicibacter lacuscaerulensis ITI-1157* | | 28,146 | 23 | 80 | 2,523 |
| *Roseobacter litoralis Och 149* | | 26,533 | 24 | 78 | 472 |
| *Citreicella sp. SE45* | | 26,233 | 25 | 78 | 2,424 |
| SAR11 | 3 | 383,928 | | | 8,500 |
| *Candidatus Pelagibacter sp. HTCC7211* | | 253,217 | 2 | 84 | 7,414 |
| *Candidatus Pelagibacter ubique HTCC1002* | | 88,822 | 3 | 80 | 300 |
| *Candidatus Pelagibacter ubique HTCC1062* | | 41,889 | 11 | 80 | 786 |
| Misc.Alphas | 208 | 825,061 | | | 37,953 |
| *Candidatus Puniceispirillum marinum IMCC1322* | | 259,512 | 1 | 76 | 18,059 |
| *alpha proteobacterium BAL199* | | 57,261 | 7 | 71 | 1,680 |
| *alpha proteobacterium HIMB114* | | 50,735 | 9 | 80 | 698 |
| *Labrenzia alexandrii DFL-11* | | 15,151 | 41 | 73 | 160 |
| *Hoeflea phototrophica DFL-43* | | 13,699 | 46 | 75 | 233 |
| **Gammaproteobacteria** | **592** | **855,576** | | | **68,776** |
| *marine gamma proteobacterium HTCC2080* | | 80,293 | 4 | 74 | 4,716 |
| *gamma proteobacterium NOR51-B* | | 51,404 | 8 | 74 | 6,161 |
| *marine gamma proteobacterium HTCC2207* | | 41,755 | 12 | 77 | 2,065 |
| *marine gamma proteobacterium HTCC2143* | | 35,745 | 15 | 71 | 1,376 |
| *marine gamma proteobacterium HTCC2148* | | 33,634 | 17 | 72 | 2,460 |
| **Betaproteobacteria** | **173** | **247,164** | | | **11,282** |
| *beta proteobacterium KB13* | | 36,854 | 14 | 85 | 1,395 |
| *Methylophilales bacterium HTCC2181* | | 29,347 | 20 | 75 | 987 |
| *Methylotenera sp. 301* | | 6,031 | 135 | 74 | 325 |
| *Methylovorus sp. SIP3-4* | | 5,467 | 147 | 73 | 407 |
| *Methylibium petroleiphilum PM1* | | 5,039 | 155 | 72 | 71 |
| **Bacteriodetes*** | **131** | **375,382** | | | **24,681** |
| *Flavobacteria bacterium MS024-2A* | | 43,539 | 10 | 76 | 2,004 |
| *Flavobacteria bacterium MS024-3C* | | 23,785 | 27 | 84 | 1,113 |
| *Zunongwangia profunda SM-A87* | | 12,861 | 53 | 73 | 1,419 |
| *Robiginitalea biformata HTCC2501* | | 12,220 | 54 | 73 | 895 |
| *Kordia algicida OT-1* | | 11,293 | 62 | 73 | 1,098 |
| **Verrucomicrobia** | **9** | **110,647** | | | **8,638** |
| *Coraliomargarita akajimensis DSM 45221* | | 29,145 | 21 | 69 | 3,718 |
| *Pedosphaera parvula Ellin514* | | 25,411 | 26 | 68 | 1,765 |
| *Verrucomicrobiae bacterium DG1235* | | 18,498 | 33 | 69 | 1,002 |
| *Opitutus terrae PB90-1* | | 9,551 | 85 | 68 | 615 |
| *Chthoniobacter flavus Ellin428* | | 8,589 | 96 | 68 | 346 |
| **Cyanobacteria** | **60** | **71,618** | | | **2,643** |
| *Synechococcus sp. WH 8109* | | 11,218 | 64 | 93 | 563 |
| *Synechococcus sp. CC9605* | | 4,139 | 183 | 90 | 171 |
| *Cyanobium sp. PCC 7001* | | 4,067 | 186 | 82 | 245 |
| *Synechococcus sp. RS9916* | | 3,813 | 202 | 86 | 172 |
| *Synechococcus sp. RCC307* | | 3,795 | 206 | 89 | 181 |
| **Archaea** | **103** | **91,028** | | | **4,028** |
| *Nitrosopumilus maritimus SCM1* | | 67,890 | 5 | 86 | 2,162 |
| *Cenarchaeum symbiosum A* | | 3,485 | 224 | 77 | 27 |
| *Sulfolobus tokodaii str. 7* | | 2,956 | 246 | 65 | 8 |
| *Aciduliprofundum boonei T469* | | 964 | 566 | 65 | 253 |
| *Pyrococcus furiosus DSM 3638* | | 687 | 695 | 67 | 10 |
| **Eukaryota** | **150** | **147,364** | | | **4,402** |
| *Micromonas sp. RCC299* | | 33,680 | 16 | 83 | 1,021 |
| *Ostreococcus lucimarinus CCE9901* | | 13,569 | 47 | 77 | 704 |
| *Floydiella terrestris* | | 11,055 | 67 | 86 | 6 |

**Table 4.2.** Summary statistics for transcripts binning to select 16 genomes.

| taxonomic affiliation[1] | genome[2] | bin size[3] | ortholog (%)[4] | genes hit (%)[5] | mean hits per gene[6] | RP%[7] | top %[8] |
|---|---|---|---|---|---|---|---|
| Alpha, SAR116 | Candidatus Puniceispirillum marinum IMCC1322 | 2,543 | 34 | (84) | 121 | 7.0 | 17 |
| Alpha | alpha proteobacterium BAL199 | 6,128 | 14 | (41) | 23 | 2.9 | 34 |
| Alpha, Roseobacter | Roseobacter AzwK-3b | 4,145 | 21 | (45) | 20 | 2.3 | 42 |
| Alpha, Roseobacter | Citreicella sp. SE45 | 5,427 | 15 | (34) | 14 | 9.2 | 24 |
| Alpha, Rickettsiales, SAR11 | Candidatus Pelagibacter sp. HTCC7211 | 1,447 | 53 | (91) | 193 | 2.9 | 38 |
| Alpha, Rickettsiales, SAR11 | Candidatus Pelagibacter ubique HTCC1002 | 1,393 | 56 | (61) | 106 | 0.3 | 64 |
| Alpha, Rickettsiales, SAR11 | Candidatus Pelagibacter ubique HTCC1062 | 1,354 | 57 | (66) | 47 | 1.9 | 43 |
| Alpha, Rickettsiales, SAR11 | alpha proteobacterium HIMB114 | 1,425 | 54 | (75) | 48 | 1.4 | 55 |
| Gamma | marine gamma proteobacterium HTCC2080 | 3,185 | 25 | (84) | 30 | 5.9 | 10 |
| Gamma | gamma proteobacterium NOR51-B | 2,930 | 28 | (67) | 26 | 12.0 | 12 |
| Gamma | marine gamma proteobacterium HTCC2207 | 2,388 | 33 | (81) | 22 | 4.9 | 26 |
| Beta | Betaproteobacterium KB13 | 1,318 | 53 | (84) | 33 | 3.8 | 58 |
| Bacteriodetes | Flavobacteria MS024-2A | 1,772 | 33 | (87) | 28 | 4.6 | 24 |
| Verrucomicrobia | *Pedosphaera parvula* Ellin514 | 6,510 | 10 | (28) | 14 | 6.9 | 16 |
| Cyanobacteria | Synechococcus WH 8109 | 2,577 | 24 | (61) | 7 | 5.0 | 10 |
| Crenarchaeota | *Nitrosopumilus maritimus* SCM1 | 1,797 | 19 | (73) | 52 | 3.1 | 45 |

[1]Taxonomic affiliation: Phylogenetic lineage of reference genome. Alpha = Alphaproteobacteria; Gamma = Gammaproteobacteria;

[2]Bin size: Total number of genes in reference genome

[3]Ortholog Conc.: Percentage of genes in the reference genome with 8 or more orthologs in the other 15 genomes.

[4]Genes hit: The number of genes hit in the reference genome. The percentage of these genes in the genome is in brackets.

[5]Mean hits: Average number of hits per gene.

[6]RP%: The proportion of reads binning to a reference genome that were annotated as ribosomal proteins.

[7]top%: The percentage the top 10 transcript recruiting genes made up of all hits to the reference genome.

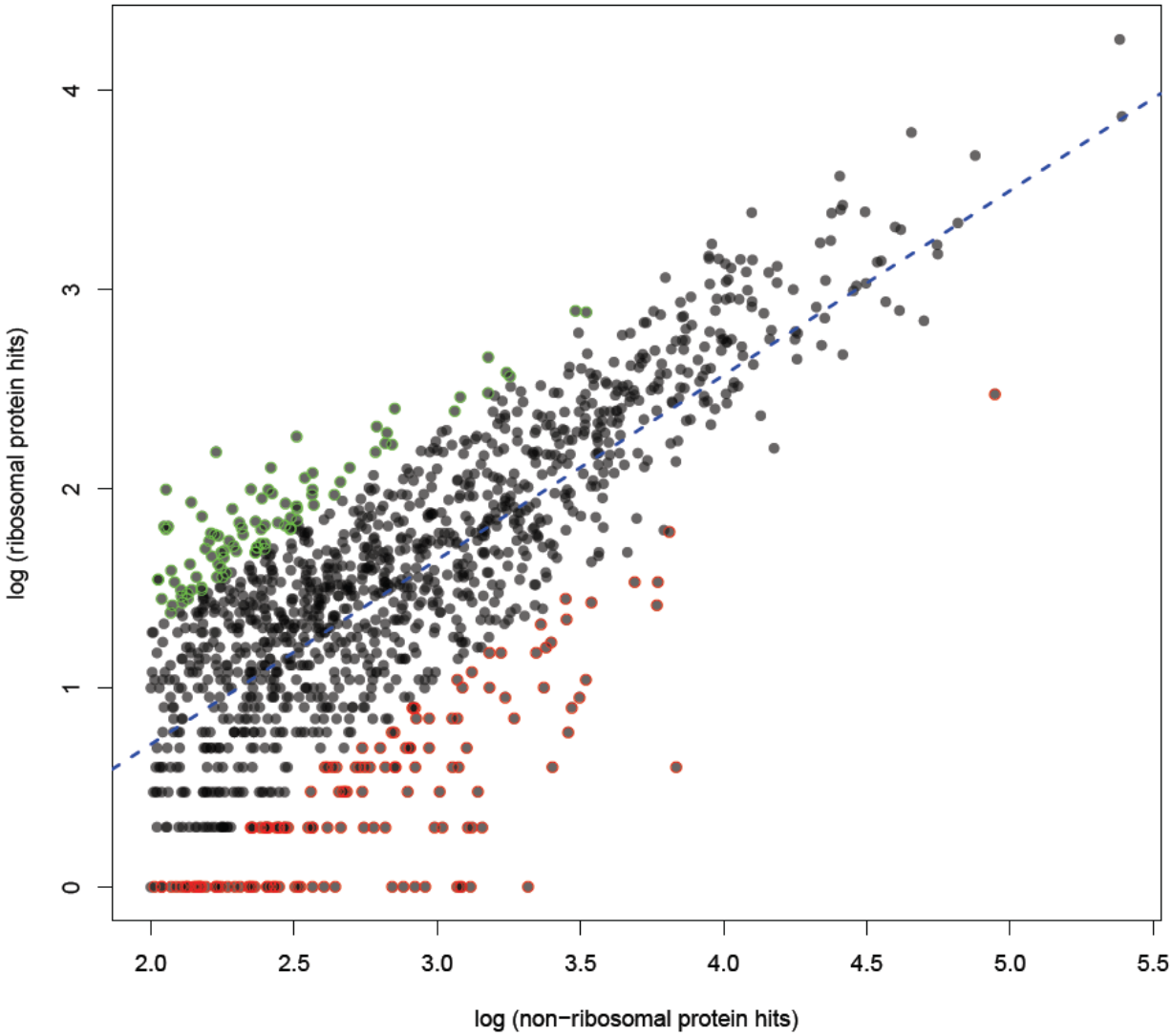



Figure 4.1. Ribosomal protein hits versus non-ribosomal protein hits for the 1,500 bacterial taxa with ≥100 hits. Reference bins with ribosomal proteins composing >20% of all hits are marked in green, and bins with <1% are marked in red. The blue line is the linear regression modeled from the log transformed values.

