MICROBIAL GENE EXPRESSION PATTERNS IN THE AMAZON RIVER AND PLUME

by

BRANDON MEYER SATINSKY

(Under the Direction of Mary Ann Moran)

ABSTRACT

The Amazon River is the world's largest river system and spans nearly 6,500 km across South America prior to discharging freshwater and nutrients into the ocean via a low salinity plume that covers 20% of the Western Tropical North Atlantic Ocean. Prokaryotes (bacteria and archaea) carry out critical ecological roles in both the marine and freshwater environments of the Amazon system, and govern critical aspects of energy production and consumption. This dissertation details improvements and expansion of meta-omics methodologies that enable studies of the composition and activity of microbial community, including protocols for the synthesis and use of internal standards. Using the improved metatranscriptomic and metagenomics method, three subsequent studies shed light on the roles of prokaryotes in the Amazon River and Plume. In one study, quantitative metatranscriptomics and metagenomics were used to generate the first fully quantitative inventories of microbial genes and transcripts in a natural ecosystem, detailing the patchiness found in the abundance and regulation of prokaryotic genes within a single water mass on the outer continental shelf. In another study, quantitative metatranscriptomics and metagenomics datasets generated from microbial communities in both free-living and particle-associated microenvironments at six Amazon Plume stations were used to detail transcriptional patterns of prokaryotic genes, finding that changes in

both gene regulation within a taxon and shifts in taxonomy among stations drive the variability in expression. Further, microbial cells associated with particulate material experience more chemically dynamic conditions spatially in the Amazon plume than do free-living cells. The final study addresses patterns of gene expression for freshwater prokaryotic communities in the lower Amazon River. High expression ratios of genes related to nitrogen cycling by Thaumarchaea taxa suggest an important role for chemoautotrophic archaea in river biogeochemistry.

INDEX WORDS: Amazon River, Amazon Plume, Ocean, Metatranscriptomics, Gene Expression, Gene Regulation, Microbial communities, RNA, Bacteria, Metagenomics, Biogeochemistry, Internal Standards

MICROBIAL GENE EXPRESSION PATTERNS IN THE AMAZON RIVER AND PLUME

by

BRANDON MEYER SATINSKY

B.S., Indiana University, 2008

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

2014

MICROBIAL GENE EXPRESSION PATTERNS IN THE AMAZON RIVER AND PLUME

by

BRANDON MEYER SATINSKY

Major Professor:	Mary Ann Moran
Committee:	Adrian B. Burd
	Jan Mrázek
	Eric V. Stabb

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
December 2014

DEDICATION

To my parents who made it all possible.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Mary Ann Moran. Her patience, encouragement, and relentless enthusiasm for her students and science know no bounds, and made this work possible and my time as a graduate student transformative and enjoyable. I could not have asked for a better mentor, and for her support and guidance I am forever grateful. I would like to thank the many members (past and present) of the Moran lab that have supported me during my time in graduate school and made coming to work everyday fun. I also thank my committee members, Dr. Adrian Burd, Dr. Eric Stabb, and Dr. Jan Mrazek who provided me with helpful advice and suggestions for developing my research. I am grateful to the many collaborators on the ANACONDAS and ROCA teams I had the pleasure of working with during my time at UGA, and whose input and assistance made this work possible. Finally, I would like to thank my friends and family for always believing in me.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

I. Microbes in the Marine Environment:

Prokaryotes in marine and freshwater environments play critical ecological roles that control fundamental aspects of energy production and consumption. The world's oceans (excluding sediments) and rivers are estimated to house a combined 1.01×10^{29} cells (Whitman *et al.,* 1998), more than 380 thousand times the estimated number of human cells on Earth. These prokaryotic cells are composed of both autotrophs and heterotrophs, with autotrophic fixation of carbon important at the global scale, while heterotrophic transformation and sequestration of organic matter determines the fate of CO_2 fixed by the primary producers, as well as the regeneration of inorganic N and P. Previous studies have identified microorganisms as the major consumers of the energy made available by primary production, showing that the smallest organisms within a community are the largest energy consumers due to their numbers and high metabolic rates (Zeuthen 1947).

In marine systems, biogeochemical transformations mediated by microorganisms take place within a poorly characterized matrix of particles, colloids, and dissolved phase materials (Azam and Malfatti 2007). Driven by variations in nutrient concentrations, light exposure, oxygen availability, and predation (Kiørboe and Jackson 2001, Stocker *et al.,* 2008), this fine-scale environmental structure impacts the taxonomy and functional gene inventories of microbial communities (Ganesh *et al.,* 2014, Smith *et al.,* 2013), and undoubtedly influences the types and rates of biogeochemical processes occurring as well. In the marine water column, the fixation of

inorganic carbon to organic matter by phytoplankton results in the flux of CO_2 from the atmosphere to the ocean. This fixed carbon may then enter the biological pump and subsequently be stored either temporarily or permanently in the deep ocean (Musat *et al.,* 2008). The biological pump works by transporting particulate organic material (POM) and dissolved organic material (DOM) originating from primary producers to deeper waters through multiple mechanisms including, but not limited to: the sinking of phytoplankton aggregates; the feeding of larger zooplankton on phytoplankton resulting in downward flux of fecal pellets; and the diel vertical migration of marine organisms that feed in surface waters but live at depth. Up in the surface waters, both POM and DOM are vulnerable to remineralization by heterotrophic bacteria that transform dissolved organic carbon (DOC) into dissolved inorganic carbon (DIC), further fueling primary production and the biological pump (Azam *et al.,* 1983, Pomeroy 1974). However, storage of carbon in the deep ocean is often temporary due to wind-driven mixing of water layers, and thermohaline circulation, both of which can carry water and carbon from depths up to the surface, and due to vertical migration of phytoplankton and zooplankton from deeper waters to the surface (Musat *et al.,* 2008). While storage in the ocean may not be permanent for most exported carbon, a fraction escapes remineralization and mixing, reaching the ocean sediment where it is stored over geological time periods in a process known as 'carbon sequestration'.

Bacteria are responsible for the bulk of primary production in the ocean, despite large contributions from eukaryotic phytoplankton in coastal and upwelling regions (Chisholm *et al.,* 1988). They are also responsible for the bulk of fixed carbon consumption (Azam 1998). However, many of the important processes underlying these carbon cycle links have been challenging to characterize because of the hundreds of uncharacterized bacterial species involved

and the diversity of links in the elemental cycles they mediate. Further, bacterial taxa can harbor the genetic capability for multiple biogeochemical roles, but expression of these roles is based on environmental conditions at a given time and place. Given the abundance and impact on global ecology of microbial communities in aquatic environments, understanding how these communities respond to a changing environment is critically important for predicting the impact of global change. Historically, our understanding of marine prokaryotes has been based on bulk measurements of heterogeneous and complex communities that obscure our ability to distinguish individual taxa and their specific functional responses to environmental conditions. More recently, significant advances in molecular biology tools have tremendously expanded our abilities to delineate the diversity and functional roles of microbial communities in the environment.

II. Metatranscriptomics as an Oceanographic Tool

Metagenomics is a methodological approach that enables characterization of the functional potential and taxonomic composition of microbial communities in the environment through sequence analysis of community genomes (Rusch *et al.,* 2007, Venter *et al.,* 2004). Metatranscriptomics is a complementary approach that measures gene expression in the environment through analysis of community transcripts (Poretsky *et al.,* 2005). Metagenomic- and metatranscriptomic-based studies have been conducted successfully in marine, freshwater, soil, gut, and other natural microbial systems (Damon, Vallon, Zimmermann, Haider, Galeote, Dequin *et al*., 2011; Dinsdale, Edwards, Hall, Angly, Breitbart, Brulc *et al*., 2008; Gifford, Sharma, Rinta-Kanto & Moran, 2011; Maurice, Haiser & Turnbaugh, 2013; Poretsky, Bano, Buchan, LeCleir, Kleikemper, Pickering *et al*., 2005; Vila-Costa, Sharma, Moran & Casamayor,

2013; Ottesen, Young, Eppley, Ryan, Chavez, Scholin *et al*., 2013), generating detailed information on community-level gene abundance and transcription patterns. In recent years, -omics-based methods have rapidly advanced by adopting high-throughput sequencing methodologies and more efficient removal of rRNA (Stewart *et al.,* 2010), resulting in a more in-depth inventory of biogeochemically relevant microbial genes and transcripts.

Most studies to date have collected meta-omics data in a relative framework (i.e., % of metagenome and % of metatranscriptome) (Campbell, Yu, Heidelberg & Kirchman, 2011; Hewson, Poretsky, Beinart, White, Shi, Bench *et al*., 2009). However, a critical limitation of relative meta-omics data from complex natural communities is that they cannot provide information on the extent or directionality of changes in any particular gene or transcript in comparative analyses. This drawback is particularly problematic for dynamic communities because a change in the abundance of one type of gene or transcript imposes a change in the percent contribution of the others, even if their actual expression levels have not changed. In the application of meta-omics technologies to ecological and biogeochemical questions in complex microbial communities, the ability to recognize which genes and transcripts are changing in absolute abundance is crucial information, requiring datasets that are not influenced by the myriad non-target processes that may be fluctuating simultaneously in a microbial cell or ecosystem.

The fine-scale information about metabolic activities of individual microbial taxa provided by -omics-based datasets is highly complementary to traditional biological and chemical process measurements. A common goal of the microbial ecology community is to integrate both into an ecosystems-level view of elemental cycling in the environment.

III. The Amazon Continuum:

The Amazon River runs nearly 6,500 km across the South American continent before emptying into the Western Tropical North Atlantic Ocean. In terms of both volume and watershed area it is the world's largest riverine system (Coles *et al.,* 2013). The Amazon basin plays a central role in global nutrient cycling, and the rainforest surrounding the river is responsible for nearly 10% of global primary production (Field *et al.,* 1998), fixing 8.5 Pg C per year (Malhi *et al.,* 2008), much of which ultimately ends up in the Amazon River. Within the river, heterotrophic bacteria rely heavily on the allochthonous input of carbon and nutrients from the surrounding rainforest and drainage basins. The organic material in the river contains a substantial quantity of both humic and fulvic acids, which together account for ~60% of riverine DOC, and are likely derived from lignin and other plant components (Ertel *et al.,* 1986). The main channel of the river is well mixed and highly turbid, and high levels of CO_2 as well as low light penetration into the water suggest an environment dominated by heterotrophic bacteria that remove, transform, and stabilize riverine organic matter during transit. These activities lead to a river outgassing of 0.5 Pg C per year (Richey *et al.,* 2002). Unlike the main channel, many of the tributaries along the river have much lower turbidity levels and may provide conditions suitable to the growth of photosynthetic prokaryotes. In addition to affecting the DOC pool, the prokaryotic community also plays an active role in the processing and cycling of nitrogen, phosphorus, and iron. However, the diversity and metabolic activity of prokaryotic communities within the Amazon River has not been well characterized.

At its mouth, the Amazon discharges water at a rate approximately 12 times that of the Mississippi River, carrying terrestrially-derived nutrients hundreds of miles offshore in a low salinity plume covering 20% of the Western Tropical North Atlantic Ocean (Richey *et al.,* 1989,

Subramaniam *et al.,* 2008). The discharge by the Amazon River accounts for 18% of the world's river input into the oceans (Richey *et al.,* 1989, Subramaniam *et al.,* 2008), and although relatively dilute compared to other rivers (Ryther *et al.,* 1967), this mixture of dissolved and particulate nitrogen, phosphate, silica, and iron that is delivered to the ocean stimulates marine microbial activity and affects both primary productivity and carbon sequestration (Subramaniam *et al.,* 2008). The fluvial export from the Amazon River amounts to 22.3 Tg yr^{-1} of DOC and 13.7 Tg yr^{-1} of POC (Richey *et al.,* 1990) and equals that of the next 8 largest rivers of the world combined (Coles *et al.,* 2013).

The community structure of the Amazon River Plume is strongly influenced by the mixing of riverine dissolved (DOM) and particulate (POM) organic material into tropical Atlantic waters. Phytoplankton take advantage of the riverine nutrient supplements, making this a hot spot for carbon sequestration via sinking cells and particles. In lower-salinity coastal waters, the phytoplankton are dominated by coastal diatoms that are supported by river-supplied inorganic nitrogen, including *Skeletonema costatum* and *Pseudonitzschia sp.* (Simon *et al.,* 2009). As turbidity and salinity decrease further out in the plume, diatoms containing the endosymbiotic diazotrophic cyanobacterium *Richelia intracellularis* (species such as *Rhizosolenia* and *Hemiaulus*) begin to thrive, fueled by the riverine input of phosphorus and iron (Goes *et al.,* 2014, Subramaniam *et al.,* 2008). Prokaryotic phytoplankton are also numerically important autotrophs in the plume (Goes *et al.,* 2014, Zehr *et al.,* 2001), and these include unicellular cyanobacteria *Synechococcus, Prochlorococcus,* and, in oceanic regions where phosphorus and silicate are limiting, N_2-fixing *Trichodesmium*. The high productivity of these autotrophic communities generates a region of the Atlantic that takes up an excess of ~15 Tg C yr^{-1} (Cooley *et al.,* 2007). Heterotrophic bacteria can remineralize organic nutrients in the plume,

further fueling primary production and increasing the flux of organic material into the deep ocean. On the benthic floor below the plume, organic carbon exported from surface waters leaves a trail as far as 1,200 km from the river mouth (Chong *et al.,* 2014), and overall the activity of these microbial communities may be responsible for the sequestration of ~28 Tg C yr^{-1} (Subramaniam *et al.,* 2008).

III. Objectives

In Chapter 2, the development of new methodologies for the use of internal standards in metagenomics and metatranscriptomics is described, expanding upon the approach first introduced by Gifford et al. (Gifford *et al.,* 2011). The chapter describes in detail the strategies to consider in the design and implementation of internal standards in –omics-based methods, and provides a detailed step-by-step protocol for the synthesis and use of the internal standard methodology. In addition, the chapter discusses the normalization of –omics-based data based on internal standard recoveries, and the advantages as well as the potential caveats associated with the use of internal standards.

In Chapter 3, metagenomic and metatranscriptomic datasets enhanced by the use of internal standards were used to assemble the first fully quantitative inventories of microbial genes and transcripts in a natural ecosystem. A highly-resolved view resulted of the gene expression driving carbon and nutrient flux through free-living and particle-associated microbes at a low-salinity, outer continental shelf site in the Amazon Plume. The level of transcription for the same gene in each microenvironment was compared; transcripts mediating key carbon and nutrient transformations were enumerated; and biogeochemical roles of transcriptionally active free-living and particle-associated cells were predicted. In addition, comparisons were made of

the importance of shifts in cell abundance versus shifts in gene regulation in determining differences in the transcript abundance of specific genes at the microspatial scale.

In Chapter 4, six Amazon Plume stations sampled across a salinity gradient were inventoried in May-June, 2010 for gene and transcript abundances of dominant microorganisms and their biogeochemically-relevant functional genes. This detailed metagenomic and metatranscriptomic dataset focused on heterotrophic activities of Amazon plume bacteria and archaea by measuring transcripts and genes encoding proteins involved in nutrient cycling, and quantitatively assessing variations in gene expression relevant to carbon, nitrogen, and phosphorus cycling across the plume. When considering what factors explain variable microbial gene expression patterns in the Amazon plume, the effects of gene regulation and taxonomic shifts on the observed differences in gene expression were analyzed. Environmental drivers in the plume were explored, and patterns in gene expression that point to environmental conditions most relevant to the activities of bacteria and archaea in the plume were investigated.

In Chapter 5, metagenomic and metatranscriptomic approaches were used to generate the first quantitative molecular inventory of freshwater microbial communities, highlighting expression patterns of environmental cycling genes at five stations in the lower reaches of the Amazon River in May 2011. This study focused on patterns in transcript abundance and expression levels of 90 biogeochemically-relevant genes covering a broad range of ecological processes. Individual populations of transcriptionally dominant Thaumarchaea were tracked through the river to investigate gene regulation patterns of individual taxa along the river and gain insights into the different environmental signals these populations respond to. To contextualize the findings in this study more broadly, transcript abundance and expression levels were compared to those in the adjoining marine Amazon plume.

References:

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.

Azam F, Malfatti F (2007). Microbial structuring of marine ecosystems. *Nat Rev Microbiol* **5:** 782-791.

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.

Chong LS, Berelson WM, McManus J, Hammond DE, Rollins NE, Yager PL (2014). Carbon and biogenic silica export influenced by the Amazon River Plume: Patterns of remineralization in deep-sea sediments. *Deep Sea Res Part 1 Oceanogr Res Pap* **85:** 124-137.

Coles VJ, Brooks MT, Hopkins J, Stukel MR, Yager PL, Hood RR (2013). The pathways and properties of the Amazon River Plume in the tropical North Atlantic Ocean. *J Geophys Res-Oceans* **118:** 6894-6913.

Cooley SR, Coles VJ, Subramaniam A, Yager PL (2007). Seasonal variations in the Amazon plume-related atmospheric carbon sink. *Global Biogeochem Cy* **21**.

Ertel JR, Hedges JI, Devol AH, Richey JE, Ribeiro MDG (1986). Dissolved Humic Substances of the Amazon River System. *Limnol Oceanogr* **31:** 739-754.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281:** 237-240.

Ganesh S, Parris DJ, DeLong EF, Stewart FJ (2014). Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J* **8:** 187-211.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* **5:** 461-472.

Goes JI, Gomes HdR, Chekalyuk AM, Carpenter EJ, Montoya JP, Coles VJ *et al.* (2014). Influence of the Amazon River discharge on the biogeography of phytoplankton communities in the western tropical north Atlantic. *Prog Oceanogr* **120:** 29-40.

Kiørboe T, Jackson GA (2001). Marine snow, organic solute plumes, and optimal chemosensory behavior of bacteria. *Limnol Oceanogr* **46:** 1309-1318.

Malhi Y, Roberts JT, Betts RA, Killeen TJ, Li W, Nobre CA (2008). Climate change, deforestation, and the fate of the Amazon. *Science* **319:** 169-172.

Musat N, Halm H, Winterholler B, Hoppe P, Peduzzi S, Hillion F *et al.* (2008). A single-cell view on the ecophysiology of anaerobic phototrophic bacteria. *Proc Natl Acad Sci U S A* **105:** 17861-17866.

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 Microbiol* **71:** 4121-4126.

Richey JE, Nobre C, Deser C (1989). Amazon river discharge and climate variability: 1903 to 1985. *Science* **246:** 101-103.

Richey JE, Hedges JI, Devol AH, Quay PD, Victoria R, Martinelli L *et al.* (1990). Biogeochemistry of Carbon in the Amazon River. *Limnol Oceanogr* **35:** 352-371.

Richey JE, Melack JM, Aufdenkampe AK, Ballester VM, Hess LL (2002). Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO2. *Nature* **416:** 617-620.

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 Biol* **5:** e77.

Ryther JH, Menzel DW, Corwin N (1967). Influence of the Amazon River outflow on the ecology of the western tropical Atlantic. *J Mar Res***:** 69-83.

Simon N, Cras AL, Foulon E, Lemee R (2009). Diversity and evolution of marine phytoplankton. *C R Biol* **332:** 159-170.

Smith MW, Zeigler Allen L, Allen AE, Herfort L, Simon HM (2013). Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. *Front Microbiol* **4**.

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.

Stocker R, Seymour JR, Samadani A, Hunt DE, Polz MF (2008). Rapid chemotactic response enables marine bacteria to exploit ephemeral microscale nutrient patches. *Proc Natl Acad Sci U S A* **105:** 4209-4214.

Subramaniam A, Yager PL, Carpenter EJ, Mahaffey C, Bjorkman K, Cooley S *et al.* (2008). Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci U S A* **105:** 10460-10465.

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.

Whitman WB, Coleman DC, Wiebe WJ (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95:** 6578-6583.

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

Zeuthen E (1947). *Body Size and Metabolic Rate in the Animal Kingdom with Special Regard to the Marine Microfauna*. H. Hagerup.

CHAPTER 2

USE OF INTERNAL STANDARDS FOR QUANTITATIVE METATRANSCRIPTOME AND METAGENOME ANALYSIS[1]

[1] Satinsky BM, Gifford SM, Crump BC, Moran MA (2013). Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. In: DeLong EF (ed). *Methods Enzymol*. Academic Press. pp 237-250; Reprinted here with permission of publisher.

Abstract

Next generation sequencing-enabled metatranscriptomic and metagenomic datasets are providing unprecedented insights into the functional diversity of microbial communities, allowing detection of the genes present in a community as well as differentiation of those being actively transcribed. An emerging challenge of meta–omics approaches is how to quantitatively compare metagenomes and metatranscriptomes collected across spatial and temporal scales, or among treatments in experimental manipulations. Here we describe the use of internal DNA and mRNA standards in meta–omics methodologies, and highlight how data collected in an absolute framework (per L or per cell) provides increased comparative power and insight into underlying causes of differences between samples.

1. Introduction

Metagenomic and metatranscriptomic methodologies have been used with great success in generating detailed information on community-level gene abundance and transcription patterns in marine, freshwater, soil, gut, and other natural microbial systems (Damon *et al*., 2011; Dinsdale *et al*., 2008; Gifford *et al*., 2011; Maurice *et al*., 2013; Poretsky *et al*., 2005; Vila-Costa *et al*., 2013; Ottesen *et al*., 2013). Most studies to date have collected meta-omics data in a relative framework, in which abundance of genes or messages is calculated as percent of the sequence library (Campbell *et al*., 2011; Hewson *et al*., 2009). However, a critical limitation of relative meta-omics data from complex natural communities is that they cannot provide information on the extent or directionality of changes in any particular gene or transcript molecule in comparative analyses. For instance, an observed decrease in the percent contribution of a transcript to the community metatranscriptome may be due to a decrease in the abundance of

that transcript or to an increase in the abundance of an unrelated transcript (Fig. 2.1). In the application of meta-omics technologies to ecological and biogeochemical questions in complex microbial communities, the ability to recognize which genes and transcript molecules are changing in absolute abundance is crucial information, requiring datasets that are not influenced by the myriad non-target processes and taxa changing simultaneously in a microbial cell or ecosystem.

To circumvent the limitations of relative metagenomic and metatranscriptomic datasets, internal genomic DNA or mRNA standards can be added at the initiation of sample processing (Gifford *et al*., 2011; Moran *et al*., 2013). Because these control molecules are mixed into and processed alongside the sample-derived nucleic acids, this allows quantification of losses throughout the preparation and analysis pipeline and, based on the number of standard molecules added at the beginning of sample processing and those recovered in the sequence library, calculation of the number of molecules of each gene or transcript in the original environment (e.g., gene copies per liter of water, or average transcripts per microbial cell). Internal standards based on a known quantity of added control molecules are used routinely in quantitative PCR studies for calculating absolute gene and transcript abundance (Church *et al*., 2005) and in microarray and RNA-seq analyses to normalize expression shifts in genes across different developmental stages or tissue types (Hannah *et al*., 2008; van de Peppel *et al*., 2003).

The benefits of quantitative meta-omics datasets can be illustrated by the following two examples. In the first, sequences binning to SAR11 member HTCC7211 in the bathypelagic waters of the Gulf of Mexico accounted for 0.84% of the bacterial metatranscriptome in natural seawater but only 0.07% in seawater exposed to oil and gas contamination from the Deepwater Horizon accident, indicating a 12-fold underrepresentation of HTCC7211 following the accident.

Yet absolute transcript numbers for this taxon calculated based on internal standard normalization revealed that transcripts were present in equal numbers in impacted and non-impacted seawater (2.8×10^{11} and 3.4×10^{11} transcripts L^{-1}) (Fig. 2.1), and that the change in percent contribution of HTCC7211 populations was due to large increases in gammaproteobacteria groups that bloomed in response to hydrocarbon inputs (Fig. 2.1) (Rivers *et al*., 2013). In a second example, expression ratios for proteorhodopsin genes in the near-shore Amazon River plume were nearly identical for the free-living and particle-associated bacteria when calculated on a relative basis (% of the metatranscriptome/% of the metagenome = ~8 for both free-living and particle-associated; Fig. 2.1). Yet on an absolute basis, the per-gene transcription level of proteorhodopsin was 2-fold higher for bacteria associated with particulate material compared to free-living cells in this ecosystem (Fig. 2.1). In these examples, normalization based on internal standard recovery provided insights into growth and regulation differences for bacteria in their natural environment, information that can be leveraged in comparative analyses across samples (for example within a time series, across a transect, or during a manipulative experiment) (Gifford *et al*., 2011; Moran *et al*., 2013) (Fig. 2.1).

To generate quantitative -omics data, internal control sequences must be readily distinguished from natural microbial community sequences during bioinformatic analyses. For metatranscriptomes, artificial mRNAs produced by *in vitro* transcription from constructed DNA templates can be used as internal standards, and preparation of mRNA standards with or without a poly(A) tail customizes them for bacterial/archaeal or eukaryotic studies. For metagenomes, genomic DNA obtained from a cultured microorganism not present in the studied environment can be added as an internal standard. In our marine and estuarine studies, DNA from the thermophilic bacterium *Thermus thermophilus* (ATCC) has served as the standard.

Calculations based on the internal standards assume that the natural nucleic acid (mRNA or genomic DNA) and the internal standards (artificial transcripts or exogenous genomic DNA) behave similarly throughout the sample and library preparation steps. However, the natural nucleic acid is enclosed in cell membranes at the initiation of processing while the internal standards are not, potentially resulting in underestimation of natural nucleic acid abundance due to incomplete cell lysis, or alternatively, underestimation of standard abundance due to longer exposure time to mechanical shearing or RNAse degradation. In the case of mRNA processing, transcript length can affect recovery because of biases against small transcripts during solid phase extraction methods and library preparation (Fig. 2.2), although this will affect both artificial and natural transcripts alike. Given an average bacterial and archaeal gene size of 924 bp (Xu *et al*., 2006), we use an internal standard of ~1000 nt to track recovery of mRNAs from typical prokaryotic genes. Internal standard length can be scaled downward if small RNAs or short transcripts are the focus of the study; scaling upward from 1000 nt does not appear to be necessary because of minimal effect on recovery for lengths >500 nt (Fig. 2.2).

2. Method Overview

The method below describes the synthesis of internal mRNA standards and then the addition and quantification of mRNA and DNA internal standards for metatranscriptome and metagenome analysis.

mRNA standards are synthesized using custom templates or commercially available plasmids that are transcribed *in vitro* to RNA. A known number of standards is added to the sample of interest and metatranscriptome processing and sequencing proceeds according to the user's protocol. The number of internal standards recovered in the sequence library is quantified via

BLAST homology searches. Data normalization is then based on the number of standards identified in a sequence library relative to the number of standards added. As a note of caution for working with RNA, care should be taken to avoid all contaminating nucleic acids and nucleases through the use of sterile technique and cleaning the working area with RNase*Zap*® or a similar reagent.

DNA standards can be prepared by purchasing or extracting DNA from a cultured microbe that is unrelated to microbes anticipated to be present in the system of interest and for which a complete genome sequence is available. A known number of genome copies is added to the sample, and metagenome processing and sequencing proceeds according to the user's protocol. The number of standard reads recovered in the sequence library is quantified via a two-step BLAST homology search and used for quantitative metagenomic analysis.

3. DNA template and Vector Design for Internal RNA Standards

Two approaches are available for obtaining the DNA template for standard synthesis. One approach involves commercially available plasmids that contain an RNA polymerase binding site. These are advantageous because of ease of use and low cost (Gifford *et al.,* 2011; Moran *et al.,* 2013), although the vectors make transcript length customization more difficult and they often contain regions of homology to functional proteins or to sequences deposited mistakenly into databases as functional proteins. This homology can make the subsequent identification of reads derived from standards more challenging in a high-throughput bioinformatics pipeline. A second approach involves the synthesis of custom DNA fragments that are inserted into plasmids. These fragments can easily be designed without homology to protein encoding genes, and provide optimal control of both length and composition.

For both template approaches, the final plasmid should contain the following components (in order): a T7 RNA polymerase promoter sequence, the internal standard sequence, and a restriction site targeting a unique site in the plasmid and preferably producing a blunt end (Fig. 2.3). For poly-A selective transcriptomes, a poly-A tail can be included in custom synthesized templates between the RNA polymerase promoter and the internal standard sequence. Whether using commercially available plasmids or custom synthesized internal standard templates, sequences should first be analyzed against relevant databases to identify regions of homology that could interfere with unambiguous identification of the standard in the sequence library.

Template size is also an important consideration because downstream processing steps during RNA processing and library preparation can lead to biases in the size of transcripts recovered. Based on addition of the six standards shown in Fig. 2.2 (representing two variations in base composition for each of three sizes: 200 nt, 500 nt, and 1000 nt), recovery efficiency in the sequence library was several orders of magnitude lower for the 200 nt mRNA standards compared to the others (Fig. 2.2). However, the duplicate standards at each size were recovered with nearly identical efficiencies, indicating that base composition is not an important factor in standard recovery. For the sequence data represented in Fig. 2.2, steps in the RNA isolation, purification, and amplification relied on solid phase extraction, while Illumina library preparation included cDNA shearing and size selection (225 bp target size), all of which could lead to size bias for both artificial and natural mRNAs. Other extraction and library preparation methods may result in different size biases, but it is not straightforward to correct for size biases since transcript length depends on operon structure rather than individual gene length. Nonetheless, an internal standard can be selected that approximates the average size of the

natural nucleic acid molecules being targeted (i.e., genomic DNA standards for metagenomes and artificial mRNAs of typical gene length for metatranscriptomes).

4. mRNA Standard Preparation

4.1 Required Materials

- *Equipment*: 4° C Micro-centrifuge, 10-, 20-, 200-, and 1000-μL pipettes, water bath, 37° C shaking incubator, thermocycler, gel electrophoresis equipment and reagents, microfluidic electrophoresis instrument or fluorometry-based instrument for measuring nucleic acid concentration.
- *Media*: LB agar, LB agar + ampicillin (100μg/mL final concentration), LB medium, LB medium + ampicillin (100μg/mL final concentration), S.O.C. medium (2% tryptone, 0.5% yeast extract, 10 mM sodium chloride, 2.5 mM potassium chloride, 10 mM magnesium chloride, 10 mM magnesium sulfate, 20 mM glucose)
- *Bacterial Cell Line*: One Shot® Top10 Chemically Competent *E. coli* (Life Technologies, Grand Island, NY)
- *Template DNA*: Custom synthesized DNA template (T7 RNA polymerase promoter, internal standard sequence, unique restriction site) inserted into a plasmid
- *Restriction Digest and End Repair:* Restriction enzyme matching unique restriction site and corresponding buffers, Mung Bean Nuclease for end repair on digests that do not produce blunt ends.
- *Commercially Available Kits*: Ambion MEGAscript® T7 Kit (Life Technologies), Quant-iT™ RiboGreen® RNA Assay Kit (Life Technologies), miniPrep plasmid extraction kit

- *Other Reagents*: phenol:chloroform:isoamyl alcohol (24:24:1)(pH=~7), citrate-saturated phenol:chloroform:isoamyl alcohol (24:24:1)(pH=4.7), sterile 2-propanol, ice cold 70% ethanol, nuclease-free 3 M sodium acetate, nuclease-free TE buffer, sterilized 100% glycerol; 1% agarose gel, nuclease-free water, RNase*Zap*® (Life Technologies)
- *Disposables*: nuclease-free 10-, 20-, 200-, and 1000-μL filter tips, nuclease-free PCR tubes, nuclease-free micro-centrifuge tubes, gloves

4.2 Plasmid amplification and stock preparation

4.2.1 Resuspension of plasmid DNA

If beginning with lyophilized plasmid DNA, spin briefly to ensure the contents are on the bottom of the tube. Resuspend the plasmid DNA in a volume of TE buffer to produce a stock concentration of 0.1 μg/μL. To prepare a working solution, add 1 μL of the stock solution to 99 μL of nuclease-free water to produce a final concentration of 1 ng/μL. The resuspended plasmid DNA can be stored at -20° C.

4.2.2 Chemical transformation of plasmid into Top10 E. coli cells

Prior to beginning the transformation ensure that all required media is prepared and sterilized. Place frozen competent cells and a pre-labeled tube on ice. Pre-warm a hot water bath to 42° C. To a tube on ice, add 2 μL of (~2 ng) the plasmid working solution to 100 μL of thawed competent cells and flick the tube gently to mix. Incubate the mixture for 30 min on ice, then heat shock in the 42° C hot water bath for 45 s. Immediately place the tube on ice for 2 min and then add 500 μL of SOC or LB liquid medium to the tube and incubate at 37° C for 1 h with shaking (~225 rpm). During this time pre-warm LB-Amp agar plates in a 37° C incubator. From the tube, pipet and spread 10 μL, 100 μL, and 200 μL on three separate LB-Amp agar plates.

Place the plates upside down in a 37° C incubator for 12-24 h. Following incubation, inoculate a single, well-isolated colony from one of the plates and place into 5 mL LB-Amp media. Grow the liquid culture at 37° C for ~8 h with vigorous shaking (~300 rpm). Remove 850 μL of the starter culture and place into a 2 mL freezer vial with 150 μL of sterilized 100% glycerol, mix thoroughly, and store at -80° C. To work from the frozen stocks, place a loopful of stock into 10 mL of LB-Amp liquid medium and grow at 37° C with vigorous shaking (~300 rpm) for 12-16 h. Harvest cells by centrifugation at 6000 x *g* for 15 min at 4° C. Discard the supernatant and recover the plasmid DNA using a commercially available plasmid mini-prep kit.

4.3 Plasmid linearization and *in vitro* transcription

4.3.1 Linearization of plasmid template

Digest two μg of plasmid with restriction enzyme targeting the site at the end of the template sequence according to the restriction enzyme protocol. Sticky ends created by non blunt-end cutting enzymes should be removed using mung bean nuclease. After digestion and end repair, bring the reaction to 100 μL by adding TE buffer, add 100 μL of phenol:chloroform:isoamyl alcohol (25:24:1, ph = ~7), and mix by vortexing. Spin the mixture for 5 min at 12,000 x *g* in a microcentrifuge. Transfer the aqueous phase to a new tube and add 0.1 volumes (~ 10 μL) of 3M sodium acetate and 0.7 volumes (~ 70 μL) of isopropanol to the tube. Mix thoroughly and incubate for 10 min at room temperature, and centrifuge for 30 min at 12,000 x *g* at 4° C. Discard supernatant and wash pellet with 200 μL of ice cold 70% ethanol. Centrifuge for 5 min and discard the supernatant being careful not to disturb the pellet. Air-dry the pellet to remove residual ethanol before resuspending the pellet in 5 μL of nuclease-free water. Transfer 2 μL of linearized plasmid into a new tube and add 2 μL of nuclease-free water. Use 1 μL of the diluted sample to check the concentration and analyze the remaining 3 μL on a 1% agarose gel to check

for complete digestion and the presence of a single-sized product. Retain the 3 μL of undiluted DNA template for subsequent steps.

4.3.2 Synthesis and purification of mRNA internal standard

Synthesis of the internal standards from a template containing a T7 promoter is completed through the use of an *in vitro* transcription reaction using the Ambion MEGAscript® High Yield T7 Kit. In a 0.2 mL tube at room temperature, combine 2μL of ATP solution, 2 μL of CTP solution, 2 μL of GTP solution, 2 μL of UTP solution, 2 μL of 10x reaction buffer, 1 μg of linearized template DNA (up to 8 μl), 2 μL of enzyme mix, and bring the total reaction volume to 20 μL with nuclease free water. Mix thoroughly by flicking and incubate the mixture at 37° C in a thermocycler with a heated lid for 16 h. Degrade the plasmid DNA by adding 1 μL of Turbo DNase to the reaction tube and incubating for 15 min at 37° C. Add 20 μL of citrate-saturated (pH 4.7) phenol:chloroform:isoamyl alcohol (25:24:1) to the tube. Vortex the mixture for 1 min and centrifuge for 2 min at 12,000 x *g* to separate the phases. Transfer the upper aqueous phase to a fresh tube and add 1 volume of chloroform:isoamyl alcohol (24:1). Vortex the mixture for 1 min and centrifuge for 2 min at 12,000 x *g*. Transfer the upper aqueous phase to a fresh tube and add 0.1 volumes of 3 M sodium acetate and 0.7 volumes of isopropanol. Mix by vortexing and incubate for 10 min at room temperature and then centrifuge for 30 min at 4° C. Carefully discard the supernatant and wash the pellet with 200 μL of ice cold 70% ethanol. Centrifuge for 5 min and carefully remove the supernatant without disturbing the pellet. Air-dry the pellet until no residual ethanol remains and resuspend the dried pellet in 50 μL of nuclease-free water. Quantify the RNA fluorometrically using a Quant-iT™ RiboGreen® RNA Assay Kit and check the transcript size using a microfluidic electrophoresis instrument (e.g., Experion Automated

Electrophoresis System, Agilent 2100 Bioanalyzer, or Agilent 2200 TapeStation). Store the mRNA internal standard stock at -80° C.

5. DNA Standard Preparation

5.1 Required Materials

- *Equipment:* Refrigerator, small tube rocker, 65° C water bath or oven, fluorometry-based instrument for measuring nucleic acid concentration, 10-, 20-, 200-, and 1000-µL pipettes.
- *Materials:* Genomic DNA from a cultured, sequenced microbe unlikely to be closely related to microbes in the natural community, e.g., *Thermus thermophilus* DSM7039 [HB27] genomic DNA (American Type Culture Collection (ATCC), Manassas, VA).
- *Commercially Available Kit*: Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies).
- *Disposables*: sterile 10-, 20-, 200-, and 1000-µL filter tips, nuclease-free micro-centrifuge tubes, gloves.

5.2 Genomic standard stock preparation

- Resuspend the genomic DNA in a volume of nuclease-free water to produce a stock concentration of 0.1 µg/µL following procedures recommended by ATCC. After rehydration incubate overnight at 4° C while rocking, and then incubate for 1 h at 65° C. To prepare a working solution, add 1 µL of the stock solution to 99 µL of nuclease-free water to produce a final concentration of 1 ng/µL. Check the DNA concentration of stocks fluorometrically using Quant-iT™ PicoGreen® dsDNA Assay Kit. The genomic DNA can be stored at -20° C.

6. Internal Standard Addition

Internal standards should be incorporated into the sample in a known amount just prior to RNA/DNA extraction. Prepare a tube with the desired lysis solution and add a known number of internal standard copies/genomes to the prepared lysis tube prior to the addition of the sample. The goal is to add an amount of internal standard sufficient for effective quantification in the sequence dataset, but not so high as to dominate the reads. This amount can be estimated from expected recovery of nucleic acids based on previous experience with the sample type. For example, if 5 μg of total RNA is expected from an extraction, the addition of 25 ng of internal standard (0.5% of the total RNA pool by weight) should be sufficient for a standard ~1000 nt in length. In our experience, a targeted ~0.5% addition has resulted in standards accounting for 0.1 – 5% of reads, depending on accuracy of our predicted RNA yield. When working with multiple standards, each standard should be added to the lysis tube independently in order to control for pipetting error.

7. Internal Standard Recoveries and Quantification

Following sequencing, the number of mRNA internal standards recovered should be quantified by a BLASTn homology search for the template sequence using a bit score cut-off of 50, equivalent to an average percent identity of 98% in our analyses. The number of genomic internal standards should be quantified by first using a BLASTn homology search against the reference genome sequence to identify all potential standard reads, and subsequently taking any hits from the initial BLASTn homology search and performing a BLASTx search against the RefSeq Protein database to identify all protein encoding reads derived from the reference genome with a bit score cut-off of 40. The second annotation step against the RefSeq Protein

database is necessary for identification of the standard reads due to a high number of false positives recruited by the BLASTn homology search. Following quantification, the internal standards should be removed from the dataset before further processing.

8. Dataset Normalization Using Internal Standards

8.1 Metatranscriptome normalization

Following identification of internal transcript standards, total transcript pool size and individual transcript abundances can be 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 = protein encoding reads in the transcriptome library
S_a = molecules of internal standard added to the sample
S_s = internal standard reads in the sequence library
T_a = total molecules of any particular transcript type in the sample. This value can be divided by the mass or volume of sample collected to calculate the transcript abundance per volume or weight.
T_s = number of transcripts of interest in the sequence library

8.2 Metagenome normalization

Following identification of internal genome standards, community gene pool size and individual gene abundances can be calculated as follows:

$$S_r = \frac{S_S}{S_P}$$

$$P_g = \frac{P_s \times S_a}{S_r} \qquad G_a = \frac{G_s \times P_g}{P_s}$$

S_r = # of molecules of internal standard genome recovered from sequencing
S_s = # of protein encoding internal standard reads in the sequence library
S_P = # of protein encoding genes in the internal standard reference genome
P_g = total # of protein encoding genes in the sample
P_s = # of protein encoding sequences in the metagenome library
S_a = # of molecules of internal standard genome added to the sample

G_a = # of molecules of any particular gene category in the sample. This can then be divided by the mass or volume of sample collected to calculate the transcript abundance per volume or weight.

ACKNOWLEDGEMENTS

This research was funded by grants from the Gordon and Betty Moore Foundation (Marine Microbiology Investigator and River-Ocean Continuum of the Amazon) and the National Science Foundation (MCB-0702125).

REFERENCES

Campbell, B. J., Yu, L., Heidelberg, J. F., & Kirchman, D. L. (2011). Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci U S A, 108*(31), 12776-12781.

Church, M. J., Short, C. M., Jenkins, B. D., Karl, D. M., & Zehr, J. P. (2005). Temporal patterns of nitrogenase gene (nifH) expression in the oligotrophic North Pacific Ocean. *Appl Environ Microbiol, 71*(9), 5362-5370.

Damon, C., Vallon, L., Zimmermann, S., Haider, M. Z., Galeote, V., Dequin, S., et al. (2011). A novel fungal family of oligopeptide transporters identified by functional metatranscriptomics of soil eukaryotes. *ISME J, 5*(12), 1871-1880.

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., et al. (2008). Functional metagenomic profiling of nine biomes. *Nature, 452*(7187), 629-632.

Gifford, S. M., Sharma, S., Rinta-Kanto, J. M., & Moran, M. A. (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J, 5*(3), 461-472.

Hannah, M. A., Redestig, H., Leisse, A., & Willmitzer, L. (2008). Global mRNA changes in microarray experiments. *Nat Biotechnol, 26*(7), 741-742.

Hewson, I., Poretsky, R. S., Beinart, R. A., White, A. E., Shi, T., Bench, S. R., et al. (2009). In situ transcriptomic analysis of the globally important keystone N2-fixing taxon Crocosphaera watsonii. *ISME J, 3*(5), 618-631.

Maurice, C. F., Haiser, H. J., & Turnbaugh, P. J. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell, 152*(1-2), 39-50.

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L. K., et al. (2013). Sizing up metatranscriptomics. *ISME J, 7*(2), 237-243.

Ottesen, E. A., Young, C. R., Eppley, J. M., Ryan, J. P., Chavez, F. P., Scholin, C. A., et al. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc Natl Acad Sci U S A, 110*(6), E488-497.

Poretsky, R. S., 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*(7), 4121-4126.