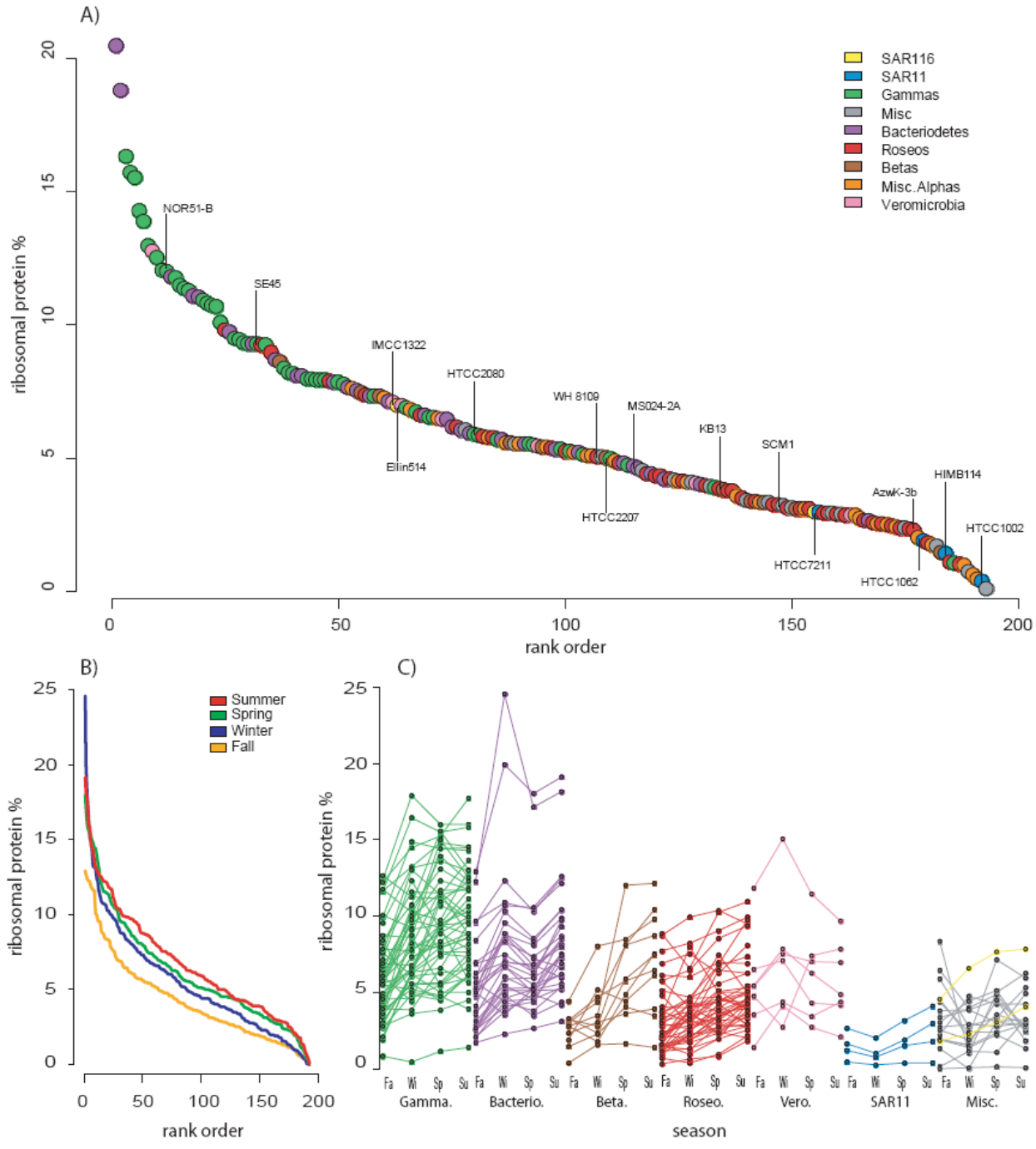



**Figure 4.2.** Relative abundance of ribosomal protein reads in the top 200 reference genomes (eukaryotic hits are not included). A) Distribution of reference genomes in rank order by the average percent ribosomal protein reads composed of all hits to a reference genome in the four seasonal datasets combined. B) Same as in A, except the four seasonal samples are plotted separately. C) Temporal trends in the proportion of ribosomal proteins for the four seasonal samples arranged by phylogenetic groupings. Fa = Fall, Wi = Winter,Sp = Spring, Su = Summer; Gamma. = Gammaproteobacteria, Bacterio. = Bacteriodetes, Beta. = Betaproteobacteria, Roseo. = Roseobacter, Vero. = Verrucomicrobia, Misc. = Miscellaneous.

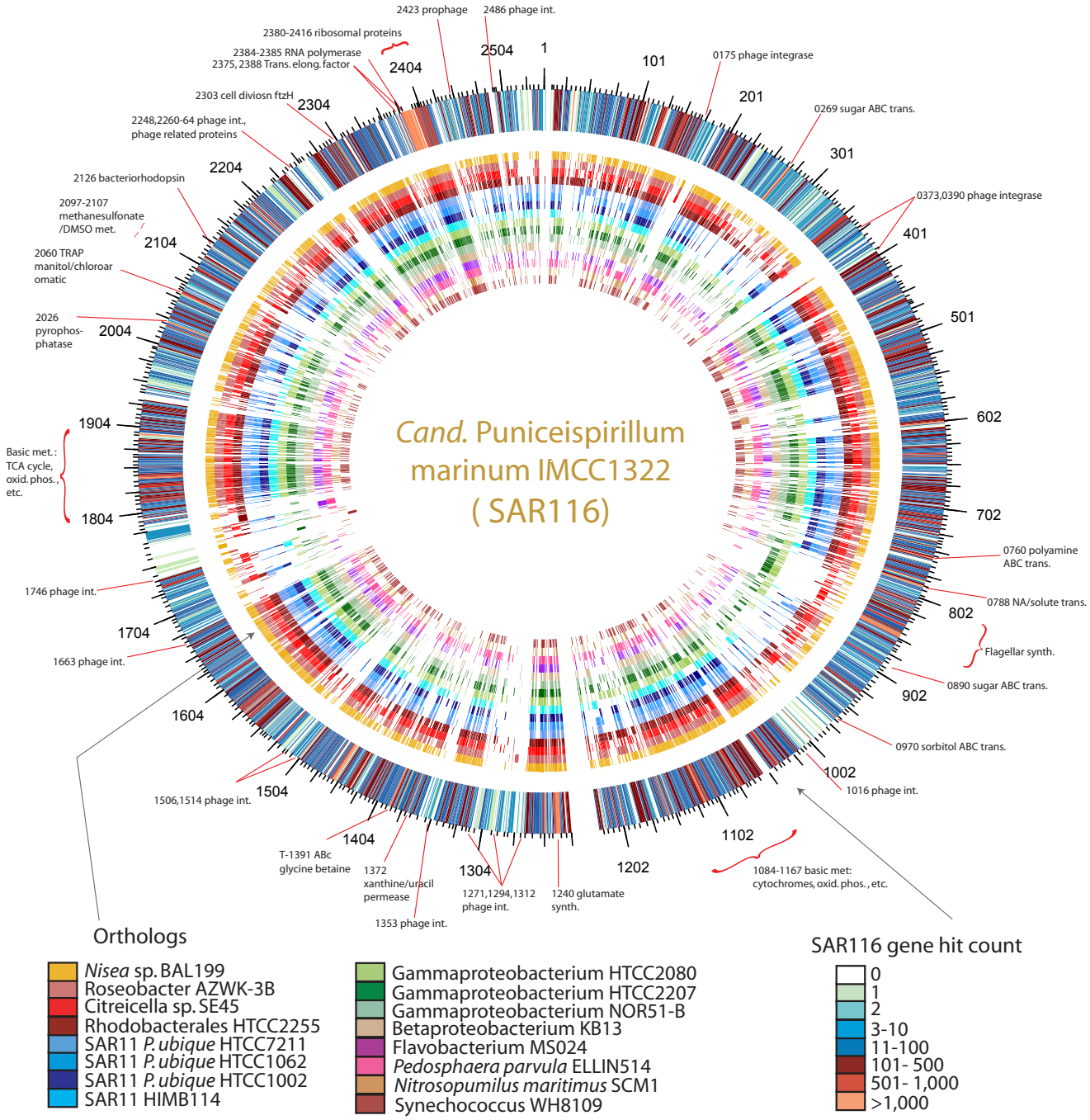



Figure 4.3. Transcriptome of SAR116 clade member *Candidatus* Puniceispirillum marinum IMCC1322. The outer colored ring shows all 2,543 genes found in the IMCC1322 genome, with each gene's color corresponding to the total number of RefSeq hits binned to it in the combined metatranscriptome. The inner colored rings denote the presence of orthologs to an IMC1322 gene in the other 15 genomes.

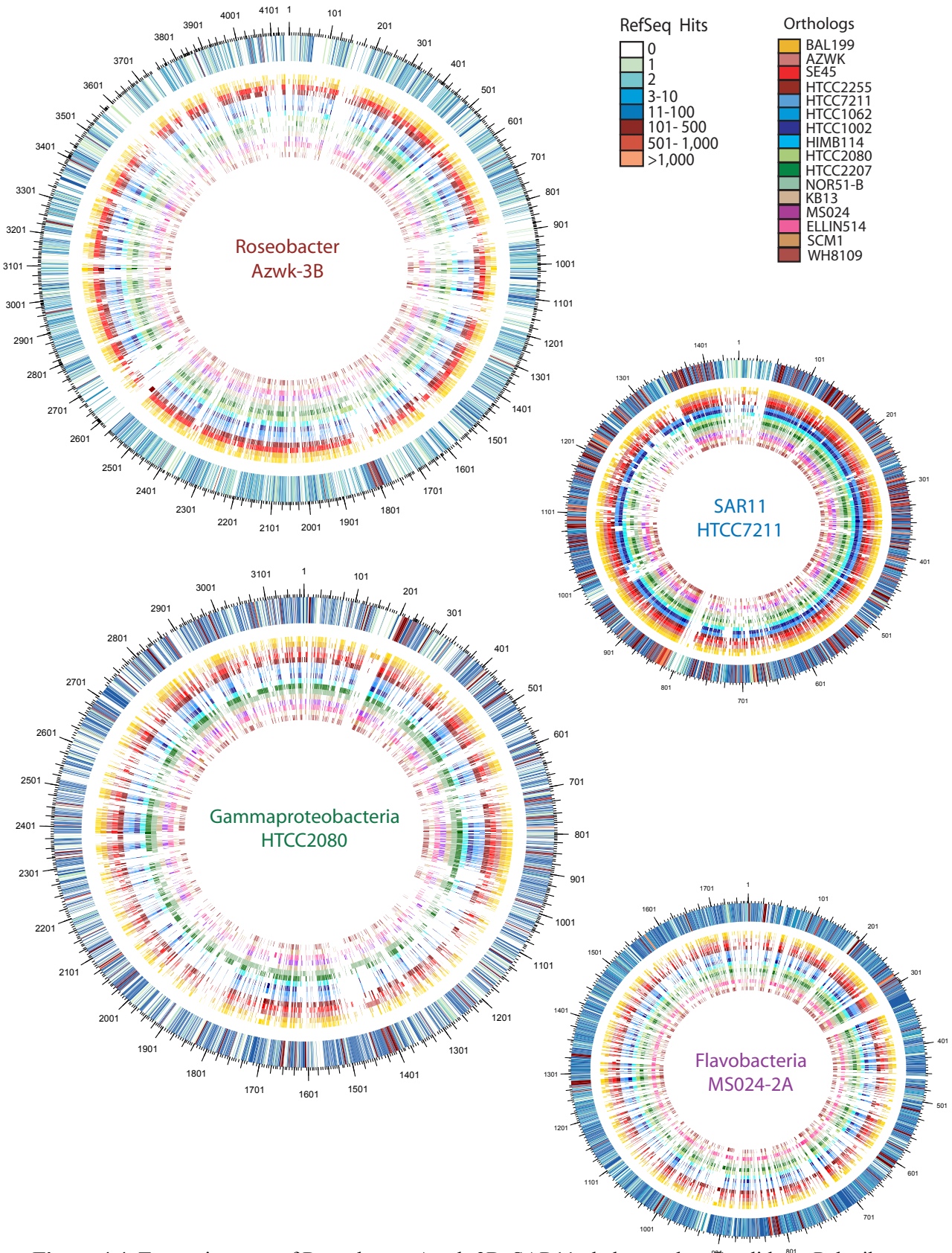



**Figure 4.4.** Transcriptomes of Roseobacter Azwk-3B, SAR11 clade member Candidatus Pelagibacter sp. HTCC7211, Gammaproteobacterium sp. HTCC2080, and Flavobacter MS024-2A. The transcriptomes are colored as in Figure 4.3.

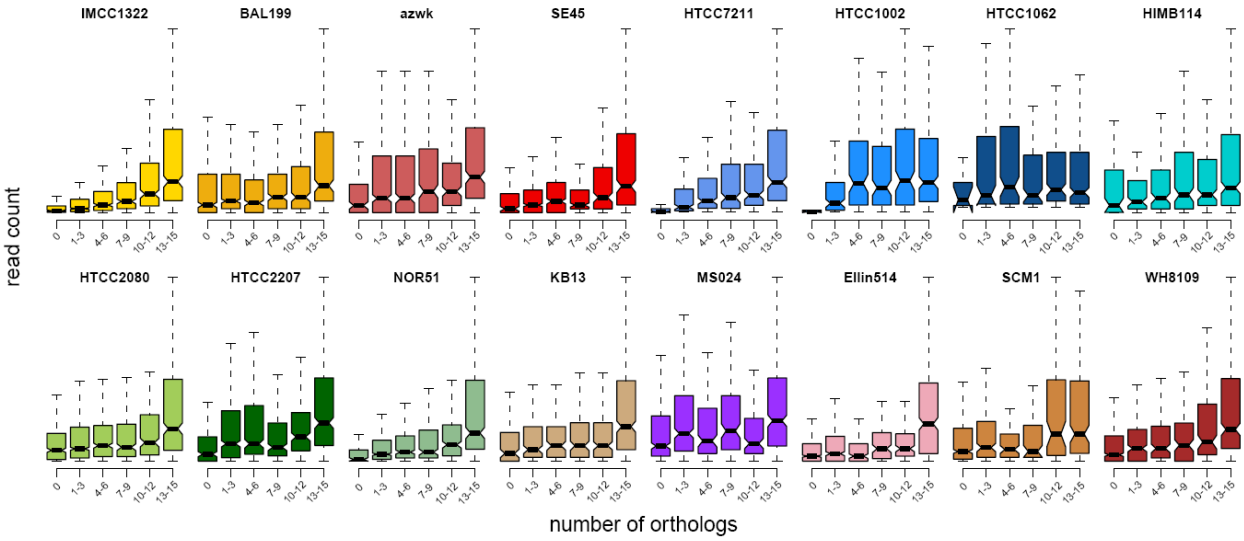
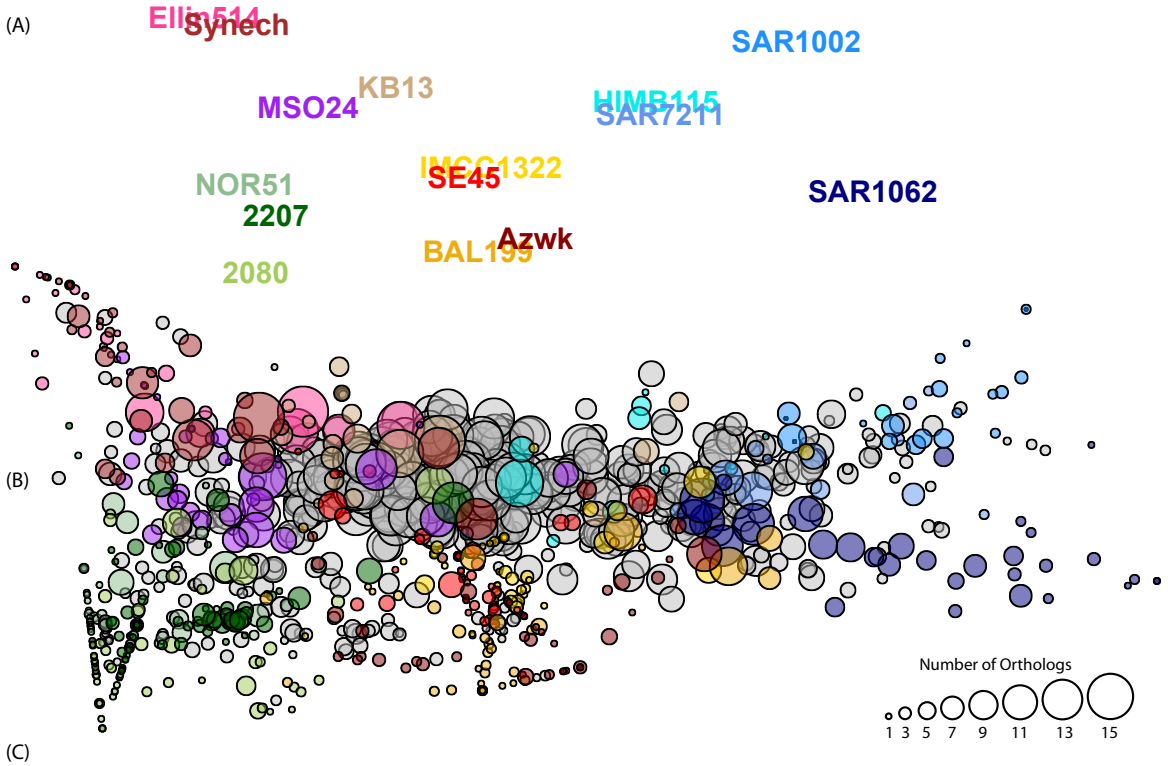



**Figure 4.5**. Gene expression as a function of orthologous relationships. The distribution of read counts for all genes in a reference genome is plotted against the number of orthologs a gene had in the other 15 genomes. The y-axis is read count, the magnitude of which varies by genome, and is not shown for simplicity. Notches in the bars indicate the median, the length of the bar is the interquartile range (IQR), and the dashed lines represent 1.5 times the IQR or the max or min value.

**Figure 4.6** Genome differentiation based on expression patterns of functional genes. A) Non-metric Multidimension Scaling (NMDS) plot of the 15 bacterial genomes (*N. maritimus* SCM1 not included) based on expression within ortholog groups. The NMDS was run with six dimensions. Only dimensions 1 and 2, which combined explained 61% of the variance, are shown here. Axis units are arbitrary and not shown for simplicity. B) The same NDMS plot shown in A, except the weighted averages of the ortholog groups are plotted instead of the genome scores. The size of the circle is relative to the number of genomes in the group. Ortholog groups that contained an indicator gene identified in the ISA analysis are colored corresponding to the genome that the gene was indicative of. C) Functional categories whose expression was indicative of a taxon.




(C)

**All**
- Transcription
  - RNA polymerase, elongation factors
- Translation
  - Ribosomal proteins, initiation and elongation factors
- Preprotein translocation
  - sec system
- Oxidative phosphorolation
  - cytochromes, ATP synthase
- Energy conservation
  - NAD+ synthetase
- Replication
  - DNA polymerase, cell division proteins

**MSO24-2A**
- TonB trans.
- Gliding motility
- Complex carbohydrate attachment and met.

**KB13**
- Methanol met.
- Propionate met.
- Bacterioferritan

**N. maritimus**
- Ammonia trans.
- Ammonia oxidation

**HTCC7211**
- Ectoine trans.
- Glycine met.
- Glycine betaine trans.
- Sulfur oxid. (APS reductase)

**Ellin514**
- Twitching Motility
- Capsular Polysacchride Production
- Secretion
- Sugar met.

**HIMB114**
- Tartrate Met.
- Phosphate trans.
- Phosphonate trans.
- Taurine met.
- Fe(II) trans.
- Tricarboxylic acid trans.