Rivers, A., Sharma, S., Tringe, S. G., Martin, J., Joye, S. B., & Moran, M. A. (2013). Transcriptional Response of Bathypelagic Marine Bacterioplankton to the Deepwater Horizon Oil Spill. *ISME J*, In Press.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D., & Holstege, F. C. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *Embo Reports, 4*(4), 387-393.

Vila-Costa, M., Sharma, S., Moran, M. A., & Casamayor, E. O. (2013). Diel gene expression profiles of a phosphorus limited mountain lake using metatranscriptomics. *Environ Microbiol, 15*(4), 1190-1203.

Xu, L., Chen. H., Hu, X., Zhang, R., Zhang, Z., & Luo, Z. W. (2006). Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Mol Biol Evol,* 23, 1107–1108.

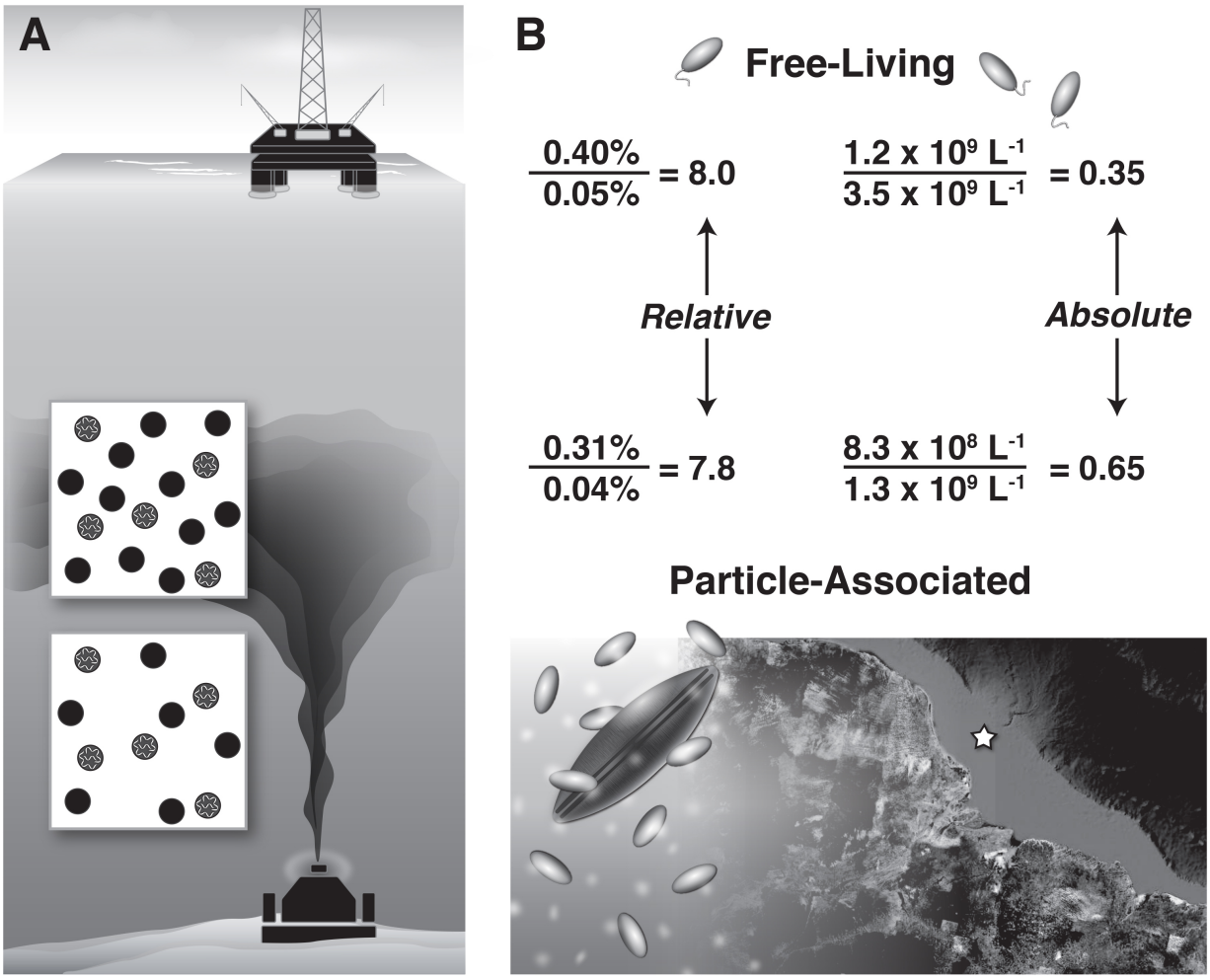


Figure 2.1. Examples of improved quantification of meta-omics data through the use of internal standards. A) Transcripts binning to SAR11 member HTCC7211 accounted for a smaller fraction of the community metatranscriptome in the oil plume caused by the Deepwater Horizon a–ccident compared to non-impacted control samples below the plume, yet the absolute number of transcripts contributed by this taxon was not different. HTCC7211 is one of several bacteria taxa dominant in the pre-spill community that did not respond to the presence of oil, whereas other taxa greatly increased in number and activity in the hydrocarbon-impacted seawater (from Rivers et al., 2013). B) Particle-associated bacteria in the Amazon River plume in June 2010 had 2-fold higher expression of proteorhodopsin genes than free-living bacteria, yet expression estimates calculated incorrectly from relative data would not have shown the differential regulation of this ecologically important gene.

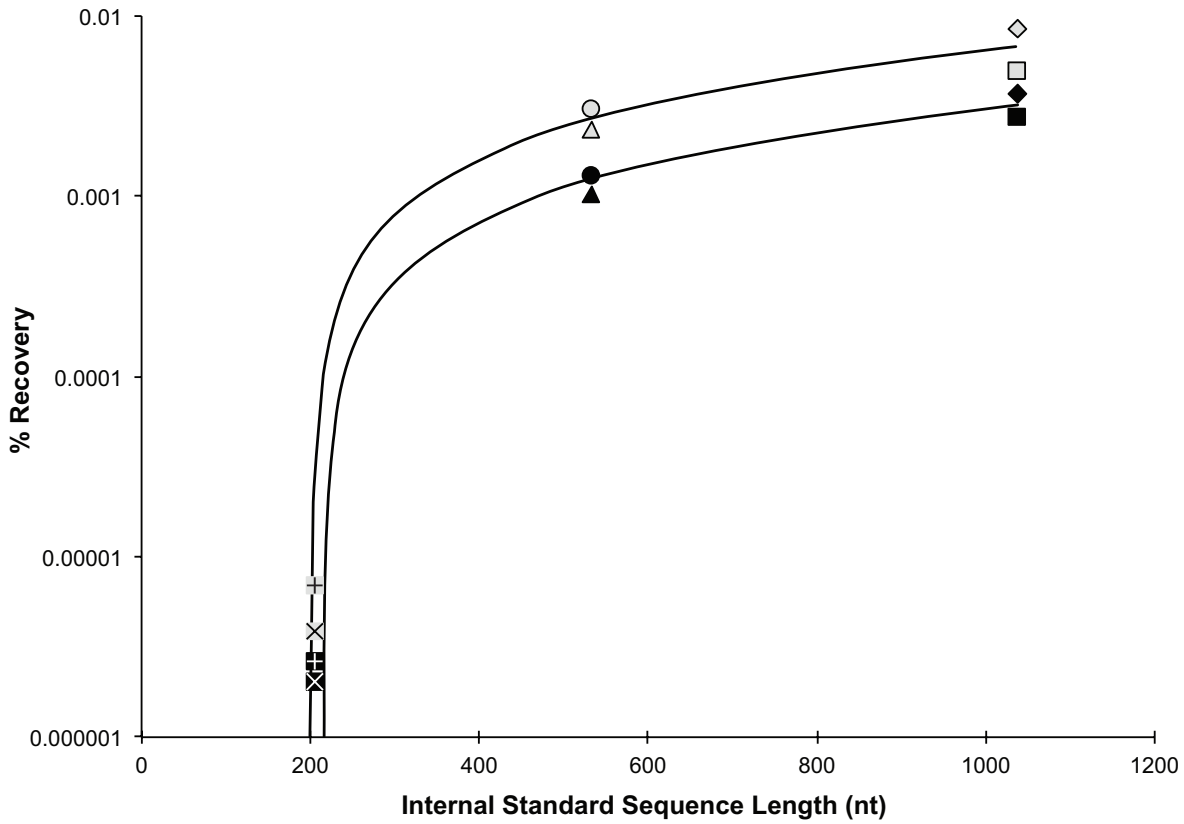


Figure 2.2. Recovery of internal mRNA standards as a function of standard length for two replicate metatranscriptome libraries (black series and gray series) from the Amazon River near Tapajos in June 2011. Two different internal standards of three different lengths (each represented as a different shaped symbol) were added to the samples at the initiation of nucleic acid extraction and percent recoveries were calculated as S_s (internal standard reads in the sequence library) *100/S_a (internal standards added to the sample).

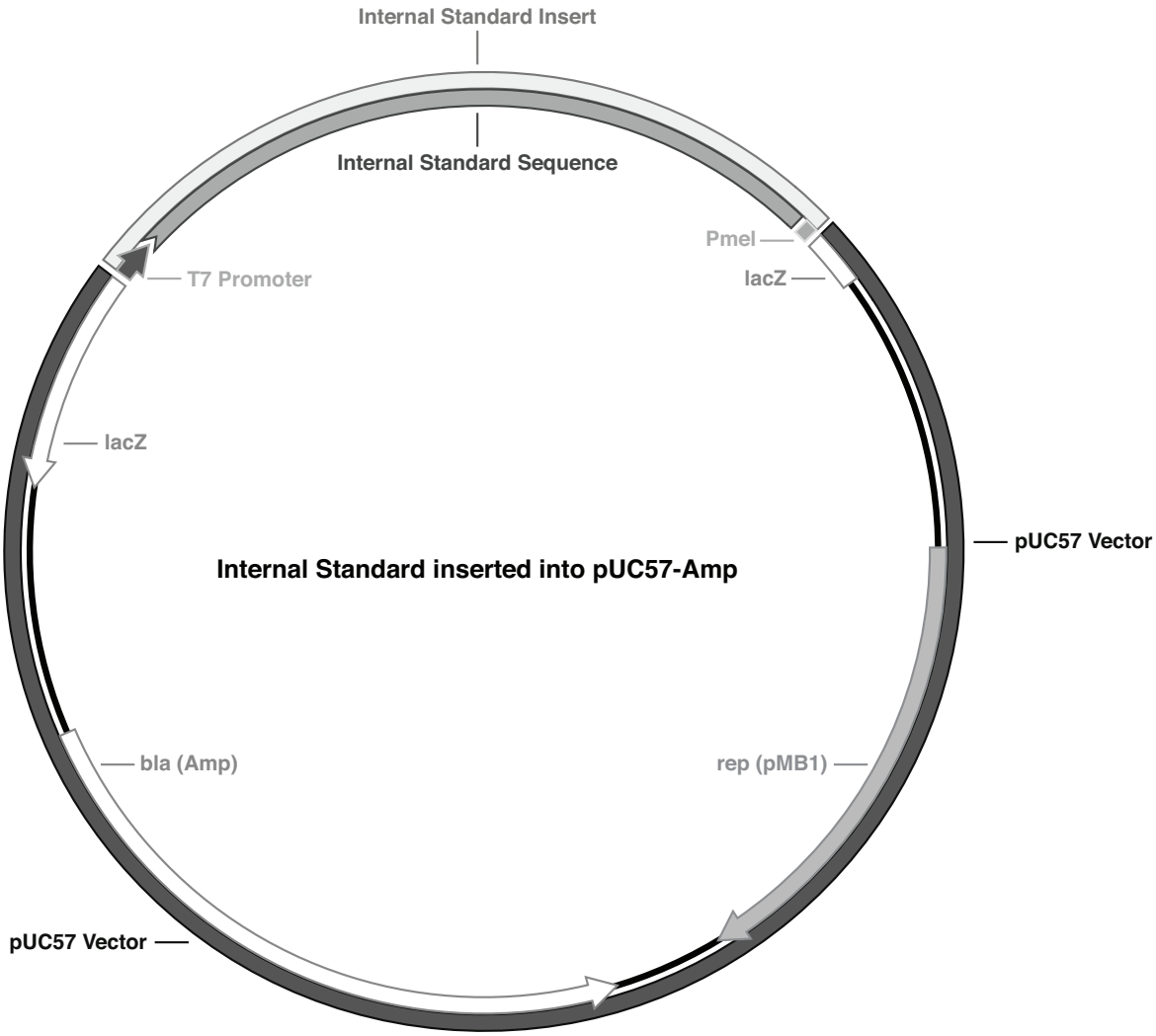


Figure 2.3. Genetic construct for *in vitro* transcription of a synthetic mRNA internal standard.

CHAPTER 3

MICROSPATIAL GENE EXPRESSION PATTERNS IN THE AMAZON RIVER PLUME[1]

[1] Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski BL, Doherty M *et al.* (2014a). Microspatial gene expression patterns in the Amazon River Plume. *Proc Natl Acad Sci U S A* **111:** 11085-11090; Reprinted here with permission of publisher.

Abstract

We investigated expression of genes mediating elemental cycling at the microspatial scale in the ocean's largest river plume using the first fully quantitative inventory of genes and transcripts. The bacterial and archaeal communities associated with a phytoplankton bloom in Amazon River plume waters at the outer continental shelf in June 2010 harbored ~1.0 x 10^{13} genes and 4.7 x 10^{11} transcripts per liter that mapped to several thousand microbial genomes. Genomes from free-living cells were more abundant than those from particle-associated cells, and they generated more transcripts per liter for carbon fixation, heterotrophy, nitrogen and phosphorus uptake, and iron acquisition, although they had lower expression ratios (transcripts $gene^{-1}$) overall. Genomes from particle-associated cells contributed more transcripts for sulfur cycling, aromatic compound degradation, and the synthesis of biologically essential vitamins, with an overall 2-fold upregulation of expression compared to free-living cells. Quantitatively, gene regulation differences were more important than genome abundance differences in explaining why microenvironment transcriptomes differed. Taxa contributing genomes to both free-living and particle-associated communities had up to 65% of their expressed genes regulated differently between the two, quantifying for the first time the extent of transcriptional plasticity in marine microbes *in situ*. In response to patchiness in carbon, nutrients, and light at the micron scale, Amazon plume microbes regulated the expression of genes relevant to biogeochemical processes at the ecosystem scale.

Significance

The microbial community of the Amazon River plume determines the fate of the world's largest input of terrestrial carbon and nutrients to the ocean. By benchmarking with internal standards during sample collection, we determined that each liter of plume seawater contains 1 trillion

genes and 50 billion transcripts from thousands of bacterial, archaeal, and eukaryotic taxa. Gene regulation by taxa inhabiting distinct microenvironments provides insights into micron-scale patterns of transformations in the marine carbon, nitrogen, phosphorus, and sulfur cycles in this globally-important ecosystem.

Introduction

Microbially-mediated biogeochemical transformations in the ocean take place within a poorly characterized matrix of particles, colloids, and dissolved phase materials (Azam and Malfatti 2007) (Fig. 3.1). This microspatial structure affects the taxonomy and functional gene inventories of microbial communities (Ganesh *et al.,* 2014, Smith *et al.,* 2013), driven by differences in nutrient concentrations, light exposure, oxygen availability, and predation (Kiørboe and Jackson 2001, Stocker *et al.,* 2008), and undoubtedly affects the types and rates of biogeochemical processes occurring as well. Bacteria associated with particulate material have higher rates of extracellular enzyme activity (Crump *et al.,* 1998, Karner and Herndl 1992) and metabolism (Delong *et al.,* 1993, Hopkinson *et al.,* 1989) compared to free-living. However, these differences remain difficult to measure and challenging to link to biogeochemistry.

The microspatial structure of the Amazon River Plume is strongly influenced by the mixing of riverine dissolved (DOC) and particulate (POC) organic carbon into the tropical Atlantic. Fluvial export from the Amazon River amounts to 22.3 Tg yr^{-1} of DOC and 13.7 Tg yr^{-1} of POC (Richey *et al.,* 1990) and equals that of the next 8 largest rivers of the world combined (Coles *et al.,* 2013). Although relatively dilute compared to other rivers (Ryther *et al.,* 1967), this mixture of dissolved and particulate nitrogen, phosphate, silica, and iron that is delivered to the

ocean stimulates marine microbial activity and affects both primary productivity and carbon sequestration at a global scale (Subramaniam *et al.,* 2008).

Here, a meta-omics methodology benchmarked with internal standards (Gifford *et al.,* 2011, Satinsky *et al.,* 2013) was used to assemble the first fully quantitative inventories of microbial genes and transcripts in a natural ecosystem, producing a highly-resolved view of gene expression driving carbon and nutrient flux through free-living and particle-associated microbes of the Amazon Plume. We compare the level of transcription for the same gene in each microenvironment; enumerate transcripts mediating key carbon and nutrient transformations; and predict the biogeochemical roles of transcriptionally active free-living and particle-associated cells.

Results and Discussion

A quantitative multi-omics dataset - Inventories of microbial genes and transcripts were obtained from particle-associated (PA) and free-living (FL) cells (defined operationally: PA, >2.0 μm; FL, 0.2 – 2.0 μm) in Amazon Plume waters at the outer continental shelf in July 2010 (Station 10). Coverage of eukaryotic transcripts was improved by additionally targeting poly(A)-tailed mRNAs in the PA fraction (Fig. 3.1). Each sequence library obtained in duplicate for the three meta–omics data types contained 1.2 - 11 x 10^6 possible protein-encoding reads averaging 190 nt. The eukaryotic community was dominated by sequences that binned to the genomes of diatoms *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, and *Odontella sinensis* and the green picoeukaryote *Micromonas* sp. RC299; direct microscopic observation confirmed a dense multi-species diatom bloom. We focus here on analysis of the genes and transcripts from Bacteria and Archaea [hereafter referred to as prokaryotes (Whitman 2009)], accounting for ~70% of the annotated reads.

Based on internal standard recoveries, the sequencing depth for each sample (% of sampled genes or transcripts sequenced) ranged from 1.5 x 10^{-6} to 8.7 x 10^{-4} (SI Appendix: Table 3.S1), leading to estimates of prokaryotic gene and transcript numbers per liter of seawater of ~1 x 10^{12} and 5 x 10^{10} (Table 3.1). Two independent measures of prokaryotic cell abundance provided a check on quantitation from internal standard recoveries. First, flow cytometric counting indicated 5.8 x 10^{9} cells L^{-1} at this station while standard-normalized metagenomic read counts estimated 3.6 x 10^{9} cells L^{-1} (Table 3.2). Second, direct microscopic counts of phycoerythrin- and chlorophyll-*a* fluorescing cells indicated 6.9 x 10^{8} cells L^{-1} for *Synechococcus* (Goes *et al.,* 2014) while calculated genome equivalents based on standard-normalized metagenomic data was 6.8 x 10^{8} L^{-1} (Table 3.2).

Natural history of Amazon Plume prokaryotes - Protein sequence comparisons against reference genomes indicated that the prokaryotic assemblage was dominated by Cyanobacteria in the Synechococcaceae family and heterotrophic bacteria in the Gammaproteobacteria, Bacteroidetes, and Alphaproteobacteria. The average percent identity between the metagenomic reads and the reference proteins they binned to ranged from 62-92% for the dominant taxa (Table 3.2). About 0.1% of metagenomic reads contained 16S rRNA gene fragments, and the taxonomic distribution of these matched the metagenomic protein binning and the patterns of PCR-amplified 16S rRNA genes (SI Appendix: Fig. 3.S1).

Prokaryotic genes in the Amazon Plume were present in higher copy number than the mRNAs produced from them. This is consistent with earlier studies of *Escherichia coli* (Taniguchi *et al.,* 2010) and marine bacterioplankton (Church *et al.,* 2010, Moran *et al.,* 2013), and reflects a small and dynamic mRNA pool in environmental cells. Genes were split 68%:32%

between the FL and PA fractions, and transcripts were split 49%:51%. Thus per-gene expression levels were 2-fold greater for PA cells than FL cells (Table 3.1), a finding in line with previous reports of higher rates of growth and enzymatic activity for attached cells (Crump *et al.,* 1998). Approximately 99% of prokaryotic genes and transcripts were bacterial and ~1% were archaeal. Compositional difference between FL and PA prokaryotes was small at Station 10. While disruption of microspatial structure during sampling could homogenize the communities, samples collected identically from other stations had distinct compositional differences between microenvironments (SI Appendix: Fig. 3.S2). There were also substantial differences in transcription patterns between size fractions within a single genome bin at this station (SI Appendix: Fig. 3.S3).

Transcript inventories of biogeochemically-relevant genes - Transcript abundance per liter of seawater (Fig. 3.1) was determined for prokaryotic genes involved in transport, fixation, and metabolism of C, N, P, and S (SI Appendix: Table 3.S2). Counts ranged from near the detection limit at ~10^5 transcripts L^{-1} for phosphonate utilization gene *phnG* up to ~4 x 10^9 transcripts L^{-1} for proteorhodopsin (Fig. 3.2). Most genes had higher transcript counts in the FL microenvironment (Fig. 3.2) due in part to the greater abundance of free-living genomes (Table 3.2). These included genes involved in heterotrophic C metabolism (12 of 15), phototrophy (5 of 10), N cycling (5 of 7), P cycling (10 of 15), and motility (5 of 5). Only a few genes had transcript inventories biased towards the PA community; these included vitamin B6 biosynthesis gene *pdxH*, S cycling genes (*cysI, soxB,* and *dmdA*), and aromatic ring cleavage gene *pcaH* (Fig. 3.2). Summing across genes with related functions, a liter of seawater at this plume station contained more transcripts mediating C and N cycling originating from genomes of free-living cells, but more transcripts for vitamin biosynthesis and S cycling from particle-associated cells.

Expression ratios of biogeochemically-relevant genes - A different calculation approach scaled transcript copies to gene copies (transcripts $gene^{-1}$) to ask whether the same gene is differentially regulated in the two microenvironments (Fig. 3.1). Heterotrophic C transformation genes were upregulated in genomes of PA cells compared to those from FL cells (Fig. 3.3), including genes for the catabolism of aminopeptidases (*pepA,L,M*), glycolysis (GAPDH), and aromatic compound metabolism (*pcaH* but not lignin-related *vanA*). Only one heterotrophy gene was significantly upregulated in FL cells, and this mediates the degradation of cellular carbon reserves (*phaZ*) and does not contradict the higher organic C availability signal for PA cells. Upregulated genes in PA prokaryotes also included those for synthesis of vitamin B6 (*pdxH*) and B1 (*thiL*), required for amino acid and central carbon metabolism (Koenigsknecht and Downs 2010). Genes for uptake of siderophore-bound Fe and Fe^{+2} were upregulated in genomes from PA cells as well, possibly linked to the role of iron-sulfur clusters and heme in respiratory chains and TCA cycle enzymes. Thus although PA prokaryotes contributed fewer transcripts to the water column, they were more transcriptionally active per gene in C heterotrophy than FL cells.

Among nutrient acquisition genes, those encoding P uptake had the greatest differences in regulation between microenvironments. Low-affinity phosphate transporter *pitA* was upregulated in the PA transcriptome and high-affinity *pstA* was upregulated in the FL transcriptome (Fig. 3.3). Together with the upregulation of phosphonate transporter gene *phnE* by FL genomes, turnover of labile P in bulk seawater is predicted to be more rapid and competition more intense compared to particle microenvironments. For N cycle genes, differences in expression levels were smaller and N availability status therefore predicted to be more similar between microenvironments. An exception was a dissimilatory nitrate reductase homolog binning exclusively to Thaumarchaeota genomes and upregulated in the FL fraction,

although the function of the *nirK* homolog in Thaumarchaeota is uncertain (Hollibaugh *et al.,* 2011). Expression ratios were higher in PA cells for genes that mediate organic and inorganic S metabolism, including DMSP and sulfate. Thus while nutrient concentrations at this station (0.2 mM NO_x, 0.4 mM PO_4) were high compared to more offshore regions of the plume, transcription patterns suggested microscale differences in their availability and turnover. Genes associated with light harvesting were also expressed differently (Fig. 3.3); FL cells had higher expression of genes for autotrophic light utilization (photosynthesis and C fixation genes *pufM*, *psbB*, *ε-CA*, and *rbcL* (IA)), while PA cells had higher expression of photoheterotrophic light utilization (proteorhodopsin). Free-living prokaryotes also had higher expression ratios of flagellar genes (Fig. 3.3).

Gene abundance versus gene regulation - Three reference genomes with high sequence coverage were used for taxon-specific determinations of whether genome abundance (genomes L^{-1}) or gene regulation (transcripts $gene^{-1}$) contributed more to differences in microenvironment transcriptomes. The potential complication that a genome bin was derived from different populations in each microenvironment was considered. However, analysis of percent identity patterns for three housekeeping genes (*rpoB, gryB*, and *recA*) indicated that while genomes may recruit more than one population, the relative population abundances did not differ substantively between FL and PA fractions for the *Synechococcus* sp. CB0205, *Pelagibacter* sp. HTCC7211, and gammaproteobacterium HTCC2070 bins (SI Appendix: Fig. 3.S4).

Regarding genome abundance differences, *Synechococcus* sp. CB0205-like populations had 1.6-fold fewer genomes L^{-1} in the PA microenvironment. Regarding gene regulation differences, all metabolic pathways in CB0205 that exhibited significant differences in expression levels were downregulated in PA cells compared to FL cells (62 pathways; Table

3.3). These included pathways related to autotrophy (photosynthesis, carotenoid biosynthesis, and C fixation), as well as biosynthesis of amino acids, fatty acids, and peptidoglycan (Table 3.3). At the individual gene level, a total of 516 were downregulated in PA cells compared to FL (SI Appendix: Table 3.S3). Among them were genes transcribed by picocyanobacteria when experiencing N limitation (*ntcA*; 4-fold downregulated) and P limitation (*ptsS*; 6-fold downregulated) (Lindell and Post 2001, Scanlan *et al.,* 1993). Only 23 individual genes were upregulated in PA cells, including stress-related proteins (universal stress protein, peroxiredoxin, and glutaredoxin) and cell wall modifying proteins (SI Appendix: Fig. 3.S5, Table 3.S3). Overall, the transcriptome of CB0205-like cells was affected equally and in the same direction by genome abundance and gene regulation factors; when averaged across the genome, PA genes were less abundant by ~1.6-fold and less transcriptionally active by ~1.6-fold (Fig. 3.4, SI Appendix: Fig. 3.S5).

The two heterotrophic taxa likewise had lower abundance of PA genomes (2.2-fold for *Pelagibacter* sp. HTCC7211 and 2.6-fold for gammaproteobacterium HTCC2080; Fig. 3.4), but different gene regulation patterns compared to the *Synechococcus* bin. For HTCC7211, all metabolic pathways with differences in expression levels (39 pathways) were upregulated in PA compared to FL cells (Table 3.3), including amino acid, nucleic acid, carboxylic acid, C1 compound, and taurine metabolism pathways (Table 3.3). For gammaproteobacterium HTCC2080, 31 metabolic pathways were upregulated in PA cells, including carboxylic and fatty acid metabolism and siderophore production, while 3 were downregulated, including flagellar assembly (Table 3.3). At the individual gene level, HTCC7211 had 84 genes upregulated in PA compared to FL cells, and HTCC2080 had 44 (SI Appendix: Tables 3.S4, 3.S5). Stress-related functions (heat shock protein, acid tolerance protein, chaperones, and glutathione S-transferase)

and transporters (ammonium, amino acids, taurine, polyamine, and glycine betaine; SI Appendix: Fig. 3.S5) were represented. Overall, genome abundance and gene regulation differences worked in opposite directions in the HTCC7211 and HTCC2080 bins; when averaged across the genome, cells assigned to these two heterotrophic taxa were less abundant by ~2.5-fold but more transcriptionally active by ~3 to 4-fold when associated with particles (Fig. 3.4).

A prokaryotic perspective on Amazon Plume biogeochemistry - High rates of C fixation and a deficit of dissolved inorganic C in the water column (~100 mmol L^{-1}) signaled a major system shift to net autotrophy at Station 10, driven by release from light limitation and stimulation by river-derived nutrients. This hot-spot of primary production was upstream of an area of high POC deposition (Chong *et al.,* 2014), indicative of sequestration of new production in nearby continental slope sediments. Surface water was dominated by dense populations of coastal diatom species, suggesting substantial contributions by these cells to the particulate material. Indeed, phytoplankton C accounted for ~80% of POC (676 mg C L^{-1} estimated from cell count and biovolume measures, compared to 842 mg C L^{-1} total POC), in agreement with $d^{13}C$ analysis and terrestrial biomarker concentrations (triterpenoids, C29-sterols, isoprenoids, and selected sugars) that suggested only 10-20% terrestrial POC.

Within this matrix of marine and terrestrial organic matter, the upregulation of photosynthesis genes in free-living cells of the dominant prokaryotic primary producers in the *Synechococcus* clade (Fig. 3.3 and SI Appendix: Fig. 3.S6) accords with their higher abundance in free-living microenvironments in this study (Table 3.2) and others (Ganesh *et al.,* 2014, Smith *et al.,* 2013). *Synechococcus* transcription patterns also suggested greater nutrient competition when free-living based on N and P stress genes *ntcA* and *pstS*, and this may reflect higher

nutrient demand in support of photosynthetic activity as well as lower nutrient availability to cells in the free-living microenvironment. Expression of nitrate and nitrite transporters in both microenvironments indicated that *Synechococcus* competed with eukaryotic primary producers for riverine-derived N (Table 3.3, SI Appendix: Table 3.S3). Although the ~1 μm-diameter cells are not typically thought to sink efficiently in the ocean, some studies suggest that picophytoplankton indeed contribute to deep flux in proportion to their biomass (Close *et al.,* 2013, Kiørboe and Jackson 2001). At this outer plume station, picocyanobacteria accounted for ~12% of prokaryotic cells (~10^9 L^{-1}) (Table 3.2), a value at the upper range for coastal oceans. In contrast, the photoheterotrophic prokaryotes showed higher expression ratios of light-capturing proteorhodopsin in the particle microenvironment (Fig. 3.3), with transcripts dominated by Flavobacteria and SAR116. The SAR11 HTCC7211 bin differed from this pattern in having similar proteorhodopsin expression ratios in both microenvironments (1.0 and 0.8 transcripts $gene^{-1}$), and previous studies agree that SAR11 proteorhodopsin genes are not highly regulated ((Steindler *et al.,* 2011). Aerobic anoxygenic phototrophy appeared insignificant here, given gene counts and expression ratios ~2 orders of magnitude lower for *bchX*, *pufL*, and *pufM* compared to proteorhodopsin and the oxygenic photosynthesis genes (Fig. 3.3). Overall transcription patterns suggest that micron-scale heterogeneity within a single water depth influences the rates and location of prokaryotic phototrophic activity.

A persistent question in microbial biogeochemistry is whether the typically higher activity of less abundant PA cells outweighs the typically lower activity of more abundant FL cells in mediating biogeochemical transformations (Crump *et al.,* 1998, Ganesh *et al.,* 2014). For the 74 biogeochemically-relevant genes assayed here, there were five answers to the question (Fig. 3.5). One gene class was upregulated sufficiently strongly in PA cells (~5-fold greater than

in FL) that they dominated the community transcriptome despite fewer genomes in this microenvironment; this group included functions related to S cycling, vitamin biosynthesis, and aromatic compound degradation. Another gene class was significantly upregulated in PA cells, but the fold-difference (3- to 4-fold) was insufficient to overcome genome abundance differences and the community transcriptomes had statistically indistinguishable representation; this group included genes for Fe acquisition, P storage, S transformation, and vitamin biosynthesis. For the three remaining outcomes, significantly more transcripts were contributed by cells in the FL microenvironment. These consisted of a gene group upregulated in PA cells but at fold-differences insufficient to make up for the greater number of FL genomes (~1.5- to 2-fold; proteorhodopsin, C heterotrophy, and other Fe and S genes); a group with no difference in expression ratios but for which there was more transcripts produced by the more numerous FL cells (phosphate and phosphonate acquisition genes); and a group upregulated in FL prokaryotes (~3-fold greater than in PA) and dominating the community transcriptome (up to 10-fold; CO_2 fixation, motility, and P acquisition; Fig. 3.5).

The importance of microspatial partitioning of microbial activities in the ocean was first suspected based on differences in taxonomic composition between free-living and particle-associated cells (Delong *et al.,* 1993) and later supported by analysis of gene inventories (Ganesh *et al.,* 2014, Smith *et al.,* 2013). At the transcriptome level, contributions are determined not just by the abundance of each taxon's genome but also by the regulation of its genes. By benchmarking with internal standards, differences in gene regulation were determined to be more important than differences in genome counts in explaining transcriptome composition at this Amazon Plume station. Up to 65% of the genes in individual taxa exhibited significantly different expression when free-living and particle-associated cells were compared (Fig. 3.4); on

average, particle-associated cells had twice as many transcripts as free-living cells (Table 3.1); and expression of biogeochemically-relevant genes varied by up to 6-fold between the microenvironments (Fig. 3.3). With the important caveat that mRNA abundance cannot be interpreted as a proxy for elemental flux (Moran *et al.,* 2013), we hypothesize that the free-living prokaryotic community contributed more to C fixation, heterotrophy, N and P uptake, and iron acquisition than the particle-associated community in this ecosystem, while the particle-associated community contributed more to S cycling, aromatic compound degradation, and the synthesis of biologically essential vitamins. The scale of spatial heterogeneity relevant to ecological processes (1) is indeed at the micron level in this ecosystem, with patchiness in abundance and regulation of prokaryotic genes within a single water mass signaling the partitioning of functions that drive elemental cycling.

Materials and Methods

Surface seawater was collected ~ 500 km north of the Amazon River mouth in June 2010. Two size fractions were obtained by sequential filtration of 156 μm pre-filtered water through 2.0 μm and 0.2 μm pore-size membrane filters. Duplicate metagenomes and metatranscriptomes were obtained from each fraction as 2 x 150 bp reads. Internal standards were added immediately prior to cell lysis in known copy numbers (SI Appendix: Table 3.S1). Assignment of reads to taxonomic bins and functions was based on Blastx searches against RefSeq or a custom database of 74 selected genes. Detailed sample collection, processing, and statistical analysis methods can be found in the supplemental methods. Sequence data are available at the NCBI SRA under accession numbers SRP039390 (metagenomes), SRP037995 (non-selective metatranscriptomes), and SRP039544 (poly(A)-selected metatranscriptomes).

Acknowledgements

We appreciate the help of A. Burd, E. Carpenter, A. Mehring, C. English, J. Mrazek, and R. Nilsen. This work was funded in part by the Gordon and Betty Moore Foundation through Grants GBMF2293 and 2928 and NSF grant OCE-0934095. Resources were provided by the University of Georgia's Georgia Advanced Computing Resource Center and CAMERA.

References

Azam F, Malfatti F (2007). Microbial structuring of marine ecosystems. *Nat Rev Microbiol* **5:** 782-791.

Chong LS, Berelson WM, McManus J, Hammond DE, Rollins NE, Yager PL (2014). Carbon and biogenic silica export influenced by the Amazon River Plume: Patterns of remineralization in deep-sea sediments. *Deep Sea Res Part 1 Oceanogr Res Pap* **85:** 124-137.

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.

Close HG, Shah SR, Ingalls AE, Diefendorf AF, Brodie EL, Hansman RL *et al.* (2013). Export of submicron particulate organic matter to mesopelagic depth in an oligotrophic gyre. *Proc Natl Acad Sci U S A* **110:** 12565-12570.

Coles VJ, Brooks MT, Hopkins J, Stukel MR, Yager PL, Hood RR (2013). The pathways and properties of the Amazon River Plume in the tropical North Atlantic Ocean. *J Geophys Res-Oceans* **118:** 6894-6913.

Crump BC, Baross JA, Simenstad CA (1998). Dominance of particle-attached bacteria in the Columbia River estuary, USA. *Aquat Microb Ecol* **14:** 7-18.

Delong EF, Franks DG, Alldredge AL (1993). Phylogenetic diversity of aggregate-attached vs free-living marine bacterial assemblages. *Limnol Oceanogr* **38:** 924-934.

Ganesh S, Parris DJ, DeLong EF, Stewart FJ (2014). Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J* **8:** 187-211.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* **5:** 461-472.

Goes JI, Gomes HdR, Chekalyuk AM, Carpenter EJ, Montoya JP, Coles VJ *et al.* (2014). Influence of the Amazon River discharge on the biogeography of phytoplankton communities in the western tropical north Atlantic. *Prog Oceanogr* **120:** 29-40.

Hollibaugh JT, Gifford S, Sharma S, Bano N, Moran MA (2011). Metatranscriptomic analysis of ammonia-oxidizing organisms in an estuarine bacterioplankton assemblage. *ISME J* **5:** 866-878.

Hopkinson CS, Sherr B, Wiebe WJ (1989). Size fractionated metabolism of coastal microbial plankton. *Mar Ecol Prog Ser* **51:** 155-166.

Karner M, Herndl GJ (1992). Extracellular enzymatic-activity and secondary production in free-living and marine-snow-associated bacteria. *Mar Biol* **113:** 341-347.

Kiørboe T, Jackson GA (2001). Marine snow, organic solute plumes, and optimal chemosensory behavior of bacteria. *Limnol Oceanogr* **46:** 1309-1318.

Koenigsknecht MJ, Downs DM (2010). Thiamine biosynthesis can be used to dissect metabolic integration. *Trends Microbiol* **18:** 240-247.

Lindell D, Post AF (2001). Ecological aspects of ntcA gene expression and its use as an indicator of the nitrogen status of marine Synechococcus spp. *Appl Environ Microbiol* **67:** 3340-3349.

Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan LK *et al.* (2013). Sizing up metatranscriptomics. *ISME J* **7:** 237-243.

Richey JE, Hedges JI, Devol AH, Quay PD, Victoria R, Martinelli L *et al.* (1990). Biogeochemistry of Carbon in the Amazon River. *Limnol Oceanogr* **35:** 352-371.

Ryther JH, Menzel DW, Corwin N (1967). Influence of the Amazon River outflow on the ecology of the western tropical Atlantic. *J Mar Res***:** 69-83.

Satinsky BM, Gifford SM, Crump BC, Moran MA (2013). Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. In: DeLong EF (ed). *Methods Enzymol*. Academic Press. pp 237-250.

Scanlan DJ, Mann NH, Carr NG (1993). The response of the picoplanktonic marine cyanobacterium Synechococcus species WH7803 to phosphate starvation involves a protein homologous to the periplasmic phosphate-binding protein of Escherichia coli. *Mol Microbiol* **10:** 181-191.

Smith MW, Zeigler Allen L, Allen AE, Herfort L, Simon HM (2013). Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. *Front Microbiol* **4**.

Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ (2011). Energy starved Candidatus Pelagibacter ubique substitutes light-mediated ATP production for endogenous carbon respiration. *PLoS One* **6:** e19725.

Stocker R, Seymour JR, Samadani A, Hunt DE, Polz MF (2008). Rapid chemotactic response enables marine bacteria to exploit ephemeral microscale nutrient patches. *Proc Natl Acad Sci U S A* **105:** 4209-4214.

Subramaniam A, Yager PL, Carpenter EJ, Mahaffey C, Bjorkman K, Cooley S *et al.* (2008). Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci U S A* **105:** 10460-10465.

Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J *et al.* (2010). Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329:** 533-538.

Whitman WB (2009). The modern concept of the procaryote. *J Bacteriol* **191:** 2000-2005; discussion 2006-2007.

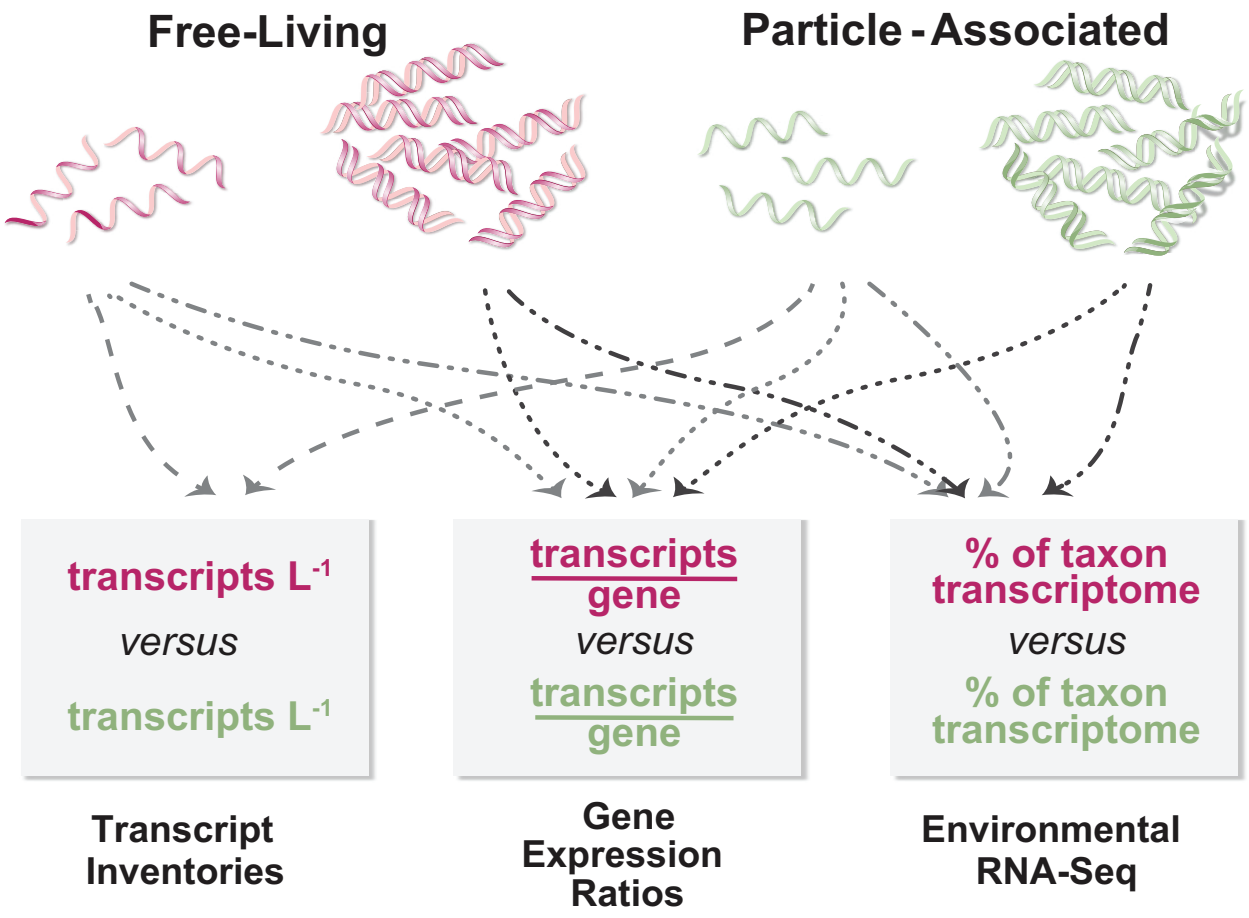


Figure 3.1 Sampling and calculation strategies for gene expression analyses. Maroon text indicates calculations for free-living cells and green for particle-associated. Approaches to expression analysis are indicated in boxes, with the first two making use of internal gene and transcript standards.

Table 3.S1. Internal standard addition and recoveries.

	Sample	Standard Copies Added	Standard Copies Recovered	Sequencing Depth (%)	Volume Filtered (L)	Normalization Factor
			Metatranscriptomes			
Standard 1	ACM11	2.10×10^{10}	19,640	9.25×10^{-5}	5.9	1.81×10^{5}
(pTXB1, 916 nt)	ACM29	2.10×10^{10}	13,769	6.54×10^{-5}	4.8	3.18×10^{5}
	ACM12	2.10×10^{10}	22,233	10.57×10^{-5}	5.9	1.60×10^{5}
	ACM30	2.10×10^{10}	7,793	3.70×10^{-5}	3.3	8.18×10^{5}
Standard 2	ACM11	1.17×10^{10}	11,350	9.68×10^{-5}	5.9	1.75×10^{5}
(pFN18A, 970 nt)	ACM29	1.17×10^{10}	6,313	5.39×10^{-5}	4.8	3.87×10^{5}
	ACM12	1.17×10^{10}	8,404	7.17×10^{-5}	5.9	2.36×10^{5}
	ACM30	1.17×10^{10}	3,139	2.68×10^{-5}	3.3	1.13×10^{6}
			Metagenomes			
Thermus thermophilus	ACM4	1.86×10^{7}	1.428	7.66×10^{-6}	4.8	2.72×10^{6}
HB8	ACM36[†]	-	-	-	5.2	-
(Genomic DNA)	ACM3	2.02×10^{7}	1.654	8.19×10^{-6}	5.2	2.35×10^{6}
	ACM37	1.86×10^{7}	0.287	1.54×10^{-6}	4.8	1.35×10^{7}
			Poly(A) Metatranscriptomes			
Standard 3	ACM8	2.00×10^{9}	17,455	8.73×10^{-4}	5.5	2.08×10^{4}
(HAP-1, 499 nt)	ACM27	2.00×10^{9}	11,362	5.68×10^{-4}	7.4	1.54×10^{3}

[†]Standards were added incorrectly to sample ACM36; gene normalization was estimated based on replicate sample ACM4.

Table 3.1. Metagenome and metatranscriptome datasets from Amazon River plume Station 10 in June 2010. Per liter calculations are based on recovery of internal standards (Table 3.S1).

	0.2 – 2.0 μm Size Fraction[¥]				≥ 2.0 μm Size Fraction					
	Metatranscriptomes		Metagenomes		Metatranscriptomes		Metagenomes		Poly(A)-Selected Metatranscriptomes	
	ACM11	ACM29	ACM4	ACM36	ACM12	ACM30	ACM3	ACM37	ACM8	ACM27
Total Reads	$3.22x10^7$	$2.04x10^7$	$1.32x10^7$	$1.44x10^7$	$4.22x10^7$	$3.08x10^7$	$1.22x10^7$	$4.66x10^6$	$3.66x10^7$	$4.46x10^7$
Paired Reads[†]	$5.39x10^6$	$3.70x10^6$	$4.66x10^6$	$1.33x10^6$	$5.08x10^6$	$5.36x10^6$	$4.21x10^6$	$1.31x10^6$	$1.16x10^7$	$8.05x10^6$
Genes or Transcripts L^{-1}										
Bacterial	$1.60x10^{11}$	$2.89x10^{11}$	$6.43x10^{12}$	$6.12x10^{12}$	$9.51x10^{10}$	$3.82x10^{11}$	$3.45x10^{12}$	$2.59x10^{12}$	-	-
Archaeal	$1.75x10^9$	$2.92x10^9$	$7.33x10^{10}$	$6.18x10^{10}$	$1.31x10^9$	$1.56x10^9$	$3.28x10^{10}$	$2.61x10^{10}$	-	-
Eukaryotic	$1.25x10^{10}$	$2.49x10^{10}$	$1.46x10^{11}$	$5.96x10^{11}$	$5.42x10^{10}$	$1.03x10^{12}$	$3.55x10^{11}$	$1.01x10^{12}$	$5.40x10^{10}$	$2.57x10^9$
Viral	$7.88x10^9$	$3.92x10^{10}$	$8.01x10^{11}$	$6.71x10^{11}$	$6.70x10^9$	$5.90x10^{10}$	$5.99x10^{11}$	$2.73x10^{11}$	-	-
Expression Ratio										
Bacterial	0.034				0.075				-	
Archaeal	0.033				0.046				-	
Eukaryotic	0.048				0.755				-	
Viral	0.030				0.072				-	

[¥]Replicate samples analyzed for each size fraction and data type are shown.

[†]Post QC sequences

Table 3.2. Contribution of the most abundant prokaryotic taxa (out of 2,999 total) to Amazon Plume metagenomes and metatranscriptomes. Data are based on replicates ACM3, ACM4, ACM11, and ACM12, and size fractions are summed.

Genome Bin	Genes per L	PA Genes (%)	% of Genes	Genes in Reference Genome	Genome Equivalents per L	% of Cell Count*	Protein % Identity	Transcripts per L	PA Transcripts (%)	% of Transcripts
Synechococcus sp. CB0205	1.60×10^{12}	39.09	15.96	2,719	5.87×10^{8}	10.13	89.5	1.90×10^{10}	16.7	7.36
Cand. Pelagibacter sp. HTCC7211	3.62×10^{11}	30.99	3.63	1,447	2.50×10^{8}	4.31	75.9	2.07×10^{9}	52.8	0.80
gammaproteobacterium HTCC2080	2.51×10^{11}	27.72	2.53	3,185	7.89×10^{7}	1.36	71.2	8.25×10^{9}	30.9	3.20
Synechococcus sp. CB0101	1.50×10^{11}	40.69	1.50	3,010	4.98×10^{7}	0.86	87.8	2.83×10^{9}	19.5	1.10
gammaproteobacterium NOR51-B	1.48×10^{11}	27.90	1.49	2,930	5.05×10^{7}	0.87	70.7	7.18×10^{9}	34.7	2.78
Synechococcus sp. RS9916	1.24×10^{11}	46.20	1.24	2,961	4.20×10^{7}	0.72	83.9	2.91×10^{9}	26.6	1.13
Muricauda ruestr. DSM13258	9.81×10^{10}	29.32	0.99	3,432	2.86×10^{7}	0.49	69.3	3.46×10^{9}	48.5	1.34
gammaproteobacterium HTCC2148	9.05×10^{10}	29.29	0.91	3,827	2.37×10^{7}	0.41	66.2	2.63×10^{9}	35.6	1.02
Fluviicola taffensis DSM 16823	7.55×10^{10}	50.56	0.75	4,033	1.87×10^{7}	0.32	64.4	7.23×10^{9}	41.9	2.80
Haliscomenobacter hy. DSM1100	7.55×10^{10}	49.87	0.75	6,858	1.10×10^{7}	0.19	61.7	1.90×10^{9}	48.2	0.74
Kordia algicida OT-1	7.46×10^{10}	34.22	0.75	4,514	1.65×10^{7}	0.29	67.5	1.91×10^{9}	48.0	0.74
gammaproteobacterium IMCC3088	6.52×10^{10}	27.83	0.65	2,855	2.28×10^{7}	0.39	67.6	1.98×10^{9}	35.3	0.77
gammaproteobacterium HTCC2207	6.34×10^{10}	27.14	0.64	2,388	2.66×10^{7}	0.46	62.4	2.26×10^{9}	23.7	0.88
Zobellia galactanivorans	6.29×10^{10}	30.77	0.63	4,732	1.33×10^{7}	0.23	68.6	1.80×10^{9}	44.6	0.70
Erythrobacter litoralis HTCC2594	6.27×10^{10}	62.52	0.62	3,011	2.08×10^{7}	0.36	76.9	3.36×10^{9}	45.3	1.30
Total			33.04		4.87×10^{8}*	21.39				26.66

*Genome equivalents for the 250 most abundant taxa (75% of prokaryotic genes) = 2.7×10^{9}. The direct cell count at Station10 was 5.80×10^{9} cells/L.

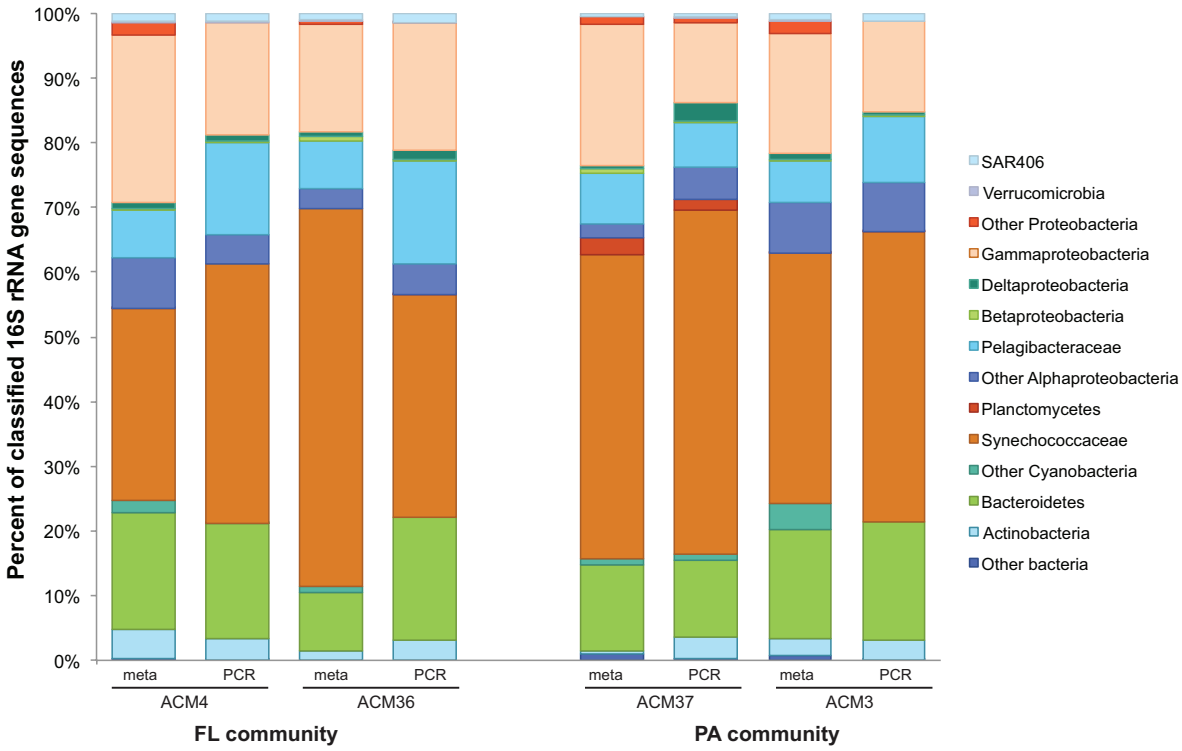


Figure 3.S1. Taxonomic composition of the prokaryotic community in the Amazon River plume in June 2010 (Station 10).

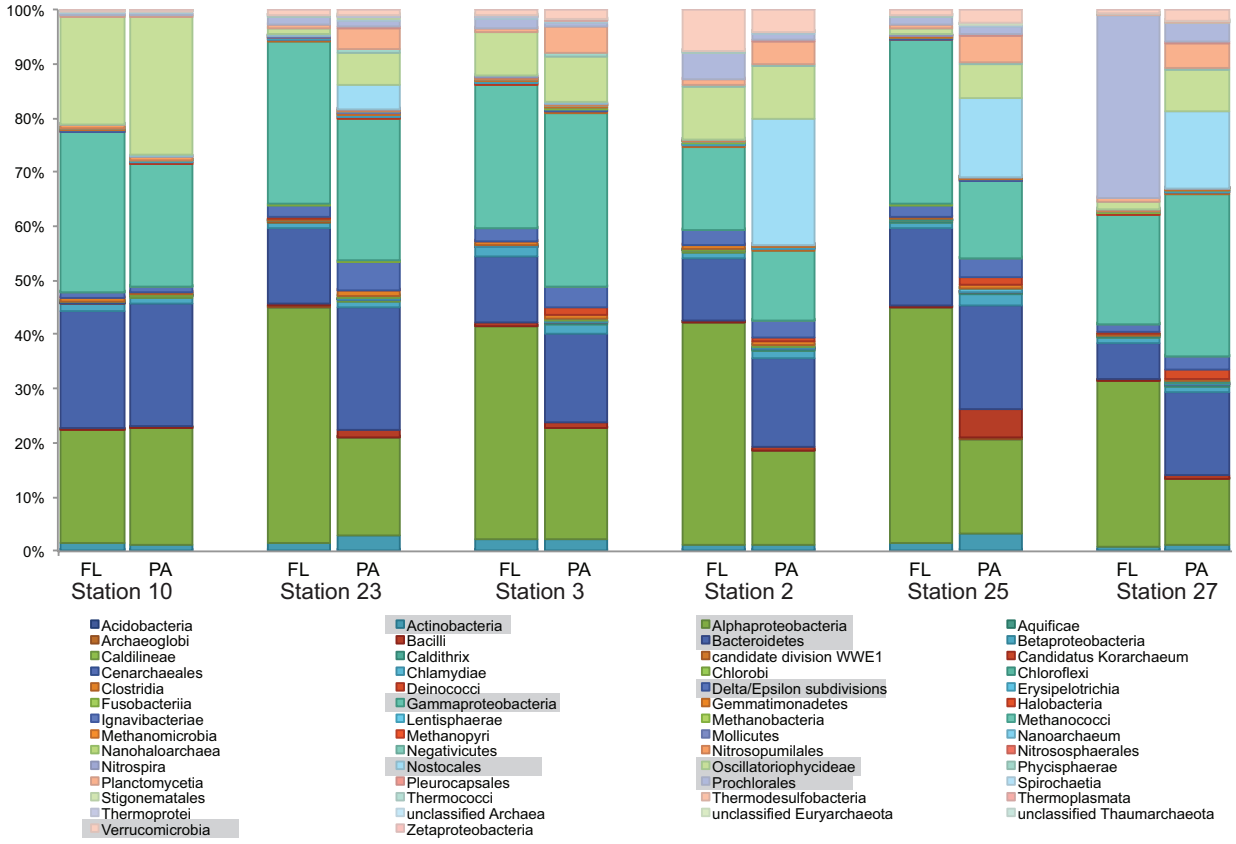


Figure 3.S2. Comparative taxonomic composition of free-living (FL) and particle-associated (PA) prokaryotic communities at six stations along the Amazon River plume in June 2010 as identified from taxonomic assignments of protein-encoding metagenomic reads

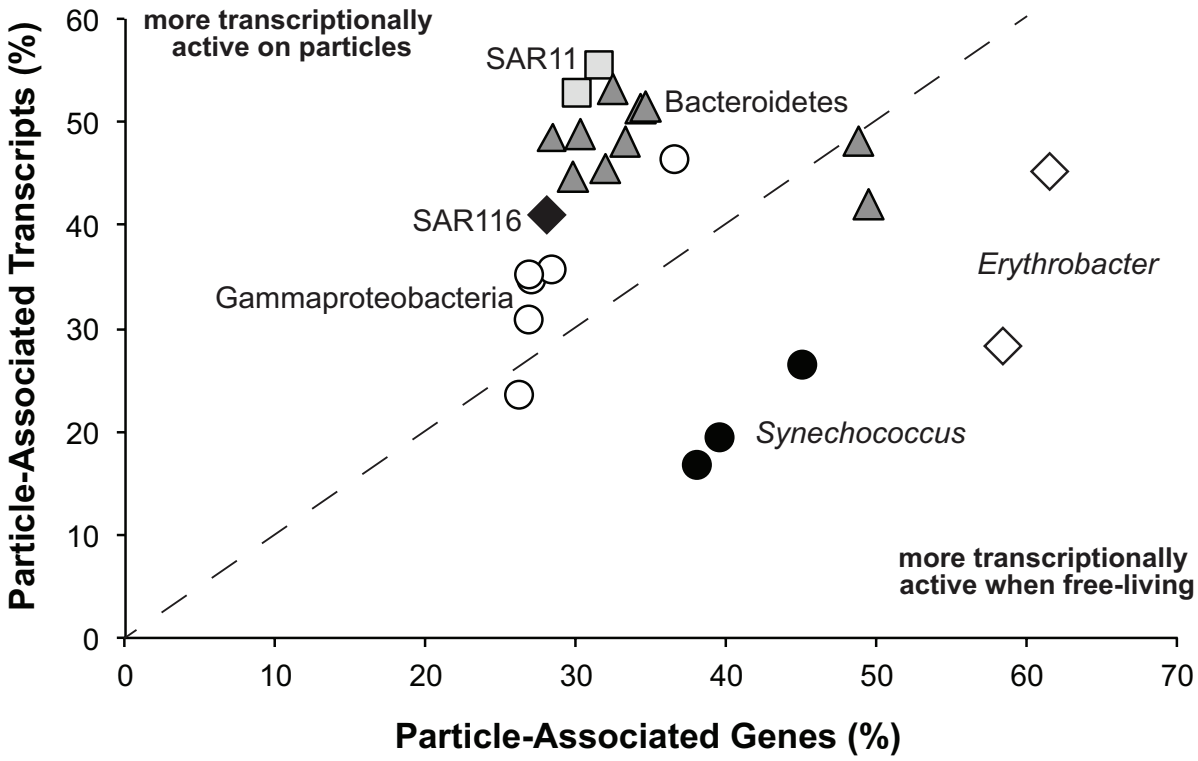


Figure 3.S3. Relative contribution of the top 24 prokaryotic taxa to the genes and transcripts in the particle-associated community of the Amazon River plume.

Table 3.S2. IDs and descriptions of 74 biogeochemically relevant genes.

ID	Gene/Protein	Description
pepA	Glutamyl aminopeptidase	Calcium-stimulated exopeptidase that selectively hydrolyze acidic amino acid residues with a preference for certain Glutamine
pepL	Leucyl aminopeptidase	Hydrolytic exopeptidase with a preference for certain Leucine and hydrophobic amino acid substrates
pepM	Methionyl aminopeptidase	Ubiquitous, essential exopeptidase that cleaves N-terminal Methionine residues from cellular proteins
pepN	Alanyl aminopeptidase	Broad specificity exopeptidase that cleaves amino acid residues from the N-terminus of peptides and protein substrates with a preference for certain Alanine
pepP/pepX	Prolyl aminopeptidase (PAP) or Xaa-Pro aminopeptidase (XAP)	PAPs preferentially cleave N-terminal proline residues from cellular proteins; XAPs are prolidases that catalyze the cleavage of Xaa-Pro dipeptides or act on aminoacyl-hydroxyproline analogs but does not act on Pro-Pro bonds
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	Catalyzes the reversible interconversion of glyceraldehyde-3-phosphate and 1,3-diphosphoglycerate
pgi	Glucose-6-phosphate isomerase	Catalyzes the reversible isomerization of glucose-6-phosphate and fructose-6-phosphate
metF	Methylenetetrahydrofolate reductase	Catalyzes the reduction of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, which is then further used in the final step of methionine biosynthesis
bglA	Beta-glucosidase	Catalyzes the hydrolysis of terminal non-reducing residues in beta-glucosides with release of glucose
pcaH	Protocatechuate 3,4-dioxygenase (3,4-PCD)	Part of the β-ketoadipate pathway that catalyzes the conversion of protocatechuate to citric acid cycle intermediates
vanA	Vanillate demethylase	Involved in vanillate (a lignin-derived monoaryl) degradation by vanillate-utilizing aerobic bacteria
phaC	Polyhydroxyalkanoate synthase	Catalyzes the polymerization of (R)-3-hydroxybutyryl-CoA to form the polyhydroxyalkanoate polymer, the last step in polyhydroxyalkanoate biosynthesis
phaP	Phasin	Coats the surface of the polyhydroxyalkanoate granules, preventing them from coalescing, in turn stabilizing the granules
phaR	Polyhydroxyalkanoate regulator	Transcriptional regulator of polyhydroxyalkanoate biosynthesis
phaZ	Polyhydroxyalkanoate depolymerase	Responsible for intracellular degradation of polyhydroxyalkanoate
amoA	Ammonia monooxygenase subunit A	Catalyzes oxidation of ammonia to hydroxylamine, the first step in the oxidation of ammonia to nitrite.
amtB	Ammonium transporter B	Membrane-bound ammonium/methylammonium transport B protein thought to be required during low [NH(x)]
cphA	Cyanophycin synthetase	Catalyzes the synthesis of cyanophycin granule polypeptide (CGP), which is used as a temporary nitrogen reserve.
cphB	Cyanophycinase	Hydrolyzes cyanophicin to the dipeptide β-Asp-Arg, the first step in making stored amino acids available to the cell

glnA	Glutamine synthetase	Catalyzes the ATP-dependent cycle whereby ammonia is incorporated into glutamate to form glutamine, the first step for ammonia assimilation into organic nitrogen
napA	Nitrate reductase subunit A	Large subunit of the Nap periplasmic nitrate reductase that catalyzes the first step of the denitrification process by conversion of nitrate to nitrite
nirK	Nitrite reductase	Key enzyme in the dissimilatory denitrification process that catalyzes the reduction of nitrite to NO
phnD	Phosphonate ABC transporter, periplasmic binding protein	Periplasmic binding protein of an ABC-type transporter system required for utilization of phosphonates and organophosphorus compounds
phnE	Phosphonate ABC transporter, integral membrane protein	Integral membrane protein of an ABC-type transporter system required for utilization of phosphonates and organophosphorus compounds
phnG	C-P lyase	Part of a membrane associated C-P lyase complex required for hydrolysis of C-P bonds to yield inorganic phosphate and the corresponding hydrocarbons.
phnH	C-P lyase	Part of a membrane associated C-P lyase complex required for hydrolysis of C-P bonds to yield inorganic phosphate and the corresponding hydrocarbons
phnM	C-P lyase	Part of a membrane associated C-P lyase complex required for hydrolysis of C-P bonds to yield inorganic phosphate and the corresponding hydrocarbons
phoA	Alkaline phosphatase	Dephosphorylates organic phosphates and is induced under phosphate starvation as a means to generate free phosphate groups for uptake and use
phoD	Alkaline phosphatase	Belongs to the Pho regulon and codes for codes for alkaline phosphatase D (APaseD), which is a secreted phosphodiesterase
phoU	Alkaline phosphatase	Serves as a signal transduction mediator, being involved in free inorganic P transport and acting as a regulator of the phosphate-specific transport system
phoX	Alkaline phosphatase	Encodes an alkaline phosphatase that uses Ca^{2+} as a cofactor and can be responsible for extracellular phosphatase activity under phosphorus limitation
pitA	Low affinity PO_4 transporter	Low-affinity inorganic phosphate transporter and when inorganic phosphate is abundant, pitA is its major uptake system
ppk1	Polyphosphate kinase	Reversibly synthesizes inorganic polyphosphate, a storage polymer made up of tens to hundreds of phosphate residues linked together by high-energy bonds
ppk2	Polyphosphate kinase	Can polymerize into an actin-like filament concurrent with its reversible synthesis of inorganic polyphosphate
pstA	Phosphate ABC transporter, permease	Membrane permease in the high-affinity phosphate-specific transport (Pst) system that facilitates the transport of phosphate across the membrane
pstC	Phosphate ABC transporter, permease	Membrane permease in the high-affinity phosphate-specific transport (Pst) system that facilitates the transport of phosphate across the membrane
pstS	Phosphate ABC transporter, periplasmic binding protein	Phosphate-binding lipoprotein found within the periplasm of the cell, it is part of the high-affinity phosphate-specific transport (Pst) system

aprA	Adenosine-5'-phosphosulfate reductase (Apr), alpha subunit	Subunit A of dissimilatory adenosine-5'-phosphosulfate (APS) reductase aprAB gene complex that catalyzes the reduction of APS to AMP and sulfite during sulfur reduction
aprB	Adenosine-5'-phosphosulfate reductase (Apr), beta subunit	Subunit B of dissimilatory adenosine-5'-phosphosulfate (APS) reductase aprAB gene complex that catalyzes the reduction of APS to AMP and sulfite during sulfur reduction
cysI	Sulfite reductase	Assimilatory sulfite reduction enzyme that catalyzes the reaction of sulfite to sulfide
cysK	Cysteine synthase	Involved in sulfur metabolism and synthesizes cysteine, the predominant mechanism by which inorganic sulfur is reduced and incorporated into organic compounds
dddD	Type III acyl coenzyme A transferase	Mediates the cleavage of DMSP forming DMS and a 3-carbon compound
dddQ	DMSP lyase	Mediates the cleavage of DMSP forming DMS and a 3-carbon compound
dmdA	DMSP demethylase	Catalyzes the first step in the DMSP demethylation pathway - cleavage of a methyl group from DMSP, eventually resulting in methionine formation and C oxidation
soxA	Cytochrome *c* (diheme)	One of the seven structural proteins involved in sulfur oxidation it combines with the SoxX protein form a cytochrome c complex that is located in the periplasm of the cell and is involved in electron transport
soxB	Sulfate thiohydrolase	One of the seven structural proteins involved in sulfur oxidation it is a type of cytochrome c protein that is located in the periplasm and is involved in the electron transport chain
fliC	Filament protein; flagellin	Structural filament protein, synthesized in the cytosol, composed of monomeric subunits that are polymerized into the long helical filament of the bacterial flagellum
fliF	MS-ring protein	Transmembrane flagellar MS-ring protein, part of the flagellar basal body, that anchors the flagellum to the cytoplasmic membrane
fliG	Flagellar motor switch protein	Essential for assembly, rotation and clockwise/counter-clockwise switching of the bacterial flagellum
motA	Flagellar motor protein	Along with MotB couples flagellar rotation to proton/sodium motive force across the membrane and forms the stator elements of the rotary flagellar machine, required for flagellar rotation
motB	Flagellar motor protein	Along with MotA couples flagellar rotation to proton/sodium motive force across the membrane and forms the stator elements of the rotary flagellar machine, required for flagellar rotation
cheA	Histidine kinase	A cytoplasmic histidine kinase that donates phosphate groups to CheY and CheB, which control flagellar responses and sensory adaptation, respectively

cheB	Methylesterase	A phosphorylation-activated response regulator involved in reversible modification of bacterial chemotaxis receptors. It is required for tumbling movement and regulates tumbling frequency based on perceived tumble-modulating signals (i.e. nutrient concentration) formed by the chemoreceptors
cheR	Methyltransferase	Involved in reversible modification of bacterial chemotaxis receptors, it plays a role in the chemosensory response and adaptation of the cell to chemical stimuli
c*heW*	Signaling protein	Plays a role in coupling methyl-accepting chemotaxis proteins, it regulates motility behavior by two distinct signals, one that stimulates and one that inhibits the intracellular phosphorylation cascade by its effect on the histidine kinase CheA
thiC	Phosphomethylpyrimidine synthase	Catalyzes the pyrimidine branch of the Thiamin biosynthesis pathway, converting 5-aminoimidazole ribonucleotide to hydroxymethylpyrimidine phosphate
thiL	Thiamin-monophosphate kinase	Catalyzes the final step of the thiamin pyrophosphate biosynthesis pathway
pdxH	Pyridoxine 5'-phosphate oxidase	Catalyzes the oxidation of pyridoxine 5'-phosphate to pyridoxal 5'-phosphate in the final step of vitamin B6 biosynthesis
pdxJ	Pyridoxine 5'-phosphate synthase	Catalyzes the condensation of 1-deoxy-d-xylulose-5-phosphate and 1-amino-3-oxo-4-(phosphohydroxy)propan-2-one to pyridoxine 5'-phosphate, a reaction involved in de novo biosynthesis of pyridoxine (vitamin B6) and pyridoxal phosphate
fecA	Ferric dicitrate transporter	TonB-ExbB-dependent ferric-siderophore specific outer membrane receptor protein. When intracellular iron is low, exogenous ferric citrate binds to the FecA receptor, which signals for and aids in translocation of ferric citrate into the cell
feoB	Fe(II) G protein-like transporter	Membrane-bound G protein-like transporter, essential for Fe(II) uptake in bacteria during conditions of low oxygen
Ftr1	High affinity Fe(II) permease	Permease component of a high-affinity Fe(II) uptake system. Expression may be increased during Fe limitation
afuA/futA/ hitA/idiA	Periplasmic Fe(III) ABC transporter	Iron-deficiency-induced, periplasmic iron-binding protein component of a ferric iron ABC-transporter system
afuB/futB	Fe(III) ABC transporter permease	Hydrophobic ferric iron ABC transporter permease protein
PR	Proteorhodopsin	Mediates light-driven proton pumps for harvesting and conversion of light into energy
bchX	Chlorophyll iron protein	Part of a photosynthetic gene cluster involved in redox reactions of the bacteriochlorophyll biosynthesis pathway
pufL	Photosynthetic reaction center subunit L	The light subunit of the photosynthetic reaction center, it helps provide the scaffolding for the chromophore in the reaction center
pufM	Photosynthetic reaction center subunit M	The medium subunit of the photosynthetic reaction center, it helps provide the scaffolding for the chromophore in the reaction center
psbB	Photosystem II CP47 chlorophyll apoprotein	Photosystem II protein that binds to chlorophyll and is found in plants, algae, and cyanobacteria

cpcD	Phycocyanin-associated linker polypeptide	Structural component of the phycobilisome
α-*ca*	α carbonic anhydrase	Zinc metalloenzyme found in bacteria, archaea, and eukaryota that participates in CO_2 diffusion, interconversion of CO_2 and HCO_3 during photosynthesis, pH homeostasis, and ion transport
ε-*ca*	ε carbonic anhydrase	Zinc metalloenzyme found in cyanobacteria carboxysomes and chemolithoautotrophs that participates in CO_2 diffusion, interconversion of CO_2 and HCO_3 during photosynthesis, pH homeostasis, and ion transport
rbcL (IA)	Ribulose 1,5-bisphosphate carboxylase/oxygenase form IA (RuBisCO IA)	Catalyzes the first, rate-limiting step of the Calvin cycle, the primary pathway for photosynthetic carbon reduction in the oceans; rbcL IA has been found in α, β, and γ-proteobacteria, cyanobacteria and prochlorales
rbcL (II)	Ribulose 1,5-bisphosphate carboxylase/oxygenase form II (RuBisCO II)	Catalyzes the first, rate-limiting step of the Calvin cycle, the primary pathway for photosynthetic carbon reduction in the oceans; rbcL II has been found in α, β, and γ-proteobacteria, and eukaryotes

Figure 3.2 Transcript inventories for biogeochemically relevant genes. Bars above the X-axis indicate more transcripts contributed by the free-living prokaryotes (maroon) and those below indicate more contributed by particle-associated (green). Asterisks indicate significant differences between fractions. Gene abbreviations: *pep*, aminopeptidases; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; *pgi*, glucose-6-phosphate isomerase; *metF*, methylenetetrahydrofolate reductase; *bglA*, beta-glucosidase; *pcaH*, protocatechuate 3,4-dioxygenase (3,4-PCD); *vanA*, vanillate demethylase; *pha*, polyhydroxyalkanoate-related; *amoA*, ammonium monooxygenase; *amtB*, ammonium transporter B; *cph,* cyanophycin-related; *glnA*, glutamine synthetase; *napA*, nitrate reductase; *nirK*, nitrite reductase; *phn*, phosphonate assimilation; *pho*, alkaline phosphatase; *pitA*, low affinity PO_4 transporter; *ppk*, polyphosphate kinase; *pst*, high affinity PO_4 transporter; *apr*, adenosine-5'-phosphosulfate reductase; *cysI*, sulfite reductase; *cysK*, cysteine synthase; *ddd*, DMSP lyase; *dmdA*, DMSP demethylase; *sox*, sulfur oxidation; *fli*, flagellar proteins; *mot,* flagellar motor; *che*, chemotaxis-related; *thi*, thiamine synthesis; *pdx*, vitamin B6 synthesis; *fecA*, Fe dicitrate transporter; *feoB*, Fe(II) G protein-like transporter; Ftr1, high affinity Fe(II) permease; *afu*, iron transport; PR, proteorhodopsin; *bchX*, chlorophyll iron protein; *puf*, photosynthetic reaction center; *psbB*, photosystem II; ca, carbonic anhydrases; *rbcL*, ribulose 1,5-bisphosphate carboxylases/oxygenases.

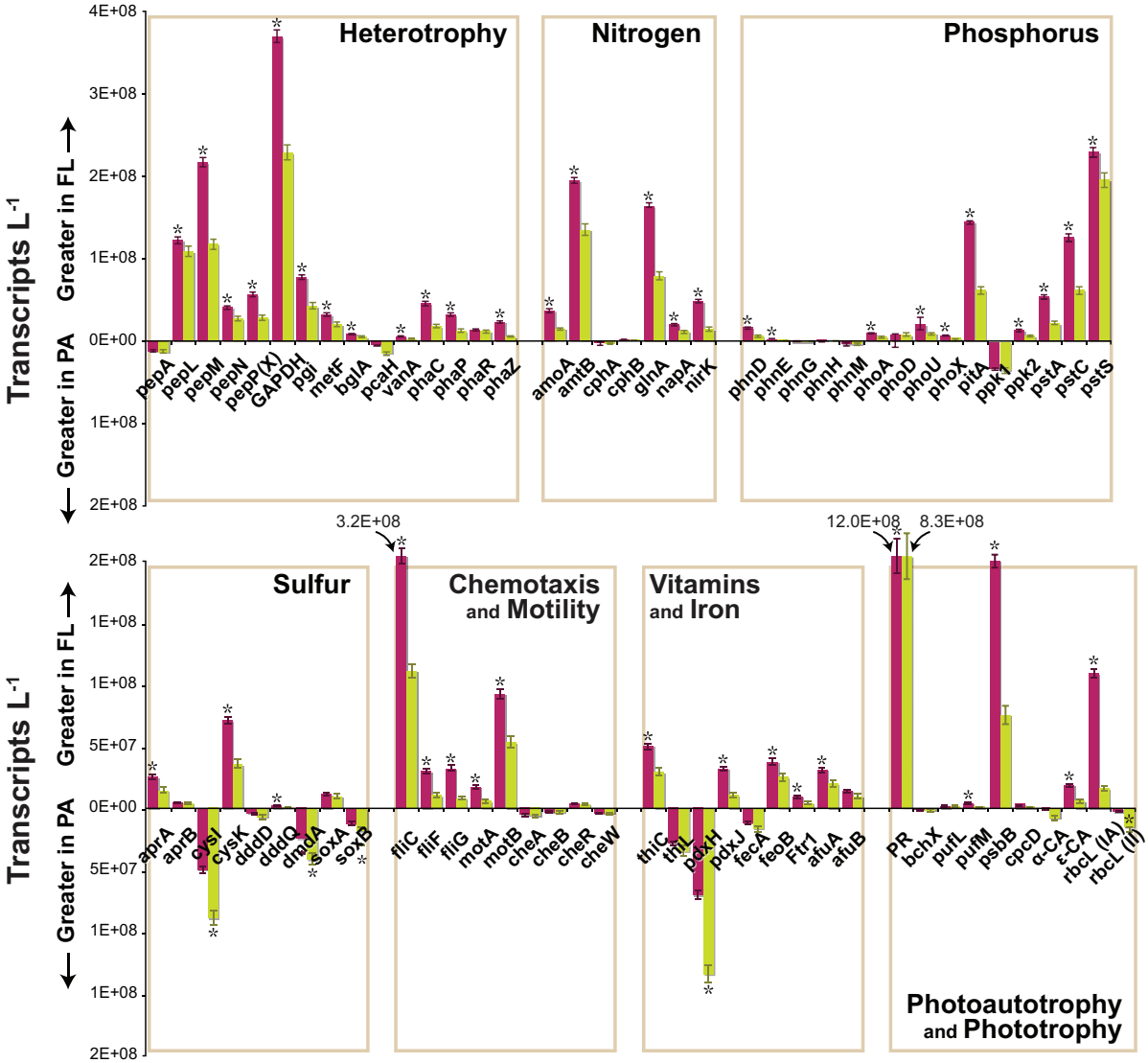

Transcripts L-1
Greater in FL
Greater in PA
4E+08
3E+08
2E+08
1E+08
0E+00
1E+08
2E+08
Heterotrophy
Nitrogen
Phosphorus
pepA
pepL
pepM
pepN
pepP(X)
GAPDH
pgi
metF
bglA
pcaH
vanA
phaC
phaP
phaR
phaZ
amoA
amtB
cphA
cphB
glnA
napA
nirK
phnD
phnE
phnG
phnH
phnM
phoA
phoD
phoU
phoX
pitA
ppk1
ppk2
pstA
pstC
pstS
3.2E+08
12.0E+08
8.3E+08
2E+08
1E+08
1E+08
5E+07
0E+00
5E+07
1E+08
1E+08
2E+08
Sulfur
Chemotaxis and Motility
Vitamins and Iron
Photoautotrophy and Phototrophy
aprA
aprB
cysI
cysK
dddD
dddQ
dmdA
soxA
soxB
fliC
fliF
fliG
motA
motB
cheA
cheB
cheR
cheW
thiC
thiL
pdxH
pdxJ
fecA
feoB
Ftr1
afuA
afuB
PR
bchX
pufL
pufM
psbB
cpcD
α-CA
ε-CA
rbcL (IA)
rbcL (II)

Figure 3.3 Gene expression ratios of biogeochemically-relevant genes. Dotted lines indicate the average expression ratio for all genes in each size fraction. Maroon, free-living; green, particle-associated. Asterisks indicate significant differences between size fractions. Gene abbreviations are as in Fig. 3.2.

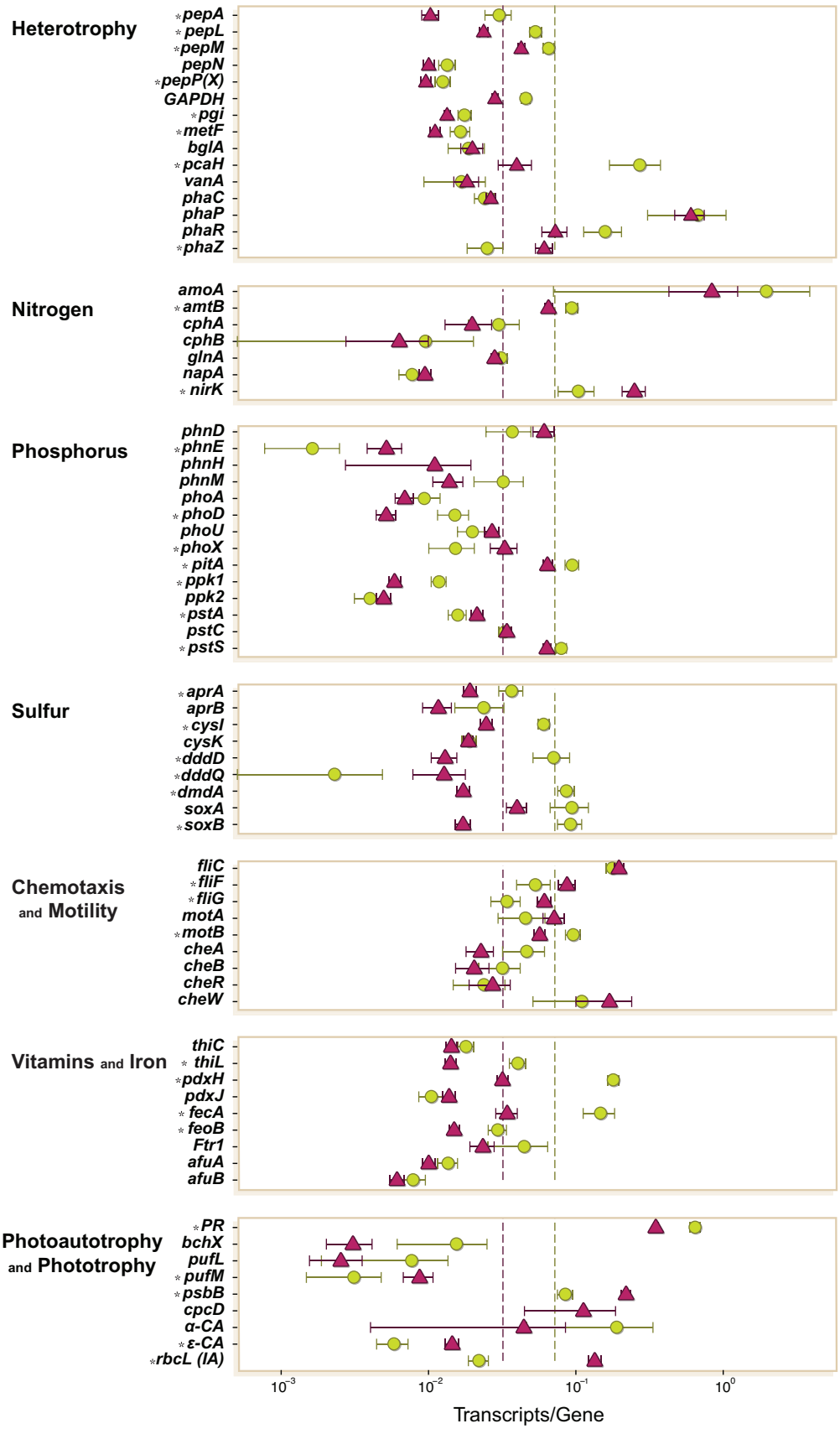
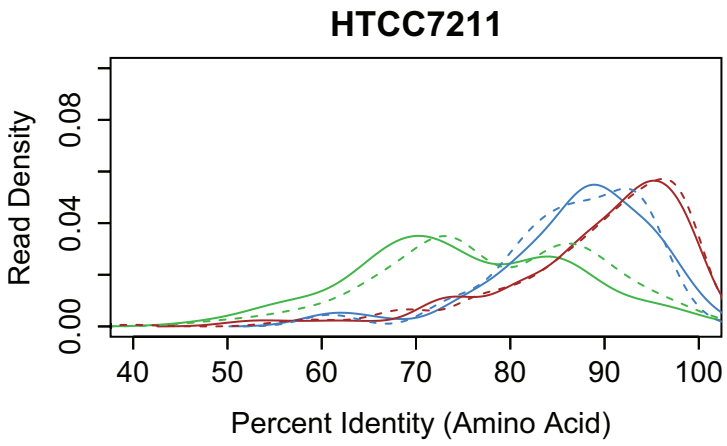
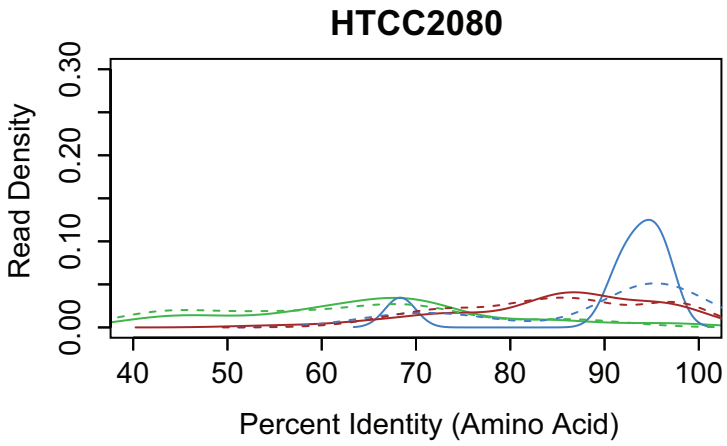

Heterotrophy
*pepA
*pepL
*pepM
pepN
*pepP(X)
GAPDH
*pgi
*metF
bglA
*pcaH
vanA
phaC
phaP
phaR
*phaZ
Nitrogen
amoA
*amtB
cphA
cphB
glnA
napA
*nirK
Phosphorus
phnD
*phnE
phnH
phnM
phoA
*phoD
phoU
*phoX
*pitA
*ppk1
ppk2
*pstA
pstC
*pstS
Sulfur
*aprA
aprB
*cysI
cysK
*dddD
*dddQ
*dmdA
soxA
*soxB
Chemotaxis and Motility
fliC
*fliF
*fliG
motA
*motB
cheA
cheB
cheR
cheW
Vitamins and Iron
thiC
*thiL
*pdxH
pdxJ
*fecA
*feoB
Ftr1
afuA
afuB
Photoautotrophy and Phototrophy
*PR
bchX
pufL
*pufM
*psbB
cpcD
α-CA
*ε-CA
*rbcL (IA)
10^-3
10^-2
10^-1
10^0
Transcripts/Gene

Figure 3.S4. Percent identity of metagenomic reads to the reference genome for universal single-copy genes rpoB (red), gyrB (green), and recA (blue) for the three most abundant taxa in the FL (dashed line) and PA (solid line) communities.