**HTCC1062**
- C6 sugar trans. and met.
- Phosphate trans.

**Syn WH8109**
- Photosynthesis
- Cell divison (*ftsH*)
- Sulfur trans. and met.
- Sulfolipid synthesis

**HTCC1002**
- Ammonia trans.
- Taurine trans.

**IMCC1322**
- Flagellar motility
- Methanesulfonate trans. and oxid.
- Protocatechuate met.

**SE45**
- Formate met.
- Urease met.
- Sulfur oxid. (sox)
- Taurine met.

**NOR51-B**
- Fatty acid met.
- TonB trans.
- AAnP

**BAL199**
- Formate met.
- Glyoxylate met.
- Fe(III) trans.
- Taurine met.

**AZWK-3B**
- AAnP
- Sulfur oxid. (sox)
- Taurine met.

**HTCC2080**
- Fatty acid met.
- TonB trans.
- AAnP

**HTCC2207**
- TonB trans.
- Flagellar motility
- Complex carbohydrate attachment and met.

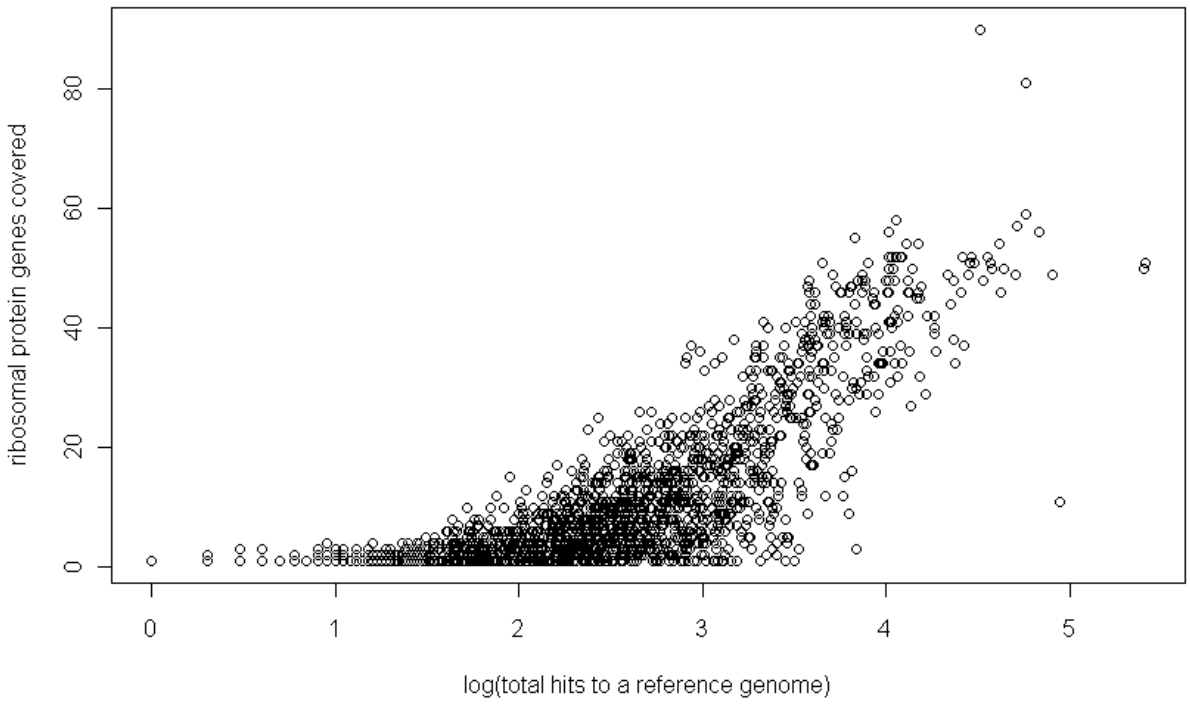



Figure 4.S1 Relationship between total hit count and the number of ribosomal genes covered in reference genomes. The number of ribosomal proteins in a genome varies, but averages around 54. Alpha proteobacterium HTCC2255 and Psychroflexus torquis ATCC 700755 had hits to 90 and 81 RP genes, respectively, which is due to the presence of contaminating sequences in these genomes.

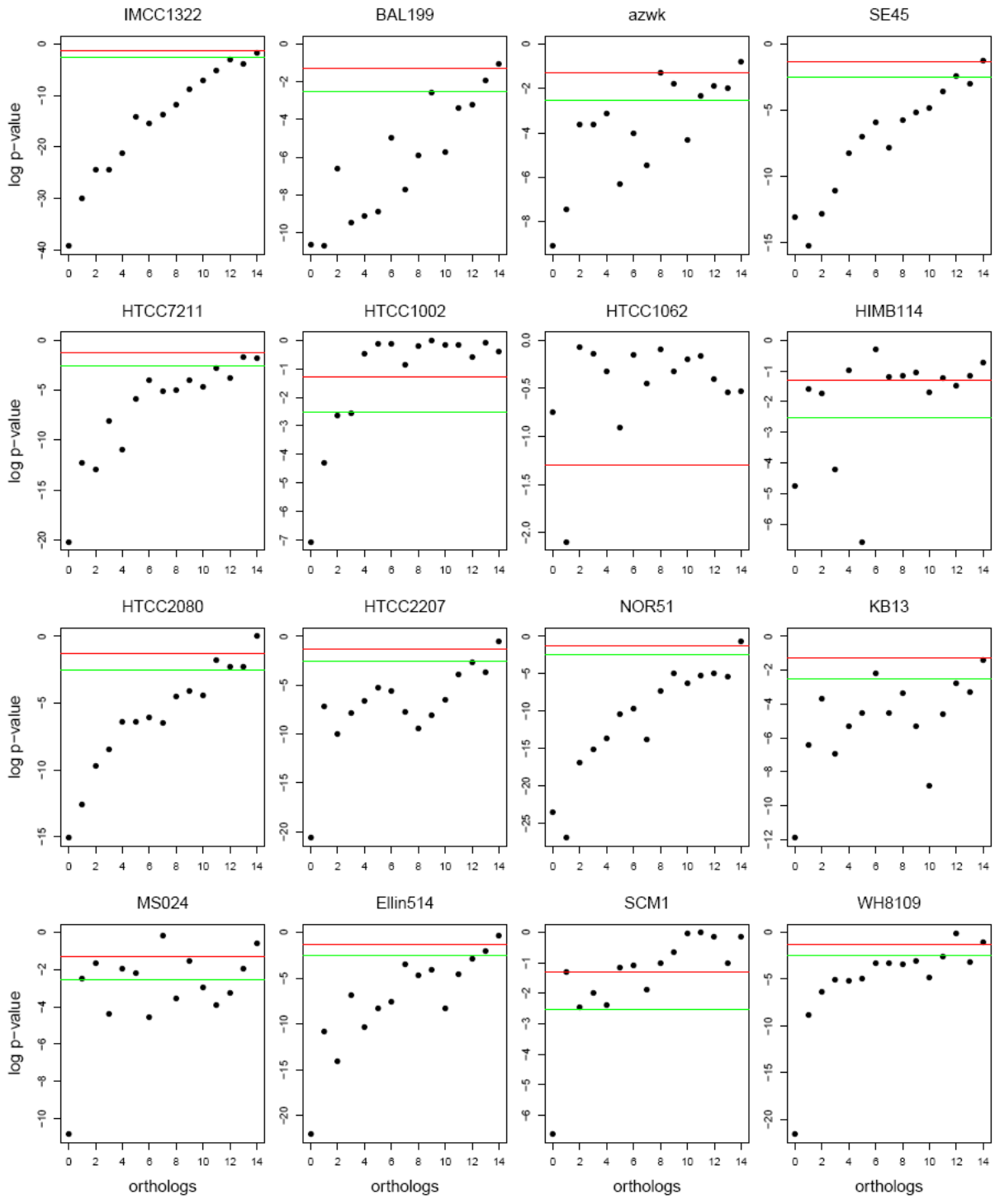



Figure 4.S2 P-values for Wilcoxon rank sum test of the differences in read counts between genes with 15 orthologs and genes with 0-14 orthologs. The red line marks a p-value of 0.05, below which any points are considered statistically significant. The Bonferroni corrected p value for multiple hypothesis testing is marked as the green line.

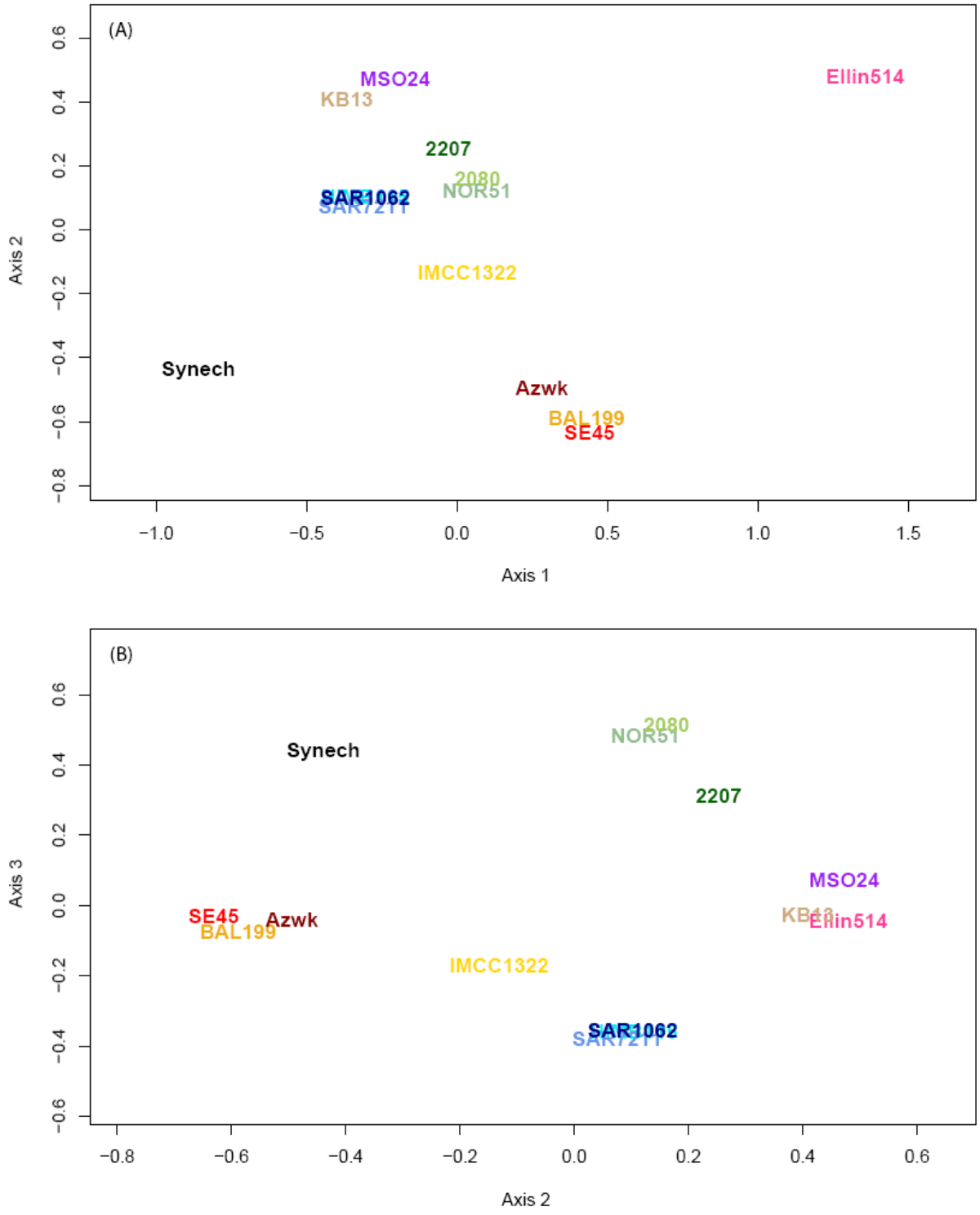



Figure 4.S3 Non-metric Multidimension Scaling (NMDS) plot of select 15 genomes (*N. maritimus* SCM1 not included) based on the presence or absence of genomes in ortholog groups. Six dimensions were included in the NDMS; A) shows dimensions one and two, while B) shows dimensions 2 and 3.

# CHAPTER 5

## SUMMARY

In the thirty plus years since the publication of Pomeroy's seminal work (1974) indicating the importance of bacteria to marine ecosystems, marine ecologists have strived to understand how bacterial processes alter the environment, and in turn, how the environment structures bacterioplankton composition and function. As we move into a 'post-genomics' era and extend our insights beyond cataloging taxa and gene data, the ability to identify expressed functional capabilities and their regulation under differing environmental conditions is increasingly important. Building upon the foundational studies of Poretskey *et al*. (2005, 2009, 2010), this dissertation used metatranscriptomic analyses to obtain insights into the ecology and biogeochemistry of coastal ecosystems by identifying active microbes and their realized functional capabilities.

Cross comparisons of samples collected at different times or locations, or under different experimental conditions, is an important process for generating and testing hypotheses. Metatranscriptomics has the potential to be a powerful tool to make such comparisons in microbial systems, but has been hampered by a poor understanding of how deeply samples are sequenced and the inability to directly compare absolute numbers of transcripts; that is, analyses typically have been limited to comparisons of proportions between samples. Furthermore, the biogeochemical value of expression data is greatly reduced if it is measured in units that do not allow for direct comparisons with other environmental parameters (rates, concentrations, etc.).

The addition of an internal standard to samples collected from coastal waters of the southeastern U.S. allowed us to make metatranscriptomics more quantitative by converting read counts into transcripts per L. The standard accounts for differences in the mRNA pool size between samples, as well as any processing and sequencing losses, allowing for direct comparisons between samples. Using an internal standard we estimated there were $10^{12}$ microbial transcripts $L^{-1}$ of coastal seawater, which agreed well with theoretical estimates based on cell abundance and RNA mass. By determining the size of the transcript pool, we were able to estimate how deeply it was sampled during sequencing. Although this was a large library relative to most previous metatranscriptome studies (~1 million protein encoding reads), it contained only ~0.00001% of the mRNAs in each sample. Even given this low coverage, we gained insights into the active microbial populations in this system and their expressed functional capabilities. Transcript abundances for 82 genes diagnostic of the marine N, P and S cycles ranged from $10^6$ (the detection limit in our study) to $10^9$ $L^{-1}$. The majority of genes had transcript abundances of $<1.5 \times 10^7$ $L^{-1}$, the level at which our analysis of statistical power revealed they would need a twofold difference in abundance between samples to be statistically different.

The most abundant transcripts in the samples ($>10^9$ transcripts $L^{-1}$) were for genes encoding the transport and oxidation of ammonia, the vast majority of which originated from an organism with high similarity to the Thaumarchaeota *Nitrosopumulis maritimus* SCM1. Indeed, a survey of the metatranscriptome revealed over 16,000 SCM1 hits (>3% of annotated reads) that covered 44% of its genome. This analysis demonstrated the utility of metatranscriptomics to detect unexpected microbial populations and determine their biogeochemical influence. Based on this information, we quantified archaeal 16S rRNA and ammonia oxidation genes over a one year period in our coastal marine study site, and found the increase in abundance to be a unique

seasonal feature of the summer samples. In addition to ammonia related genes, we were able to examine the abundance and sequence diversity of SCM1 transcripts for carbon fixation proteins, reactive oxygen species stress proteins, and many iron-sulfur and copper containing proteins. The high expression of several hypothetical proteins indicates they are ideal targets for future studies targeting SCM1 genes of environmental importance. In contrast to SCM1, transcripts for ammonia oxidizing bacteria were in relatively low abundance (<0.4% of annotated reads), with no transcripts detected for ammonia oxidation or carbon fixation. Overall, the results provided insights into the temporal variation of ammonia oxidizing microbes and their activities.

The *N. maritimus* study demonstrated that members of the same functional guild can have different responses in the same environment, providing insights into how the niche space is divided between microbial groups. We next examined how such niche diversification occurs within the broader community, a particularly important question since genome and metagenome studies have begun to indicate extensive overlap in functional potential between even very distantly related microbial taxa. One way that taxa with similar functional capabilities may differ is in the amount of transcriptional effort devoted to growth. Using the more than 200,000 ribosomal proteins sequenced in the metatranscriptomes as a proxy for relative growth rate, we detected clear differences in activity, with members of the Gammaproteobacteria and Bacteroidetes highly active and members of SAR11 clade the least active. In contrast, an examination of the most highly expressed functional genes among these populations showed them to have extensive overlap, even among distantly related taxa. This was reflected in a strong positive relationship between expression level and functional redundancy (as measured by the number of orthologs a gene had in the other 15 genomes). By examining genes that were outliers for expression level within an ortholog group, we were able to identify genes indicative of niche

specialization. This included genes for the uptake and metabolism of a variety of substrates, strategies for energy generation, and capabilities for physical interactions with other organisms or particles in the environment. These results shed light on the different genetic and regulatory tactics that allows for the active coexistence of so many microbial taxa in the marine plankton.

**References**

Pomeroy LR (1974). The ocean's food web, a changing paradigm. *Bioscience* **24:** 499-504.

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M *et al* (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Micro* **71:** 4121-4126.

Poretsky RS, Hewson I, Sun SL, Allen AE, Zehr JP, Moran MA (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11:** 1358-1375.

Poretsky RS, Sun S, Mou X, Moran MA (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* **12:** 616-627.

# APPENDIX A

# QUANTITATIVE MICROBIAL METATRANSCRIPTOMICS[1]

[1]Gifford, S.M., Satinsky, B., and Moran, M.A. Submitted to *Methods in Molecular Biology: Environmental Microbiology (2nd Ed.).*

**Abstract**

The direct retrieval and sequencing of environmental RNA is emerging as a powerful technique to elucidate the *in situ* activities of microbial communities. Here we provide a metatranscriptomic protocol describing environmental sample collection, rRNA depletion, mRNA amplification, cDNA synthesis, and bioinformatic analysis. In addition, the preparation of internal RNA standards and their addition to the sample is described, providing a method by which transcript numbers can be expressed as absolute abundances in the environment and more readily compared to other biogeochemical and ecological measurements.

**Introduction**

Advances in molecular techniques have revolutionized the field of microbial ecology, particularly in revealing the extraordinary phylogenetic and functional diversity contained within microbial communities. A major contemporary challenge is identifying which components of this complex functional gene pool are actively being expressed and how that expression varies over time and space. The direct collection and sequencing of RNA from the environment (termed metatranscriptomics) fulfills this need by providing a measure of a community's instantaneous transcriptional response to its surrounding environment. The development of this method in parallel with advances in next generation sequencing technologies have made metatranscriptomics a powerful approach for analyzing *in situ* microbial expression in a wide variety of habitats.

The metatranscriptomics approach was first described by Porestky *et al*. (2005), and while there have been several modifications since then, it consists largely of the same modules. Cellular biomass is rapidly collected from the environment in a manner that disturbs ambient conditions as little as possible. RNA is extracted from the samples and treated with DNase to

remove any residual DNA (Fig. A.1). As ribosomal RNA (rRNA) makes up the majority of total cellular RNA, steps are taken to decrease rRNA abundance in order to increase the yield of protein-encoding sequences in the resulting libraries. Partially due to rRNA reduction, there is a significant decrease in total RNA mass and the sample is linearly amplified to produce sufficient material for sequencing. Finally, the amplified RNA (aRNA) is converted to double stranded cDNA, which can be sequenced through a variety of methods (Sanger, 454 pyrosequencing, Illumina, etc.). Using this approach, Poretsky *et al*. (2005, 2009, 2010) and others (Frias-Lopez *et al*., 2008; Gilbert *et al*., 2008; Urich *et al*., 2008; Shi *et al.*, 2009; Hewson *et al.*, 2009) were able to successfully characterize metatranscriptomes from a variety of environments.

The most significant methodological challenge for metatranscriptomics has been efficient removal of rRNA. Poretsky *et al.* (2005, 2009, 2010) used a dual removal approach based on two commercially available kits. In the first round, rRNA is enzymatically digested by an exonuclease that targets the 5' monophosphates found on rRNA, leaving mRNAs, which have a 5' triphosphate, intact. In the second round, biotinylated probes are hybridized to the rRNA and are bound to streptavidin coated magnetic beads, allowing for physical separation via a magnetic stand. Typically, this dual approach removes ~50% of contaminating rRNA (see chapter 2 and Poretsky *et al.*, 2009), although concern has been raised that the first round may cause bias in the resulting transcript library (He *et al.*, 2010). More recently, Stewart *et al.* (2010) improved rRNA removal efficiency by using only the second (hybridization based) approach but creating custom hybridization probes to target the rRNAs in each individual sample. This method has been shown to decrease the proportion of rRNA reads to 10 to 30% of total sequences (Stewart *et al.*, 2010; A. Rivers and B. Satinsky, unpublished data).

A second challenge has been the interpretation of transcript abundances, which have traditionally been measured only as relative proportions within a sample (see chapter 2). The ability to make quantitative interpretations, including cross sample comparisons, is limited when only proportional data is available. For example, a change in the abundance of one transcript category in a metatranscriptome causes the other categories' proportional representations to change also, even if the absolute abundance of those other types remains constant. This limitation can be overcome by the addition of a standard (an artificial mRNA) just prior to starting the sample extraction (chapter 2). Since both the amount of standard added and the amount of standard recovered is known, one can calculate the depth of sequencing and absolute copy number of a transcript category in a more ecologically relevant unit, such as copies volume $^{-1}$ or copies mass$^{-1}$.

Here we present an updated version of the Poretsky *et al*. (2005) protocol, using the custom subtractive hybridization protocol developed by Stewart *et al*. (2010) for rRNA removal and the addition of internal standards to obtain absolute copy numbers in the environment from chapter 2. The method takes advantage of several commercial kits, and the reader should thoroughly familiarize him/herself with each kit's manual. A number of steps use spin cartridges for purification, which efficiently capture mRNA sized fragments (>200 nt), but small RNAs, including many regulatory RNAs, are likely not retained. Quantification is carried out spectrophotometrically (e.g., Nanodrop spectrophotometer) or with a fluorescence assay (e.g., PicoGreen for DNA and RiboGreen for RNA). Nucleic acid size distributions are visualized with an Agilent Bioanalyzer or Experion automated gel electrophoresis system.

RNAs have short half lives and are quickly degraded by ubiquitous RNAses. Wearing gloves, working in a clean lab space such as a PCR hood with a UV lamp, and being sure to

clean all pipettes and surfaces with an RNase degrading solution (ex: RNaseZap, Ambion) improve success rates. When not actively working with the RNA sample, it should be kept either on ice or frozen at -80°C. Only plastic wear that has been certified as nuclease free should be used. ART barrier tips are recommend for pipetting.

**1. Environmental Collection**

While a variety of methods can be used to collect biomass for RNA extraction, important points to consider are that the sample is collected as quickly as possible to prevent turnover of the mRNA pool and to keep the sampling conditions as close to ambient as possible to reduce transcriptional response to changes during collection. The sample should be preserved immediately after collection, either by snap freezing in liquid $N_2$ or by addition of an appropriate preservative. For optimal downstream processing, attempt to collect enough biomass to yield 5-20 μg of total RNA. Here we describe a collection method for aquatic environments.

Materials:

Peristaltic or vacuum pump
Tubing (preferably acid washed)
Pre-filter (if desired, for example a 3 μm pore-size) and collection filter (typically 0.22 μm pore size; we recommend Supor (Pall, Port Washington, NY))
Filter housings
Liquid nitrogen or RNALater (Applied Biosystems, Austin, TX)
Graduated 10 or 20 L carboy
Sterile forceps
Whirl-Pak® bags (Nasco, Fort Atkinson, WI)

1.1 Setup the filtration system consisting of tubing, pre-filter (optional), 0.22 μm filter, and a graduated carboy (Fig. A.2).