HTCC7211
Read Density
0.00
0.04
0.08
40
50
60
70
80
90
100
Percent Identity (Amino Acid)


HTCC2080
Read Density
0.00
0.10
0.20
0.30
40
50
60
70
80
90
100
Percent Identity (Amino Acid)

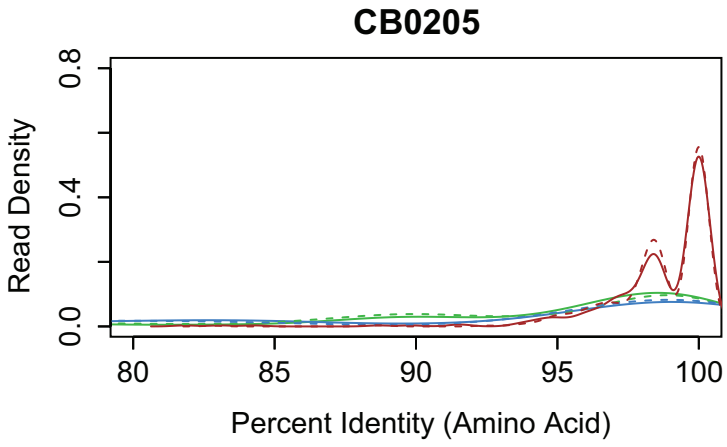

CB0205
Read Density
0.0
0.4
0.8
80
85
90
95
100
Percent Identity (Amino Acid)

Table 3.3. Selected pathways from the KEGG database with different expression ratios (transcripts gene^{-1}) between microenvironments for *Synechococcus* sp. CB0205, *Pelagibacter* sp. HTCC7211, and Gammaproteobacterium HTCC2080. FL, greater in free-living cells; PA, greater in particle-associated cells.

Pathway	Description	CB0205	HTCC7211	HTCC2080
ko00010	Glycolysis/gluconeogenesis	FL	PA	PA
ko00030	Pentose phosphate path.	FL		PA
ko00051	Fructose, mannose metab.	FL	PA	
ko00071	Fatty acid metab.		PA	PA
ko00190	Oxidative phosphorylation	FL	PA	PA
ko00195	Photosynthesis	FL		
ko00196	Photosynthesis - antenna	FL		
ko00230	Purine metabolism	FL	PA	PA
ko00240	Pyrimidine metabolism	FL	PA	PA
ko00253	Tetracycline biosyn.	FL		PA
ko00260	Glycine, serine, threonine metab.		PA	
ko00280	Val, leu, isoleu degrad.	FL	PA	PA
ko00310	Lysine degrad.	FL		
ko00330	Arginine, proline metab.	FL	PA	PA
ko00340	Histidine metab.	FL	PA	
ko00360	Phenylalanine metab.			PA
ko00380	Tryptophan metab.	FL		
ko00430	Taurine metab.		PA	
ko00480	Glutathione metab.	FL	PA	PA
ko00500	Starch, sucrose metab.	FL		FL
ko00521	Streptomycin biosyn.	FL		
ko00523	Polyketide biosyn.	FL		
ko00620	Pyruvate metabolism	FL	PA	PA
ko00630	Glyoxylate, dicarboxylate metab.	FL	PA	PA
ko00640	Propanoate metab.		PA	
ko00650	Butanoate metab.	FL		PA
ko00670	One carbon pool by folate	FL	PA	
ko00710	Carbon fixation	FL		
ko00730	Thiamine metab.	FL		
ko00740	Riboflavin metab.	FL		
ko00760	Nicotinate metab.	FL	PA	PA
ko00770	Pantothenate, CoA biosyn.	FL	PA	
ko00780	Biotin metab.	FL		
ko00790	Folate biosyn.	FL	PA	
ko00860	Porphyrin, chlorophyll metab.	FL	PA	PA
ko00910	Nitrogen metab.	FL	PA	PA
ko00920	Sulfur metab.	FL	PA	
ko01053	Biosyn. siderophore NRPs			PA
ko01055	Biosyn. vancomycin	FL		
ko02020	Two-component systems	FL	PA	PA
ko02040	Flagellar assembly			FL
ko03030	DNA replication	FL	PA	PA
ko03060	Protein export	FL	PA	
ko03430	Mismatch repair	FL	PA	PA
ko03440	Homologous recomb.	FL	PA	PA

Table S3.3. Gene expression ratios in the *Synechococcus* sp. CB0205 bin.

NCBI GI	Protein Description	Mean Expression Ratio (FL)	Mean Expression Ratio (PA)	PA Regulation
497998808	hypothetical protein SCB02_00319	0.0159	0.0511	Up
497998935	hypothetical protein SCB02_00599	0.0039	0.016	Up
497999395	hypothetical protein SCB02_01633	0.0223	0.1099	Up
497999396	hypothetical protein SCB02_01638	0.0421	0.1981	Up
498000149	tRNA nucleotidyltransferase/poly(A) polymerase	0.014	0.0326	Up
498000159	glycosyltransferase	0.002	0.011	Up
498000180	hypothetical protein SCB02_03348	0.0304	0.1832	Up
498000306	queuine tRNA-ribosyltransferase	0.0105	0.0296	Up
498000620	hypothetical protein SCB02_04388	0.0309	0.051	Up
498000622	dihydrobiliverdin:ferredoxin oxidoreductase	0.0305	0.072	Up
498000774	NAD(P)H-quinone oxidoreductase subunit N	0.0103	0.0485	Up
498000927	putative enolase-phosphatase E-1	0.0086	0.0233	Up
498001071	hypothetical protein SCB02_05303	0.3047	0.7016	Up
498001275	outer envelope membrane protein	0.0193	0.0296	Up
498001991	hypothetical protein SCB02_07563	0.0106	0.0325	Up
498002581	sugar kinase	0.0242	0.0526	Up
498002588	hypothetical protein SCB02_09140	0.0031	0.0116	Up
498002720	serine:pyruvate/alanine:glyoxylate aminotransferase	0.1162	0.2943	Up
498002799	thioredoxin peroxidase	0.0792	0.2075	Up
498003008	hypothetical protein SCB02_10519	0.0041	0.0404	Up
498003662	alpha-glucosidase	0.01	0.0388	Up
498003678	superfamily II DNA/RNA helicase	0.0876	0.1566	Up
498003936	putative L-cysteine/cystine lyase	0.023	0.0477	Up
497992572	hypothetical protein SCB02_00062	0.0039	0.0015	Down
497998706	protein phosphatase 2C	0.0088	0.003	Down
497998744	30S ribosomal protein S14	0.0903	0.0132	Down
497998754	polynucleotide phosphorylase/polyadenylase	0.017	0.0051	Down
497998755	hypothetical protein SCB02_00354	0.0159	0.0056	Down
497998831	hypothetical protein SCB02_00369	0.01	0.0031	Down
497998836	methionine sulfoxide reductase B	0.0263	0.0072	Down
497998838	putative lipidA disaccharide synthetase	0.0134	0.0035	Down
497998840	acetyl-CoA carboxylase biotin carboxylase subunit	0.0127	0.0033	Down
497998842	hypothetical protein SCB02_00389	0.0148	0.004	Down
497998855	ATPase	0.0201	0.0099	Down
497998861	FeS assembly ATPase SufC	0.0196	0.0059	Down
497998888	ferredoxin-thioredoxin reductase catalytic chain	0.0203	0.0088	Down
497998890	transcriptional regulator	0.0176	0.0046	Down
497998893	phycobilisome rod-core linker polypeptide cpcG	0.0612	0.0169	Down
497998906	Rho termination factor domain-containing protein	0.0246	0.0088	Down
497998908	hypothetical protein SCB02_00569	0.0138	0.0034	Down
497998918	peroxiredoxin	0.0076	0.0024	Down
497998942	peroxiredoxin	0.0386	0.0167	Down
497998956	periplasmic trypsin-like serine protease	0.0301	0.0048	Down
497998977	RND family outer membrane efflux protein	0.0109	0.0049	Down
497999018	photosystem II complex extrinsic protein precursor U	0.0269	0.0168	Down
497999025	hypothetical protein SCB02_00786	0.0503	0.0175	Down
497999029	2-methylthioadenine synthetase	0.0049	0.0022	Down
497999031	major facilitator superfamily permease	0.0084	0.002	Down
497999034	Fe-S oxidoreductase	0.0108	0.0068	Down
497999038	6-pyruvoyl tetrahydropterin synthase	0.0231	0.0053	Down
497999047	zeta-carotene desaturase	0.0166	0.0058	Down
497999080	porin	0.0277	0.0136	Down
497999098	porin	0.0505	0.0211	Down
497999102	glycosyl transferase family protein	0.0045	0.0015	Down
497999105	glycosyltransferase	0.0069	0.0034	Down
497999112	hypothetical protein SCB02_00966	0.0083	0.0024	Down
497999113	photosystem II D2 protein (photosystem q(a) protein)	0.043	0.0195	Down
497999117	copper-transporting ATPase	0.006	0.0015	Down
497999144	RNA-binding protein	0.0412	0.0168	Down
497999166	phytoene synthase	0.0186	0.0041	Down
497999169	phytoene dehydrogenase	0.0156	0.0019	Down
497999172	NAD(P)H-quinone oxidoreductase subunit F	0.0383	0.0084	Down
497999182	NAD(P)H-quinone oxidoreductase subunit 4	0.0112	0.004	Down
497999184	nucleoside-diphosphate-sugar transferase	0.0206	0.0069	Down

497999190	methylenetetrahydrofolate reductase	0.0107	0.0023	Down
497999193	NADH:ubiquinone oxidoreductase subunit J	0.0489	0.0082	Down
497999206	NADH:ubiquinone oxidoreductase subunit H	0.0332	0.0109	Down
497999210	citrate synthase	0.0112	0.0031	Down
497999212	NAD(P)H-quinone oxidoreductase subunit H	0.0214	0.0076	Down
497999253	TM2 domain-containing protein	0.0062	0.0027	Down
497999262	glutathione synthetase	0.0128	0.0022	Down
497999283	arginyl-tRNA synthetase	0.0114	0.0016	Down
497999304	guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	0.0126	0.0047	Down
497999312	two-component response regulator	0.0068	0.002	Down
497999338	50S ribosomal protein L10	0.0666	0.0183	Down
497999369	50S ribosomal protein L1	0.0394	0.0072	Down
497999371	50S ribosomal protein L11	0.1236	0.0245	Down
497999373	transcription antitermination protein NusG	0.0385	0.0065	Down
497999374	putative kinase	0.0075	0.0034	Down
497999379	aspartyl/glutamyl-tRNA amidotransferase subunit B	0.0091	0.002	Down
497999386	arginine decarboxylase	0.0041	0.0022	Down
497999391	putative flavoprotein	0.01	0.0023	Down
497999419	hypothetical protein SCB02_01718	0.2035	0.0677	Down
497999425	glycine dehydrogenase	0.0111	0.0018	Down
497999427	glycine cleavage system protein H	0.0104	0.0037	Down
497999429	50S ribosomal protein L9	0.0245	0.0053	Down
497999439	DnaB replicative helicase	0.0076	0.003	Down
497999442	putative membrane protein of ABC transport system	0.012	0.0023	Down
497999576	putative urea ABC transporter, substrate binding protein	0.0335	0.0057	Down
497999578	hypothetical protein SCB02_02052	0.0503	0.0038	Down
497999603	uroporphyrin-III c-methyltransferase	0.0259	0.0022	Down
497999605	ferredoxin-nitrite reductase	0.025	0.0084	Down
497999612	formate and nitrite transporters	0.0995	0.0158	Down
497999614	hypothetical protein SCB02_02092	0.0153	0.0051	Down
497999620	assimilatory nitrate reductase (ferredoxin) precursor	0.0065	0.0017	Down
497999627	nitrate permease NapA	0.0123	0.0016	Down
497999629	major facilitator superfamily permease	0.0041	0.0013	Down
497999648	polyphosphate kinase	0.0025	0.0008	Down
497999650	formyltetrahydrofolate deformylase	0.0104	0.0027	Down
497999669	molecular chaperone DnaK	0.0127	0.0024	Down
497999678	argininosuccinate synthase	0.0205	0.0016	Down
497999690	excinuclease ABC subunit B	0.0057	0.0009	Down
497999730	trigger factor	0.0084	0.0028	Down
497999746	ATP-dependent Clp protease proteolytic subunit ClpP	0.0117	0.003	Down
497999748	glycosyl transferase, group 2 family protein	0.0246	0.0057	Down
497999757	sulfolipid biosynthesis protein (UDP-sulfoquinovose synthase)	0.0458	0.0071	Down
497999772	putative penicillin-binding protein	0.0038	0.001	Down
497999776	GMP synthase	0.0286	0.0025	Down
497999781	elongation factor P	0.0301	0.0065	Down
497999813	putative cyclophilin-type peptidyl-prolyl cis-trans isomerase	0.0071	0.0033	Down
497999815	glyceraldehyde 3-phosphate dehydrogenase	0.0422	0.0112	Down
497999820	UDP-N-acetylmuramate--L-alanine ligase	0.0072	0.0013	Down
497999822	chaperone protein DnaJ	0.0079	0.002	Down
497999832	Na+/H+ antiporter	0.0418	0.0114	Down
497999848	phosphorylase	0.008	0.0041	Down
497999852	acyl carrier protein	0.0217	0.0066	Down
497999932	3-oxoacyl-acyl-carrier-protein] synthase II	0.0706	0.01	Down
497999934	transketolase	0.0321	0.0041	Down
497999936	thiamine biosynthesis protein ThiC	0.0232	0.014	Down
497999938	hypothetical protein SCB02_02826	0.0233	0.0053	Down
497999940	cysteine synthase A	0.0081	0.0018	Down
497999957	lysyl-tRNA synthetase	0.0068	0.0011	Down
497999971	signal recognition particle-docking protein FtsY	0.0118	0.0031	Down
497999977	hypothetical protein SCB02_02916	0.0352	0.0082	Down
497999980	hypothetical protein SCB02_02921	0.0167	0.0031	Down
497999982	DNA gyrase/topoisomerase IV, subunit A	0.0065	0.0024	Down
497999988	excinuclease ABC subunit A	0.0066	0.0022	Down
497999990	DNA repair ATPase RecN	0.0043	0.0013	Down
497999992	DNA polymerase III subunit beta	0.0067	0.004	Down
498000001	phosphoribosylformylglycinamidine synthase II	0.0169	0.0051	Down
498000004	amidophosphoribosyltransferase	0.0207	0.0061	Down
498000007	photosystem II reaction center protein Z	0.0449	0.0142	Down
498000019	preprotein translocase subunit SecA	0.011	0.0038	Down
498000035	Mg2+ transporter	0.0113	0.0058	Down

498000058	S-adenosyl-L-homocysteine hydrolase	0.0151	0.0022	Down
498000087	rod shape-determining protein MreB	0.0051	0.0018	Down
498000093	argininosuccinate lyase	0.0092	0.0023	Down
498000103	hypothetical protein SCB02_03193	0.013	0.0023	Down
498000109	glycosyl transferase family protein	0.0089	0.0031	Down
498000114	amino acid permease-associated region	0.0086	0.002	Down
498000123	carbamoyltransferase	0.0127	0.0045	Down
498000128	hypothetical protein SCB02_03238	0.021	0.0032	Down
498000131	isocitrate dehydrogenase	0.0175	0.0049	Down
498000135	Heme oxygenase (decyclizing)	0.0505	0.0226	Down
498000140	glycosyl transferase family protein	0.0181	0.0081	Down
498000142	ABC-type multidrug transport system, ATPase and permease components	0.0154	0.0067	Down
498000145	glycosyltransferase	0.0087	0.0039	Down
498000172	hypothetical protein SCB02_03393	0.008	0.0033	Down
498000199	hypothetical protein SCB02_03418	0.013	0.0035	Down
498000210	GDP-mannose 4,6-dehydratase	0.0129	0.0051	Down
498000212	dTDP-glucose 4-6-dehydratase-like protein	0.0233	0.006	Down
498000216	UDP-glucose 6-dehydrogenase	0.0087	0.0021	Down
498000219	photosystem II reaction center protein J	0.0666	0.0133	Down
498000221	photosystem II reaction center L	0.0997	0.0101	Down
498000224	cytochrome b559 subunit beta	0.0402	0.0162	Down
498000226	Ycf48-like protein	0.0214	0.0113	Down
498000229	rubredoxin	0.0167	0.0039	Down
498000231	NADH:ubiquinone oxidoreductase subunit A	0.0319	0.0051	Down
498000233	NADH dehydrogenase subunit B	0.052	0.007	Down
498000235	magnesium chelatase, ATPase subunit D	0.0063	0.0024	Down
498000249	DegT/DnrJ/EryC1/StrS aminotransferase family protein	0.0085	0.0025	Down
498000257	imidazoleglycerol-phosphate dehydratase	0.014	0.0044	Down
498000269	lignostilbene-alpha,beta-dioxygenase and related enzyme	0.004	0.0016	Down
498000271	hypothetical protein SCB02_03573	0.0178	0.0048	Down
498000276	phosphoglucomutase/phosphomannomutase family protein	0.0079	0.0025	Down
498000281	hypothetical protein SCB02_03653	0.174	0.0597	Down
498000312	hypothetical protein SCB02_03668	0.0267	0.0095	Down
498000316	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	0.01	0.0017	Down
498000318	serine hydroxymethyltransferase	0.0134	0.0032	Down
498000325	glycosyltransferase	0.0099	0.004	Down
498000326	isopropylmalate isomerase large subunit	0.0177	0.0038	Down
498000328	alpha mannosidase	0.0053	0.0028	Down
498000331	Sec-independent protein translocase protein TatA	0.0601	0.0215	Down
498000334	hypothetical protein SCB02_03788	0.0051	0.0021	Down
498000338	CRP family global nitrogen regulatory protein	0.0131	0.0032	Down
498000341	putative inorganic carbon transporter/0-antigen polymerase (ICT/OAP) family protein	0.0051	0.0018	Down
498000366	hypothetical protein SCB02_03863	0.019	0.0099	Down
498000372	beta carotene hydroxylase	0.0176	0.0037	Down
498000381	coenzyme A biosynthesis bifunctional protein CoaBC (DNA/pantothenate metabolism flavoprotein)	0.0101	0.0026	Down
498000401	photosystem II manganese-stabilizing protein	0.034	0.0091	Down
498000406	Sulfate adenylyltransferase	0.031	0.0083	Down
498000408	cell division protein FtsH	0.033	0.0085	Down
498000410	chorismate synthase	0.0192	0.0038	Down
498000415	Fe-S oxidoreductase	0.0115	0.0049	Down
498000432	ATP-dependent Clp protease adaptor protein ClpS	0.0168	0.0034	Down
498000435	L,L-diaminopimelate aminotransferase	0.0065	0.0014	Down
498000437	ribonucleases G and E	0.0254	0.0032	Down
498000443	cyclopropane-fatty-acyl-phospholipid synthase family protein	0.013	0.0023	Down
498000460	Lon protease domain-containing protein	0.0238	0.0049	Down
498000463	elongation factor Tu	0.1445	0.0376	Down
498000466	elongation factor G	0.1601	0.0211	Down
498000468	30S ribosomal protein S7	0.0486	0.0057	Down
498000472	hypothetical protein SCB02_04086	0.004	0.002	Down
498000478	ferredoxin-dependent glutamate synthase	0.0243	0.0041	Down
498000481	lipoyl synthase	0.0097	0.0048	Down
498000485	photosystem I P700 chlorophyll a apoprotein A1	0.0941	0.0518	Down
498000493	photosystem I P700 chlorophyll a apoprotein A1	0.1211	0.0508	Down
498000503	photosystem I P700 chlorophyll a apoprotein A2	0.0839	0.0468	Down
498000505	universal stress protein	0.0212	0.0116	Down
498000517	photosystem II PsbB protein	0.1362	0.0488	Down
498000524	30S ribosomal protein S1	0.0674	0.0091	Down

498000533	S-adenosylmethionine synthetase	0.0191	0.0053	Down
498000537	phycobilisome rod-core linker polypeptide (L-RC 28.5)	0.0741	0.0141	Down
498000559	hypothetical protein SCB02_04258	0.062	0.0092	Down
498000561	phycobilisome linker polypeptide, C-phycoerythrin class I-associated protein	0.0557	0.0101	Down
498000563	phycobilisome linker polypeptide, C-phycoerythrin-associated protein	0.0309	0.0048	Down
498000565	phycobilisome linker polypeptide, C-phycoerythrin class I-associated protein	0.044	0.0072	Down
498000567	phycoerythrin linker gene region	0.0043	0.0016	Down
498000574	hypothetical protein SCB02_04318	0.0059	0.0014	Down
498000587	bilin biosynthesis protein CpeZ	0.013	0.0016	Down
498000609	C-phycoerythrin class I alpha chain	0.31	0.1243	Down
498000613	C-phycoerythrin class I beta chain	0.3171	0.0625	Down
498000615	phycoerythrobilin:ferredoxin oxidoreductase	0.0109	0.0016	Down
498000624	R-phycocyanin II beta chain	0.0537	0.0154	Down
498000633	phycocyanin, alpha subunit	0.1074	0.0657	Down
498000635	HEAT repeat-containing PBS lyase	0.0096	0.0028	Down
498000638	phosphoribosylaminoimidazole synthetase	0.0229	0.003	Down
498000647	hypothetical protein SCB02_04445	0.0253	0.0095	Down
498000650	hypothetical protein SCB02_04525	0.0182	0.0074	Down
498000693	cell division protein	0.0119	0.0067	Down
498000702	photosystem I protein PsaD	0.0802	0.0258	Down
498000704	anthranilate synthase component I	0.0099	0.0018	Down
498000705	phosphoenolpyruvate carboxylase	0.0129	0.0017	Down
498000707	hypothetical protein SCB02_04605	0.0124	0.0053	Down
498000716	hypothetical protein SCB02_04620	0.0277	0.0083	Down
498000722	hypothetical protein SCB02_04659	0.0374	0.0149	Down
498000753	recombinase A	0.0077	0.003	Down
498000756	50S ribosomal protein L3	0.1743	0.0191	Down
498000776	30S ribosomal protein S3	0.0864	0.012	Down
498000787	50S ribosomal protein L29	0.0987	0.0277	Down
498000790	50S ribosomal protein L5	0.0403	0.0066	Down
498000796	30S ribosomal protein S8	0.0558	0.0161	Down
498000798	50S ribosomal protein L6	0.1213	0.0128	Down
498000800	50S ribosomal protein L15	0.0536	0.0228	Down
498000808	preprotein translocase subunit SecY	0.0331	0.0075	Down
498000812	adenylate kinase	0.0279	0.004	Down
498000814	30S ribosomal protein S13	0.1065	0.0301	Down
498000833	DNA-directed RNA polymerase subunit alpha	0.0368	0.0069	Down
498000837	50S ribosomal protein L13	0.0908	0.0204	Down
498000842	30S ribosomal protein S9	0.0752	0.0099	Down
498000843	alanine racemase	0.0069	0.0028	Down
498000854	photosystem I reaction centre subunit XI	0.0653	0.0091	Down
498000870	carboxyl-terminal protease	0.0043	0.0012	Down
498000907	cytochrome b6	0.0625	0.0259	Down
498000909	cytochrome b6-f complex subunit IV	0.0792	0.0509	Down
498000911	Calcium/calmodulin dependent protein kinase II association-domain protein	0.0041	0.0016	Down
498000925	photosystem I reaction center subunit IV	0.0448	0.0171	Down
498000931	cell division protein sepF	0.0347	0.0116	Down
498000980	putative transporter, membrane component	0.0065	0.002	Down
498000989	GPH family sugar transporter	0.0095	0.0021	Down
498000991	DNA-binding protein HU	0.042	0.0161	Down
498001008	iron ABC transporter ATP-binding protein	0.0044	0.0019	Down
498001020	GTPase CgtA	0.0075	0.0035	Down
498001022	translation initiation factor IF-2	0.0226	0.0066	Down
498001058	trypsin-like serine protease	0.0207	0.0082	Down
498001079	ribose-5-phosphate isomerase A	0.0363	0.0037	Down
498001081	DNA-directed RNA polymerase subunit beta	0.0417	0.0059	Down
498001088	DNA-directed RNA polymerase subunit gamma	0.0263	0.0051	Down
498001089	DNA-directed RNA polymerase subunit beta'	0.0316	0.0074	Down
498001091	ribosomal RNA large subunit methyltransferase N	0.0199	0.0038	Down
498001095	hypothetical protein SCB02_05393	0.0472	0.0186	Down
498001101	hypothetical protein SCB02_05398	0.06	0.0101	Down
498001103	DEAD/DEAH box helicase-like protein	0.0067	0.0029	Down
498001126	adenylosuccinate lyase	0.0107	0.004	Down
498001131	nitrogen regulatory protein P-II	0.0297	0.0108	Down
498001137	phycobilisome linker polypeptide, allophycocyanin-associated, core (LC 7.7)	0.0536	0.0136	Down
498001147	allophycocyanin subunit beta	0.0653	0.025	Down

498001150	allophycocyanin alpha chain	0.2001	0.0351	Down
498001152	anchor polypeptide LCM	0.0408	0.0083	Down
498001154	H+-transporting ATP synthase	0.3173	0.0461	Down
498001158	F0F1 ATP synthase subunit A	0.1588	0.0272	Down
498001160	F0F1 ATP synthase subunit B'	0.0835	0.0131	Down
498001164	F0F1 ATP synthase subunit B	0.0473	0.0085	Down
498001165	F0F1 ATP synthase subunit alpha	0.0547	0.0162	Down
498001167	F0F1 ATP synthase subunit epsilon	0.0341	0.0088	Down
498001203	F0F1 ATP synthase subunit beta	0.0551	0.0099	Down
498001204	chaperonin GroEL	0.0131	0.003	Down
498001206	phosphoglyceromutase	0.0152	0.0054	Down
498001212	hypothetical protein SCB02_05778	0.0089	0.004	Down
498001232	hypothetical protein SCB02_05853	0.0071	0.0032	Down
498001258	50S ribosomal protein L21	0.0617	0.0124	Down
498001260	phosphoribosylamine--glycine ligase	0.0066	0.0033	Down
498001271	UDP-3-O-3-hydroxymyristoyl] N-acetylglucosamine deacetylase	0.0117	0.0018	Down
498001277	(3R)-hydroxymyristoyl-ACP dehydratase	0.0336	0.0086	Down
498001279	peptide methionine sulfoxide reductase MsrA	0.0083	0.0017	Down
498001285	hypothetical protein SCB02_05933	0.0151	0.0052	Down
498001287	multifunctional aminopeptidase A	0.0077	0.0027	Down
498001291	tyrosyl-tRNA synthetase	0.0127	0.0037	Down
498001301	long-chain acyl-CoA synthetase	0.017	0.0025	Down
498001319	branched-chain alpha-keto acid dehydrogenase subunit E2	0.0212	0.0097	Down
498001323	branched-chain alpha-keto acid dehydrogenase subunit E2	0.0225	0.0038	Down
498001335	apolipoprotein N-acyltransferase	0.0092	0.0036	Down
498001350	hypothetical protein SCB02_06085	0.0079	0.0021	Down
498001353	hypothetical protein SCB02_06100	0.0046	0.0022	Down
498001358	cell division protein FtsH3	0.0106	0.004	Down
498001362	putative riboflavin kinase/FAD synthase	0.0231	0.0117	Down
498001409	thiamine-phosphate pyrophosphorylase	0.0078	0.002	Down
498001412	hypothetical protein SCB02_06212	0.0282	0.0146	Down
498001417	30S ribosomal protein S16	0.153	0.0205	Down
498001452	pyruvate dehydrogenase E1 alpha subunit	0.0247	0.0068	Down
498001458	tRNA-specific 2-thiouridylase MnmA	0.0103	0.003	Down
498001477	photosystem II CP43 protein	0.0652	0.0254	Down
498001497	peptidyl-prolyl cis-trans isomerase	0.0261	0.0076	Down
498001511	3-beta hydroxysteroid dehydrogenase/isomerase family protein	0.0481	0.0143	Down
498001523	hypothetical protein SCB02_06467	0.017	0.0043	Down
498001539	thioredoxin reductase	0.0095	0.0046	Down
498001561	hypothetical protein SCB02_06507	0.0368	0.0177	Down
498001563	hypothetical protein SCB02_06522	0.0162	0.0037	Down
498001570	putative nicotinamide nucleotide transhydrogenase, subunit alpha 2	0.0141	0.003	Down
498001573	putative nicotinamide nucleotide transhydrogenase, subunit beta	0.0403	0.012	Down
498001575	hypothetical protein SCB02_06557	0.022	0.0044	Down
498001582	glutamyl-tRNA synthetase	0.0062	0.0032	Down
498001607	50S ribosomal protein L19	0.182	0.0241	Down
498001609	methionine aminopeptidase	0.0167	0.0061	Down
498001614	hypothetical protein SCB02_06642	0.0224	0.0046	Down
498001618	TRAP-T family tripartite transporter	0.0127	0.0021	Down
498001630	glutamate-1-semialdehyde aminotransferase	0.0365	0.0092	Down
498001636	putative glycolate oxidase subunit GlcD	0.0084	0.0022	Down
498001638	hypothetical protein SCB02_06712	0.0145	0.0035	Down
498001645	permease	0.0069	0.0015	Down
498001686	hypothetical protein SCB02_06806	0.0127	0.0043	Down
498001688	bifunctional phosphoribosyl-AMP cyclohydrolase/phosphoribosyl-ATP pyrophosphatase protein	0.024	0.0118	Down
498001696	RND family multidrug efflux protein	0.0026	0.0007	Down
498001701	putative membrane protein	0.0099	0.0051	Down
498001706	hypothetical protein SCB02_06886	0.0364	0.0062	Down
498001723	AraC-type regulatory protein, putative	0.0081	0.0009	Down
498001768	Methyltransferase type 11	0.0119	0.0031	Down
498001854	hypothetical protein SCB02_07278	0.0033	0.0014	Down
498001878	Rho termination factor domain-containing protein	0.0454	0.0189	Down
498001886	hypothetical protein SCB02_07333	0.0696	0.0052	Down
498001900	hypothetical protein SCB02_07428	0.0132	0.0051	Down
498001937	hypothetical protein SCB02_07468	0.0537	0.0127	Down
498001954	hypothetical protein SCB02_07518	0.0098	0.0023	Down
498001974	hypothetical protein SCB02_07583	0.0109	0.0044	Down
498002000	ABC-type multidrug transport system, ATPase and permease components	0.0195	0.0113	Down

498002084	ATP phosphoribosyltransferase catalytic subunit	0.0163	0.0081	Down
498002335	CO2 hydration protein ChpX	0.0101	0.0044	Down
498002355	NAD(P)H-quinone oxidoreductase chain 4	0.0366	0.0079	Down
498002357	NAD(P)H-quinone oxidoreductase subunit F	0.03	0.006	Down
498002359	putative carboxysome shell polypeptide CsoS3	0.0098	0.003	Down
498002378	carboxysome shell polypeptide, CsoS2	0.0262	0.0044	Down
498002379	ribulose bisphosphate carboxylase, small chain	0.2919	0.0674	Down
498002381	ribulose bisophosphate carboxylase	0.1115	0.0202	Down
498002383	carboxysome shell peptide	0.1185	0.0157	Down
498002398	MFS superfamily sulfate permease	0.0256	0.0065	Down
498002401	Ca2+/Na+ antiporter	0.0223	0.0039	Down
498002407	ammonium transporter	0.1559	0.0512	Down
498002412	di/tricarboxylate transporter	0.0167	0.0046	Down
498002414	hypothetical protein SCB02_08605	0.0086	0.0032	Down
498002415	Trk family sodium transporter	0.0136	0.0032	Down
498002418	light-independent protochlorophyllide reductase subunit B	0.0194	0.0077	Down
498002419	protochlorophyllide reductase iron-sulfur ATP-binding protein	0.1673	0.069	Down
498002438	Zn-dependent membrane associated protease	0.0337	0.0056	Down
498002443	GTP cyclohydrolase I	0.0369	0.0134	Down
498002445	short chain dehydrogenase	0.0089	0.0028	Down
498002449	acyl-ACP reductase	0.0131	0.0054	Down
498002451	aldehyde decarbonylase	0.027	0.0071	Down
498002456	small subunit ribosomal protein S1	0.0305	0.005	Down
498002457	hypothetical protein SCB02_08730	0.0145	0.0033	Down
498002459	hypothetical protein SCB02_08735	0.0081	0.0026	Down
498002464	hypothetical protein SCB02_08760	0.0261	0.009	Down
498002469	acetolactate synthase 3 catalytic subunit	0.028	0.0051	Down
498002486	cob(I)alamin adenosyltransferase	0.0117	0.006	Down
498002496	hypothetical protein SCB02_08820	0.0062	0.0025	Down
498002510	ferredoxin	0.0465	0.0097	Down
498002522	carbohydrate kinase	0.007	0.0037	Down
498002540	glycosyl transferase family 39	0.0122	0.0075	Down
498002568	cytochrome C6	0.0215	0.0061	Down
498002569	nucleoside-diphosphate-sugar epimerase	0.0262	0.0044	Down
498002571	glycogen branching enzyme	0.0177	0.0018	Down
498002573	acyl esterase	0.0041	0.0016	Down
498002577	hypothetical protein SCB02_09100	0.0181	0.0067	Down
498002579	hypothetical protein SCB02_09105	0.0046	0.0009	Down
498002590	phosphofructokinase	0.0057	0.0026	Down
498002594	homoserine kinase	0.0135	0.0019	Down
498002598	threonyl-tRNA synthetase	0.0098	0.002	Down
498002605	cation efflux system protein	0.0096	0.0031	Down
498002606	membrane associated GTPase	0.0108	0.0018	Down
498002614	hypothetical protein SCB02_09240	0.018	0.0084	Down
498002616	menaquinone biosynthesis protein MenD (2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate	0.0056	0.0014	Down
498002623	glycogen synthase	0.0136	0.0041	Down
498002627	3-phosphoshikimate 1-carboxyvinyltransferase	0.0124	0.0026	Down
498002631	3-octaprenyl-4-hydroxybenzoate decarboxylase	0.0064	0.0015	Down
498002633	nitrilase	0.0117	0.0032	Down
498002636	solanesyl diphosphate synthase	0.023	0.0047	Down
498002638	acetyl-coenzyme A synthetase	0.0078	0.0032	Down
498002649	heavy metal ABC transporter (HMT) family permease/ATP-binding protein	0.0128	0.0033	Down
498002654	carbamoyl phosphate synthase small subunit	0.018	0.0022	Down
498002658	tRNA/rRNA methyltransferase (SpoU):RNA methyltransferase TrmH	0.0202	0.0034	Down
498002660	aspartyl/glutamyl-tRNA amidotransferase subunit A	0.0234	0.0037	Down
498002661	DNA polymerase III subunit alpha	0.006	0.0024	Down
498002662	hypothetical protein SCB02_09450	0.0415	0.0139	Down
498002663	30S ribosomal protein S15	0.084	0.0155	Down
498002689	hypothetical protein SCB02_09535	0.0101	0.004	Down
498002697	preprotein translocase subunit SecD	0.0152	0.0034	Down
498002705	permease	0.0108	0.0014	Down
498002716	methyltransferase, UbiE/COQ5 family protein	0.0054	0.0009	Down
498002718	glutamine synthetase, glutamate--ammonia ligase	0.0136	0.0048	Down
498002723	pyridoxal-dependent decarboxylase family protein	0.0051	0.0015	Down
498002730	insulinase family protein	0.0044	0.0017	Down
498002736	ABC transporter	0.013	0.0034	Down
498002737	ABC-transporter, membrane spanning component	0.0066	0.0011	Down
498002753	glycosyltransferase	0.0164	0.004	Down

498002754	30S ribosomal protein S2	0.0897	0.0177	Down
498002755	translation elongation factor Ts	0.0223	0.0059	Down
498002759	sulfite reductase subunit beta	0.0141	0.0045	Down
498002760	glycyl-tRNA synthetase beta subunit	0.0162	0.0039	Down
498002761	fatty acid desaturase	0.0236	0.0029	Down
498002763	geranylgeranyl hydrogenase	0.0243	0.0063	Down
498002766	GTP-binding protein TypA	0.0226	0.0086	Down
498002770	permease	0.0135	0.002	Down
498002772	cytochrome c assembly protein	0.0237	0.0053	Down
498002775	fructose 1,6-bisphosphatase II	0.022	0.0082	Down
498002777	glutamyl-tRNA reductase	0.0121	0.0023	Down
498002779	glucose-1-phosphate adenylyltransferase	0.0192	0.0068	Down
498002789	Uracil phosphoribosyltransferase	0.0135	0.0022	Down
498002792	pentapeptide repeat-containing protein	0.0094	0.0028	Down
498002793	hydrogenase accessory membrane protein	0.0515	0.0101	Down
498002797	phosphoribosylformylglycinamidine synthase I	0.0388	0.0056	Down
498002803	two component transcriptional regulator, winged helix family protein	0.0505	0.0114	Down
498002806	hypothetical protein SCB02_09907	0.0284	0.0099	Down
498002812	zinc metallopeptidase	0.0135	0.0043	Down
498002815	proton extrusion protein PcxA	0.0095	0.0033	Down
498002821	methionyl-tRNA synthetase	0.0062	0.0014	Down
498002827	30S ribosomal protein S18	0.2506	0.0542	Down
498002829	50S ribosomal protein L33	0.077	0.0307	Down
498002831	phenylalanyl-tRNA synthetase subunit beta	0.0085	0.0017	Down
498002835	allophycocyanin alpha, B subunit	0.0323	0.0185	Down
498002836	DnaJ domain-containing protein	0.0053	0.0011	Down
498002848	hypothetical protein SCB02_10052	0.0139	0.005	Down
498002852	methionine synthase (5-methyltetrahydrofolate--homocysteine methyltransferase)	0.0061	0.0008	Down
498002858	hypothetical protein SCB02_10077	0.0542	0.0252	Down
498002879	excinuclease ABC subunit C	0.0081	0.0034	Down
498002881	flavin reductase-like domain-containing protein	0.0527	0.021	Down
498002885	hypothetical protein SCB02_10124	0.0079	0.0015	Down
498002887	D-alanyl-D-alanine carboxypeptidase	0.0067	0.0021	Down
498002888	hypothetical protein SCB02_10134	0.0286	0.0065	Down
498002894	diaminopimelate epimerase	0.0179	0.0054	Down
498002896	leucyl-tRNA synthetase	0.0053	0.0018	Down
498002897	glucose-6-phosphate isomerase	0.0086	0.0024	Down
498002900	protease	0.005	0.0027	Down
498002905	N-acetyl-gamma-glutamyl-phosphate reductase	0.0142	0.0027	Down
498002914	molecular chaperone DnaK	0.0063	0.0022	Down
498002915	putative phosphate ABC transporter	0.0393	0.0067	Down
498002916	putative phosphate ABC transporter	0.0211	0.0055	Down
498002918	phosphate ABC transporter ATP-binding protein	0.0141	0.0023	Down
498002922	inositol-1(or 4)-monophosphatase	0.0077	0.0024	Down
498002924	ATP phosphoribosyltransferase regulatory subunit	0.0069	0.0027	Down
498002928	hypothetical protein SCB02_10269	0.0324	0.0159	Down
498002937	glucosylglycerol-phosphate synthase	0.005	0.0017	Down
498002953	50S ribosomal protein L28	0.0407	0.0086	Down
498003040	glutaredoxin	0.0133	0.0015	Down
498003042	rhomboid family protein	0.0148	0.0052	Down
498003079	1-deoxy-D-xylulose-5-phosphate synthase	0.0271	0.0047	Down
498003083	hypothetical protein SCB02_10731	0.0191	0.0049	Down
498003085	pyruvate kinase	0.0181	0.0035	Down
498003086	peptide ABC transporter permease	0.0085	0.0017	Down
498003090	ATP-dependent Clp protease proteolytic subunit	0.0201	0.0069	Down
498003092	Thf1-like protein	0.0239	0.0061	Down
498003100	cytochrome b6-f complex subunit PetN	0.0663	0.0189	Down
498003175	endolysin	0.0146	0.0044	Down
498003271	hypothetical protein SCB02_11216	0.0097	0.0038	Down
498003304	fructose-1,6-bisphosphate aldolase	0.0286	0.0057	Down
498003306	hypothetical protein SCB02_11286	0.0029	0.0009	Down
498003314	cell division protein FtsH4	0.0102	0.0043	Down
498003317	hypothetical protein SCB02_11321	0.0115	0.0036	Down
498003320	Fe-S cluster protein	0.0207	0.0022	Down
498003331	hypothetical protein SCB02_11371	0.0079	0.0027	Down
498003332	magnesium-protoporphyrin IX monomethyl ester cyclase	0.0197	0.0042	Down
498003333	putative modulator of DNA gyrase; TldD	0.0087	0.0018	Down
498003345	transcription-repair coupling factor	0.0033	0.0006	Down
498003348	carboxyl-terminal processing protease	0.0083	0.002	Down