1.2 Place one end of the tubing in the water and draw water through the filter system, measuring the volume filtered by its accumulation in the collection carboy. The appropriate volume to filter will depend on the environmental cell concentration. For

coastal or limnological samples, 5-10 L is often sufficient. Oligotrophic samples may require higher volumes. Total collection times should be kept as short as possible, optimally finishing the collection in 5 to 10 min, and no longer than 30 min.

1.3 After the desired volume has been filtered, allow any water remaining in the line to pass through the filters. For optimal RNA yield, the surface of the filter should be nearly dry.

1.4 Fold the 0.22 μm filter and place into a Whirl-Pak® bag. Remove any air from the Whirl-Pak® by squeezing it out with your gloved fingers. Place the Whirl-Pak® into a liquid $N_2$ dewar. Alternatively, preserve the filter by submerging it in a tube containing 10 ml RNAlater.

1.5 Repeat the process to collect an additional filter to be used for the DNA extraction; this is needed for the custom probe rRNA reduction protocol (Stewart *et al.*, 2010).

**2. Internal Standard Synthesis**

Construction of the internal RNA standards is done by *in vitro* transcription of DNA templates; these can be either commercially available plasmids (such as is commonly used for cloning) or synthesized DNA that is inserted into a plasmid. The use of a template that is already part of commercially available plasmids is attractive for its ease of use and low cost (see chapter 2). However, these plasmids make size customization difficult and often contain regions of homology to functional proteins. An alternative approach is to create a custom sequence, which is then synthesized and inserted into a plasmid, providing optimal control over sequence length and composition. For either approach, the final plasmid should contain the following components (in order): an RNA polymerase promoter sequence, the internal standard sequence, and a restriction site (targeting a unique site in the vector and preferably producing a blunt end). Candidate internal standard sequences should be compared to relevant databases to identify any

regions of homology that could interfere with unambiguous identification of the added standard in the sequence library. Multiple standards of different length and sequence composition should be designed. The addition of multiple standards to the sample helps to control for pipetting errors and size selection biases that may decrease the accuracy of the final quantification estimate. The appropriate amount of internal standard added to a sample would ideally be based on the expected total RNA mass yield. An addition of 0.5% proportion of internal standard mass to expected sample mass of RNA is appropriate for next generation sequencing. Here we provide a general outline of a protocol starting with a custom designed standard. For standard construction using commercially available plasmids see chapter 2.

Materials:

Plasmid containing internal standard sequence
RNA polymerase and buffers
100% Ethanol
Cloning system
Restriction enzyme and buffers
miniPrep plasmid extraction kit

*In vitro* transcription of standard containing plasmid

2.1 Amplify the plasmid containing the internal standard sequence by cloning it into *Escherichia coli* or other appropriate vector.

2.2 Purify the amplified plasmid with a miniPrep kit.

2.3 Linearize the plasmid with the restriction enzyme targeting the restriction site at the end of the template sequence. If sticky ends were generated remove with mung bean nuclease.

2.4 Purify the linearized plasmid with a phenol:chloroformisoamyl alcohol extraction.

2.5 *In vitro* transcribe the plasmid with an RNA polymerase matching the template promoter.

2.6 Degrade the plasmid DNA using DNase.

2.7 Purify the RNA standard with a phenol:chloroformisoamyl alcohol extraction.

2.8 Quantify the RNA standard fluorometrically with Ribogreen, and confirm the standard is a single fragment of expected size using an Experion or Agilent electrophoresis system.

**3. RNA extraction**

Many different methods are available for RNA extraction depending upon the environment of interest (aquatic, terrestrial, tissue, etc.) Here, we describe a modified approach based on Qiagen's RNeasy kit.

Materials:

Vortex station with 50 ml tube adapter (MO BIO Laboratories, Carlsbad, CA)
Rubber mallet and scissors
50 and 15 ml Falcon tubes
RNeasy RNA extraction kit (Qiagen, Valencia, CA)
Extra RLT buffer (Qiagen, Valencia, CA)
β-mercaptoethanol
30 ml syringe and 18-21 gauge needles
Centrifuge for both large (50 and 15 ml) and small (1.5 ml) tubes
100% molecular grade ethanol
Vacuum manifold
0.2 mm low-binding zirconium beads (OPS Diagnostics, Lebanon, NJ). Sterilized by heating at 500 °C overnight in a combustion oven.

3.1 Prepare a 50 ml Falcon tube with 8 ml RLT buffer (β-mercaptoethanol added) and 2 ml beads.

3.2 Add the internal RNA standards to each Falcon tube. Each standard should be added independently (i.e. not as a pooled master mix) so that pipetting errors will be included in variance estimates.

3.1 Remove the filter from liquid nitrogen or -80 °C storage, break up to expose the most filter surface, and add to the Falcon tube. Many filters are brittle when frozen and can be easily shattered with a mallet. Alternatively, the filters can be cut up with sterilized scissors. After adding the filter to the Falcon tube, cap tightly and seal with parafilm. If

the samples were preserved using RNALater, the filters should be removed from the RNALater solution and any excess RNALater allowed to drip off the filter by gently squeezing with sterile forceps. The filter should then placed into a Whirl-Pak bag, snap frozen in liquid nitrogen, and processed as described above.

3.3 Place the Flacon tubes on a vortex adapter and vortex at maximum speed for 10 min.

3.4 Centrifuge at 5000 rpm for 1 min.

3.5 Using a 1000 µl pipette, transfer the liquid to a clean 15 ml Falcon tube. Ideally, 80-90% of the original volume should be recovered.

3.6 Centrifuge at 5000 rpm for 5 min.

3.7 Gently pour the supernatant into a clean 50 ml Falcon tube, being careful not to disturb the pellet. At this point, the supernatant should be free of all beads and filter material.

3.8 Add 1X volume of 100% ethanol.

3.9 Shear the sample by drawing the ethanol-lysis mixture up into the 30 ml syringe with an 18 to 21 gauge needle and then expel. Repeat three times, then draw up the solution and keep it in the 30 ml syringe.

3.10 Place an RNeasy spin cartridge on the vacuum manifold. Turn on the vacuum and slowly expel the lysis mixture from the syringe into the cartridge. Depending on how much biomass was on the original filter, it may take several minutes to pass all of the lysate through the column. (If a vacuum manifold is unavailable, the lysate can be passed through the spin column using multiple centrifugations). After all the lysate has been filtered, remove the column from the manifold, placing it back in the collection tube, and centrifuge at 11,000 rpm for 1 min to remove any residual lysate solution.

3.11 Continue by following the standard RNeasy protocol as described in the kit manual. Briefly, wash once with 700 µl RW1, and twice with 500 µl RPE. Conduct a final centrifugation to remove any residual solutions. Place in a new collection tube and elute with two separate aliquots of 50 µl RNase free water. Place on ice.

3.12 Quantify the RNA yield with either a Nanodrop spectrophotometer or RiboGreen-based fluorometric technique.

This is a potential stopping point. The eluted RNA can be frozen at -80°C. However, it is a good idea to keep the number of freeze/thaws to a minimum to reduce RNA degradation, so if possible continue on with the DNA removal step.

**4. Removal of Residual DNA**

A double treatment with TurboDNase is highly effective in digesting contaminant DNA in the RNA preparation. Note, for this and all other air incubations, place the tube(s) in a rack that allows ample air movement around the tube. For many of the reactions, it is important that temperature is uniform around the tube.

Materials:

Turbo DNA-free (Applied Biosystems, Austin, TX)
Centrifuge
Incubator

4.1 The sample should be in 90 µl of nuclease-free water.

4.2 Add 10 µl DNase buffer and 3 µl TurboDNase.

4.3 Incubate at 37°C for 20 min in an incubator.

4.4 Remove the mixture from the incubator and add an additional 3 µl of TurboDNase.

4.5 Return to the incubator for another 20 min.

4.6 Add 20 µl inactivation reagent and incubate at room temperature for 2 min, vortexing every 20 or 30 seconds.

4.7 Centrifuge at max rpm (typically 14,000 rpm) for 1 min.

4.8 Being careful not to disturb the inactivation reagent at the bottom of the tube, transfer the supernatant (~90 to 100 µl) to a new tube and place on ice.

This is a potential stopping point. Store the sample at -80°C

**5. Ribosomal RNA Reduction**

Here we provide a brief overview of the custom rRNA depletion protocol and direct the reader to the original description by Stewart et al. (2010) for specific details. This method uses universal primers to PCR amplify rRNA genes from a DNA sample collected in parallel with the RNA samples (the DNA filter must be extracted prior to starting the rRNA subtraction protocol). Several independent amplifications are carried out, depending on the rRNA targeted for removal (i.e. 16S/23S Bacteria, 16S/23S Archaea, 18S/28S Eukaryotes). The universal primers are modified to incorporate a T7 promoter into the PCR products. The PCR amplified rDNA templates are then transcribed *in vitro* to make anti-sense rRNA probes containing biotinylated nucleotides. The probes are hybridized to the sample rRNA, bound to streptavidin magnetic beads, and physically separated from the rest of the sample via a magnetic stand.

Materials:

T7 modified PCR primers (see Stewart et al. 2010 for primer design)
Herculase II Fusion Polymerase (Agilent Technologies, Santa Clara, CA)
QIAquick PCR purification kit (Qiagen, Valencia, CA)
MEGAscript Transcription Kit (Applied Biosystems, Austin, TX)
MEGAclear Kit (Applied Biosystems, Austin, TX)

Biotin-11-CTP (10mM) (Roche Applied Science, Indianapolis, IN)
Biotin-16-UTP (10mM) (Roche Applied Science, Indianapolis, IN)
SUPERase•In RNase Inhibitor (Applied Biosystems, Austin, TX)
RNeasy MinElute Cleanup Kit (Qiagen, Valencia, CA)
Streptavidin-coated Magnetic Beads (New England Biolabs, Ipswich, MA)
20X Sodium Chloride-Citrate (SSC) Buffer (RNase-free) (Applied Biosystems, Austin, TX)
DynaMag Spin Magnet (Invitrogen, Carlsbad, CA)
Formamide (100%)
0.1M NaOH (nuclease-free)

5.1 PCR Amplification of rRNA Genes

5.1.1 For each rRNA gene amplification (16S Bacterial, 23S Bacterial, 16S Archaeal, 23S Archaeal, 18S Eukaryotic, and 28S Eukaryotic), prepare 4-5 individual 50 µL PCR reactions in 0.2 mL tubes on ice. For each reaction combine 5 to 100 ng template DNA, 10 µL Herculase 5X Buffer, 0.5 µL dNTP (100 mM), 1.25 µL forward primer (10 µM), 1.25 µL reverse primer (10 µM), 1 µL Herculase II Fusion Polymerase, and nuclease-free water to 50 µL reaction volume. Mix samples and briefly centrifuge.

5.1.2 Place the reactions in a thermal cycler and run with one of the two following protocols. For all targets other than Bacterial 23S, denature at 92°C for 2 min, run 35 to 40 cycles 95°C for 20 s, 55°C for 20 s, 72°C for 2 min, and end with a final extension at 72°C for 3 min. For Bacterial 23S targets, denature at 92°C for 2 min, run 35 to 40 cycles of 95°C for 20 s, 39°C for 20 s, 72°C for 90 s, and end with a final extension at 72°C for 3 min.

5.1.3 Pool the replicate 50 µL reactions into a single microcentrifuge tube.

5.1.4 Clean up the reactions using a QIAquick PCR purification kit, eluting in 30 to 50 µL of elution buffer (EB).

5.1.5 Quantify the PCR products with Nanodrop spectrophotometer or PicoGreen-based fluorometric method. It is critical to obtain 250 to 500 ng $\mu L^{-1}$ of pooled PCR products before proceeding to the *in vitro* transcription. The PCR products should also be run on a gel to confirm the correct product amplification

5.2 Biotin-labeled Anti-sense RNA Probe Creation

Anti-sense rRNA probes are synthesized via *in vitro* transcription with T7 RNA polymerase using the MEGAscript High Yield Transcription kit. Prepare separate 20 µl reactions for each rRNA probe type (16S, 18S, etc.). *In vitro* transcription reaction volumes can be doubled to increase yield if necessary.

5.2.1 For each rRNA gene product (16S, 18S, etc.), combine the following in order in a 0.2 mL PCR tube: 1 µL PCR amplicons (250-500 ng) from previous amplification, 2 µL ATP (75mM), 2 µL GTP (75 mM), 1.5 µL CTP (75 mM), 1.5 µL UTP (75 mM), 3.75 µL Biotin-11-CTP (10 mM), 3.75 µL Biotin-16-UTP (10 mM), 2 µL 10X buffer, 0.5 µL RNase Inhibitor (Ambion), 2 µL T7 polymerase.

5.2.2 Incubate in a thermal cycler at 37°C overnight (heated lid set to 105°C).

5.2.3 Add 1 µL DNase I to each reaction and incubate for 30 min at 37°C in a thermal cycler.

5.2.4 Clean up the reaction with a MEGAclear kit, eluting in 50 µL of elution buffer.

5.2.5 Quantify probe concentration using either a Nanodrop spectrophotometer or RiboGreen-based fluorescence method. A good transcription will result in >50 to 75 fold increase over the input DNA mass.

5.3 Subtractive Hybridization

5.3.1 Determine the input quantities of sample RNA template and biotinylated-rRNA probes. Ideally, 250 to 500 ng of sample RNA (i.e. the original RNA pool containing mRNA and rRNA) will be used in the rRNA reduction processes. However, subtraction can be successfully conducted using lower template quantities if necessary. Each individual probe should be added at a probe:template ratio of 2:1. For example, if 500 ng of sample RNA is added to the reaction, then 1000 ng of each unique probe should be added to the same reaction. The final total RNA (sample + rRNA probes) in the depletion reaction is calculated as:

(sample RNA mass) + [ (number unique rRNA probes) X (2 X sample RNA mass) ]

5.3.2 Calculate the volume of streptavidin bead suspension required and prepare by washing. A volume of 100 μL of streptavidin beads can be used with up to 2,000 ng total RNA (rRNA probes + RNA sample). Based on the total RNA calculated in 5.3.1, add the appropriate volume of streptavidin beads needed into a 1.5 ml tube. Place the tube in a magnetic stand and let sit for 3 min. Discard the supernatant. Remove the tube from the stand and resuspend the beads in an equal volume of 0.1M NaOH. Place back on the stand, bind the beads, and discard the supernatant. Remove from the stand and add an equal volume of 1X SSC buffer to the beads, mixing thoroughly to resuspend. Again separate the beads and discard the supernatant. Repeat the 1X SSC wash twice and on the third wash leave the beads in the SSC buffer and place on ice.

5.3.3 In a 0.5 mL tube combine: RNA sample and each rRNA probe (volumes determined in 5.3.1), 1 μL RNase inhibitor, 2.5 μL 20X SSC buffer, and 10 μL 100% Formamide. Bring the volume up to 50 μL with water.

5.3.4 Incubate in a thermal cycler for 5 minutes at 70°C followed by ramping down to 25°C using 5°C increments for 1 min each.

5.3.5 Remove from the thermal cycler and incubate for 5 min at room temperature. During this period it is useful to continue on with the bead dry-down step below (5.3.6).

5.3.6 Place the washed streptavidin beads (5.3.2) on the magnetic stand and allow the beads to separate for 3 min. Discard the supernatant.

5.3.7 To the hybridization reaction tube add 1X SSC-20% Formamide solution so that the end volume of the hybridization reaction is equal to the initial aliquoted bead volume (5.3.2). For example, if the initial volume of beads aliquoted to deplete an individual reaction was equal to 200 μL, then add 150 μL of 1X SSC-20% Formamide solution to the 50μL hybridization reaction.

5.3.8 Add the hybridization reaction mix from 5.3.7 to the tube containing the dried beads (5.3.6). Incubate at room temperature for 10 minutes, occasionally flicking to mix.

5.3.9 Place the tube in a magnetic stand and allow the beads to separate for 3 min.

5.3.10 Transfer the supernatant (containing the purified RNA sample) into a clean 1.5 mL collection tube.

5.3.11 Resuspend the beads with 1X SSC, matching the original volume of the bead suspension (5.3.2). Return the beads to the stand and incubate for 3 min. Transfer the supernatant to the tube containing the first aliquot of supernatant (5.3.9).

5.3.12 Clean up and concentrate the depleted RNA using an RNeasy MinElute kit (Qiagen).

5.3.13 Quantify the enriched mRNA and confirm rRNA reduction with a Bioanalyzer or Experion system. This is potential stopping point. Store at -80 C.

**6. Amplification**

To obtain enough material for sequencing, the enriched mRNA sample is linearly amplified using the MessageAmp™ Amplification Kit, consisting of four main steps: polyadenylation, reverse transcription to single stranded cDNA, second strand cDNA synthesis, and *in vitro* transcription to anti-sense aRNA. The user should closely read and follow the protocol described in the kit manual. Here we only provide a brief overview.

Materials:

MessageAmp™ II-Bacteria aRNA Amplification Kit (Applied Biosystems, Austin, TX)
Thermal cycler
Incubator
Tabletop centrifuge (all centrifugations are conducted at ~10,000 RPM)
100% ethanol

6.1 Polyadenylation

6.1.1 Add 10-200 ng of mRNA in a total volume of 5 µl of water to a 0.5 ml tube.

6.1.2 Denature sample in a thermal cycler for 10 min at 70°C.

6.1.3 Assemble polyadenylation master mix using the online calculator. Gently vortex and centrifuge.

6.1.4 Add 5 µl polyadenylation master mix to each sample.

6.1.5 Incubate at 37°C for 15 min. During this incubation, you may want to prepare the first strand synthesis master mix (see 6.2 below)

6.1.6 Remove the samples from the incubator, place on ice, and proceed immediately to the next step.

6.2 First Strand Synthesis

6.2.1 Prepare the first strand master mix using the online calculator. Gently vortex and centrifuge.

6.2.2 Add 10 µl of the master mix to each sample. Gently vortex and centrifuge.

6.2.3 Incubate for 2 hr at 42°C. Then place on ice and proceed immediately to the second strand cDNA synthesis.

6.3 Second Strand Synthesis

6.3.1 On ice, assemble the second strand master mix using the online calculator. Gently vortex and centrifuge.

6.3.2 Add 80 µl of the master mix to each sample. Gently vortex and centrifuge.

6.3.3 Incubate in a thermal cycler pre-cooled to 16°C for 2 hrs (the lid temperature should either match or be turned off). During this incubation, bring the bottle of nuclease-free water to 50°C.

6.3.4 When the incubation is finished, place the samples on ice and proceed to the cDNA clean up.

6.4 cDNA clean up

6.4.1 Add 250 µl cDNA binding buffer and transfer to a cDNA clean up spin cartridge.

6.4.2 Centrifuge for 1 min. Discard the flow through.

6.4.3 Pipette 500 µl wash buffer onto the cartridge. Centrifuge for 1 min. Discard flow through.

6.4.4 Centrifuge for an additional minute to remove any trace amounts of ethanol.

6.4.5 Transfer the cartridge to a clean cDNA elution tube.

6.4.6 Elute by adding 18 µl preheated 50°C nuclease-free water to the cartridge. Incubate at room temperature for 2 min. Centrifuge for 1 min.

6.4.7 Discard the cartridge and place the samples on ice.

6.5 *In vitro* transcription

6.5.1 Prepare the *in vitro* transcription master mix using the online calculator. Gently vortex and centrifuge.

6.5.2 Add 24 µl of the master mix to each sample. Gently vortex and centrifuge.

6.5.3 Incubate at 37°C. A 14 h incubation time is recommended to maximize aRNA yield.

6.5.4 Add 60 µl nuclease-free water to bring the final volume up to 100 µl and place on ice.

6.6 aRNA purification

6.6.1 At least 30 min before starting the purification incubate the nuclease-free water at 55°C.

6.6.2 Add 350 µl aRNA binding buffer to each sample.

6.6.3 Add 250 µl 100% ethanol. Mix by pipetting up and down.

6.6.4 Transfer the mixture to an aRNA filter column. Centrifuge for 1 min. Discard flow through.

6.6.5 Apply 650 µl wash buffer to the column. Centrifuge for 1 min. Discard flow through.

6.6.6 Centrifuge for an additional 2 min to remove any trace amounts of ethanol.

6.6.7 Transfer the cartridge to a clean collection tube.

6.6.8 Add 200 µl of the preheated 55°C water to the center of the column. Place the column in incubator set at 55°C for 10 min.

6.6.9 Centrifuge for 1.5 min. Discard the flow through. There should now be ~200 µl of purified aRNA.

6.6.10 Quantify using either a Nanodrop spectrophotometer or RiboGreen-based fluorescence detection. This is a potential stopping point. Store at -80°C.

**7. cDNA synthesis.**

Single stranded RNA is converted to cDNA via the Universal RiboClone cDNA Synthesis System using random primers. We typically use 10 µg of RNA in the cDNA synthesis to obtain a final mass of ~5-8 µg cDNA. The amount of aRNA used can be varied depending on the particular requirements for sequencing. In the protocol below the steps are described without reagent volumes, as they depend on the amount of input RNA used. For example, a 10 µg amount requires scaling the reagents in each step up by 5X. The appropriate volumes can be found in the kit manual. Pay close attention to the kit reagent concentrations as they are apt to change between lots. We have found it easiest to conduct the second strand synthesis in a refrigerated incubator (cooled to 14°C at least an hour before using), as it does not require splitting up a single sample into multiple 0.5 ml tubes. However, if a reliable refrigerated incubator cannot be found, a thermal cycler can be used. During the cleanup, the cDNA can be eluted in either nuclease free water or TE depending on the downstream requirements of sequencing.

Materials:

Universal RiboClone cDNA Synthesis System (Promega, Madison, WI)
Incubator or hybridization oven
Refrigerated incubator or thermal cycler
0.1 mM nuclease-free EDTA (Applied Biosystems, Austin, TX)
QiaQuick PCR cleanup kit (Qiagen, Valencia, CA)
Vacuum manifold
Centrifuge

Method:

7.1 First strand synthesis

7.1.1 Aliquot out the volume of sample needed for 10 µg aRNA. If the volume is < 65 µl, bring up to 65 µl with nuclease-free water. If the aRNA concentration is > 65 µl, concentrate either via speed vacuum or ethanol precipitation.