498003350	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase	0.0144	0.0034	Down
498003375	Hsp33-like chaperonin	0.0069	0.0022	Down
498003378	hypothetical protein SCB02_11536	0.0255	0.0072	Down
498003380	peptide chain release factor 3	0.0092	0.0016	Down
498003415	hypothetical protein SCB02_11631	0.0095	0.0024	Down
498003563	sirohydrochlorin cobaltochelatase	0.0037	0.0012	Down
498003568	hypothetical protein SCB02_12028	0.0114	0.0019	Down
498003573	carbamoyl phosphate synthase large subunit	0.0101	0.0022	Down
498003577	Sodium:alanine symporter family protein	0.0118	0.0039	Down
498003597	hypothetical protein SCB02_12153	0.019	0.0102	Down
498003607	hypothetical protein SCB02_12205	0.0406	0.0129	Down
498003614	ABC-type phosphate transport system, substrate binding protein	0.0377	0.006	Down
498003615	putative phosphate ABC transporter	0.0128	0.0017	Down
498003616	putative phosphate ABC transporter	0.0064	0.0024	Down
498003619	chromate transporter	0.004	0.0014	Down
498003623	transcriptional regulator	0.0514	0.0081	Down
498003625	two component transcriptional regulator	0.0089	0.004	Down
498003628	hypothetical protein SCB02_12310	0.033	0.007	Down
498003633	oxidoreductase, FAD-dependent	0.0048	0.001	Down
498003637	glutamine synthetase catalytic region	0.0113	0.0042	Down
498003642	ANL40	0.0311	0.0067	Down
498003653	fatty acid desaturase, type 2	0.015	0.0058	Down
498003659	tRNA (uracil-5-)-methyltransferase Gid	0.0122	0.002	Down
498003665	two-component response regulator	0.021	0.0085	Down
498003680	RNA-binding protein RbpD	0.0488	0.0101	Down
498003683	membrane bound transcriptional regulator-like protein	0.0299	0.0052	Down
498003700	biotin synthase	0.0158	0.0022	Down
498003702	hypothetical protein SCB02_12692	0.0215	0.003	Down
498003703	diaminopimelate decarboxylase	0.0188	0.0026	Down
498003705	ATPase	0.0291	0.0086	Down
498003709	glyceraldehyde-3-phosphate dehydrogenase	0.0099	0.0024	Down
498003715	phosphatidate cytidylyltransferase	0.0136	0.0011	Down
498003726	aminopeptidase N	0.0086	0.0016	Down
498003731	ferredoxin-NADP oxidoreductase	0.0197	0.0027	Down
498003756	2-isopropylmalate synthase	0.014	0.005	Down
498003758	hypothetical protein SCB02_12972	0.0061	0.0023	Down
498003764	inosine 5-monophosphate dehydrogenase	0.042	0.0128	Down
498003780	putative beta-lactamase	0.0042	0.0014	Down
498003786	hypothetical protein SCB02_13074	0.0122	0.0043	Down
498003790	homoserine dehydrogenase	0.0113	0.0043	Down
498003797	fused dienelactone hydrolase/uncharacterized domain	0.0052	0.0026	Down
498003812	GTP-binding protein YchF	0.0109	0.0049	Down
498003815	membrane fusion protein	0.0104	0.0046	Down
498003818	UGMP family protein	0.0532	0.0043	Down
498003819	photosystem I reaction center subunit III	0.056	0.0188	Down
498003823	Sec-independent protein secretion pathway component TatC	0.0109	0.0032	Down
498003825	cytochrome b6-f complex iron-sulfur subunit	0.0506	0.008	Down
498003828	apocytochrome f	0.0322	0.0096	Down
498003830	prolipoprotein diacylglyceryl transferase	0.0149	0.0019	Down
498003841	putative exopolyphosphatase	0.0059	0.0016	Down
498003858	chaperonin GroEL	0.008	0.0023	Down
498003859	hypothetical protein SCB02_13301	0.0273	0.0084	Down
498003899	hypothetical protein SCB02_13396	0.0184	0.0046	Down
498003905	NAD(P)H-quinone oxidoreductase subunit 2	0.0851	0.0092	Down
498003913	glycosyltransferase of family UDP-glucose:tetrahydrobiopterin glucosyltransferase	0.0153	0.0034	Down
498003915	DMT family permease	0.027	0.0084	Down
498003919	hypothetical protein SCB02_13476	0.0216	0.0089	Down
498003926	AAA family ATPase	0.0049	0.0019	Down
498003943	indole-3-glycerol-phosphate synthase	0.0183	0.0037	Down
498003945	dihydrolipoamide dehydrogenase	0.0181	0.0019	Down
498003950	UDP-N-acetylglucosamine 1-carboxyvinyltransferase	0.0101	0.0027	Down
498003965	D-alanine--D-alanine ligase	0.0083	0.0029	Down
498003973	cell division protein FtsZ	0.0469	0.0125	Down
498003977	integral membrane protein	0.0118	0.0032	Down
498003978	ATP-dependent Clp protease-like protein	0.0108	0.0019	Down
498003979	ATP-dependent Clp protease proteolytic subunit	0.0171	0.0055	Down
498004011	kinase	0.0184	0.0066	Down
497998709	menaquinone biosynthesis methyltransferase	0.0036	0.0025	-
497998725	glycosyltransferase	0.0038	0.0028	-

497998745	GTP-binding protein LepA	0.0142	0.0142	-
497998833	hypothetical protein SCB02_00359	0.0101	0.0043	-
497998862	HIT (histidine triad) family protein	0.0332	0.0306	-
497998910	membrane protein-related protein	0.0168	0.0078	-
497998912	phosphatidylcholine-hydrolyzing phospholipase D family protein	0.0038	0.0067	-
497998917	hypothetical protein SCB02_00564	0.0062	0.0116	-
497998933	RND family multidrug efflux transporter	0.0007	0.0006	-
497998951	HAD superfamily hydrolase	0.0028	0.0028	-
497998975	hypothetical protein SCB02_00676	0.0302	0.0237	-
497999016	inositol monophosphate family protein	0.0109	0.0073	-
497999045	pyrimidine reductase, riboflavin biosynthesis	0.0111	0.0084	-
497999053	hypothetical protein SCB02_00846	0.0032	0.0046	-
497999069	ribosome-binding factor A	0.0051	0.0047	-
497999096	porin	0.0089	0.0056	-
497999136	two-component system response regulator	0.0066	0.0042	-
497999164	tRNA nucleotidyltransferase/poly(A) polymerase	0.0056	0.0026	-
497999186	hypothetical protein SCB02_01121	0.0164	0.0106	-
497999222	ABC-type multidrug transport system, ATPase and permease components	0.0024	0.0011	-
497999229	adenylylsulfate kinase	0.0049	0.0027	-
497999241	two component LuxR family transcriptional regulator	0.008	0.0043	-
497999243	hypothetical protein SCB02_01256	0.0079	0.0044	-
497999256	hypothetical protein SCB02_01286	0.0065	0.0062	-
497999259	thioesterase	0.0041	0.0016	-
497999281	Ion transport protein	0.0041	0.004	-
497999285	glutaredoxin	0.0055	0.013	-
497999314	ABC transporter ATP-binding protein	0.0032	0.0024	-
497999336	two-component sensor histidine kinase	0.0024	0.0018	-
497999352	general (type II) secretion pathway protein D precursor	0.0029	0.0041	-
497999399	GntR family regulatory protein	0.0034	0.0053	-
497999423	hypothetical protein SCB02_01713	0.0101	0.0084	-
497999434	Delta-9 acyl-phospholipid desaturase	0.0267	0.0169	-
497999459	hypothetical protein SCB02_01798	0.0096	0.0075	-
497999482	hypothetical protein SCB02_01828	0.0046	0.0055	-
497999484	hypothetical protein SCB02_01833	0.009	0.0056	-
497999487	hypothetical protein SCB02_01843	0.008	0.004	-
497999561	CTP synthetase	0.011	0.0115	-
497999582	urease accessory protein UreG	0.0078	0.0042	-
497999607	hypothetical protein SCB02_02062	0.008	0.0059	-
497999610	hypothetical protein SCB02_02067	0.0041	0.0083	-
497999632	molybdenum cofactor biosynthesis protein A	0.0034	0.002	-
497999665	Fe-S oxidoreductas	0.0055	0.003	-
497999676	ABC transporter, substrate binding protein, phosphate	0.0031	0.0012	-
497999694	hypothetical protein SCB02_02267	0.0107	0.0062	-
497999698	hypothetical protein SCB02_02277	0.0127	0.0085	-
497999704	phage integrase family protein	0.0036	0.0082	-
497999726	putative cell envelope-related function transcriptional attenuator	0.0024	0.0022	-
497999750	ATP-dependent protease ATP-binding subunit ClpX	0.0046	0.0032	-
497999767	hypothetical protein SCB02_02417	0.0092	0.0046	-
497999789	hypothetical protein SCB02_02477	0.0147	0.0175	-
497999836	general secretion pathway protein E	0.01	0.0077	-
497999838	twitching motility protein	0.0026	0.0013	-
497999859	putative ribonuclease III	0.0071	0.0031	-
497999946	peptidase, M20D family protein	0.0052	0.0025	-
497999954	serine/threonine protein kinase	0.0022	0.0019	-
497999965	hypothetical protein SCB02_02881	0.0101	0.004	-
497999998	threonine synthase	0.0715	0.0685	-
498000002	hypothetical protein SCB02_02971	0.0057	0.0082	-
498000017	DNA mismatch repair protein MutS	0.0022	0.0013	-
498000044	translation initiation factor IF-3	0.017	0.0136	-
498000046	tRNA delta(2)-isopentenylpyrophosphate transferase	0.0101	0.0109	-
498000048	DNA gyrase subunit B	0.027	0.0298	-
498000052	hypothetical protein SCB02_03086	0.0143	0.0462	-
498000055	crcB protein	0.0202	0.0404	-
498000057	glutathione peroxidase	0.023	0.0113	-
498000080	A/G-specific adenine glycosylase	0.0067	0.0076	-
498000147	UvrD/REP helicase	0.0125	0.0154	-
498000158	glycosly transferase	0.0044	0.0097	-
498000161	lipopolysaccharide synthesis sugar transferase	0.0076	0.0035	-
498000163	sugar transferase	0.0211	0.0145	-

498000166	lipolytic protein G-D-S-L family	0.0045	0.0019	-
498000169	glycosyltransferase	0.0098	0.0073	-
498000174	glycosyltransferase	0.0052	0.0051	-
498000184	hypothetical protein SCB02_03358	0.0195	0.0411	-
498000189	glycosyltransferase	0.008	0.0115	-
498000195	glycosyl transferase family 2	0.0119	0.0082	-
498000197	hypothetical protein SCB02_03388	0.01	0.0103	-
498000202	methyltransferase FkbM	0.0082	0.0094	-
498000206	hypothetical protein SCB02_03408	0.0031	0.0038	-
498000208	hypothetical protein SCB02_03413	0.0023	0.0038	-
498000214	putative GDP-L-fucose synthetase	0.0132	0.0143	-
498000251	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	0.004	0.0033	-
498000254	deoxyribodipyrimidine photolyase	0.0064	0.0037	-
498000262	hypothetical protein SCB02_03543	0.0016	0.004	-
498000296	oxidoreductase	0.0105	0.0159	-
498000330	pentapeptide repeat-containing protein	0.0156	0.0248	-
498000333	photosystem II reaction center protein H	0.1507	0.0682	-
498000345	cob(I)alamin adenosyltransferase	0.0055	0.004	-
498000347	deoxycytidine triphosphate deaminase	0.0061	0.0027	-
498000355	ferredoxin	0.0024	0.003	-
498000360	soluble lytic transglycosylase	0.0034	0.0038	-
498000362	phosphoglucosamine mutase	0.0063	0.0059	-
498000476	hypothetical protein SCB02_04081	0.0278	0.0173	-
498000491	hypothetical protein SCB02_04116	0.0018	0.0047	-
498000530	transcriptional regulator NrdR	0.0156	0.0093	-
498000552	hypothetical protein SCB02_04238	0.0075	0.0076	-
498000582	hypothetical protein SCB02_04308	0.0077	0.0124	-
498000589	hypothetical protein SCB02_04323	0.0019	0.0017	-
498000598	hypothetical protein SCB02_04343	0.0027	0.0017	-
498000617	hypothetical protein SCB02_04383	0.0115	0.0064	-
498000645	rare lipoprotein A	0.0082	0.011	-
498000701	ATPase	0.0083	0.0043	-
498000748	hypothetical protein SCB02_04654	0.0168	0.0098	-
498000772	putative ldpA protein	0.0148	0.0192	-
498000930	formamidopyrimidine-DNA glycosylase	0.0034	0.0022	-
498000937	peptidoglycan-binding LysM	0.0042	0.003	-
498000939	putative aldehyde dehydrogenase	0.0037	0.002	-
498001033	putative glutathione S-transferase	0.0019	0.0025	-
498001034	aspartoacylase	0.003	0.0042	-
498001097	hypothetical protein SCB02_05383	0.0105	0.0065	-
498001145	cell division protein FtsW	0.0051	0.0034	-
498001169	hypothetical protein SCB02_05573	0.0048	0.0048	-
498001172	hypothetical protein SCB02_05588	0.0014	0.0025	-
498001196	forkhead-associated protein	0.0255	0.0458	-
498001200	FHA modulated glycosyl transferase/transpeptidase	0.006	0.0069	-
498001218	DnaK family protein	0.0124	0.0188	-
498001227	3-hydroxyisobutyrate dehydrogenase related protein	0.0021	0.002	-
498001241	ribosomal protein L11 methyltransferase	0.0085	0.005	-
498001245	cytochrome c-550	0.0299	0.018	-
498001247	ribonuclease Z	0.0072	0.008	-
498001306	transporter, major facilitator family protein	0.0072	0.0049	-
498001339	superoxide dismutase	0.0091	0.0051	-
498001371	hypothetical protein SCB02_06140	0.0142	0.0189	-
498001372	hypothetical protein SCB02_06145	0.0109	0.0188	-
498001380	two-component sytem response regulator	0.0103	0.0117	-
498001439	prohibitin family protein	0.0046	0.0019	-
498001448	hypothetical protein SCB02_06272	0.0157	0.0201	-
498001464	Type II alternative RNA polymerase sigma factor, sigma-70 family protein	0.0012	0.0013	-
498001537	hypothetical protein SCB02_06462	0.0042	0.0045	-
498001605	CPA1 family Na+/H+ antiporter	0.0063	0.0066	-
498001632	TRAP-T family tripartite transporter	0.0075	0.0037	-
498001658	hypothetical protein SCB02_06747	0.0063	0.0096	-
498001671	ABC transporter	0.006	0.0036	-
498001673	hypothetical protein SCB02_06782	0.0028	0.0035	-
498001708	RNA polymerase sigma factor, sigma-70 family protein	0.0017	0.001	-
498001710	hypothetical protein SCB02_06856	0.0115	0.0073	-
498001725	putative DNA repair ATPase	0.0016	0.001	-
498001726	DNA repair exonuclease	0.0036	0.0017	-

498001746	cytochrome c oxidase subunit I	0.0012	0.0016	-
498001750	flavoprotein related to choline dehydrogenase	0.0029	0.0025	-
498001783	hypothetical protein SCB02_07041	0.002	0.0029	-
498001786	hypothetical protein SCB02_07046	0.001	0.0014	-
498001790	neuromedin U	0.0029	0.0012	-
498001797	regulatory proteins, Crp family protein	0.0235	0.037	-
498001801	ferritin	0.0068	0.0058	-
498001806	Beta-lactamase-like protein	0.0032	0.0014	-
498001808	SAM-dependent methyltransferase	0.0039	0.0082	-
498001810	transcriptional regulator	0.0054	0.0081	-
498001812	hypothetical protein SCB02_07126	0.0336	0.0221	-
498001814	hypothetical protein SCB02_07131	0.004	0.0039	-
498001857	metallo-beta-lactamase superfamily hydrolase	0.0061	0.0045	-
498001867	asparagine synthase	0.0027	0.0016	-
498001923	hypothetical protein SCB02_07393	0.0033	0.0045	-
498001960	dienelactone hydrolase	0.0049	0.0037	-
498002007	putative cytochrome P450	0.0068	0.0192	-
498002045	porin-like protein	0.0757	0.1048	-
498002386	hypothetical protein SCB02_08510	0.008	0.0082	-
498002405	hypothetical protein SCB02_08575	0.0442	0.0452	-
498002416	VIC family potassium channel protein	0.0119	0.0081	-
498002417	light-independent protochlorophyllide reductase subunit N	0.0175	0.0128	-
498002428	hypothetical protein SCB02_08655	0.0049	0.0033	-
498002454	creatininase	0.0033	0.004	-
498002463	hypothetical protein SCB02_08755	0.0183	0.0134	-
498002493	ribosome recycling factor	0.036	0.0202	-
498002506	hypothetical protein SCB02_08860	0.0025	0.0013	-
498002512	hypothetical protein SCB02_08875	0.0081	0.0072	-
498002520	putative photosystem II reaction center Psb27 protein	0.0264	0.0193	-
498002525	single-stranded DNA-binding protein	0.0109	0.0105	-
498002542	hypothetical protein SCB02_08965	0.0103	0.0071	-
498002543	hypothetical protein SCB02_08970	0.0092	0.008	-
498002558	hypothetical protein SCB02_09025	0.0083	0.014	-
498002559	putative pterin-4-alpha-carbinolamine dehydratase	0.0094	0.0095	-
498002560	asparagine synthase (glutamine-hydrolyzing)	0.0121	0.0158	-
498002561	ATP-dependent Clp protease, Hsp 100, ATP-binding subunit ClpB	0.004	0.0052	-
498002574	hypothetical protein SCB02_09090	0.0071	0.0029	-
498002607	photosystem II protein D1	0.0746	0.0655	-
498002612	zinc transporter ZupT	0.0085	0.0047	-
498002613	hypothetical protein SCB02_09235	0.0056	0.0032	-
498002647	methionine-S-sulfoxide reductase	0.0084	0.0075	-
498002650	hypothetical protein SCB02_09400	0.0036	0.0069	-
498002652	anthranilate phosphoribosyltransferase	0.0101	0.0058	-
498002656	hypothetical protein SCB02_09420	0.0067	0.0081	-
498002679	hypothetical protein SCB02_09505	0.0035	0.0027	-
498002694	pyruvate dehydrogenase E1 beta subunit	0.0199	0.0147	-
498002717	hypothetical protein SCB02_09622	0.0096	0.0063	-
498002719	allophycocyanin subunit beta	0.0252	0.0133	-
498002739	DevA-like ABC transporter ATPase component	0.0058	0.0049	-
498002756	hypothetical protein SCB02_09727	0.0068	0.0074	-
498002764	putative carboxypeptidase	0.0035	0.0021	-
498002767	hypothetical protein SCB02_09777	0.0077	0.004	-
498002780	6-phosphogluconate dehydrogenase	0.0026	0.0027	-
498002782	putative 6-phosphogluconolactonase (DevB, Pgl)	0.0038	0.0027	-
498002783	hypothetical protein SCB02_09832	0.0655	0.1034	-
498002801	hypothetical protein SCB02_09897	0.0316	0.0355	-
498002809	uridine kinase	0.0024	0.0029	-
498002814	bifunctional adenosylcobalamin biosynthesis protein (adenosylcobinamide kinase /	0.0084	0.0102	-
498002816	hypothetical protein SCB02_09947	0.0168	0.0141	-
498002819	hypothetical protein SCB02_09962	0.0121	0.0126	-
498002842	epimerase, PhzC/PhzF-like protein	0.0029	0.0015	-
498002844	NAD(P)H-dependent glycerol-3-phosphate dehydrogenase	0.0063	0.0078	-
498002846	hypothetical protein SCB02_10047	0.0179	0.0392	-
498002898	DnaB-like helicase	0.0066	0.0038	-
498002909	peptidylprolyl isomerase	0.0089	0.0038	-
498002910	5'-methylthioadenosine phosphorylase	0.0052	0.0033	-
498002912	hypothetical protein SCB02_10219	0.0041	0.0047	-
498002931	hypothetical protein SCB02_10274	0.0212	0.0152	-
498002935	hypothetical protein SCB02_10284	0.0035	0.0087	-

498002936	heat shock protein 90	0.0063	0.0068	-
498002946	hypothetical protein SCB02_10319	0.0035	0.0021	-
498002951	glycerol kinase	0.0031	0.0017	-
498002952	small mechanosensitive ion channel	0.0053	0.0046	-
498002963	glycosyl transferase, group 2 family protein	0.0036	0.0022	-
498002966	hypothetical protein SCB02_10389	0.0037	0.0042	-
498002988	hypothetical protein SCB02_10459	0.0033	0.0024	-
498002991	hypothetical protein SCB02_10474	0.0028	0.0028	-
498003031	hypothetical protein SCB02_10556	0.0016	0.0016	-
498003033	hypothetical protein SCB02_10561	0.0038	0.0044	-
498003043	putative Zn peptidase	0.0048	0.0025	-
498003050	ATPase	0.0032	0.0032	-
498003059	hypothetical protein SCB02_10651	0.008	0.0042	-
498003061	SAM-dependent methyltransferase	0.0016	0.0012	-
498003064	hypothetical protein SCB02_10671	0.0103	0.0059	-
498003084	hypothetical protein SCB02_10736	0.0126	0.0207	-
498003088	FtsH ATP-dependent protease-like protein	0.0095	0.0059	-
498003098	hypothetical protein SCB02_10781	0.1285	0.0826	-
498003102	ATP-dependent Clp protease adaptor	0.011	0.013	-
498003156	hypothetical protein SCB02_10906	0.0034	0.0066	-
498003157	hypothetical protein SCB02_10911	0.0032	0.0028	-
498003159	hypothetical protein SCB02_10921	0.005	0.0024	-
498003163	phosphoribosylaminoimidazole carboxylase ATPase subunit	0.0033	0.0015	-
498003168	hypothetical protein SCB02_10951	0.0233	0.0259	-
498003169	hypothetical protein SCB02_10956	0.0174	0.0134	-
498003229	CBS	0.0125	0.0063	-
498003234	nucleotide sugar epimerase	0.0067	0.0035	-
498003243	hypothetical protein SCB02_11111	0.0042	0.0032	-
498003245	fumarate hydratase	0.0048	0.006	-
498003278	3-isopropylmalate dehydrogenase	0.0108	0.0101	-
498003324	hypothetical protein SCB02_11346	0.0147	0.006	-
498003334	putative modulator of DNA gyrase	0.004	0.0025	-
498003373	hypothetical protein SCB02_11511	0.0047	0.0043	-
498003374	ABC-type cobalt transport system, ATPase component	0.0036	0.0029	-
498003377	hypothetical protein SCB02_11531	0.0181	0.0074	-
498003384	small-conductance mechanosensitive channel	0.0024	0.0018	-
498003387	ribonucleotide reductase	0.0035	0.0033	-
498003397	hypothetical protein SCB02_11581	0.0043	0.0093	-
498003399	choloylglycine hydrolase	0.0025	0.0061	-
498003530	hypothetical protein SCB02_11898	0.0032	0.0072	-
498003532	hypothetical protein SCB02_11903	0.0027	0.0043	-
498003543	ATLS1-like light-inducible protein	0.0088	0.0048	-
498003576	hypothetical protein SCB02_12058	0.008	0.0107	-
498003579	glutathione reductase (NADPH)	0.0061	0.0048	-
498003582	glutathione S-transferase	0.0033	0.0035	-
498003591	putative universal stress protein f	0.0016	0.0052	-
498003592	hypothetical protein SCB02_12128	0.0034	0.0036	-
498003593	hypothetical protein SCB02_12133	0.004	0.0073	-
498003603	selenophosphate synthase/FAD/NAD(P)-binding domain-containing protein	0.002	0.0017	-
498003609	hypothetical protein SCB02_12215	0.0039	0.0075	-
498003611	major facilitator superfamily permease / multidrug efflux transporter	0.0041	0.0056	-
498003622	CHAD domain containing protein	0.0051	0.0044	-
498003627	porin	0.0476	0.0613	-
498003632	rubredoxin	0.0123	0.0205	-
498003634	reductase	0.0386	0.0479	-
498003635	hypothetical protein SCB02_12345	0.0221	0.0388	-
498003638	cation transport ATPase	0.0022	0.0019	-
498003661	carotenoid isomerase	0.0068	0.0046	-
498003663	hypothetical protein SCB02_12487	0.0213	0.0178	-
498003664	hypothetical protein SCB02_12492	0.0084	0.0066	-
498003666	hypothetical protein SCB02_12502	0.0267	0.0228	-
498003692	PsbP	0.0169	0.0086	-
498003707	hypothetical protein SCB02_12717	0.0033	0.0087	-
498003755	putative sn-glycerol-3-phosphate ABC transporter, permease protein	0.0037	0.0106	-
498003757	hypothetical protein SCB02_12967	0.0094	0.0081	-
498003759	Fe-S oxidoreductase	0.0018	0.0014	-
498003760	lycopene cyclase (CrtL-type)	0.0033	0.0036	-
498003765	thioredoxin	0.018	0.0084	-
498003777	cytochrome cM	0.0116	0.0101	-

498003787	SufE protein	0.0045	0.0027	-
498003799	hypothetical protein SCB02_13119	0.0102	0.0175	-
498003817	hypothetical protein SCB02_13194	0.0262	0.014	-
498003864	putative multidrug efflux ABC transporter	0.0216	0.027	-
498003866	protoheme IX farnesyltransferase	0.0179	0.0234	-
498003870	putative cytochrome c oxidase, subunit 2	0.0215	0.0178	-
498003872	cytochrome c oxidase, subunit I	0.0181	0.0136	-
498003906	ABC transporter ATP-binding protein	0.0065	0.0048	-
498003923	putative inner membrane protein translocase component YidC	0.0122	0.0104	-
498003928	spectrin repeat-containing protein	0.0153	0.0105	-
498003929	hypothetical protein SCB02_13521	0.0084	0.0027	-
498003937	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase	0.0043	0.0019	-
498003953	pentapeptide repeat-containing protein	0.0045	0.0038	-
498003963	(dimethylallyl)adenosine tRNA methylthiotransferase	0.0052	0.0039	-
498003987	hypothetical protein SCB02_13711	0.0013	0.0084	-
498003992	glycyl-tRNA synthetase subunit alpha	0.0166	0.0118	-
498004010	hypothetical protein SCB02_13759	0.0118	0.0068	-

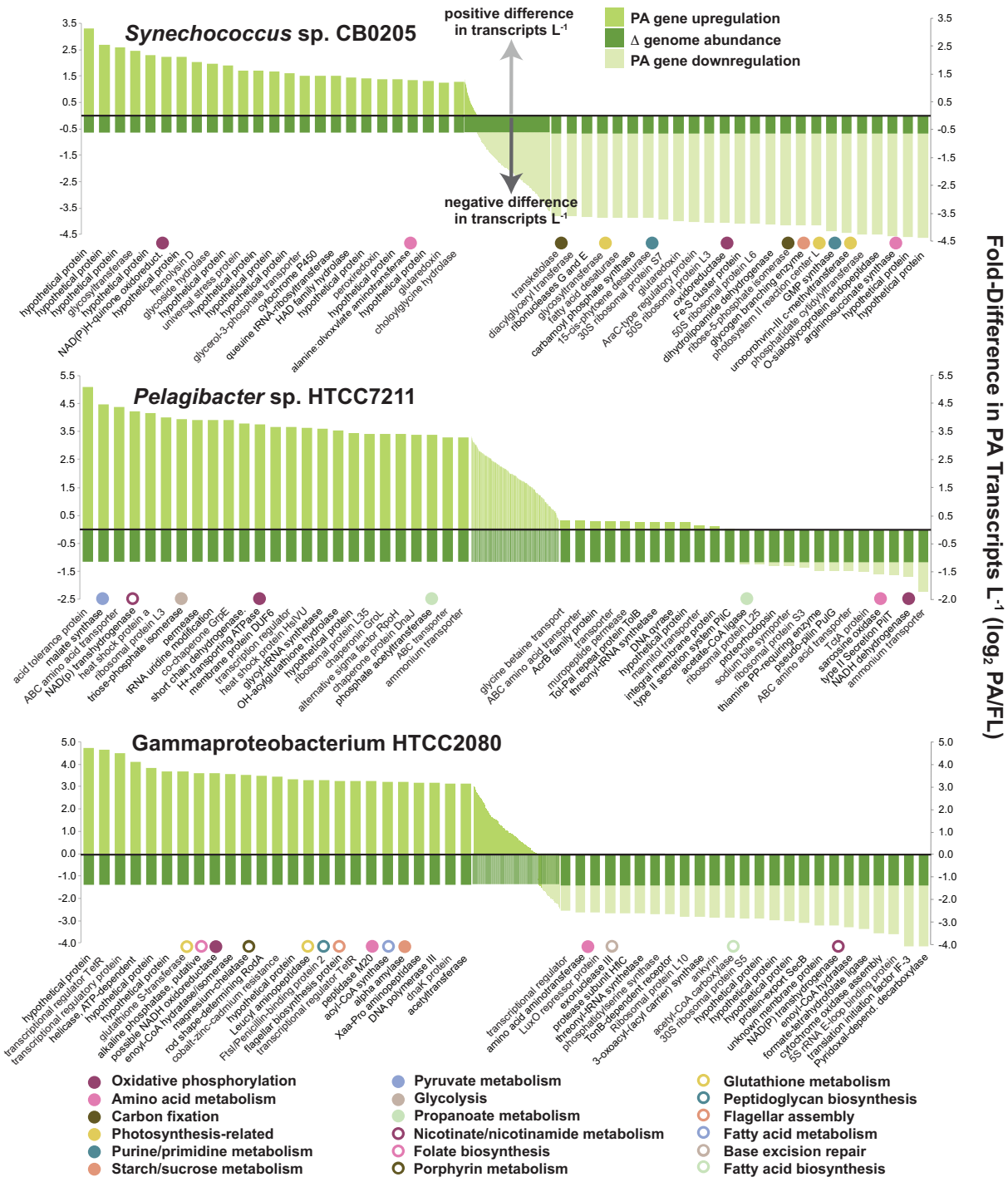


Figure 3.S5. The importance of cell abundance versus gene regulation to the differential contribution of PA cells to the Amazon Plume metatranscriptome for three major bacterial taxa. For each taxon, the direction and magnitude of the difference in PA cells compared to FL cells that is attributed to each factor is shown for the 25 genes with the most positive and most negative differences, and KEGG pathway assignments are indicated by color-coded circles.

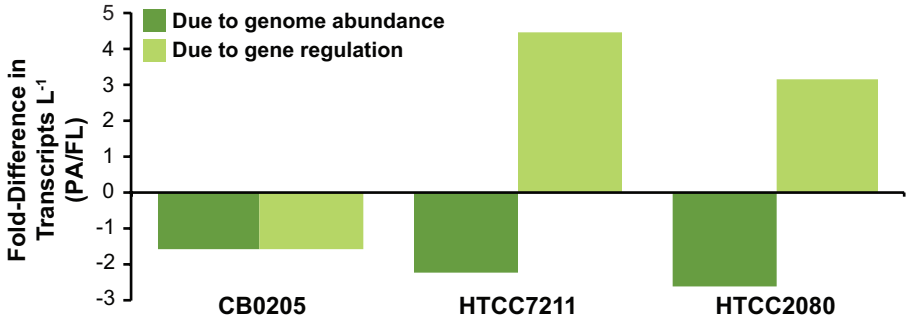


Figure 3.4 Role of cell abundance versus gene regulation in the differential contribution of PA cells to the Amazon Plume metatranscriptome.

Table 3.S4. Gene expression ratios in the *Pelagibacter* sp. HTCC7211 bin.

NCBI GI	Protein Description	Mean Expression Ratio (FL)	Mean Expression Ratio (PA)	PA Regulation
495819520	DNA-directed RNA polymerase, beta' subunit	0.0114	0.0456	Up
495819535	cytochrome c oxidase, subunit II	0.0183	0.0947	Up
495819590	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA	0.0537	0.8066	Up
495819592	glycine betaine transport ATP-binding protein opuAA	0.0026	0.0196	Up
495819601	general amino acid ABC transporter, ATP-binding protein	0.0037	0.0778	Up
495819605	ATP-dependent Zn protease	0.0142	0.0398	Up
495819621	Receptor family ligand binding domain protein	0.058	0.1563	Up
495819635	2-isopropylmalate synthase	0.3554	3.2775	Up
495819639	chaperone protein DnaK	0.0231	0.1487	Up
495819692	3-hydroxyacyl-CoA dehydrogenase	0.0058	0.0458	Up
495819722	integral membrane protein	0.0692	0.5633	Up
495819728	serine hydroxymethyltransferase	0.0075	0.0468	Up
495819738	HflK protein	0.0227	0.1614	Up
495819741	acyl carrier protein	0.0928	0.293	Up
495819758	translation elongation factor G	0.0175	0.0527	Up
495819798	cytochrome-c oxidase	0.0817	0.5174	Up
495819802	alanyl-tRNA synthetase	0.0013	0.0063	Up
495819841	glutamine synthetase, catalytic domain, putative	0.035	0.0827	Up
495819850	trap dicarboxylate transporter - dctp subunit	0.0678	0.4403	Up
495819874	citrate (si)-synthase	0.0067	0.0329	Up
495819902	ribosomal protein L2	0.0121	0.0645	Up
495819920	pirin	0.0094	0.0766	Up
495819927	DNA-directed RNA polymerase, alpha subunit	0.0201	0.068	Up
495819970	glutamate synthase large subunit	0.0154	0.0478	Up
495819989	acid tolerance regulatory protein actr	0.0111	0.3769	Up
495820016	cell division protein FtsZ	0.0097	0.0334	Up
495820052	heat shock protein HslVU, ATPase subunit HslU	0.003	0.0376	Up
495820074	Bacterial extracellular solute-binding protein, family 7	0.0507	0.1481	Up
495820146	GTP cyclohydrolase I	0.0104	0.0589	Up
495820147	NAD(p) transhydrogenase subunit beta	0.1718	3.2063	Up
495820166	alanine dehydrogenase	0.0091	0.0805	Up
495820193	translation elongation factor Tu	0.1686	0.331	Up
495820205	H+-transporting two-sector ATPase (subunit b)	0.0041	0.0556	Up
495820243	import inner membrane translocase, subunit Tim44	0.0049	0.0455	Up
495820250	taurine transport system periplasmic protein	0.0515	0.221	Up
495820272	OmpA family protein	0.055	0.4371	Up
495820285	malate synthase	0.0009	0.02	Up
495820304	ABC proline/glycine betaine transporter, periplasmic substrate-binding protein	0.0179	0.075	Up
495820329	heAt shock protein a	0.0752	1.3574	Up
495820360	substrate-binding region of ABC-type glycine betaine transport system	0.0625	0.2092	Up
495820365	ABC transporter	0.0526	0.1471	Up
495820376	spermidine/putrescine-binding periplasmic protein	0.1081	0.3382	Up
495820390	translation initiation factor IF-2	0.0029	0.021	Up
495820391	ATP-dependent protease La	0.0047	0.032	Up
495820433	DNA-directed RNA polymerase, beta subunit	0.0113	0.0274	Up
495820456	ABC transporter, QAT family, substrate-binding protein, putative	0.0124	0.0423	Up
495820460	carbamoyl-phosphate synthase, large subunit	0.0031	0.0115	Up
495820606	glutamine synthetase, type III	0.0034	0.0172	Up
495820631	leucyl aminopeptidase	0.0042	0.0186	Up
495820666	NADH dehydrogenase (quinone), D subunit	0.0104	0.063	Up
495820681	ribosomal protein S11	0.0211	0.0957	Up
495820706	betaine-aldehyde dehydrogenase	0.0014	0.0083	Up
495820716	co-chaperone GrpE	0.0052	0.0779	Up
495820729	Integral membrane protein DUF6	0.0685	0.8739	Up
495820782	non-specific DNA-binding protein HBsu	0.1866	0.611	Up
495820856	thioredoxin-disulfide reductase	0.0107	0.0638	Up
495820892	permease	0.0024	0.0362	Up
495820976	triose-phosphate isomerase	0.02	0.3099	Up
495821028	extracellular solute-binding protein, family 5	0.0159	0.0739	Up
495821044	putative permease, major facilitator superfamily	0.004	0.0156	Up
495821048	protease DO	0.0105	0.0469	Up

495821054	30S ribosomal protein S1	0.015	0.0487	Up
495821062	protein RecA	0.0188	0.1162	Up
495821077	ATP synthase F1, alpha subunit	0.0188	0.0984	Up
495821094	H+-transporting two-sector ATPase (subunit b)	0.0388	0.1869	Up
495821103	xanthine/uracil/vitamin C permease family protein	0.109	0.2955	Up
495821111	smc protein	0.0021	0.0128	Up
495821156	ATP-binding component of ABC transporter	0.0058	0.0572	Up
495821159	chaperonin GroL	0.0431	0.4632	Up
495821160	ribosomal protein S4	0.0072	0.032	Up
495821187	ABC transporter	0.1852	0.7251	Up
495821262	oxidoreductase, aldo/keto reductase family	0.0021	0.0171	Up
495821376	tartrate dehydrogenase	0.0033	0.0278	Up
495821387	catalase/peroxidase HPI	0.0202	0.1227	Up
495821393	formate dehydrogenase, alpha subunit	0.0064	0.019	Up
495821415	polyribonucleotide nucleotidyltransferase	0.0119	0.0488	Up
495821419	conserved hypothetical protein	0.0111	0.0651	Up
495821420	alternative sigma factor RpoH	0.0044	0.0465	Up
495821440	cytochrome B/c1	0.0109	0.0422	Up
495821487	transcription termination factor Rho	0.0067	0.026	Up
495821508	aminomethyltransferase, glycin cleavage system T protein	0.0462	0.1439	Up
495821524	V-type H(+)-translocating pyrophosphatase	0.0698	0.132	Up
495821537	chaperone protein DnaJ	0.0048	0.0505	Up
495821544	ammonium transporter family protein	0.0633	0.6174	Up
495819481	ATP synthase F0, A subunit	0.0139	0.0366	-
495819485	gxgxg motif-containing protein	0.0109	0.0219	-
495819490	FolC bifunctional protein	0.011	0.0576	-
495819493	Na+/H+ antiporter NhaA	0.007	0.0119	-
495819498	glutamyl-tRNA(Gln) amidotransferase subunit A	0.0043	0.014	-
495819500	ribosomal protein L14	0.0123	0.0383	-
495819503	acetyl-CoA carboxylase, biotin carboxylase	0.0073	0.0103	-
495819511	spermidine/putrescine transport system permease protein potb	0.0043	0.0215	-
495819525	phosphoribosylformylglycinamidine synthase II	0.0026	0.0051	-
495819529	histidyl-tRNA synthetase	0.0038	0.0094	-
495819532	ABC-type branched-chain amino acid transport system, permease protein I	0.006	0.0367	-
495819536	phenylalanyl-tRNA synthetase, beta subunit	0.0008	0.0061	-
495819537	nucleotidyltransferase/DNA polymerase involved in DNA repair/SOS mutagenesis and repair	0.0012	0.0028	-
495819541	NAD+ synthetase	0.0043	0.0214	-
495819551	ribosomal protein L20	0.0075	0.0174	-
495819557	NADH dehydrogenase i chain a	0.0168	0.1133	-
495819573	ABC transporter, membrane spanning protein (glycine betaine)	0.0055	0.0182	-
495819597	putative coenzyme F420-reducing hydrogenase, beta subunit family protein	0.0019	0.0025	-
495819598	prolyl-tRNA synthetase	0.003	0.0054	-
495819617	conserved hypothetical protein TIGR00701	0.0011	0.0073	-
495819640	acetyl-CoA carboxylase, carboxyl transferase, alpha subunit	0.0029	0.0054	-
495819641	pyruvate, phosphate dikinase	0.0022	0.0075	-
495819643	glycine betaine/L-proline transport system permease	0.0043	0.0054	-
495819675	putative 3-hydroxyisobutyrate dehydrogenase	0.0014	0.0073	-
495819684	tryptophanyl-tRNA synthetase	0.0039	0.0128	-
495819686	glutamate synthase, large subunit	0.0014	0.0032	-
495819688	glutamine synthetase, type I	0.0134	0.0327	-
495819691	type II Secretion PilQ	0.0259	0.0853	-
495819702	signal recognition particle protein	0.0035	0.0085	-
495819705	methylenetetrahydrofolate reductase	0.0235	0.037	-
495819720	ATP synthase F1, gamma subunit	0.004	0.007	-
495819724	ammonium transporter family protein	0.026	0.0124	-
495819727	branched-chain amino acid transport system/permease component	0.0046	0.0087	-
495819731	preprotein translocase SecG subunit	0.0197	0.0251	-
495819753	5-methylcytosine-specific restriction enzyme A	0.0038	0.0196	-
495819755	pyruvate dehydrogenase (acetyl-transferring), homodimeric type	0.0033	0.0058	-
495819765	trigger factor	0.0022	0.0069	-
495819766	DNA primase	0.0016	0.0044	-
495819768	flavin-containing monooxygenase FMO	0.0048	0.0169	-
495819780	2-isopropylmalate synthase/homocitrate synthase family protein	0.0014	0.0081	-
495819788	type 4 fimbrial biogenesis protein PilP	0.0775	0.1379	-
495819799	succinate dehydrogenase, flavoprotein subunit	0.0039	0.0096	-
495819822	ribosomal protein S21	0.0149	0.0428	-
495819823	amidophosphoribosyltransferase	0.0029	0.0057	-

495819826	choline dehydrogenase	0.0016	0.0077	-
495819827	electron transfer flavoprotein-ubiquinone oxidoreductase	0.0034	0.0043	-
495819854	formate--tetrahydrofolate ligase	0.0043	0.018	-
495819856	ribonucleoside-diphosphate reductase, beta subunit	0.0148	0.0242	-
495819857	glutamyl-tRNA synthetase	0.0019	0.0042	-
495819876	galactarate dehydratase, putative	0.03	0.0465	-
495819901	30S ribosomal protein S16	0.0187	0.0279	-
495819904	tRNA (guanine-N1)-methyltransferase	0.0092	0.0364	-
495819931	RNA polymerase sigma factor	0.0058	0.0091	-
495819935	glycine amidinotransferase	0.0018	0.0097	-
495819939	segregation and condensation protein B	0.0025	0.0209	-
495819948	nitrogen-fixing NifU domain protein	0.0032	0.0229	-
495819949	acetate--CoA ligase	0.0049	0.0047	-
495819951	dna-directed rna polymerase omega subunit	0.0114	0.036	-
495819966	hydroxyacylglutathione hydrolase	0.0028	0.0326	-
495819968	mannitol transporter	0.0182	0.0202	-
495819974	conserved hypothetical protein	0.0065	0.0447	-
495819983	ribosomal protein S3	0.0085	0.0074	-
495819987	3-demethylubiquinone-9 3-O-methyltransferase	0.0017	0.0109	-
495819988	probable NADH-ubiquinone oxireductase	0.0012	0.0047	-
495820000	ribonuclease E/G	0.0012	0.0079	-
495820003	ribosomal protein S2	0.0092	0.0359	-
495820035	histidine triad protein	0.0014	0.0088	-
495820048	gamma-butyrobetaine,2-oxoglutarate dioxygenase, putative	0.001	0.0058	-
495820064	transporter	0.0405	0.0496	-
495820075	nucleoside diphosphate kinase	0.0025	0.0101	-
495820076	ribonucleoside-diphosphate reductase, alpha subunit	0.0186	0.0324	-
495820077	thiamine pyrophosphate-requiring enzyme	0.0187	0.0152	-
495820079	anthranilate phosphoribosyltransferase	0.0019	0.0096	-
495820099	argininosuccinate synthase	0.0012	0.0059	-
495820101	ribosomal protein S13p/S18e	0.0412	0.2372	-
495820103	CarD-like transcriptional regulator family protein	0.0063	0.0236	-
495820105	D-galactarate dehydratase/Altronate	0.0151	0.0331	-
495820109	methylenetetrahydrofolate dehydrogenase	0.006	0.0152	-
495820111	malate dehydrogenase, NAD-dependent	0.005	0.0324	-
495820118	transporter, major facilitator family	0.0009	0.0055	-
495820126	adenylylsulfate kinase	0.0269	0.0361	-
495820129	phosphoribosylformylglycinamidine synthase I	0.0046	0.0087	-
495820131	conserved hypothetical protein	0.0035	0.008	-
495820133	protein TolQ	0.0158	0.034	-
495820134	phosphate acetyltransferase	0.0017	0.0176	-
495820137	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	0.0058	0.0152	-
495820138	exodeoxyribonuclease I	0.0023	0.0108	-
495820140	sodium Bile symporter family protein	0.0089	0.0081	-
495820141	glycine betaine/L-proline transport system permease protein	0.0055	0.0073	-
495820148	type II Secretion PilT	0.0327	0.0237	-
495820162	chromosome partitioning protein	0.0041	0.0231	-
495820168	preprotein translocase, SecY subunit	0.0087	0.0142	-
495820175	ATP-dependent Clp protease, ATP-binding subunit ClpX	0.0085	0.0177	-
495820177	UDP-3-O-(3-hydroxymyristoyl) N-acetylglucosamine deacetylase	0.0038	0.0118	-
495820199	adenylylsulfate reductase, alpha subunit	0.0227	0.0398	-
495820201	selenocysteine lyase chain A	0.0028	0.0079	-
495820212	Tol-Pal system beta propeller repeat protein TolB	0.0152	0.0185	-
495820215	isoleucyl-tRNA synthetase	0.0024	0.009	-
495820216	Na+/solute symporter, Ssf family	0.3479	0.5602	-
495820229	conserved hypothetical protein	0.0172	0.065	-
495820232	ATP synthase F1, beta subunit	0.0327	0.0893	-
495820235	acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II	0.002	0.0086	-
495820242	ribosomal protein L1	0.0043	0.0247	-
495820249	LexA repressor	0.0099	0.021	-
495820278	ribosomal protein S6	0.0051	0.0166	-
495820288	ribosomal protein L16	0.0157	0.084	-
495820297	DNA uptake lipoprotein	0.0007	0.0059	-
495820299	phosphoribosylaminoimidazole-succinocarboxamide synthase	0.0023	0.0137	-
495820303	conserved hypothetical integral membrane protein	0.0051	0.0256	-
495820307	3-oxoacyl-(acyl-carrier-protein) synthase	0.0045	0.025	-
495820310	protein-export membrane protein SecD	0.0021	0.0137	-
495820312	creatinase	0.0011	0.0043	-
495820326	glycyl-tRNA synthetase, beta subunit	0.0015	0.0183	-
495820335	bacteriorhodopsin	0.8933	0.8271	-