7.1.2 Add random primers to the aRNA. Gently mix and centrifuge.

7.1.3 In a preheated thermal cycler, denature the RNA-primer mixture at 70°C for 10 min. Immediately after, place the tubes on ice for 5 min.

7.1.4 Transfer the mixture to a 1.5 ml tube. This size is necessary to account for the increase in volume in the coming steps.

7.1.5 Add first strand 5X buffer and RNasin ribonuclease inhibitor. Gently mix the reaction and briefly centrifuge.

7.1.6 Incubate mixture in an incubator at 37°C for 5 min.

7.1.7 Add sodium pyrophosphate, AMV reverse transcriptase, and nuclease-free water. Gently mix and centrifuge.

7.1.8 Incubate mixture in incubator at 37°C for 1 h. Afterward, place on ice and proceed directly to second strand synthesis.

7.2 Second strand synthesis

7.2.1 On ice, add the following components to the first strand reaction: second strand 5X buffer, BSA, DNA polymerase, RNase H, and nuclease-free water. Gently mix and centrifuge.

7.2.1 Incubate at 14°C for 2 h.

7.2.2 Remove the second strand reaction from the incubator or thermal cycler and add T4 DNA polymerase.

7.2.3 Return to the incubator or thermal cycler set at 14°C and incubate for another 10 min. The temperature in this step deviates from the kit protocol.

7.2.4 Add 10 µl of 0.1 mM EDTA per µg input RNA to stop the reaction and place the mixture on ice.

7.3 QiaQuick Clean up

The volumes below are based on a cDNA synthesis of 10 µg, for which the final volume in 7.2.4 is 550 µl.

7.3.1 To increase elution efficiency and reduce guanidinium thiocyanate carry over, warm the PE buffer to 37°C at least 2 h before using.

7.3.2 Divide the sample into two 275 µl aliquots placed in 2 ml tubes.

7.3.3 Add 688 µl PB buffer and mix thoroughly by vortexing. The mixture should be yellow. If orange or violet, the pH is not correct and will need to be adjusted (see kit manual).

7.3.4 Place a mini column on the vacuum manifold and start the vacuum. Pipette the mixture from both tubes onto the column until the entire volume has passed through. Remove suction.

7.3.5 Remove the PE buffer from the 37°C incubator and add 750 µl PE to the column. Restore the vacuum until the buffer has passed through. Repeat the wash with another 750 µl PE.

7.3.6 Remove the cartridge from the manifold and place in a collection tube. Centrifuge at 10,000 rpm for 2 min to remove any residual wash solution. Transfer to a clean 1.5 ml tube.

7.3.7 Add 50 µl of nuclease-free water or TE buffer (see note above) and let stand at room temperature for 1 min. Centrifuge at 10,000 rpm for 2 min. Discard the cartridge. The cDNA is now ready for sequencing.

**8. Bioinformatics analysis**

Processing of the resulting sequence reads involves quality trimming, internal standard quantification, residual rRNA identification and removal, and finally functional annotation of the protein encoding reads. Several of these processing steps can be carried out using platforms freely available through CAMERA (http://camera.calit2.net) or MG-RAST (http://metagenomics.anl.gov).

Next generation reads can produce both systematic and random sequencing errors specific to the platform used. A quality metric (such as Phred) should be used to identify and remove low quality regions or entire sequences. The number of internal standards recovered by the sequencing should be quantified by a BLASTn homology search and removed from further processing. Inevitably, some rRNAs will escape the rRNA reduction process and need to be removed to prevent misleading functional annotations of these sequences. A BLASTn homology search against the SILVA large and small rRNA subunit database (www.arb-silva.de) can be used to identify bacterial, archaeal, and eukaryotic rRNA sequences. Once identified, these sequences should be removed from further processing. Finally, potential protein-encoding reads can be annotated based on homology to databases that span a wide range of functional

resolution from broad functions (COGs) to strain specific proteins (RefSeq). The calculations for total transcript pool size and individual transcript abundance are calculated as follows:

$$P_a = \frac{P_s \times S_a}{S_s} \qquad T_a = \frac{T_s \times P_a}{P_s}$$

$P_a$ = total transcripts in the sample

$P_s$ = potential protein encoding sequences (total number of sequences – rRNA sequences – $S_s$)

$S_a$ = molecules of internal standard added to the sample

$S_s$ = internal standard sequences

$T_a$ = molecules of any particular transcript category in the sample. This can then be divided by the mass or volume of sample collected to calculate the transcript abundance on a per environmental unit basis.

**References**

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al.* (2008). Microbial community gene expression in ocean surface waters. *PNAS* 105: 3805-3810.

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

He S, Wurtzel O, Singh K, Froula JL,Yilmaz S, Tringe SG, Wang Z, et al. (2010) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods* 7:807-812

Hewson I, Poretsky RS, Dyhrman ST, Zielinski B, White AE, Tripp HJ *et al.* (2009b). Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J* 3: 1286-1300.

Poretsky RS, Bano N, Buchan A, LeCleir G, Kleikemper J, Pickering M *et al.* (2005). Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol* 71: 4121-4126.

Poretsky RS, Hewson I, Sun SL, Allen AE, Zehr JP, Moran MA (2009b). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* 11: 1358-1375.

Poretsky RS, Sun S, Mou X, Moran MA (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* 12: 616-627.

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

Stewart FJ, Ottesen EA, DeLong EF (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J* 4: 896-907.

Urich TA, Lanzen J, Qi DH, Huson DH, Schleper C, Schuster SC (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* 3.

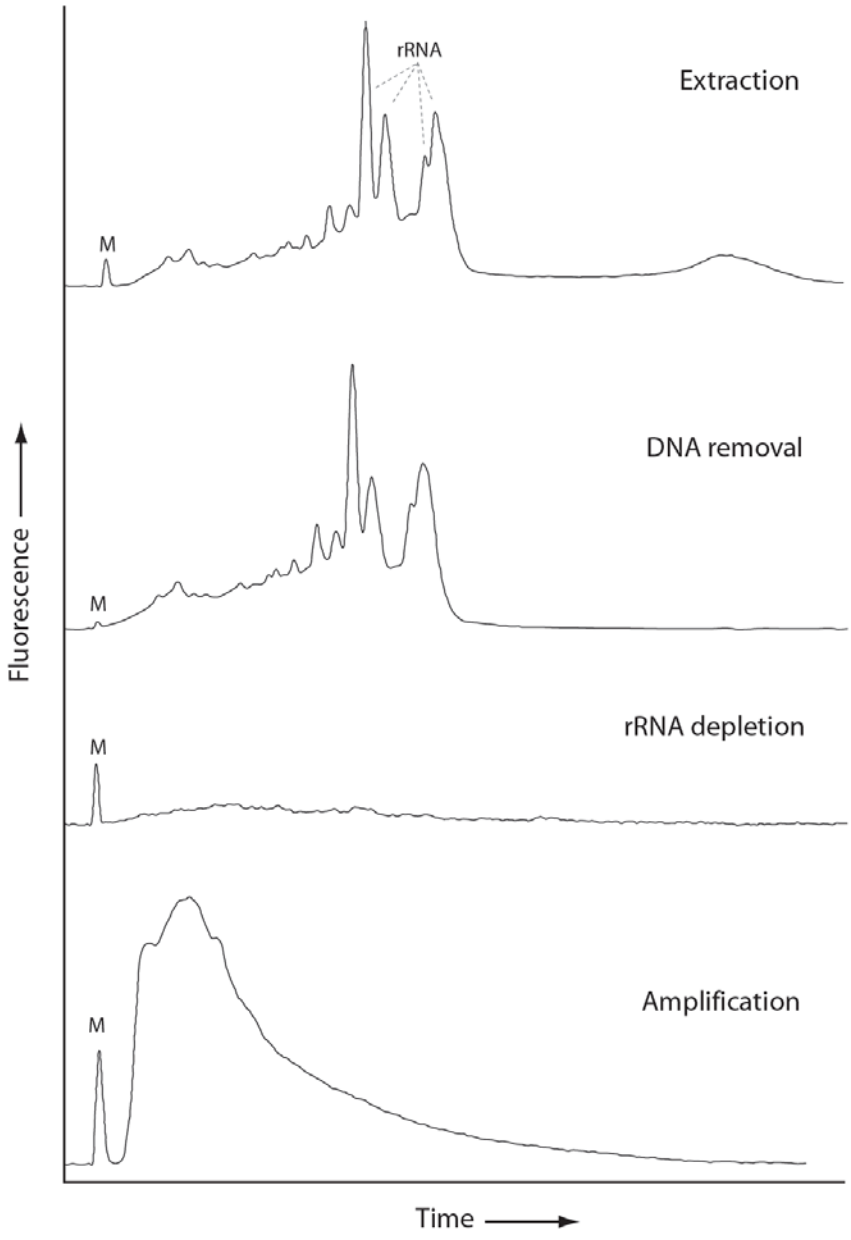



**Figure A.1**. Size distributions of sample RNA at different stages of processing visualized with the Experion automated gel electrophoresis system. Fragment length and abundance are proportional to run time and fluorescence, respectively. The gel marker is labeled 'M". The distinct rRNA peaks dominant the total RNA pool in the extracted and TurboDNase treated samples, but are greatly diminished after subtractive hybridization.

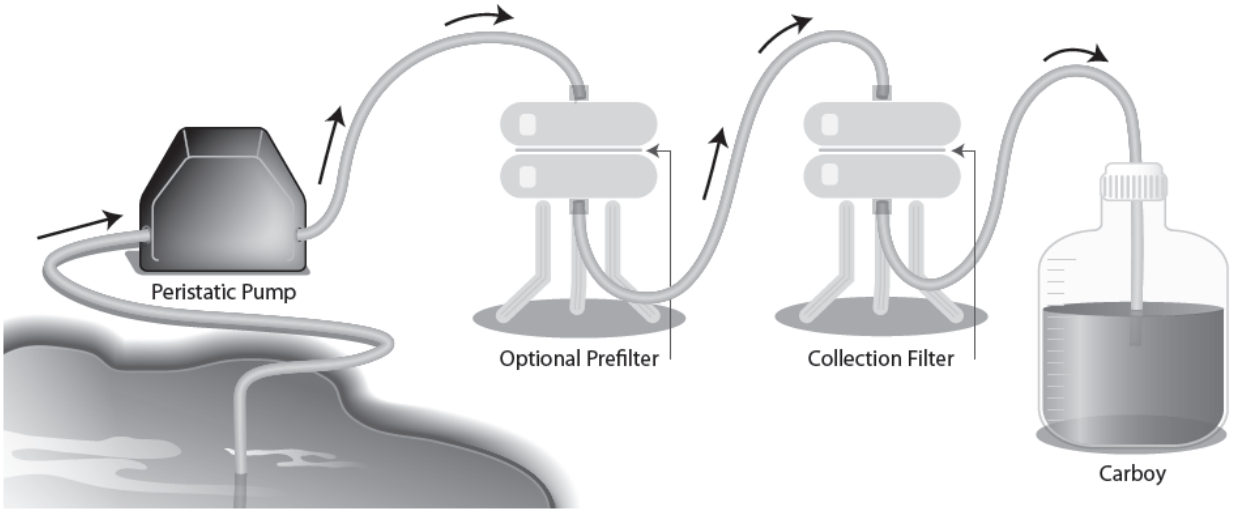



**Figure A.2**. Filtration setup for direct cell collection from an aquatic environment.