495820356	putative monomeric sarcosine oxidase	0.0168	0.028	-
495820369	dihydroxy-acid dehydratase	0.0087	0.0168	-
495820374	6-O-methylguanine DNA methyltransferase	0.0046	0.0225	-
495820377	spermine/spermidine synthase	0.0052	0.0128	-
495820400	lipoprotein	0.0318	0.085	-
495820415	integral membrane protein MviN	0.0016	0.0108	-
495820435	50S ribosomal protein L5	0.0083	0.066	-
495820449	DNA topoisomerase I	0.0018	0.0048	-
495820452	ribosomal protein L7/L12	0.0424	0.1562	-
495820455	homoserine O-acetyltransferase	0.0025	0.0077	-
495820467	potassium uptake protein	0.0042	0.0136	-
495820469	UDP-N-acetylmuramoylalanyl-D-glutamate	0.0015	0.0057	-
495820550	dimethylglycine dehydrogenase	0.0083	0.0174	-
495820564	ketol-acid reductoisomerase	0.0048	0.0184	-
495820567	short chain dehydrogenase	0.0017	0.0235	-
495820577	mannitol transporter	0.0033	0.0042	-
495820587	threonyl-tRNA synthetase	0.0048	0.0058	-
495820588	methionine adenosyltransferase	0.0033	0.0108	-
495820593	ureidoglycolate dehydrogenase	0.0035	0.0105	-
495820599	D-amino-acid dehydrogenase small chain	0.003	0.0049	-
495820605	DNA gyrase, B subunit	0.0035	0.0117	-
495820615	3-isopropylmalate dehydratase, large subunit	0.0023	0.0087	-
495820620	ribosomal protein L11	0.0137	0.0728	-
495820637	50S ribosomal protein L19	0.0115	0.0654	-
495820640	lysine/ornithine decarboxylase	0.002	0.0158	-
495820642	LSU ribosomal protein L10P	0.0182	0.0518	-
495820662	mechanosensitive ion channel	0.0033	0.007	-
495820672	replicative DNA helicase	0.0031	0.0091	-
495820702	ribosomal protein S12	0.0209	0.0843	-
495820704	rod shape-determining protein MreB	0.0109	0.0289	-
495820731	argininosuccinate lyase	0.002	0.0067	-
495820732	methylmalonate-semialdehyde dehydrogenase	0.0044	0.0197	-
495820740	conserved hypothetical protein	0.0013	0.0127	-
495820747	type II Secretion PilY1	0.0087	0.0144	-
495820749	ABC transporter, permease protein	0.0027	0.0059	-
495820752	acetolactate synthase, large subunit, biosynthetic type	0.0013	0.0058	-
495820765	excinuclease ABC, B subunit	0.0016	0.0059	-
495820766	DNA gyrase, A subunit	0.0048	0.0058	-
495820774	extracellular solute-binding protein, family 1	0.0813	0.3771	-
495820778	HesB/YadR/YfhF	0.0031	0.0136	-
495820791	sarcosine dehydrogenase	0.0007	0.0027	-
495820800	phosphopantothenoylcysteine decarboxylase/phosphopantothenate--cysteine ligase	0.0028	0.0122	-
495820807	phenylalanyl-tRNA synthetase, alpha subunit	0.0018	0.0107	-
495820814	glycine dehydrogenase	0.0039	0.0096	-
495820816	L-lactate dehydrogenase	0.0016	0.0074	-
495820822	ribosomal protein L3	0.0026	0.0415	-
495820835	excinuclease ABC, A subunit	0.0021	0.0109	-
495820838	putative tricarboxylic transport TctC	0.0482	0.1568	-
495820842	glycine betaine transporter, ATP-binding protein	0.0032	0.0075	-
495820843	TctA protein	0.0128	0.0101	-
495820847	ammonium transporter	0.0537	0.193	-
495820849	cytochrome c oxidase, subunit I	0.0425	0.0635	-
495820869	Integral membrane protein DUF6	0.0012	0.0045	-
495820881	putative pseudo-pilin PulG	0.0914	0.0737	-
495820884	DNA-directed DNA polymerase gamma/tau subunit	0.0028	0.0102	-
495820895	flavin Mononucleotide Binding Protein	0.0021	0.005	-
495820896	glutamyl-tRNA(gln) amidotransferase chain B	0.0023	0.0111	-
495820902	amino acid ABC transporter	0.0133	0.0107	-
495820907	histone deacetylase family protein	0.0017	0.008	-
495820912	quinolinate synthetase complex, A subunit	0.0043	0.0149	-
495820918	ribosomal protein L6	0.0054	0.0178	-
495820919	tail-specific proteinase	0.0019	0.0084	-
495820932	hypothetical protein PB7211_1033	0.0038	0.0068	-
495820939	chromosomal replication initiator protein DnaA	0.002	0.0096	-
495820965	adenosylhomocysteinase	0.0075	0.0473	-
495820978	CTP synthase	0.0036	0.007	-
495820980	ribosomal protein L25, Ctc-form	0.0236	0.0226	-
495820991	glycerol kinase	0.0016	0.0035	-
495820993	putative hemimethylated DNA binding domain protein	0.0096	0.0337	-

495821000	NADH dehydrogenase i chain b	0.0334	0.023	-
495821002	NADH dehydrogenase (quinone), G subunit	0.0036	0.0135	-
495821021	proline dipeptidase	0.0015	0.008	-
495821027	transcription termination factor	0.0074	0.0118	-
495821030	penicillin-binding protein 1A	0.0046	0.0076	-
495821035	adenosylmethionine-8-amino-7-oxononanoate aminotransferase	0.0058	0.0306	-
495821051	conserved hypothetical protein	0.0042	0.0458	-
495821059	adenylylsulfate reductase membrane anchor	0.0423	0.0785	-
495821061	GTP-binding protein LepA	0.0016	0.0049	-
495821075	adenylosuccinate lyase	0.0037	0.0192	-
495821081	ATP-dependent protease	0.0083	0.031	-
495821095	single-strand binding protein family protein	0.0119	0.0347	-
495821097	sarcosine oxidase alpha subunit	0.0044	0.0102	-
495821104	amino acid ABC transporter, permease protein	0.013	0.0163	-
495821116	acetyl-coenzyme A carboxylase carboxyl transferase subunit beta	0.0058	0.0164	-
495821128	aconitate hydratase 1	0.0024	0.0069	-
495821130	glycine cleavage system H protein	0.0093	0.0199	-
495821131	ubiquinol-cytochrome c reductase, iron-sulfur subunit	0.0055	0.038	-
495821132	delta-aminolevulinic acid dehydratase	0.0037	0.0085	-
495821133	conserved hypothetical protein	0.0073	0.0308	-
495821134	transcription regulator	0.0039	0.0493	-
495821137	ribosomal protein S20	0.0068	0.0417	-
495821157	NADP-dependent malic enzyme	0.0042	0.0128	-
495821189	imidazoleglycerol-phosphate dehydratase	0.0072	0.0231	-
495821197	30S ribosomal protein S14	0.0146	0.0611	-
495821209	O-acetylhomoserine (thiol)-lyase	0.0096	0.02	-
495821230	ribosomal protein L13	0.0078	0.0514	-
495821233	conserved hypothetical protein	0.0618	0.0818	-
495821248	ribosomal protein L35	0.0048	0.0517	-
495821253	translation elongation factor Ts	0.0043	0.0251	-
495821273	preprotein translocase, SecA subunit	0.0018	0.0103	-
495821291	DNA polymerase III, alpha subunit	0.0014	0.0065	-
495821294	2-polyprenylphenol 6-hydroxylase	0.0008	0.0068	-
495821302	acetolactate synthase, small subunit	0.0036	0.0086	-
495821309	isocitrate lyase	0.0023	0.0192	-
495821311	outer membrane protein TolC	0.0014	0.0095	-
495821319	conserved hypothetical protein	0.0054	0.0366	-
495821325	muropeptide permease	0.0073	0.0089	-
495821331	morn repeat protein	0.0018	0.0158	-
495821333	homocysteine S-methyltransferase	0.002	0.0029	-
495821335	octaprenyl-diphosphate synthase	0.0034	0.0154	-
495821355	glutamate synthase subunit 1	0.0132	0.0608	-
495821364	tol-pal system protein YbgF, putative	0.0089	0.0349	-
495821391	acetyl-CoA carboxylase, biotin carboxyl carrier protein	0.0138	0.0485	-
495821403	dihydrodipicolinate synthase	0.0033	0.0145	-
495821406	cysteinyl-tRNA synthetase	0.0015	0.0112	-
495821409	cytochrome c oxidase assembly protein CtaG	0.0051	0.023	-
495821412	sarcosine oxidase	0.0061	0.0176	-
495821433	3-deoxy-7-phosphoheptulonate synthase	0.011	0.0477	-
495821442	NADH oxidoreductase (quinone), F subunit	0.0021	0.0059	-
495821454	AcrB/AcrD/AcrF family protein	0.0047	0.0058	-
495821481	chaperonin GroS	0.0543	0.2835	-
495821507	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	0.002	0.0035	-
495821518	membrane protein	0.0238	0.0371	-
495821519	RIP metalloprotease RseP	0.006	0.0103	-
495821522	hypothetical protein PB7211_1417	0.017	0.0204	-
495821531	probable integral membrane protein	0.0043	0.0047	-
495821538	conserved hypothetical protein	0.0084	0.0138	-
495821542	sarcosine oxidase beta subunit	0.0072	0.0053	-
495821548	NADH dehydrogenase i chain l	0.0043	0.0064	-
495821550	glycine betaine transporter, ATP-binding protein	0.0042	0.009	-
495821560	type II Secretion System PilC	0.018	0.0184	-
495821568	metallo-beta-lactamase family protein	0.002	0.0091	-

Table 3.S5. Gene expression ratios in the gammaproteobacterium HTCC2080 bin.

NCBI GI	Protein Description	Mean Expression Ratio (FL)	Mean Expression Ratio (PA)	PA Regulation
494440354	hypothetical protein	0.0154	0.0868	Up
494440403	dnaK protein	0.0403	0.3517	Up
494440409	bifunctional isocitrate dehydrogenase kinase/phosphatase protein	0.0375	0.3177	Up
494440423	Secreted Zn-dependent peptidase, insulinase family protein	0.0529	0.2275	Up
494440451	probable acid-CoA ligase	0.0478	0.3195	Up
494440531	Alpha amylase, catalytic subdomain	0.1094	1.0107	Up
494440576	C-terminal processing peptidase	0.0221	0.1555	Up
494440618	Leucyl aminopeptidase	0.0399	0.3909	Up
494440739	isoquinoline 1-oxidoreductase, beta subunit	0.4149	2.4925	Up
494440758	enoyl-CoA hydratase/isomerase family protein	0.061	0.7257	Up
494441092	signal peptide peptidase SppA, 67K type	0.023	0.0509	Up
494441139	acyl-CoA synthase	0.025	0.1787	Up
494441143	acyl-CoA synthase	0.1803	1.6682	Up
494441150	transcriptional regulator, TetR family protein	0.156	1.4948	Up
494441152	putative lipoprotein	0.1305	0.3226	Up
494441170	putative TonB-dependent receptor	0.0562	0.1313	Up
494441263	invasion protein IbeA	0.0198	0.121	Up
494441273	peptidase M20	0.006	0.0568	Up
494441340	putative acyl coenzyme A dehydrogenase (HcaD-like) protein	0.0524	0.2198	Up
494441342	hypothetical protein	0.0209	0.2073	Up
494441355	aminotransferase, class III	0.0608	0.1833	Up
494441435	Magnesium-chelatase, subunit H	0.0199	0.2316	Up
494441467	cobalt-zinc-cadmium resistance protein (CzcA)-like	0.0113	0.1214	Up
494441545	cobalt-zinc-cadmium resistance protein (CzcA)-like	0.0775	0.3368	Up
494441547	phosphoenolpyruvate carboxykinase (ATP)	0.1487	0.9292	Up
494441555	ribonucleotide-diphosphate reductase alpha subunit	0.0284	0.0945	Up
494441566	Cell division protein FtsI/penicillin-binding protein 2	1.188	11.6239	Up
494441632	oxidoreductase, short chain dehydrogenase/reductase family protein	0.105	0.6064	Up
494441786	acyl-CoA dehydrogenase-like protein	0.028	0.0857	Up
494441804	putative indolepyruvate ferredoxin oxidoreductase alpha subunit	0.0154	0.079	Up
494441857	TonB-dependent receptor	0.0247	0.1271	Up
494442038	2,4-dienoyl-CoA reductase FadH1	0.2144	1.5905	Up
494442120	glutathione S-transferase	0.0032	0.0405	Up
494442294	putative transcriptional regulator, Fis family protein	0.0489	0.1822	Up
494442514	Acetyltransferase	0.0083	0.0724	Up
494442681	helicase, ATP-dependent	0.0068	0.1158	Up
494442774	hypothetical protein	0.0252	0.0736	Up
494442862	Glutamate synthase domain 2	0.048	0.1487	Up
494442911	putative type 4 fimbrial biogenesis pily1-related protein signal peptide	0.0367	0.1738	Up
494443014	L-carnitine dehydratase/bile acid-inducible protein F	0.0301	0.1995	Up
494443015	5-oxoprolinase (ATP-hydrolyzing)	0.0525	0.4344	Up
494443109	TPR domain protein	0.012	0.0414	Up
494443148	acyl-CoA dehydrogenase-like protein	0.0061	0.0316	Up
494443189	acriflavin resistance protein	0.0089	0.0685	Up
494440127	protease subunit HflC	0.0559	0.0238	Down
494440167	Beta-glucosidase	0.1017	0.0507	Down
494440245	Ribosomal protein L10	1.4544	0.5637	Down
494440247	DNA-directed RNA polymerase subunit beta	0.2435	0.1589	Down
494440267	50S ribosomal protein L5	0.4236	0.1986	Down
494440272	30S ribosomal protein S5	0.8106	0.2938	Down
494440275	protein translocase subunit SecY	0.5032	0.2822	Down
494440652	Ankyrin	0.2788	0.1054	Down
494440764	aminotransferase, class III	0.152	0.0712	Down
494440790	putative unknown membrane associated protein	0.0873	0.0257	Down
494440793	Branched-chain amino acid aminotransferase I	0.2981	0.1329	Down
494441146	Outer membrane protein	0.0547	0.031	Down
494441371	Pyridoxal-dependent decarboxylase	0.2288	0.0358	Down
494441741	Cytochrome oxidase assembly protein	0.067	0.0155	Down
494442055	formate--tetrahydrofolate ligase	0.0599	0.016	Down
494442274	putative TonB-dependent outer membrane receptor	0.6055	0.2498	Down
494442351	enoyl-CoA hydratase/isomerase family protein	0.0753	0.0212	Down
494442352	acetyl-CoA carboxylase carboxyltransferase	0.0674	0.0251	Down

494442353	hypothetical protein	0.07	0.024	Down
494442365	Flagellar biosynthesis protein FliR	0.2362	0.1502	Down
494442593	translation initiation factor IF-3	0.6376	0.0998	Down
494442597	threonyl-tRNA synthetase	0.0645	0.0274	Down
494442812	protein-export protein SecB	0.1159	0.0375	Down
494442833	NAD(P) transhydrogenase subunit alpha	0.1068	0.0314	Down
494442915	Inorganic diphosphatase	0.1243	0.0813	Down
494443095	Na(+)-translocating NADH-quinone reductase subunit F	0.1295	0.0695	Down
494443140	ribosomal 5S rRNA E-loop binding protein Ctc/L25/TL5	0.2073	0.0469	Down
494440042	ribulose-phosphate 3-epimerase	0.035	0.0274	-
494440089	metallo-beta-lactamase family protein	0.0231	0.0242	-
494440093	hypothetical protein	0.0134	0.0438	-
494440094	hypothetical protein	0.0087	0.0127	-
494440108	Gamma-glutamyltransferase	0.0119	0.0677	-
494440168	putative sodium/hexose cotransport protein	0.0521	0.0306	-
494440170	Glycine hydroxymethyltransferase	0.0923	0.0965	-
494440248	DNA-directed RNA polymerase subunit beta'	0.3469	0.4348	-
494440251	translation elongation factor G	0.1887	0.3033	-
494440252	conserved repeat domain protein	0.0099	0.0119	-
494440261	30S ribosomal protein S3	0.8761	0.789	-
494440262	50S ribosomal protein L16	0.119	0.2637	-
494440279	DNA-directed RNA polymerase subunit alpha	0.2874	0.2363	-
494440293	UDP-glucose pyrophosphorylase	0.0441	0.0378	-
494440303	acyl-CoA dehydrogenase-like protein	0.014	0.0245	-
494440310	Oar-like outer membrane protein protein, OmpA family	0.1152	0.1737	-
494440380	hypothetical protein	0.0192	0.0263	-
494440393	hypothetical protein	0.0438	0.0315	-
494440413	transcriptional regulator, TetR family protein	0.0024	0.0602	-
494440424	pyruvate dehydrogenase subunit E1	0.0704	0.0576	-
494440428	acyl-CoA synthase	0.0223	0.0122	-
494440448	TonB-dependent receptor	0.0157	0.0285	-
494440463	putative esterase	0.0186	0.0196	-
494440471	hypothetical protein	0.018	0.0638	-
494440472	putative ferredoxin	0.0188	0.0437	-
494440474	Transcriptional regulator	0.01	0.0349	-
494440489	Cytochrome c-type biogenesis protein CcmF	0.0328	0.0639	-
494440493	chromosome segregation protein	0.0265	0.0298	-
494440496	transcriptional regulatory protein	0.0052	0.1165	-
494440501	citrate synthase	0.0363	0.1259	-
494440504	succinate dehydrogenase flavoprotein subunit	0.0666	0.0642	-
494440507	dihydrolipoamide acetyltransferase	0.1104	0.0719	-
494440511	Succinyl-CoA synthetase, alpha subunit	0.2277	0.4655	-
494440530	Multi-sensor Hybrid Histidine Kinase	0.015	0.078	-
494440535	major facilitator family transporter	0.0056	0.0393	-
494440566	hypothetical protein	0.0734	0.0535	-
494440570	thioredoxin reductase 1	0.0424	0.0755	-
494440586	hypothetical protein	0.0009	0.0114	-
494440593	acyl-CoA synthase	0.005	0.0122	-
494440606	sodium/proton antiporter	0.0392	0.0437	-
494440623	hypothetical protein	0.0515	0.0712	-
494440631	long-chain-fatty-acid--CoA ligase	0.0233	0.0405	-
494440650	ATP-dependent protease ATP-binding subunit	0.0694	0.1366	-
494440658	3-oxoacyl-(acyl carrier protein) synthase	0.2047	0.0787	-
494440666	DNA polymerase III delta prime subunit	0.1052	0.0713	-
494440687	sensor histidine kinase	0.0385	0.0522	-
494440715	hypothetical protein	0.0057	0.0113	-
494440752	probable choline transporter	0.0171	0.0163	-
494440761	hypothetical protein	0.0208	0.0159	-
494440791	periplasmic protein TonB	0.1021	0.1436	-
494440794	Lipid A export ATP-binding/permease protein MsbA	0.0925	0.1682	-
494440795	TonB system biopolymer transport component	0.2686	0.1925	-
494440796	TonB system biopolymer transport component	0.1305	0.104	-
494440804	Ubiquinone biosynthesis hydroxylase, UbiH/UbiF/VisC/COQ6	0.0433	0.2344	-
494440806	aminopeptidase P II	0.0154	0.034	-
494440814	D-erythrose-4-phosphate dehydrogenase	0.0336	0.1945	-
494440818	TonB-dependent receptor	0.3984	0.2619	-
494440824	S-adenosyl-L-homocysteine hydrolase	0.0568	0.0634	-
494440844	RNA polymerase sigma factor	0.2212	0.1325	-
494440852	50S ribosomal subunit protein L28	0.1896	0.1055	-
494440860	Exonuclease III	0.0567	0.0245	-

494440886	Sodium-transporting two-sector ATPase	0.2831	0.2967	-
494440889	F-type H+-transporting ATPase c chain	0.7616	0.473	-
494440890	F0F1 ATP synthase subunit A	0.532	0.282	-
494440893	Putative ParA family protein	0.0523	0.0365	-
494440897	putative inner membrane protein translocase component YidC	0.0155	0.0318	-
494440905	DNA gyrase subunit B	0.0228	0.0485	-
494440908	glycyl-tRNA synthetase subunit alpha	0.0268	0.0819	-
494440911	potassium uptake protein TrkH	0.013	0.0388	-
494440912	potassium transporter peripheral membrane component	0.0147	0.0161	-
494440923	cytochrome c oxidase, subunit II	0.3062	0.5784	-
494440924	Cytochrome-c oxidase	0.2258	0.3464	-
494440926	cytochrome c oxidase, subunit III	0.1771	0.1709	-
494440936	hypothetical protein	0.0655	0.1077	-
494440942	N-acetylglutamate synthase	0.0159	0.0309	-
494440961	TPR domain protein	0.003	0.0082	-
494441008	flagellar biosynthesis protein	0.0519	0.4975	-
494441030	transcription termination factor Rho	0.065	0.135	-
494441060	deoxyribodipyrimidine photolyase, putative	0.0061	0.0078	-
494441063	dihydroorotate dehydrogenase/oxidoreductase, FAD-binding protein	0.0359	0.0457	-
494441066	hypothetical protein	0.0395	0.1249	-
494441070	hypothetical protein	0.0279	0.1783	-
494441078	hypothetical protein	0.0308	0.0315	-
494441085	hypothetical protein	0.0048	0.0105	-
494441138	Amidohydrolase family protein	0.0118	0.024	-
494441144	hypothetical protein	0.0238	0.029	-
494441145	hypothetical protein	0.0718	0.0737	-
494441153	TonB-dependent receptor	0.0857	0.0749	-
494441188	hypothetical protein	0.0087	0.0084	-
494441189	methionine synthase I	0.0068	0.0271	-
494441191	methionine synthase I	0.0362	0.033	-
494441216	hypothetical protein	0.0056	0.0074	-
494441219	alcohol dehydrogenase, zinc-containing	0.0086	0.0352	-
494441226	Alcohol dehydrogenase, class IV	0.0146	0.0246	-
494441251	Alcohol dehydrogenase large subunit	0.0217	0.0307	-
494441260	hypothetical protein	0.0078	0.0042	-
494441261	TPR domain protein	0.0092	0.027	-
494441262	TonB-dependent receptor	0.0848	0.1119	-
494441265	hypothetical protein	0.0075	0.0146	-
494441274	MFS permease	0.0568	0.0609	-
494441287	UvrD/REP helicase	0.0023	0.0154	-
494441290	rarD protein	0.0376	0.0725	-
494441312	GTP-binding protein LepA	0.1105	0.0709	-
494441339	hypothetical protein	0.0087	0.0076	-
494441389	aminomethyl transferase family protein	0.0067	0.009	-
494441408	phosphoenolpyruvate carboxylase	0.032	0.0151	-
494441412	glucose-6-phosphate 1-dehydrogenase	0.0846	0.1687	-
494441414	glucokinase	0.1339	0.099	-
494441439	Photosynthetic reaction center H-chain	0.0038	0.0152	-
494441544	Secretion protein HlyD	0.1055	0.1084	-
494441560	Bacterioferritin	0.0098	0.0364	-
494441567	putative rod shape-determining protein RodA	1.782	19.8767	-
494441570	D-alanyl-D-alanine carboxypeptidase	0.0139	0.0172	-
494441573	peptide chain release factor RF-3	0.0454	0.1303	-
494441576	glutamine synthetase, putative	0.0791	0.1553	-
494441607	Kynureninase	0.0098	0.0272	-
494441621	Cysteine synthase	0.0435	0.0438	-
494441638	Acyl-CoA hydrolase-like protein	0.0412	0.0734	-
494441639	short chain dehydrogenase	0.0617	0.0457	-
494441640	acetyl-coenzyme A synthetase	0.0202	0.0294	-
494441641	Putative cyclase	0.0108	0.0321	-
494441642	taurine dioxygenase	0.0231	0.0597	-
494441643	sodium/alanine symporter	0.0176	0.037	-
494441646	Catalase	0.1254	1.049	-
494441659	alcohol dehydrogenase, iron-containing	0.0574	0.15	-
494441669	ATP-dependent metalloprotease FtsH	0.093	0.1107	-
494441685	SmpB protein	0.1181	0.1418	-
494441688	30S ribosomal protein S15	0.5528	0.3424	-
494441696	DNA polymerase III, epsilon subunit	0.0101	0.0903	-
494441708	putative integral membrane protein	0.0196	0.0594	-

494441716	monooxygenase, flavin-binding family protein	0.0261	0.026	-
494441730	leucyl-tRNA synthetase	0.0112	0.0958	-
494441756	LuxO repressor protein	0.0318	0.0139	-
494441824	amidophosphoribosyltransferase	0.0578	0.0436	-
494441830	tryptophan synthase subunit beta	0.0541	0.0673	-
494441856	acetoin dehydrogenase complex, E1 component, alpha subunit	0.0137	0.0467	-
494441870	cytochrome P450	0.0119	0.021	-
494441872	glutathione-dependent formaldehyde-activating, GFA	0.0103	0.0211	-
494441898	putative PQQ-dependent polyvinyl alcohol dehydrogenase precursor	0.0096	0.0228	-
494441911	putative TonB-dependent receptor	0.0347	0.039	-
494441918	cyclohexanone monooxygenase	0.0098	0.0144	-
494441942	hypothetical protein	0.0146	0.1202	-
494441945	putative integral membrane transport protein	0.006	0.018	-
494441946	cytochrome P450 family protein	0.0134	0.0541	-
494441963	Phosphoesterase, PA-phosphatase related protein	0.036	0.0475	-
494441970	probable oxidoreductase	0.015	0.0179	-
494441971	hypothetical protein	1.1194	5.6645	-
494441999	arylsulfatase B precursor	0.0065	0.0041	-
494442009	hypothetical protein	0.0257	0.0325	-
494442010	hypothetical protein	0.0212	0.0482	-
494442013	putative beta-ketoacyl synthase	0.0488	0.1003	-
494442054	N-acyl-D-glutamate amidohydrolase	0.0175	0.0232	-
494442059	formate dehydrogenase alpha subunit	0.034	0.0245	-
494442061	sarcosine dehydrogenase	0.0372	0.0307	-
494442071	hypothetical protein	0.0658	0.0237	-
494442074	hypothetical protein	0.0331	0.0825	-
494442088	putative saccharopine dehydrogenase	0.02	0.0175	-
494442089	4-aminobutyrate aminotransferase	0.0074	0.0163	-
494442093	YfmJ	0.0127	0.0845	-
494442097	hypothetical protein	0.0159	0.037	-
494442101	sarcosine oxidase, alpha subunit family protein	0.0064	0.0166	-
494442106	hypothetical protein	0.014	0.0373	-
494442110	TonB-dependent receptor	0.0685	0.041	-
494442112	TonB-dependent receptor	0.0139	0.0334	-
494442122	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II	0.0046	0.0223	-
494442149	RND family efflux transporter	0.0759	0.0433	-
494442150	proton/peptide symporter family protein	0.0076	0.0192	-
494442171	hypothetical protein	0.0122	0.0222	-
494442186	ammonium transporter	0.0878	0.1078	-
494442198	sulfatase family protein	0.0231	0.0744	-
494442226	Xaa-Pro aminopeptidase family enzyme	0.0296	0.2694	-
494442232	probable long chain fatty acid CoA ligase	0.0211	0.0389	-
494442260	hypothetical protein	0.0416	0.0326	-
494442261	hypothetical protein	0.0134	0.0231	-
494442286	Spermidine/putrescine ABC transporter ATP-binding subunit	0.0482	0.0883	-
494442290	hypothetical protein	0.0093	0.0391	-
494442300	tRNA nucleotidyltransferase, putative	0.0158	0.1236	-
494442308	Parvulin-like peptidyl-prolyl isomerase	0.048	0.035	-
494442354	transcriptional regulator	0.0277	0.0128	-
494442371	DNA gyrase subunit A	0.0584	0.1392	-
494442407	Helix-turn-helix, AraC type	0.025	0.0473	-
494442433	translocase	0.0181	0.0315	-
494442444	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase	0.0428	0.1319	-
494442459	ubiquinol--cytochrome c reductase, cytochrome b	0.1952	0.1547	-
494442464	hypothetical protein	0.1431	0.0486	-
494442489	Twin-arginine translocation protein TatB	0.0376	0.0476	-
494442491	Glutamyl-tRNA(Gln) amidotransferase A subunit	0.0157	0.0823	-
494442502	hypothetical protein	0.0068	0.0106	-
494442516	possible NADH oxidoreductase	0.0169	0.2043	-
494442528	cAMP-binding protein - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinase	0.0646	0.0358	-
494442529	adenylate cyclase PLUS two component hybrid sensor and regulator	0.0237	0.0927	-
494442590	phenylalanyl-tRNA synthetase alpha subunit	0.1198	0.1435	-
494442595	hypothetical protein	0.0101	0.0196	-
494442636	hypothetical protein	0.0035	0.0204	-
494442668	oligopeptide ABC transporter periplasmic oligopeptide-binding protein	0.0068	0.0067	-
494442669	Phosphoglycerate mutase 1	0.0216	0.069	-

494442674	alkaline phosphatase, putative	0.0047	0.0571	-
494442704	isocitrate lyase	0.2878	0.2803	-
494442711	hypothetical protein	0.0803	0.057	-
494442714	chaperone protein HscA	0.0814	0.0533	-
494442718	cysteine desulfurase	0.1292	0.1793	-
494442724	phosphoribosylformylglycinamidine synthase	0.0272	0.0534	-
494442728	Alpha-methylacyl-CoA racemase	0.0051	0.0132	-
494442740	protein kinase C inhibitor	0.0299	0.0429	-
494442749	probable phosphate transporter	0.0895	0.1058	-
494442750	hypothetical protein	0.0537	0.0954	-
494442752	Phosphate-selective porin O and P	0.0214	0.0179	-
494442788	hypothetical protein	0.1741	0.1196	-
494442799	efflux transporter, RND family, MFP subunit	0.0287	0.0367	-
494442816	Glutamate--ammonia ligase	0.0225	0.0373	-
494442821	DNA topoisomerase IV subunit A	0.0177	0.0497	-
494442824	translation elongation factor P	0.0446	0.0304	-
494442832	anti-sigm factor, ChrR	0.1323	0.1085	-
494442836	ABC transporter, ATP-binding/permease protein	0.0141	0.0565	-
494442883	L-threonine 3-dehydrogenase	0.035	0.0499	-
494442897	NAD synthase	0.024	0.0391	-
494442898	competence protein ComL	0.0226	0.0328	-
494442913	diguanylate cyclase/phosphodiesterase (GGDEF & EAL domains)	0.0769	0.0773	-
494442917	phosphatidylserine synthase	0.0519	0.0218	-
494442952	Bacterioferritin	0.0563	0.0845	-
494442984	uridylate kinase	0.0854	0.0876	-
494442985	ribosome recycling factor	0.1446	0.0755	-
494442990	probable outer membrane protein	0.065	0.0418	-
494442991	hypothetical protein	0.1416	0.4198	-
494443006	putative citrate lyase beta subunit	0.0161	0.0259	-
494443007	MmgE/PrpD	0.0079	0.0149	-
494443008	4-hydroxyphenylacetate 3-hydroxylase family protein	0.0067	0.0075	-
494443013	3-oxoadipate enol-lactone hydrolase	0.0136	0.0239	-
494443016	5-oxoprolinase (ATP-hydrolyzing)	0.0121	0.0242	-
494443026	Methenyltetrahydrofolate cyclohydrolase	0.0204	0.0501	-
494443027	major facilitator superfamily MFS_1	0.0117	0.089	-
494443062	hypothetical protein	0.0146	0.0464	-
494443085	lipid A export ATP-binding/permease protein MsbA	0.0257	0.027	-
494443087	MotA/TolQ/ExbB proton channel	0.0238	0.0747	-
494443098	hypothetical protein	0.0576	0.0315	-
494443100	Bacterial ring hydroxylating dioxygenase, alpha subunit:Immunoglobulin/major histocompatibility	0.0092	0.0267	-
494443101	aminomethyltransferase	0.0242	0.0353	-
494443103	phytoene dehydrogenase	0.0114	0.0117	-
494443108	hypothetical protein	0.0018	0.0256	-
494443110	hypothetical protein	0.0134	0.058	-
494443111	TonB-dependent receptor	0.0535	0.0596	-
494443119	hypothetical protein	0.0084	0.0233	-
494443150	hypothetical protein	0.0103	0.0105	-
494443158	sulfatase family protein	0.0275	0.106	-
494443166	Na(+)-translocating NADH-quinone reductase subunit B	0.1354	0.1162	-
494443168	glyceraldehyde-3-phosphate dehydrogenase	0.0498	0.0284	-
494443175	DNA topoisomerase I	0.0153	0.0309	-
494443180	hypothetical protein	0.0355	0.2272	-
494443181	hypothetical protein	0.0316	0.227	-
494443188	hypothetical protein	0.0192	0.5064	-
494443204	hypothetical protein	0.0397	0.0468	-
494443205	Isoleucyl-tRNA synthetase, class Ia	0.034	0.037	-

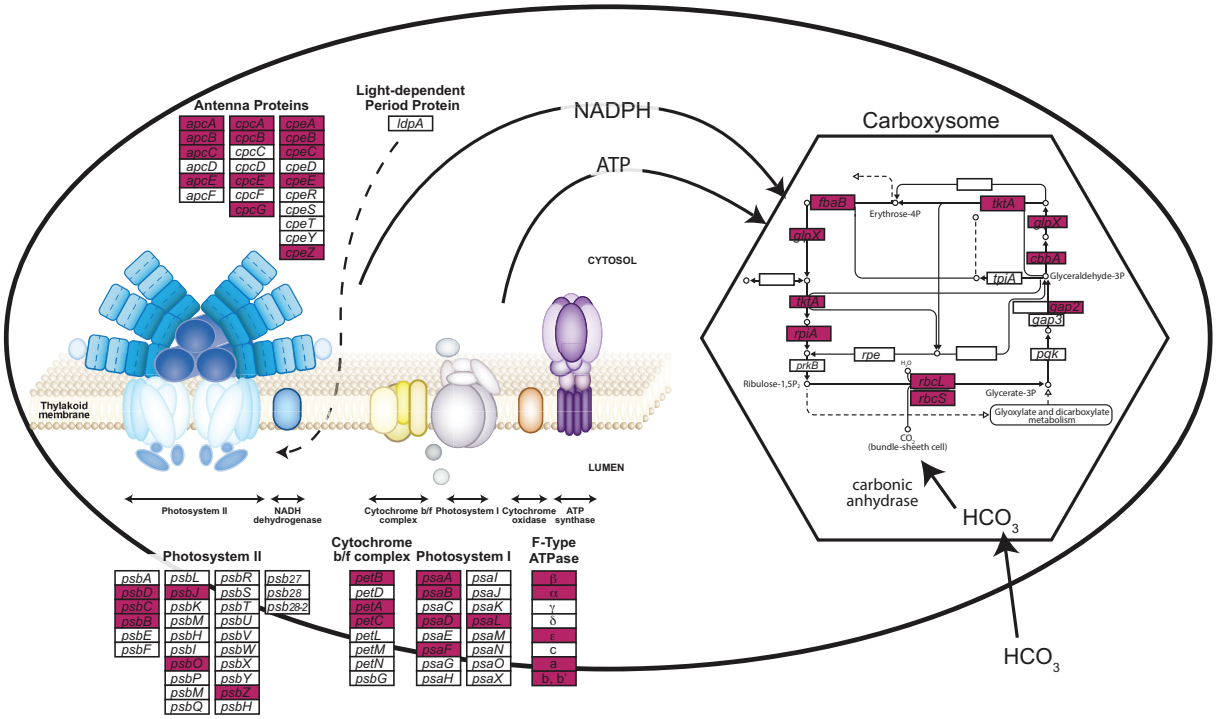


Figure 3.S6. Photosynthesis-related genes in the Synechococcus sp. CB0205 bin exhibiting significantly different regulation between microenvironments. Gene functional annotations are based on KEGG pathways ko00195, ko00196, ko00710. Maroon = upregulated in free-living cells. No photosynthesis genes were upregulated in particle-associated cells.

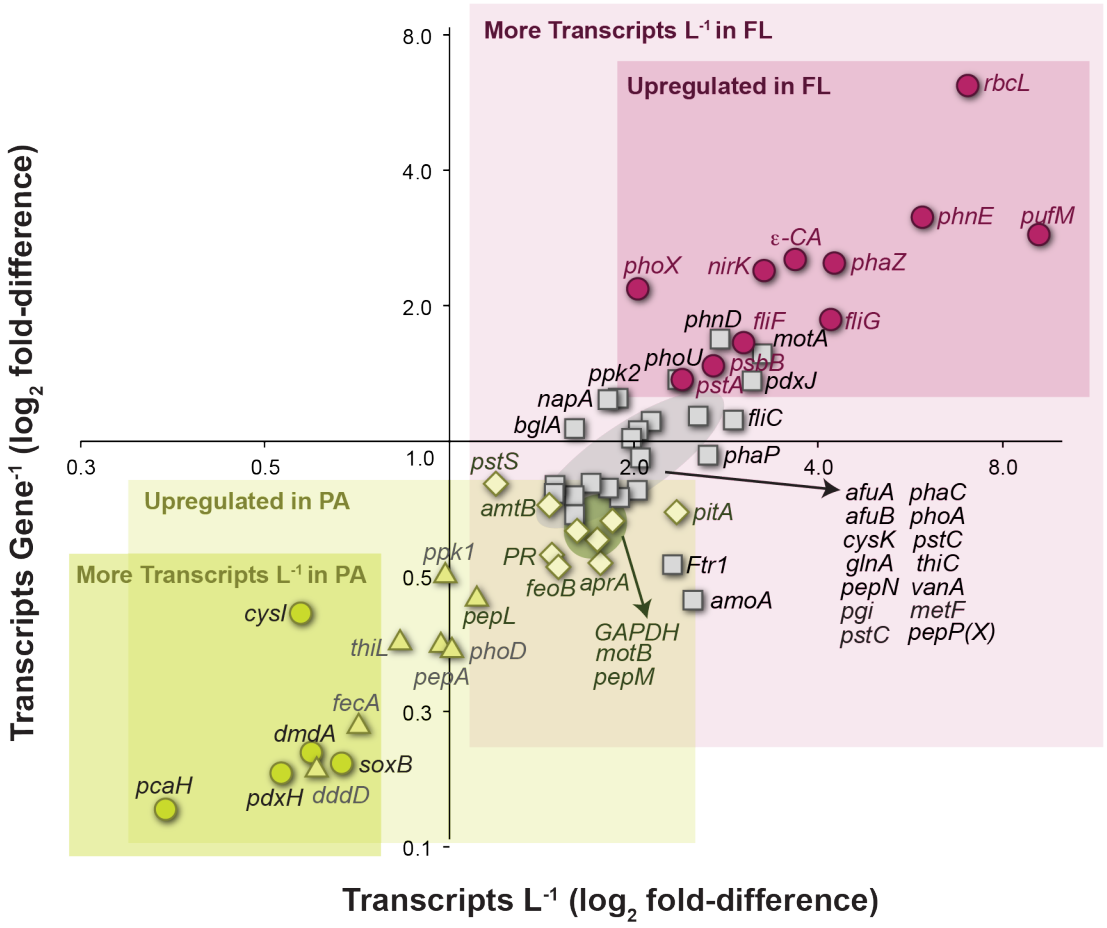


Figure 3.5 Outcome of differential transcript inventories and expression ratios for biogeochemically-relevant genes. Inventory ratios are plotted on the X-axis, gene expression ratios on the Y-axis. Green symbols: genes upregulated in the PA fraction and in greater abundance in the PA community (circles), no difference in abundance between PA and FL communities (triangles), or greater abundance in the FL community (diamonds). Gray symbols: genes not upregulated in either fraction, but in greater abundance in the FL community. Maroon symbols: genes both upregulated and more abundant in the FL community. Gene abbreviations are as in Fig. 3.2.

Supplemental Methods:

Sample Collection

Duplicate surface samples were collected from the Amazon River plume aboard the R/V Knorr in June 2010 (4 52.71'N, 51 21.59'W) during a period of high river discharge. The collection site (Station 10, 4° 52.71'N, 51° 21.59'W; S = 21.0; T = 29.6°C), located ~ 500 Km to the north of the Amazon River mouth, was characterized by the presence of coastal diatoms in the top 8 m of the water column. Sampling was conducted between 0700 and 0900 local time by gently impeller pumping (modified Rule 1800 submersible sump pump) surface water through 10 m of tygon tubing (3 cm) to the ship's deck where it then flowed through a 156 μm mesh into 20 L carboys. In the lab, cells were partitioned into two size fractions by sequential filtration (using a Masterflex peristaltic pump) of the pre-filtered seawater through a 2.0 μm pore-size, 142 mm diameter polycarbonate (PCTE) membrane filter (Sterlitech Corporation, Kent, CWA) and a 0.22 μm pore-size, 142 mm diameter Supor membrane filter (Pall, Port Washington, NY). Metagenomic and non-selective metatranscriptomic analyses were conducted on both pore-size filters; poly(A)-selected (eukaryote-dominated) metatranscriptomic analyses were conducted only on the larger pore-size filter (2.0 μm pore-size). All filters were immediately submerged in RNAlater (Applied Biosystems, Austin, TX) in sterile 50 mL conical tubes, incubated at room temperature overnight and then stored at -80^{o}C until extraction. Filtration and stabilization of each sample was completed within 30 min of water collection. Additionally, samples were collected and preserved for flow cytometry. After staining with 1X SYBR Green II and incubating in the dark for at least 10 minutes, samples were analyzed using a Beckman Coulter Cyan flow cytometer with an excitation wavelength of 488nm and emission detection filter of

530nm. Bacterial cells were quantified using a combination of light scatter and fluorescence detection with a bead standard of known concentration used to calculate the volume counted by the machine.

Total Community RNA Processing for Metatranscriptomes

Prior to RNA extraction, the filters were thawed, removed from the preservative solution, placed in Whirl-Pak® bags (Nasco, Fort Artkinson, WI), and flash-frozen in liquid nitrogen. RNA extraction and DNA removal were carried out as previously described (Gifford *et al.,* 2011, Poretsky *et al.,* 2009a, Poretsky *et al.,* 2009b). In brief, a lysis tube was prepared for each sample consisting of a sterile 50 mL conical tube containing 8 mL of RLT Lysis Solution (Qiagen, Valencia, CA), 3 grams of RNA PowerSoil beads (Mo-Bio, Carlsbad, CA), and internal standards (described below). Filters inside the bags were broken into small pieces using a rubber mallet and transferred to the lysis tubes. Tubes were vortexed for 10 min to lyse cells, and RNA was purified from cell lysate using an RNeasy Kit (Qiagen, Valencia, CA) followed by two successive treatments with the Turbo DNA-free kit (Invitrogen, Carlsbad, CA) to completely remove residual DNA. Ribosomal RNA (rRNA) was selectively removed using community-specific biotinylated-rRNA probes prepared with DNA from a sample collected simultaneously (Stewart *et al.,* 2010). To maximize the removal of rRNA, probes were created for Bacterial and Archaeal 16S and 23S rRNA and Eukaryotic 18S and 28S rRNA. Probe-bound rRNA was removed via hybridization to streptavidin-coated magnetic beads (New England Biolabs, Ipswich, MA), and successful removal of rRNA from the samples was confirmed using either an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA) or a Bioanalyzer (Agilent Technologies, Santa Clara, CA). rRNA-depleted samples were linearly

amplified using the MessageAmp II-Bacteria Kit (Applied Biosystems, Austin, TX), and amplified mRNA was converted into cDNA using the Superscript III First Strand synthesis system (Invitrogen, Carlsbad, CA) with random primers, followed by the NEBnext mRNA second strand synthesis module (New England Biolabs, Ipswich, MA), both according to manufacturer protocols. Synthesized cDNA was purified using the QIAquick PCR purification kit (Qiagen, Valencia, CA) followed by EtOH precipitation, resuspended in 100 μL of TE buffer and stored at -80^{o} C until library preparation for sequencing.

Poly(A)-tail Selected RNA Processing for Metatranscriptomes

To ensure sufficient coverage of eukaryotic transcriptomes, a second metatranscriptome protocol was used that selectively sequenced messages with poly(A) tails; this was carried out for the >2.0 μm pore-size filter only. Samples were prepared as described above with the following exceptions. Each lysis tube was prepared with 9 mL of RLT Lysis Solution, 250 μL of zirconium beads (OPS Diagnostics, Lebanon, NJ, USA), and an internal poly(A)-tailed mRNA standard (see below). Following lysis, poly(A)-tailed mRNA was isolated from total RNA using an Oligotex mRNA kit (Qiagen, Valencia, CA), and mRNA was linearly amplified with a MessageAmp II-aRNA Amplification Kit (Applied Biosystems, Austin, TX). Double stranded cDNA was prepared as described above except without ethanol precipitation.

DNA Processing for Metagenomes

DNA was extracted and purified as previously described (Crump *et al.,* 1999, Crump *et al.,* 2003, Zhou *et al.,* 1996) with some modification. Briefly, each filter was thawed, removed from the preservative solution, and rinsed three times in autoclaved, filter-sterilized, 0.1%

phosphate-buffered saline (PBS) to remove any residual RNAlater. Each filter was shattered as described above and placed in a tube containing DNA extraction buffer [DEB: 0.1 M Tris-HCl (pH 8), 0.1 M Na-EDTA (pH 8), 0.1 M $Na_2H_2PO_4$ (pH 8), 1.5 M NaCl, 5% CTAB]. All liquid from the rinses as well as the original RNAlater was pushed through a 0.2μm Sterivex-GP filter capsule (EMD Millipore, Billerica, MA), which was subsequently rinsed 3 times to salvage any lost cells. The capsule was opened and the filter sliced into pieces and added to the tube with the original membrane filter and an internal genomic DNA standard (described below). Following treatments with proteinase-K, lysozyme, and sodium dodecyl sulfate, DNA was purified via phenol:chloroform extraction and isopropanol precipitation.

Sequencing

cDNA and DNA was sheared ultrasonically to ~200-250 base pair fragments and TruSeq libraries (Illumina Inc., San Diego, CA) were constructed for paired-end (2 x 150) sequencing using the Illumina Genome Analyzer IIx sequencing platform (Illumina Inc., San Diego, CA).

Internal Standards

Omics processing included the addition of internal standards (Satinsky *et al.,* 2013) to allow for calculation of volume-based absolute copy numbers for each gene or transcript type, rather than just relative quantification (i.e., counts L^{-1} rather than percent of sequence library). Two mRNA standards without poly(A) tails (to mimic prokaryotic and organelle mRNAs) were synthesized by *in vitro* transcription using a method modified from (Gifford *et al.,* 2011). The standards were constructed by linearizing a pTXB1 vector (New England Biolabs, Ipswich, MA) with *Nco*I restriction enzyme (New England Biolabs, Ipswich, MA) or pFN18A Halotag T7

Flexi Vector (Promega, Madison, WI) with *Bam*HI restriction enzyme (New England Biolabs, Ipswich, MA). Each was purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. The 5' nucleotide overhangs were removed using Mung Bean Nuclease (New England Biolabs, Ipswich, MA), followed by purification via phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. Complete digestion of the vector was confirmed on a 1% agarose gel. The DNA fragment was then transcribed *in vitro* using the Riboprobe *in vitro* Transcription System (Promega, Madison, WI) according to the manufacturer's protocol using a T7 RNA polymerase to create 916 nt (pTXB1 standard) or 970 nt (pFN18A) artificial transcripts. Residual DNA was removed using RQ1 RNase-Free DNase and the RNA was purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. The RNA standards were quantified using the Quant-iT Ribogreen RNA Reagent and Kit (Invitrogen, Carlsbad, CA), and RNA nucleotide length was confirmed with an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA). A known number of each standard (pTXB1 = 2.104 x 10^{10} copies; pFN18A = 1.172 x 10^{10} copies) was added independently to each lysis tube immediately prior to the addition of the sample filter.

An mRNA standard with a poly(A) tail (to mimic eukaryotic nuclear mRNA) was created from an HAP-1 Protelomerase viral gene. To create the standard, a 544 bp amplicon containing a poly(A) tail and a T7 promoter was produced from the template DNA through PCR. The PCR amplicons were then used as the template DNA for an *in vitro* transcription reaction to produce the resulting 499 nucleotide poly(A)-tailed mRNA. The addition of 2.0 x 10^9 copies of the standard occurred immediately prior to cell lysis.

The genomic internal standard for metagenomic samples consisted of *Thermus thermophilus* DSM7039 [HB8] genomic DNA (American Type Culture Collection, Manassas,

VA) added immediately prior to cell lysis. The amount of DNA standard added was estimated to be ~ 1% (8.4 ng per liter filtered) of sample DNA based on yields of DNA from a typical filter (Table S1).

16S and 18S rRNA tag sequencing

DNA was PCR-amplified in four replicate 20 µl reactions using bacterial 16S ribosomal RNA gene primers. The bacteria-specific forward primer began at 27F in the V2 region and included a 454B FLX linker (GCCTTGCCAGCCCGCTCAG *TC* AGRGTTTGATYMTGGCTCAG). The reverse primer began at 338R and included a 454A linker and a unique 8 base pair barcode (denoted by N in primer sequence; GCCTCCCTCGCGCCATCAG NNNNNNN *CA* TGCWGCCWCCCGTAGGWGT) (Modified from (Hamady *et al.,* 2008)). Products from replicate amplifications were combined and purified with the S.N.A.P. UV-Free Gel Purification Kit (Invitrogen, Carlsbad, CA) by gel isolation from a 0.8% agarose gel. Purified samples were sent for pyrosequencing on a Roche-454 FLX Pyrosequencer at Engencore at the University of South Carolina using titanium chemistry (http://engencore.sc.edu/).

Bioinformatics

Following sequencing, paired-end Illumina reads were joined using the She-ra program (Rodrigue *et al.,* 2010) with the default parameters and a quality metric score of 0.5. Paired reads were trimmed using Seqtrim (Falgueras *et al.,* 2010) with default settings. To remove rRNA, tRNA, and internal standard sequences from the metatranscriptome reads, a Blastn search was performed against a database containing representative ribosomal RNA and tRNA sequences

along with the internal standard sequences. All reads with a bit score greater ≥50 to one of the sequences in the database were removed (Gifford *et al.,* 2011). To remove internal standard sequences from the metagenome reads, a Blastn search against the *T. thermophilus* HB8 genome was carried out, and any hits from the metagenome sequences with a bit score ≥50 were queried against the RefSeq protein database using a Blastx search; hits matching *Thermus thermophilus* proteins with a bit score ≥40 were designated as internal standard.

Reads representing genes or transcripts of 74 selected biogeochemically-relevant genes were identified using a Blastx search against a custom database consisting of multiple reference sequences from diverse taxa for each gene, along with paralogs having sequences most similar to the gene of interest. This gene-specific reference database was tested on a subset of Amazon reads using a bit score ≥40, and re-analysis of the positive reads against the RefSeq protein database was used to adjust the composition of the database. To obtain an estimate of error that was propagated through the calculations, a Poisson distribution was generated for each gene using the rpois function in R with the lamba parameter equal to the count of reads and 10,000 randomly sampled values returned. The Poisson distributions were converted to normalized count distributions (copies L^{-1}) and then averaged together for replicate samples. The resulting distributions were used to calculate expression ratios for each gene by dividing the transcripts L^{-1} distribution by the genes L^{-1} distribution. Expression ratios between size fractions were considered significantly different when separated by 2 standard deviations. For comparison, a one-tailed statistical comparison was also performed using Welch's T-test, an adaptation of the standard T-test for two samples with unequal variances. Of the 55 genes identified to have significantly different ratios based on 2 standard deviations of the Poisson distribution, 50 of

these (91%) had p-values of 0.1 or less with the Welsh's T-test, while the remaining 5 had p-values of <0.14.

Cell numbers of the transcriptionally dominant prokaryotic organisms (those contributing the most transcripts to samples from this station) were estimated based on the metagenomic coverage of each reference genome. The number of protein encoding reads identified in the metagenomes (ACM3 and ACM4) for each taxon (found with a Blastx analysis of the metagenomic reads against the Refseq protein database) was divided by the number of protein encoding genes present in the reference genome for corresponding taxon. This average fold-coverage of the genes was used as an index of genome copy number, and we assumed one genome copy per cell.

Three of the most dominant prokaryotic organisms were used for a genome-wide comparative expression analysis between the two size fractions. Protein encoding reads binning to each of the three taxonomic bins were identified using Blastx against the Refseq protein database. KEGG GENE assignments for reads binning to CB0205, HTCC7211, and HTCC2080 were obtained by annotating the reference proteins against KEGG GENES using the KEGG Automatic Annotation Server (KAAS) (Moriya *et al.,* 2007) and KEGG pathway reconstruction was performed using MinPath (Ye and Doak 2009). For each individual gene or pathway in a taxonomic bin, Poisson distributions were randomly generated using the rpois function in R with the lambda parameter equal to the count of reads, with 10,000 randomly sampled values returned. The resulting distributions were then averaged between replicate samples, and the mean and standard deviation from the resulting distributions were determined. These calculations were performed on each of the two size fractions, and each gene within a taxonomic bin was

compared between the fractions. Genes or pathways were designated as differentially expressed when at least 2 standard deviations existed between the average values for each size fraction.

Ribosomal RNA gene sequences in the metagenomes were identified by mapping paired reads to 16S and 18S rRNA sequence databases with Bowtie2 (2.0.0-beta5) using local alignment mode with default options. Hits were classified with the classify.seqs program in mothur (1.24.0) (Schloss *et al.,* 2009) using a minimum bootstrap cutoff of 80%. Each end of the paired-end reads was also mapped individually to 16S and 18S rRNA sequence databases using the Bowtie2 end-to-end alignment mode with default options. Both members of the paired-end reads were classified with RTAX (v0.983) (Soergel *et al.,* 2012) in forward and reverse directions, and the read with the higher RTAX value of the pair was retained. The Greengenes 16S rRNA 99% OTUs (v.12.10) database (McDonald *et al.,* 2012) was used for classification of paired reads using mothur, the Greengenes 16S rRNA 97% OTUs (v.12.10) was used for unpaired reads using RTAX (due to limitations of the program), and a custom database of 18S rRNA genes modified from the Silva-euks (v.108) database (Quast *et al.,* 2013) after removal of sequences with more than 5 ambiguous bases (1953 seqs), homopolymers greater than 10 bp (527 seqs), and long polymers of 6bp repeats (7 seqs). Eukaryotic taxonomy strings in this database were manually reduced to Kingdom, Phylum, Class, Order, Family, Genus, and Species.

Ribosomal RNA gene sequences from the tag sequencing were processed on the Data Intensive Academic Grid (DIAG) shared computational cloud at the University of Maryland School of Medicine Institute for Genome Sciences (IGS) with the AmpliconNoise pipeline (Quince *et al.,* 2011), using recommended procedures for quality control (CleanMinMax.pl, PyroNoiseT, SeqDistT, SeqNoiseT). Maximum sequence length was set to 250 base pairs

(Parse.pl), and chimeras were identified and removed (PerseusD). Sequences were clustered into operational taxonomic units (OTUs) using Qiime, and sequences from each sample were unweighted (unweight_fasta.py), concatenated, and subjected to primers removal. OTUs were identified using uclust (pick_otus.py), and representative sequences were selected (pick_rep_set.py). The taxonomy of OTUs was determined in MacQiime (assign_taxonomy.py), retraining the RDP Classifier to use the October 2012 Greengenes taxonomic database (McDonald *et al.,* 2012).

Supplemental Methods References:

Crump BC, Armbrust EV, Baross JA (1999). Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia river, its estuary, and the adjacent coastal ocean. *Appl Environ Microbiol* **65:** 3192-3204.

Crump BC, Kling GW, Bahr M, Hobbie JE (2003). Bacterioplankton community shifts in an arctic lake correlate with seasonal changes in organic matter source. *Appl Environ Microbiol* **69:** 2253-2268.

Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC bioinformatics* **11:** 38.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* **5:** 461-472.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5:** 235-237.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A *et al.* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6:** 610-618.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35:** W182-185.

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

Poretsky RS, Hewson I, Sun S, 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.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P *et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41:** D590-596.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011). Removing noise from pyrosequenced amplicons. *BMC bioinformatics* **12:** 38.

Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ *et al.* (2010). Unlocking short read sequencing for metagenomics. *PLoS One* **5:** e11840.

Satinsky BM, Gifford SM, Crump BC, Moran MA (2013). Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. In: DeLong EF (ed). *Methods Enzymol*. Academic Press. pp 237-250.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75:** 7537-7541.

Soergel DA, Dey N, Knight R, Brenner SE (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J* **6:** 1440-1444.

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.

Ye Y, Doak TG (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* **5:** e1000465.

Zhou J, Bruns MA, Tiedje JM (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62:** 316-322.

CHAPTER 4

TRANSCRIPTIONAL REGULATION OF ELEMENTAL CYCLING GENES BY AMAZON PLUME PROKARYOTES[1]

[1] Satinsky, B.M., Crump, B.C., Smith, C.B., Sharma, S., and Moran, M.A. To be submitted to *International Society of Microbial Ecology Journal.*

Abstract

Bacteria and archaea fill critical ecological roles in the oceans by controlling fundamental aspects of energy production and consumption. Here, quantitative metatranscriptomics and metagenomics were used to investigate expression patterns of prokaryotic genes at six stations in the Amazon River Plume. Prokaryotic communities harbored an average of 3.3×10^{11} transcripts L^{-1}, and taxa dominating the transcriptomes included Gammaproteobacteria, Flavobacteria, and the cyanobacterium *Synechococcus* at the lower salinity stations; Verrucomicrobia and the endosymbiotic diazotroph *Richelia intracellularis* dominated in the mesohaline regions; and *Prochlorococcus* and *Alteromonas* at the higher salinity oceanic station. Across the plume, cells in the particle-associated microenvironment had higher expression levels (transcripts gene copy^{-1}) than in the free-living microenvironment for most genes mediating carbon, nitrogen, and phosphorus transport and metabolism. Expression levels of genes involved in nitrogen fixation and phosphorus acquisition were the most spatially variable across the plume. As Amazon Plume waters mix into the western tropical North Atlantic, expression levels of key biogeochemical genes in free-living cells remain relatively constant, while those in particle associated cells change many fold. This high variability in expression by particle-associated cells was driven both by changes in gene regulation within a taxon and by shifts in taxonomy among stations.

Introduction

The Amazon River is the world's largest in terms of volume and area, traversing 6,500 km through the South American continent before mixing into seawater of the Western Tropical North Atlantic Ocean (WTNA) (Coles *et al.,* 2013). At its mouth, the Amazon discharges water at a rate approximately 12 times that of the Mississippi River, carrying terrestrially-derived nutrients offshore in a low salinity plume that is sufficient in scope to impact marine primary productivity and carbon sequestration on a global scale (Richey *et al.,* 1989, Subramaniam *et al.,* 2008). On the benthic floor below the plume, organic carbon exported from surface waters leaves a trail as far as 1,200 km from the river mouth (Chong *et al.,* 2014).

Previous studies of the Amazon plume have shown that primary production in lower salinity waters is dominated by coastal diatoms, typically *Skeletonema* and *Pseudonitzschia*, and their activity is supported largely by the river-supplied inorganic nitrogen (Goes *et al.,* 2014, Simon *et al.,* 2009). As turbidity decreases and salinity increases along the plume and nitrogen becomes limited, *Rhizosolenia* and *Hemiaulus* become the dominant diatoms (Goes *et al.,* 2014, Subramaniam *et al.,* 2008); these cells carry an endosymbiotic cyanobacterium *Richelia intracellularis* that provides fixed nitrogen in an arrangement referred to as a diatom-diazotroph association (DDA). Prokaryotic phytoplankton are also numerically important autotrophs in the plume (Goes *et al.,* 2014, Zehr *et al.,* 2001), and these include unicellular cyanobacteria *Synechococcus, Prochlorococcus,* and in oceanic regions where phosphorus and silicate are limiting, N_2-fixing *Trichodesmium*. The high productivity of these autotrophic communities generates a region of the WTNA that takes up an excess of ~15 Tg C per year (Cooley *et al.,* 2007).

Heterotrophic bacteria have a major influence over the fate of CO_2 fixed by the plume primary producers through their roles in transformation and sequestration of organic carbon and by the regeneration of inorganic N and P. These important processes have been more difficult to characterize compared to the activities of autotrophs, however, because of the hundreds of largely uncharacterized bacterial species involved and the diversity of links in the elemental cycles they mediate. Further, bacterial taxa can harbor the genetic capability for multiple biogeochemical roles, but whether these capabilities are expressed and, if so, at what level is based on environmental conditions at a given time and place.

Molecular ecology studies have begun to provide new details on microbes in the Amazon plume. Results from this body of work revealed insights into the diazotrophic organisms, including vertical and horizontal distributions of the nitrogenase gene (*nifH*) (Foster *et al.,* 2007), abundances of diazotrophic cyanobacteria (Goebel *et al.,* 2010), and identity of nutrients limiting nitrogen fixation (Turk-Kubo *et al.,* 2012). More recently, metagenomic and metatranscriptomic methods have described microspatial patterns in bacterial and archaeal gene expression in the lower salinity region of the plume (Satinsky *et al.,* 2014a), diversity and gene expression of diazotrophs (Hilton *et al.,* 2014), and gene expression by microeukaryotic communities as they relate to biogeochemical patterns (Zielinski *et al.,* 2014). For this study, six Amazon Plume stations varying in salinity from 23 to 36 were inventoried in May-June, 2010 for gene and transcript abundances of 37 biogeochemically-relevant functional genes. This detailed metagenomic and metatranscriptomic dataset offers an unprecedented view of activities of Amazon plume bacteria and archaea by focusing on variations in gene expression relevant to carbon, nitrogen, and phosphorus cycling.

Methods

Surface plume metagenomic and metatranscriptomic samples were obtained in duplicate for two discrete size fractions at 6 stations that spanned a range of salinities, nutrient concentrations, and microbial communities (Table 4.1). Collection and processing were carried out as described previously (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b). In brief, microbial cells from two size fractions [2 - 156 μm, designated as particle-associated (PA); and 0.2 – 2 μm, designated as free-living (FL)], were collected by filtration and preserved in RNAlater (Applied Biosystems, Austin, TX.). Subsequent processing for metatranscriptomic analysis included extraction of RNA, removal of residual DNA, and depletion of rRNA. Ribosomal RNA (rRNA) - depleted RNAs were linearly amplified and converted to cDNAs. Processing for metagenomic analysis included extraction of DNA and removal of residual proteins and RNA. For both nucleic acid types, internal standards in the form of artificial mRNAs (metatranscriptomic samples) or genomic DNA from *Thermus thermophilus* HB8 (metagenomic samples) were added immediately prior to cell lysis to allow calculation of volume-based copy numbers (i.e., copies L^{-1}) (Satinsky *et al.,* 2013).

Following processing, nucleic acid preparations were sheared ultrasonically to ~200-250 bp fragments and libraries were constructed for Illumina paired-end sequencing (2 x 150 bp; Illumina Inc., San Diego, CA). A custom bioinformatic pipeline was established to carry out quality control of the Illumina reads, join and trim paired-end sequences, and quantify and remove internal standard sequences (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b). rRNA reads were identified in the metatranscriptomic sequences by performing a blastn search against a database containing representative ribosomal RNA sequences. Protein-encoding reads were identified by using either a RAPSearch2 (Zhao *et al.,* 2012) query against the RefSeq protein

database or a blastx query against a custom database consisting of biogeochemically relevant gene sequences. Per liter copy numbers of each gene and transcript type were calculated based on internal standard recovery (Table 4.S1) by comparing the abundance of standard sequences in the Illumina libraries to the known number of standards added into the cell lysis buffer at the initiation of filter processing (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b). Biological replicates were averaged and expression levels were calculated as previously described by Satinsky et al. (Satinsky *et al.,* 2014a). For any gene or taxonomic bin with at least one replicate transcriptome containing reads, calculations of minimum possible transcript per liter and expression levels were made by setting any missing metagenome sample counts equal to the limit of detection and any missing metatranscriptome sample counts equal to zero.

To analyze population structure in reference genome bins, ribosomal protein reads from five taxa abundant in the metatranscriptomes were identified through a similarity search using RAPsearch2 against the NCBI RefSeq protein database. Reads hitting selected ribosomal proteins were pooled within a genome bin from all stations and size fractions and clustered at 95% nucleotide identity using CD-HIT (Huang *et al.,* 2010) to generate OTU clusters. Reads from each individual station and size fraction were then separately classified into the representative OTUs using blastn with a minimum alignment length of 50 nt. Singleton OTUs (those with only a single hit across all samples) were removed from subsequent analyses, as were any stations/fractions with fewer than 20 reads across all OTU bins. Distributions of remaining OTUs at each station/fraction were compared using a Bray-Curtis dissimilarity analysis and clustered hierarchically to identify genome bins with highly similar population structure, defined operationally here as those with dissimilarity scores of $\leq$30%. Two abundant ribosomal proteins

were analyzed for each taxonomic bin, and the final dissimilarity score was based on the average of the two proteins.

Results

Genes and transcripts were collected in May-June 2010 from free-living (FL; operationally defined as 0.2 - 2.0 μm) and particle-associated (PA; 2.0 - 156 μm) microbial communities at six stations in the Amazon River plume (Table 4.1) (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b). All samples were collected from surface waters (0-5 m) between the hours of 0700 to 1000 to control for diel patterns in expression. The stations fell into three classes established previously for the plume based on salinity (Subramaniam *et al.,* 2008): low salinity stations 10 and 23 (salinity <30), mesohaline stations 3, 2, and 25 (salinity 30 - 35), and oceanic station 27 (salinity >35) (Fig. 4.1, Table 4.1). Each sample was collected in duplicate, yielding 24 metatranscriptomic libraries (averaging ~7.0 ± 4.0 million paired, quality controlled reads each) and 24 metagenomic libraries (averaging ~4.0 ± 2.0 reads each) with a mean length of 197 nt. Reads were binned taxonomically and functionally based on best-hit assignments to reference genomes. We focus here on the patterns of bacterial and archaeal genes and transcripts along the plume.

The most transcriptionally dominant taxonomic bin at a station accounted for an average of ~2.5 x 10^{10} transcripts L^{-1} or ~8% of the prokaryotic transcript pool (Fig. 4.1), with this percentage dropping to an average of ~1.5% by the 10^{th} most transcriptionally dominant taxon. Highly-recruiting heterotrophic bins at the low salinity stations included the marine Gammaproteobacteria HIMB30 and HIMB55, with SAR86-like transcripts particularly abundant at Station 10 and SAR116- and SAR324-like transcripts abundant at Station 23 (Fig. 4.1). At

mesohaline Stations 3, 2, and 25, more transcripts binned to heterotrophic Verrucomicrobium *Coraliomargarita akajimensis* DSM45221 than to any other genome, accounting for as much as 17% of the prokaryotic transcriptome at Station 2 (Fig. 4.1). Some taxonomic bins overlapped with those at the low salinity stations (SAR116, SAR324, and Gammaproteobacteria HIMB30, HIMB55, and HTCC2207) and new groups also appeared (Gammaproteobacteria in the Alteromonadaceae family, Verrucomicrobium DG1235, several Roseobacter taxa, and SAR11 members HIMB5, HIMB59, and HTCC7211). At oceanic station 27, transcripts recruiting to SAR116, Roseobacter, *Alteromonas*, and *C. akajimensis* reference genomes represented the dominant heterotrophic groups (Fig. 4.1).

Along with the heterotrophs, sequences from cyanobacterium *Synechococcus* were important in the transcriptome at most stations (binning to CB0205, WH8109, CC9605, RS9916, and WH8102 Fig. 4.1). However, at Stations 2 and 25, the dominant cyanobacterial transcript bin was the endosymbiotic diazotroph *Richelia intracellularis*, which formed a diatom-diazotroph association (DDA) with eukaryote *Hemiaulus hauckii*. As predicted from the endosymbiotic nature of the association, the vast majority of *Richelia*-like transcripts were found in the particle-associated microbial community (~96%) where the *Hemiaulus* chains were collected. At oceanic Station 27, transcripts related to *Prochlorococcus* replaced *Synechococcus* as the most abundant prokaryotic primary producer. Across all the bacteria and archaea in the plume, transcripts were present in lower numbers than genes by an average of ~25-fold at each station (Table 4.1). Free-living cells were responsible for an average of ~90% of the community metagenome but only ~60% of the community metatranscriptome (Table 4.1), and thus had lower overall gene expression levels.

Community Patterns in Gene Expression

We inventoried transcripts and genes encoding 37 proteins involved in carbon, nitrogen, and phosphorus cycling, and calculated three indexes to assess patterns in predicted microbial activity across the plume: 1) median expression levels (transcripts gene copy^{-1}) for each gene across the plume; 2) expression levels for each gene individually by station and size fraction; and 3) transcript abundance (transcripts L^{-1}) for each gene.

Among the 16 surveyed genes involved in heterotrophic C metabolism, carbon storage gene *phaP* (involved in polyhydroxyalkanoate synthesis) had the highest median expression level (0.5 transcripts gene copy^{-1}; Fig. 4.2A), followed by the regulatory gene for polyhydroxyalkanoate synthesis (*phaR*), a tetrahydromethanopterin-linked C1 metabolism protein (*fae*, formaldehyde activating enzyme), and an aminopeptidase that cleaves N-terminal methionine from proteins (*pepM*). When calculated for individual stations, expression levels of most genes varied considerably among the particle-associated communities, but were lower and relatively constant among the free-living bacterial communities (Fig. 4.2B). When analyzed in terms of transcript abundance, glycolysis gene GAPDH (encoding glyceraldehyde-3-phosphate dehydrogenase) and *pepM* contributed the greatest number of transcripts per liter of seawater (Fig. 4.2C), followed by genes for carbohydrate degradation (*bglA*, encoding terminal glucose cleavage via beta-glucosidase) and glycolysis enzyme *pgi* (encoding glucose-6-phosphate isomerase). Transcripts for genes encoding aromatic compound degradation (*pcaH*, *vanA*) and carbon storage as polyhydroxyalkanoate (*phaC,P,R,Z*) were at low to average levels in plume waters and were consistent within each functional gene group (Fig. 4.2C).

Among the 9 surveyed genes involved in nitrogen cycling, those encoding nitrogenase (*nifH*) and ammonia monooxygenase (*amoA*) had the highest median expression levels (~0.7

transcripts gene copy^{-1}; Fig. 4.3A), although both were present at only a subset of stations [*nifH* at Stations 2 and 25 where the DDAs were found (Hilton *et al.,* 2014), *amoA* at Stations 10 and 3 where ammonia-oxidizing Thaumarcheota were found]. Genes encoding enzymes for nitrate reduction (*narG* and *napA*) and nitrogen storage (*cphAB*, cyanophycin synthesis and hydrolysis respectively) had among the lowest median expression levels. When calculated for individual stations, expression level patterns were similar to those for the carbon genes surveyed in that the particle-associated communities exhibited high variability in expression levels among stations for most genes, but free-living bacterial communities showed lower and more constant expression levels across all stations (Fig. 4.3B). Per liter of seawater, transcripts for genes involved in ammonium uptake (*amtB*, ammonium transporter) and assimilation (*glnA*, glutamate synthetase) were present in highest numbers, followed by transcripts for a *nirK*-like gene (nitrate reductase) and *amoA* binning to the Thaumarcheota genome sequences, although these were found only at Stations 10 and 3 (Fig. 4.3C).

For the 12 surveyed genes mediating phosphorus cycling, median expression levels were highest for phosphonate uptake gene *phnD* (periplasmic binding protein; 0.05 transcripts gene copy^{-1}), the regulator and membrane protein of the high-affinity phosphate transporter (*pstS* and *pstC*), and the low-affinity phosphate transporter *pitA* (Fig. 4.4A). As was the case for carbon and nitrogen genes, expression levels varied considerably by station for the particle-associated community but less so for free-living cells (Fig. 4.4B). When calculated as inventories per liter of seawater, transcripts from *pstA, pstC, and pstS* were the most abundant phosphorus-related transcripts, followed by the phosphorus acquisition regulator *phoU* (Fig. 4.4C).

Genes sharing similar patterns of expression were identified by correlation analysis followed by hierarchical clustering. The tightest clusters of gene expression patterns were

typically within a size fraction, with genes in the PA community typically grouping together and those in the FL community grouping together. Several genes involved in high affinity phosphate transport clustered into the same expression pattern, including the permease components *pstA* and *pstC* genes and the regulator *phoU* (Fig. 4.5). Expression patterns of genes related to carbon storage also correlated within a size fraction (*phaC, phaP,* and *phaZ* in free-living cells; *phaC, phaR,* and *phaZ* in particle-associated cells) but not between fractions (Fig. 4.5).

Taxon-Specific Patterns in Gene Expression

By keeping taxonomic composition constant, the role of gene regulation in explaining differences in plume transcript gene^{-1} ratios could be explored. To do this, expression levels were calculated for single reference genome bins and then compared across the plume to look for evidence of up- or down-regulation within a bin. Five genome bins with high coverage in the metatranscriptomes (Figs. 4.1 and 4.S1) were first assayed for population structure to eliminate cases in which populations were inconsistent among stations. For each reference bin, sequences from two highly-expressed ribosomal proteins were used to generate operational taxonomic units (OTUs) at 95% nucleotide identity. Reads from each individual station and size fraction were then assigned to the OTUs, and resulting distributions were clustered hierarchically to generate population fingerprints (Fig. 4.S1). We note that because the OTUs were based on read lengths of 225 nt and mapped to varying regions of the full-length gene, even a completely homogenous population would produce multiple OTUs. Thus this method looked for bins with similar patterns, but did not provide information on whether multiple populations were clustered within those bins. For two of the reference bins (SAR116-related *Candidatus puniceispirillum* IMCC1322 and marine Gammaproteobacteria HIMB55 bins), populations had variable

fingerprints through the plume stations and therefore were not considered further. For three of the reference genome bins (*Coraliomargarita akajimensis* DSM45221, marine Gammaproteobacteria HIMB30, and SAR324 cluster bacterium JCVI-SC AAA005), population structure was similar across five or more samples (defined operationally as ≤30% dissimilarity; Fig. 4.S1), and these were subsequently used in analysis of taxon-specific gene expression patterns.

The same major shifts in gene expression as observed in the community analysis also emerged for the three individual reference genome bins (Fig. 4.6). *Coraliomargarita akajimensis* populations had highest transcripts gene copy^{-1} ratios and greatest variability between stations when associated with particulate material in the plume, with the exception of *pstC*. HIMB30 and SAR324 genes also followed this pattern. All taxa showed variability by gene in terms of which station exhibited the highest expression levels. Thus as for the community analyses, gene expression levels at the individual taxon level were higher and more variable when members were part of the particle-associated community, and lower and less dynamic when they were part of the free-living community (Fig. 4.6).

Discussion

In considering what factors explain variable microbial gene expression patterns in the Amazon plume, one possibility is that cells regulate gene transcription differently in response to perception of environmental conditions. In this case, the functional annotations of the genes exhibiting the largest expression differences would be informative regarding the specific environmental parameters most affecting microbial activity in the plume. A second possible explanation is that different expression levels across the plume results from variation in

taxonomic composition of the microbial community. Under this scenario, differences in genome composition and/or regulatory capacity of the dominant taxa at each plume location will influence expression levels, in which case expression level changes will be less informative of the environmental drivers of microbial activity.

To tease these factors apart, we held taxonomy constant by analyzing single genome bins across samples. The requirement for high sequence coverage and consistent within-bin population structure constrained these taxon-specific analyses to just three species and nine genes. The observed gene expression patterns were similar to the community-level analyses, however, and established that bacterial taxa at the species and population levels regulated how much they expressed genes of biogeochemical relevance across locations (station) and microenvironments (particle-associated or free-living) in the plume (Fig. 4.6). In the most extreme example, the *Coraliomargarita* bacterial bin altered gene expression in the free-living fractions of the high-affinity phosphate transporter gene *pstC* from barely measurable at Station 23 to over 1,500 transcripts gene copy^{-1} at Station 27 (Fig. 4.6).

Nonetheless, taxonomy of the dominant bacteria and archaea also differed by station and microenvironment and could not be ruled out as an influence on shifting expression levels (Fig. 4.S2). For example, genes from SAR11 and SAR86 taxa together accounted for 28% of the free-living metagenome but only 7% of the particle-associated metagenome when averaged across the six stations (Fig. 4.S2). This raises the possibility that smaller genomes with fewer regulatory systems typical of streamlined taxa (Giovannoni *et al.,* 2005, Luo *et al.,* 2014) leads to inherent differences in transcriptional responses to environmental conditions. When we compared the range of expression levels for pairs of streamlined and non-streamlined taxa in the Alphaprotobacteria (SAR11 bins versus Roseobacter bins) and Gammaproteobacteria (SAR86

bins versus Alteromonadales bins), we found that the streamlined group of the pair had a smaller expression range by 2-fold when comparing cells in the particle microenvironment (expression varied over a range of 3.2 transcripts gene copy^{-1} for streamlined taxa versus 6.3 transcripts gene copy^{-1} for non-streamlined) and 18-fold when comparing cells in the free-living microenvironment (over a range of 0.02 transcripts gene copy^{-1} for streamlined taxa versus 0.32 transcripts gene copy^{-1} for non-streamlined) (Fig. 4.7). Further, the station with the highest taxonomic similarity between free-living and particle-associated communities (Station 10; Fig. 4.S2) also had the most similar gene expression levels between size fractions (Figs. 4.2-4.4). Thus taxonomic differences between stations and microenvironments also likely explained a portion of the expression variation.

Recognizing a role for both regulation and taxonomy in the observed differences in expression levels, we asked what the expression patterns infer about environmental conditions experienced by microbes in the plume. The most consistent observation was that particle-associated cells had higher and more variable gene expression levels compared to free-living, and this was the case for all three of the categories of elemental cycling genes surveyed (carbon, nitrogen, phosphorus) (Fig. 4.8). A similar difference in gene expression levels between these two microenvironments was reported in a previous analysis of Station 10 (Satinsky *et al.,* 2014a), and the additional five plume stations included here makes this trend is even more pronounced. Across all 37 genes, the average transcript gene^{-1} ratio was 1.3 for particle-associated cells and 0.07 for free-living cells, for an ~18-fold difference in transcript numbers per gene copy. In the three mesohaline stations (Stations 3, 2, and 25), this difference was particularly pronounced, with an average difference of ~27-fold between particle-associated and free-living cells. These expression patterns suggest less variability in environmental conditions in

bulk seawater compared to particulate material along the axis of the plume, and this was reflected in the extent of gene regulation by the cells, or the taxonomic composition of the successful taxa, or both.

The gene with the greatest community-wide variation in expression was *nifH* in the particle-associated community (Fig. 4.8), which ranged from no measurable expression at Station 10 (even though the genes were present) to 8 and 35 transcripts gene copy^{-1} at Stations 2 and 25 (among the highest of any gene surveyed). Stations 2 and 25 were characterized by DDA blooms (Yeung et al. 2012), and cyanobacterial endosymbiont *Richelia intracellularis* populations accounted for 71% (Station 2) and 87% (Station 25) of all *nifH* transcripts. *nifH* transcripts were present in lower numbers at Station 27, and in this case *Trichodesmium* contributed 86% of the transcript inventory. Despite *nifH* being the most variable of all genes surveyed, dissolved nitrogen concentrations were low throughout the plume (Table 4.S2) and below detection levels at most surface stations (Goes et al. 2014), suggesting that any inputs of new nitrogen were fully assimilated by the microbial community.

Phosphorus acquisition genes harbored by particle-associated cells accounted for five of the next nine genes showing the greatest variation in expression, including phosphate transporter genes (*pstA*, *pstC*, *pitA*) and alkaline phosphatases (*phoU*, *phoX*) (Fig. 4.8). Highest gene expression levels of phosphorus acquisition genes occurred in the particulate community at Stations 2 and 25, coincident with the location of nitrogen-fixing DDAs and the lowest measured dissolved phosphate concentrations (0.11 mM at Station 2 and below detection at Station 25; Table 4.S2). Transcripts mapping to *Richelia intracellularis* accounted for 21% of the *pstC* transcripts and 19% of the *pstA* transcripts at Station 25, supporting the idea that upregulation of phosphorus acquisition at these stations reflected heightened phosphorus demand linked to

nitrogen fixation. Thus phosphorus availability was apparently one of the most dynamic signals to which prokaryotic cells responded along the plume. Carbon cycle genes with high variability in expression included glycolysis gene GAPDH whose expression level peaked at Station 25, protein degradation gene *pepM*, which peaked at Station 2, and aromatic ring cleavage gene *pcaH*, which peaked at Station 3. Free-living cells had highest variance in expression levels of genes for carbon storage as polyhydroxyalkanoate (*phaP* and *phaZ*), peaking in expression at Station 10 where a high chlorophyll inventory and bacterial production rate (Table 4.S2) suggested greater carbon availability to heterotrophic bacteria (Fig. 4.8).

The Amazon Plume represents a region of the western tropical North Atlantic that is characterized by reduced sea surface salinity and elevated chlorophyll *a* concentrations (Yeung *et al.,* 2012). Activities of plume heterotrophic bacteria are critical to the transformation and sequestration of carbon and the cycling of nitrogen and phosphorus in this system, but are challenging to characterize because of taxonomic diversity and the multiplicity of biogeochemical pathways they mediate. In an effort to gain insights into the roles of bacteria and archaea along the plume, we asked whether patterns in gene expression could point to the environmental conditions most relevant to microbial cells at each station and microenvironment. We conclude that particle-associated microenvironments present more chemically dynamic conditions for their bacterial and archaeal inhabitants, and that conditions experienced by attached cells varies much more along the Amazon plume than do conditions experienced by free-living cells. Expression levels of nitrogen fixation, as expected from the DDA populations, phosphorus acquisition as free or organically-bound phosphate, and heterotrophic carbon processing all peaked for particle-associated cells at the mesohaline stations, providing additional insight for interpreting the low levels of nutrients measured in seawater at these stations. In the

free-living microenvironment, carbon storage was the most dynamic activity and peaked closest to the river mouth. Such high variation in transcript gene^{-1} ratios in bacterial and archaeal cells indicates that relative gene abundance is not a faithful reflection of the biogeochemical activity of plume microbes, with expression levels of the same gene varying up to several orders of magnitude along the Amazon Plume gradient.

Acknowledgements

We appreciate advice and assistance from Jan Mrazek and Roger Nilsen. This work was funded by the Gordon and Betty Moore Foundation through grants GBMF2293, GBMF2928, and GBMF538.01. Resources and technical expertise were provided by the University of Georgia's Georgia Advanced Computing Resource Center

References

Chong LS, Berelson WM, McManus J, Hammond DE, Rollins NE, Yager PL (2014). Carbon and biogenic silica export influenced by the Amazon River Plume: Patterns of remineralization in deep-sea sediments. *Deep Sea Res Part 1 Oceanogr Res Pap* **85:** 124-137.

Coles VJ, Brooks MT, Hopkins J, Stukel MR, Yager PL, Hood RR (2013). The pathways and properties of the Amazon River Plume in the tropical North Atlantic Ocean. *J Geophys Res-Oceans* **118:** 6894-6913.

Cooley SR, Coles VJ, Subramaniam A, Yager PL (2007). Seasonal variations in the Amazon plume-related atmospheric carbon sink. *Global Biogeochem Cy* **21**.

Foster RA, Subramaniam A, Mahaffey C, Carpenter EJ, Capone DG, Zehr JP (2007). Influence of the Amazon River plume on distributions of free-living and symbiotic cyanobacteria in the western tropical north Atlantic Ocean. *Limnol Oceanogr* **52:** 517-532.

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.

Goebel NL, Turk KA, Achilles KM, Paerl R, Hewson I, Morrison AE *et al.* (2010). Abundance and distribution of major groups of diazotrophic cyanobacteria and their potential contribution to N(2) fixation in the tropical Atlantic Ocean. *Environ Microbiol* **12:** 3272-3289.

Goes JI, Gomes HdR, Chekalyuk AM, Carpenter EJ, Montoya JP, Coles VJ *et al.* (2014). Influence of the Amazon River discharge on the biogeography of phytoplankton communities in the western tropical north Atlantic. *Prog Oceanogr* **120:** 29-40.

Hilton JA, Satinsky BM, Doherty M, Zielinski BL, Zehr JP (2014). Metatranscriptomics of N_2-fixing cyanobacteria in the Amazon River plume. *ISME J* **In Revision**.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26:** 680-682.

Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA (2014). Evolutionary analysis of a streamlined lineage of surface ocean Roseobacters. *ISME J* **8:** 1428-1439.

Richey JE, Nobre C, Deser C (1989). Amazon river discharge and climate variability: 1903 to 1985. *Science* **246:** 101-103.

Satinsky BM, Gifford SM, Crump BC, Moran MA (2013). Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. In: DeLong EF (ed). *Methods Enzymol*. Academic Press. pp 237-250.

Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski BL, Doherty M *et al.* (2014a). Microspatial gene expression patterns in the Amazon River Plume. *Proc Natl Acad Sci U S A* **111:** 11085-11090.

Satinsky BM, Zielinski BL, Doherty M, Smith CB, Sharma S, Paul JH *et al.* (2014b). The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. *Microbiome* **2:** 17.

Simon N, Cras AL, Foulon E, Lemee R (2009). Diversity and evolution of marine phytoplankton. *C R Biol* **332:** 159-170.

Subramaniam A, Yager PL, Carpenter EJ, Mahaffey C, Bjorkman K, Cooley S *et al.* (2008). Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci U S A* **105:** 10460-10465.

Turk-Kubo KA, Achilles KM, Serros TR, Ochiai M, Montoya JP, Zehr JP (2012). Nitrogenase (nifH) gene expression in diazotrophic cyanobacteria in the Tropical North Atlantic in response to nutrient amendments. *Front Microbiol* **3:** 386.

Yeung LY, Berelson WM, Young ED, Prokopenko MG, Rollins N, Coles VJ *et al.* (2012). Impact of diatom-diazotroph associations on carbon export in the Amazon River plume. *Geophys Res Lett* **39**.

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

Zhao Y, Tang H, Ye Y (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28:** 125-126.

Zielinski BL, Allen AE, Carpenter EJ, Coles VJ, Crump BC, Doherty M *et al.* (2014). Microeukaryotic Gene Expression Parallels Amazon River Plume Biogeochemistry. *ISME J* **In Review**.

Table 4.1. Metagenome and metatranscriptome data summaries for 6 Amazon River plume stations in June 2010. Duplicate samples were collected for each of two size fractions for each data type at all six stations. Per liter calculations are based on recovery of internal standards (Table 4.S1).

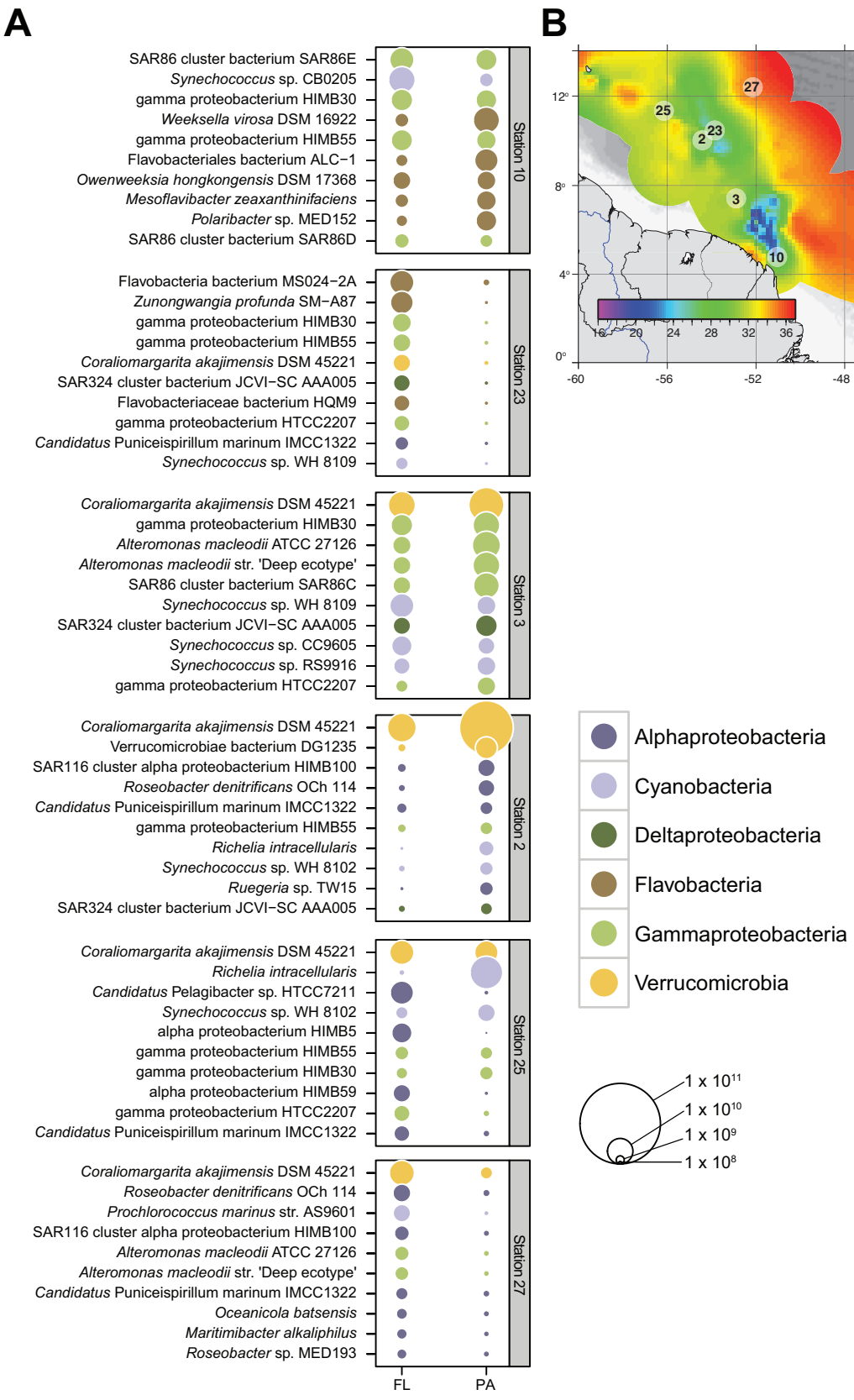
	Station 10	Station 23	Station 3	Station 2	Station 25	Station 27
Location (Lat, Long)	4.8818, -51.3608	10.6821, -54.4213	7.2875, -53.0005	10.2885, -54.512	11.3123, -56.4266	12.4145, -52.2197
Depth (m)	4.26	3.64	3.76	4.47	3.93	3.89
Salinity (PSU)	22.55	26.49	30.80	31.80	31.87	36.03
Raw reads						
Metagenomic	4.44×10^{7}	5.69×10^{7}	7.39×10^{7}	5.08×10^{7}	7.87×10^{7}	6.01×10^{7}
Metatranscriptomic	1.25×10^{8}	1.38×10^{8}	1.00×10^{8}	1.27×10^{8}	1.27×10^{8}	1.95×10^{8}
Joined reads post QC						
Metagenomic	1.15×10^{7}	1.24×10^{7}	1.39×10^{7}	1.72×10^{7}	1.88×10^{7}	1.64×10^{7}
Metatranscriptomic	1.95×10^{7}	3.07×10^{7}	1.77×10^{7}	2.49×10^{7}	3.48×10^{7}	3.49×10^{7}
Mean read length (bp)	192	200	197	191	205	196
Protein Encoding Reads						
Metagenomic	4.98×10^{6}	2.31×10^{6}	5.55×10^{6}	6.38×10^{6}	5.86×10^{6}	6.77×10^{6}
Metatranscriptomic	4.08×10^{6}	3.21×10^{6}	2.36×10^{6}	3.53×10^{6}	2.41×10^{6}	6.52×10^{6}
Prokaryotic Gene Copies L^{-1}	8.77×10^{12}	1.47×10^{13}	3.42×10^{12}	1.61×10^{12}	8.67×10^{12}	2.21×10^{12}
% FL	68.25%	98.18%	97.28%	92.98%	99.21%	92.12%
% PA	31.75%	1.82%	2.72%	7.02%	0.79%	7.88%
Prokaryotic Transcripts L^{-1}	4.13×10^{11}	1.79×10^{11}	5.60×10^{11}	3.39×10^{11}	3.40×10^{11}	1.30×10^{11}
% FL	49.25%	93.79%	39.69%	21.47%	64.44%	84.32%
% PA	50.75%	6.21%	60.31%	78.53%	35.56%	15.68%

Table 4.S1. Metagenome and metatranscriptome dataset descriptions.

Sample	Site	Sample Type	Size Fraction	Collection Time	Volume Filtered (L)	Paired Reads	Avg. Read Length (nt)	Normalization Factor
ACM1	St. 2	DNA	2.0 -156 µm	5/25/10 7:55 AM	10.8	3,650,345	187	326,737
ACM2	St. 2	DNA	0.2 - 2.0 µm	5/25/10 7:55 AM	10.8	5,355,957	193	824,111
ACM3	St. 10	DNA	2.0 -156 µm	6/5/10 7:49 AM	5.2	4,211,422	187	2,347,636
ACM4	St. 10	DNA	0.2 - 2.0 µm	6/5/10 8:50 AM	4.8	4,656,473	188	2,718,586
ACM5	St. 27	DNA	2.0 -156 µm	6/21/10 8:43 AM	10.1	3,212,733	183	200,129
ACM6	St. 27	DNA	0.2 - 2.0 µm	6/21/10 7:55 AM	10	6,343,198	194	606,156
ACM11	St. 10	RNA	0.2 - 2.0 µm	6/5/10 7:00 AM	5.9	5,393,725	185	168,983
ACM12	St. 10	RNA	2.0 -156 µm	6/5/10 7:00 AM	5.9	5,084,135	181	188,525
ACM13	St. 23	RNA	2.0 -156 µm	6/16/10 9:04 AM	9.9	9,701,872	181	12,896
ACM14	St. 2	RNA	0.2 - 2.0 µm	5/25/10 7:10 AM	9.4	9,750,279	182	85,944
ACM15	St. 2	RNA	2.0 -156 µm	5/25/10 7:10 AM	9.4	845,395	182	2,132,430
ACM16	St. 27	RNA	0.2 - 2.0 µm	6/21/10 9:28 AM	10.3	14,288,947	186	20,944
ACM17	St. 27	RNA	2.0 -156 µm	6/21/10 9:28 AM	10.3	10,530,220	186	16,386
ACM18	St. 3	DNA	0.2 - 2.0 µm	5/26/10 9:10 AM	10.2	6,972,957	212	1,475,896
ACM19	St. 3	DNA	2.0 -156 µm	5/26/10 8:23 AM	10.4	1,813,399	203	349,248
ACM20	St. 23	DNA	0.2 - 2.0 µm	6/16/10 8:06 AM	10	1,330,783	209	11,447,432
ACM21	St. 23	DNA	2.0 -156 µm	6/16/10 10:15 AM	10.8	6,652,135	198	146,836
ACM22	St. 25	DNA	0.2 - 2.0 µm	6/18/10 7:57 AM	9.4	9,713,864	209	337,421
ACM23	St. 25	DNA	2.0 -156 µm	6/18/10 8:45 AM	4.85	3,366,035	218	589,720
ACM24	St. 2	DNA	0.2 - 2.0 µm	5/25/10 7:10 AM	4.85	4,514,764	219	603,802
ACM25[†]	St. 2	DNA	2.0 -156 µm	5/25/10 8:39 AM	11.1	3,719,777	186	-
ACM29	St. 10	RNA	0.2 - 2.0 µm	6/5/10 7:49 AM	4.8	3,698,936	196	334,631
ACM30	St. 10	RNA	2.0 -156 µm	6/5/10 8:50 AM	3.3	5,358,597	198	925,928
ACM31	St. 3	RNA	0.2 - 2.0 µm	5/26/10 10:00 AM	10.3	1,929,806	189	423,090
ACM32	St. 3	RNA	2.0 -156 µm	5/26/10 8:23 AM	10.4	4,172,258	189	42,790
ACM33	St. 23	RNA	0.2 - 2.0 µm	6/16/10 9:04 AM	10.3	2,059,460	188	654,064
ACM34	St. 25	RNA	0.2 - 2.0 µm	6/18/10 7:57 AM	8	11,466,966	209	70,399
ACM35	St. 25	RNA	2.0 -156 µm	6/18/10 7:57 AM	8	11,301,324	202	45,963
ACM36[†]	St. 10	DNA	0.2 - 2.0 µm	6/5/10 7:49 AM	5.2	1,325,814	194	-
ACM37	St. 10	DNA	2.0 -156 µm	6/5/10 8:50 AM	4.8	1,313,669	207	13,533,646
ACM38	St. 23	DNA	0.2 - 2.0 µm	6/16/10 10:15 AM	10.8	280,840	213	206,871,443
ACM39	St. 23	DNA	2.0 -156 µm	6/16/10 8:06 AM	10	4,154,716	209	2,387,283
ACM40	St. 3	DNA	0.2 - 2.0 µm	5/26/10 10:00 AM	10.4	3,324,817	216	1,393,074
ACM41[†]	St. 3	DNA	2.0 -156 µm	5/26/10 10:00 AM	10.4	1,753,412	205	-
ACM46	St. 2	RNA	0.2 - 2.0 µm	5/25/10 8:39 AM	10.6	6,690,564	178	60,426
ACM47	St. 2	RNA	2.0 -156 µm	5/25/10 8:39 AM	10.6	7,644,381	199	496,871
ACM48	St. 3	RNA	0.2 - 2.0 µm	5/26/10 8:23 AM	10.4	5,346,824	180	520,152
ACM49	St. 3	RNA	2.0 -156 µm	5/26/10 9:10 AM	10	6,271,644	180	1,125,329
ACM50	St. 23	RNA	0.2 - 2.0 µm	6/16/10 9:04 AM	9.9	5,327,535	192	412,247
ACM51	St. 23	RNA	2.0 -156 µm	6/16/10 9:04 AM	10.3	13,564,669	210	11,031
ACM52	St. 25	RNA	2.0 -156 µm	6/18/10 9:30 AM	9.5	8,772,347	193	1,328,987
ACM53	St. 25	DNA	0.2 - 2.0 µm	6/18/10 8:45 AM	9.8	263,233	215	144,810,010
ACM54	St. 25	DNA	2.0 -156 µm	6/18/10 9:30 AM	5.05	5,468,342	213	104,368
ACM55[†]	St. 27	DNA	0.2 - 2.0 µm	6/21/10 8:43 AM	13.4	4,217,346	239	56,055,488
ACM56	St. 27	DNA	2.0 -156 µm	6/21/10 9:28 AM	12.8	2,665,169	198	1,899,200
ACM58	St. 25	RNA	0.2 - 2.0 µm	6/18/10 8:45 AM	9.7	3,242,424	184	852,842
ACM59	St. 27	RNA	0.2 - 2.0 µm	6/21/10 8:43 AM	12.5	6,128,209	195	158,696
ACM60	St. 27	RNA	2.0 -156 µm	6/21/10 7:55 AM	9.6	3,905,567	186	20,515

[†]Standards were added incorrectly to sample ACM25, 36, 41, and 55; gene normalization was estimated based on replicate sample for each.

Figure 4.1 (a) Transcript inventories (mRNAs L^{-1}) for the ten most transcriptionally abundant prokaryotic taxonomic bins at each plume station. Bubbles are colored by taxonomic group (Alphaproteobacteria, dark purple; Cyanobacteria, light purple; Deltaproteobacteria, dark green; Flavobacteria, brown; Gammaproteobacteria, light green; Verrucomicrobia, yellow) (b) Map of the May-June 2010 Amazon plume showing the locations of the stations superimposed on salinity contours.


A
B
SAR86 cluster bacterium SAR86E
Synechococcus sp. CB0205
gamma proteobacterium HIMB30
Weeksella virosa DSM 16922
gamma proteobacterium HIMB55
Flavobacteriales bacterium ALC−1
Owenweeksia hongkongensis DSM 17368
Mesoflavibacter zeaxanthinifaciens
Polaribacter sp. MED152
SAR86 cluster bacterium SAR86D
Station 10
Flavobacteria bacterium MS024−2A
Zunongwangia profunda SM−A87
gamma proteobacterium HIMB30
gamma proteobacterium HIMB55
Coraliomargarita akajimensis DSM 45221
SAR324 cluster bacterium JCVI−SC AAA005
Flavobacteriaceae bacterium HQM9
gamma proteobacterium HTCC2207
Candidatus Puniceispirillum marinum IMCC1322
Synechococcus sp. WH 8109
Station 23
Coraliomargarita akajimensis DSM 45221
gamma proteobacterium HIMB30
Alteromonas macleodii ATCC 27126
Alteromonas macleodii str. 'Deep ecotype'
SAR86 cluster bacterium SAR86C
Synechococcus sp. WH 8109
SAR324 cluster bacterium JCVI−SC AAA005
Synechococcus sp. CC9605
Synechococcus sp. RS9916
gamma proteobacterium HTCC2207
Station 3
Coraliomargarita akajimensis DSM 45221
Verrucomicrobiae bacterium DG1235
SAR116 cluster alpha proteobacterium HIMB100
Roseobacter denitrificans OCh 114
Candidatus Puniceispirillum marinum IMCC1322
gamma proteobacterium HIMB55
Richelia intracellularis
Synechococcus sp. WH 8102
Ruegeria sp. TW15
SAR324 cluster bacterium JCVI−SC AAA005
Station 2
Coraliomargarita akajimensis DSM 45221
Richelia intracellularis
Candidatus Pelagibacter sp. HTCC7211
Synechococcus sp. WH 8102
alpha proteobacterium HIMB5
gamma proteobacterium HIMB55
gamma proteobacterium HIMB30
alpha proteobacterium HIMB59
gamma proteobacterium HTCC2207
Candidatus Puniceispirillum marinum IMCC1322
Station 25
Coraliomargarita akajimensis DSM 45221
Roseobacter denitrificans OCh 114
Prochlorococcus marinus str. AS9601
SAR116 cluster alpha proteobacterium HIMB100
Alteromonas macleodii ATCC 27126
Alteromonas macleodii str. 'Deep ecotype'
Candidatus Puniceispirillum marinum IMCC1322
Oceanicola batsensis
Maritimibacter alkaliphilus
Roseobacter sp. MED193
Station 27
FL
PA
27
25
2
23
3
10
12°
8°
4°
0°
-60
-56
-52
-48
-42
16
20
24
28
32
36
Alphaproteobacteria
Cyanobacteria
Deltaproteobacteria
Flavobacteria
Gammaproteobacteria
Verrucomicrobia
1 x 10^11
1 x 10^10
1 x 10^9
1 x 10^8

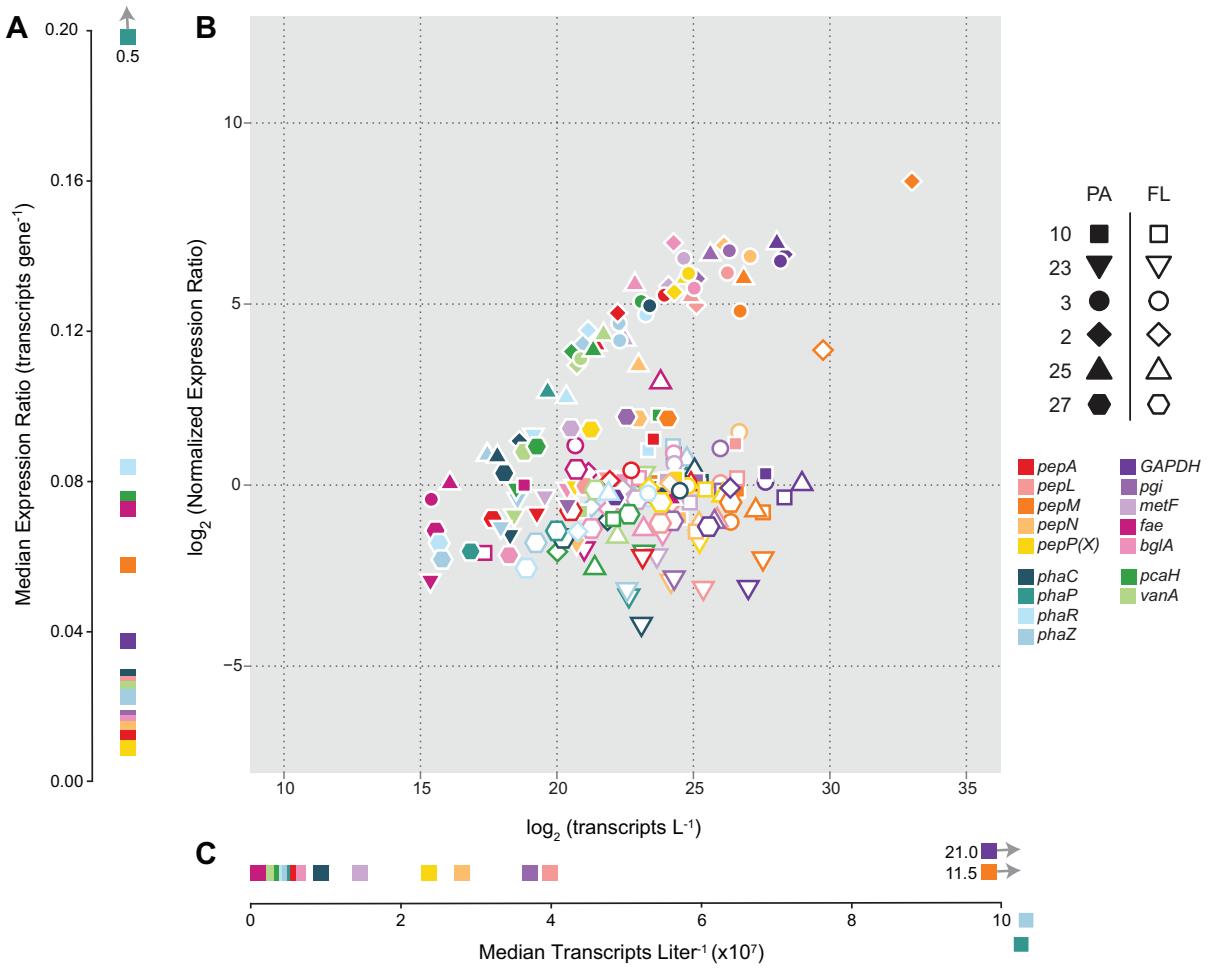


Figure 4.2 (a) Median expression levels (transcripts gene copy^{-1}) of 16 prokaryotic genes involved in heterotrophic carbon metabolism across all samples; (b) expression levels normalized to the median plotted against transcript abundance; (c) median transcript abundance.

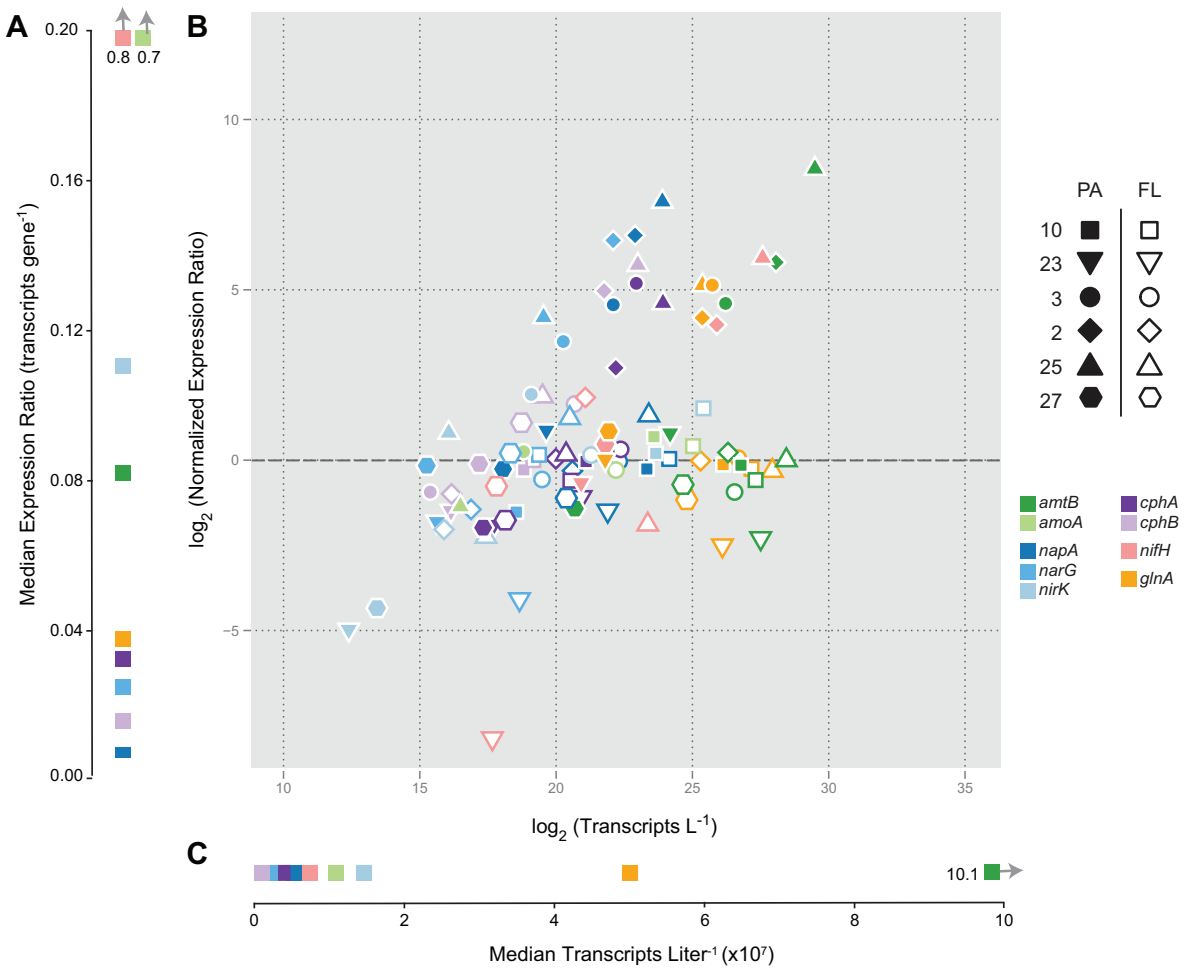


Figure 4.3 (a) Median expression levels (transcripts gene copy^{-1}) of 9 prokaryotic genes involved in nitrogen transport and metabolism across all samples; (b) expression level normalized to the median plotted against transcript abundance; (c) median transcript abundance.

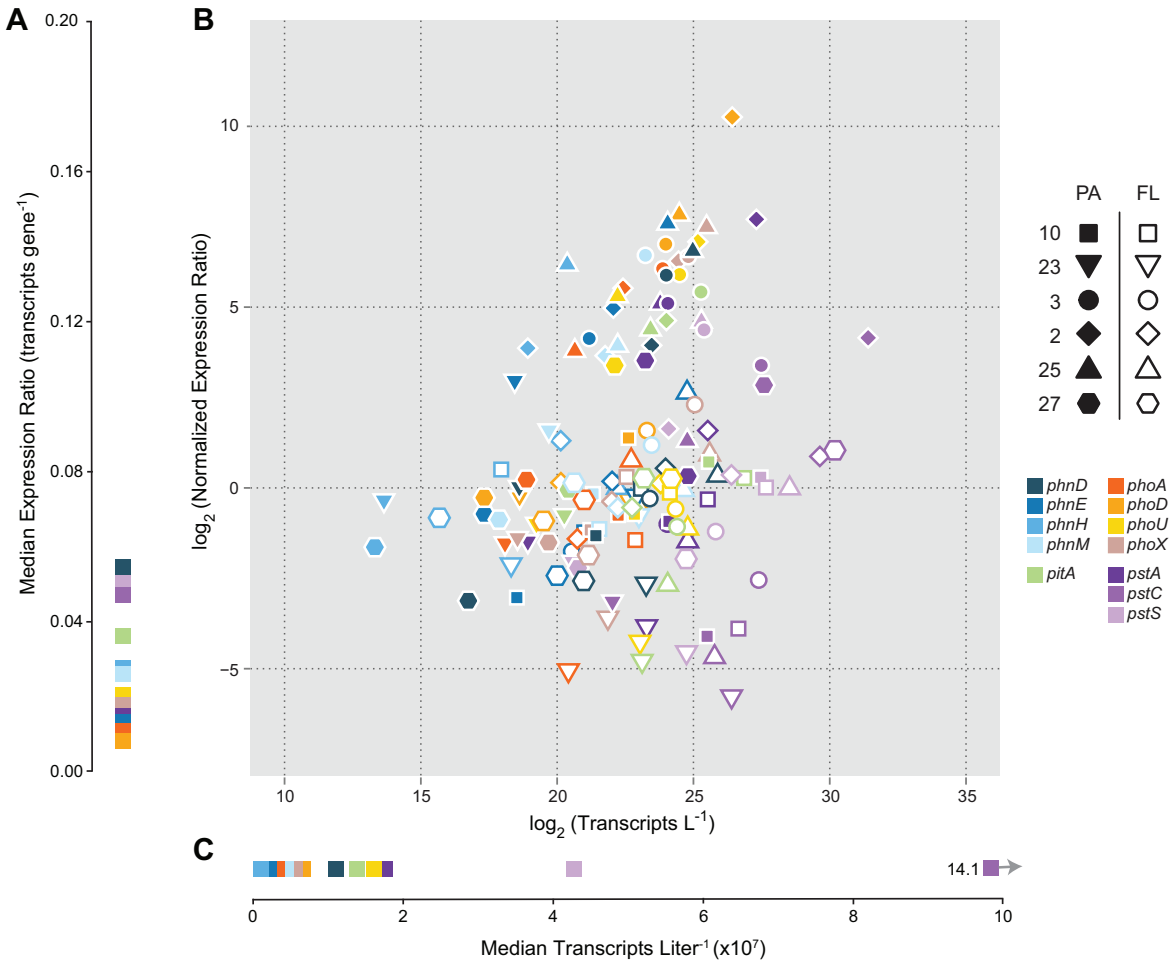


Figure 4.4 (a) Median expression levels (transcripts gene copy^{-1}) of 12 prokaryotic genes involved in phosphorus transport and metabolism across all samples; (b) expression level normalized to the median plotted against transcript abundance; (c) Median transcripts abundance.

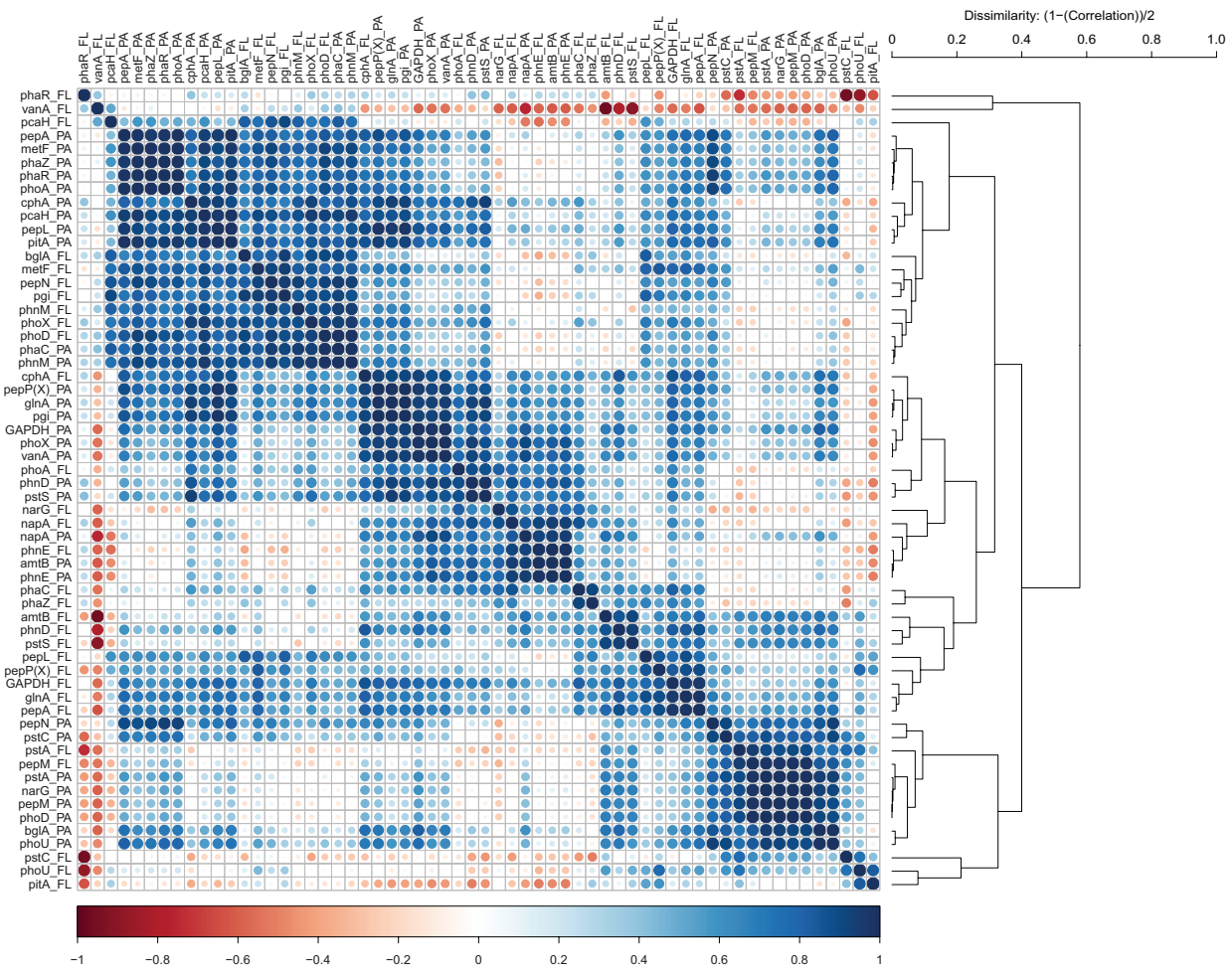


Figure 4.5 Correlation analysis and hierarchical clustering of biogeochemically-relevant genes in each size fraction. Blue represents positive correlations, red indicates negative correlations, and circle area is representative of the absolute value of the correlation coefficient.

Figure 4.S1 OTU fingerprints and hierarchical clustering for two individual ribosomal proteins for each of 5 reference genome bins. Individual stations and size fractions identified as containing the same population structure for a given gene and reference genome are marked with red circles.

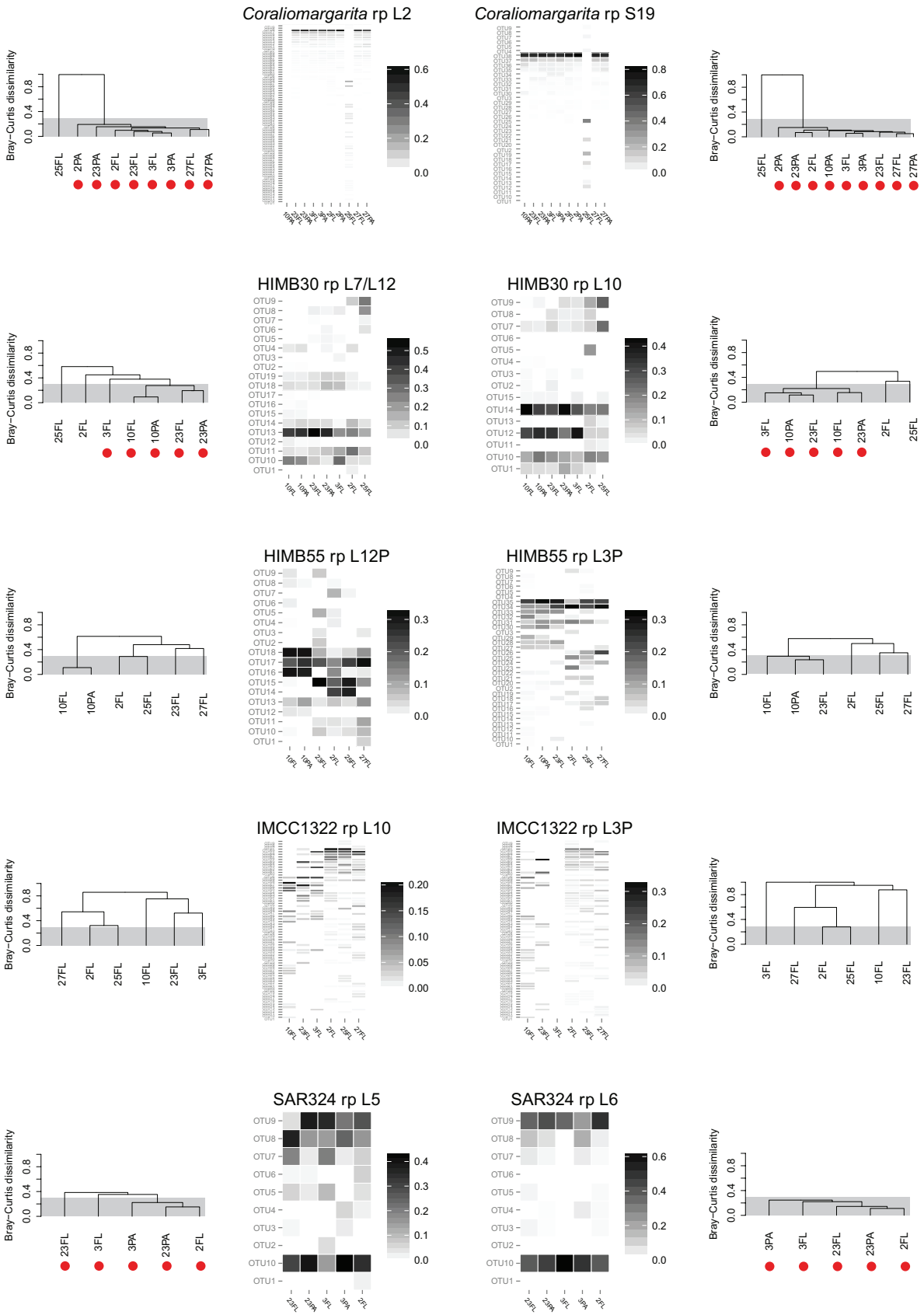

Coraliomargarita rp L2
Coraliomargarita rp S19
HIMB30 rp L7/L12
HIMB30 rp L10
HIMB55 rp L12P
HIMB55 rp L3P
IMCC1322 rp L10
IMCC1322 rp L3P
SAR324 rp L5
SAR324 rp L6
Bray−Curtis dissimilarity

Figure 4.6 Transcript inventories and expression levels for 9 biogeochemically-relevant genes in 3 individual reference genome bins (*Coraliomargarita akajimensis* DSM 45221, yellow; gammaproteobacterium HIMB30, light green; SAR324 cluster bacterium JCVI−SC AAA005, dark green) across plume stations containing the same population structure within a bin.

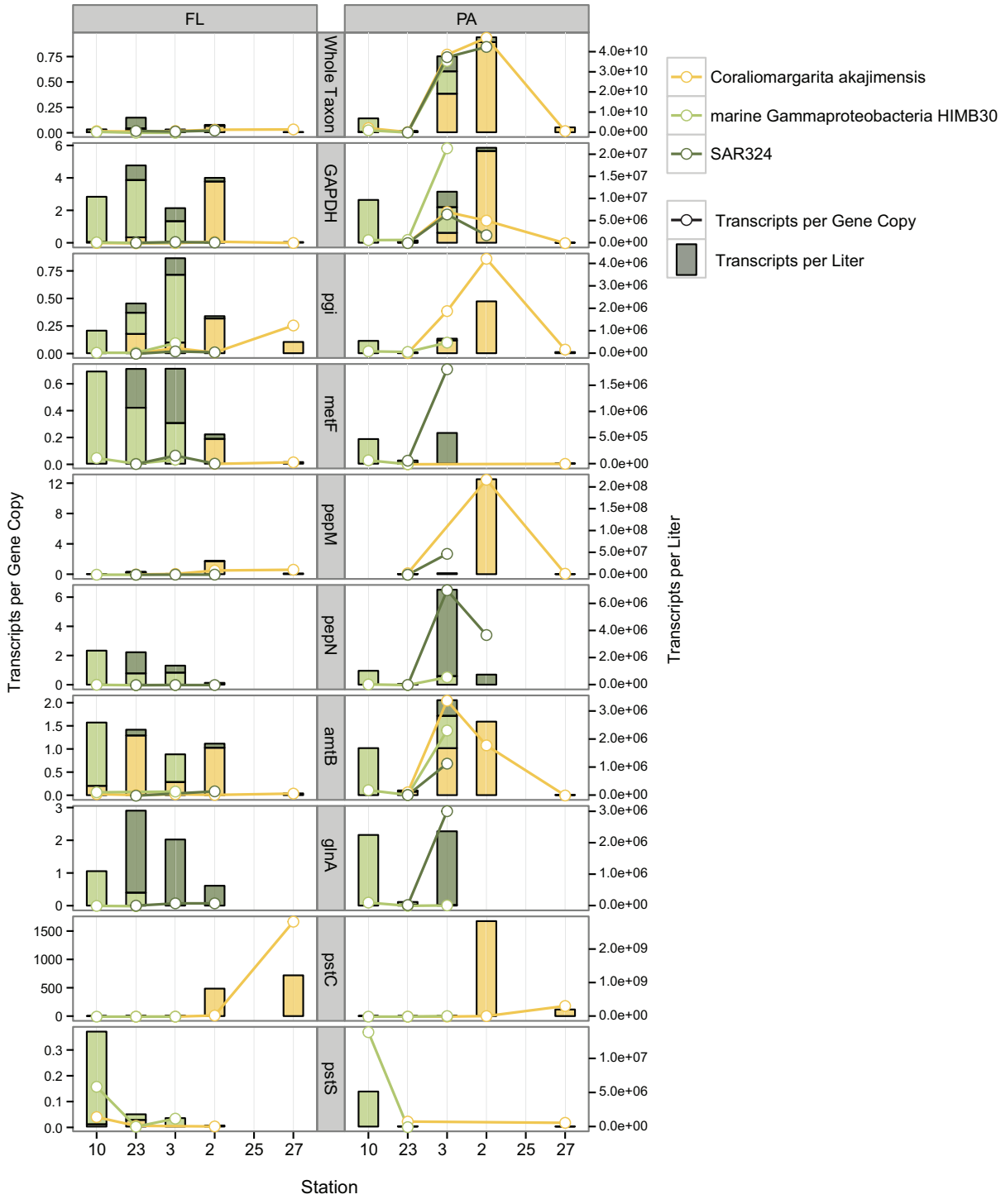
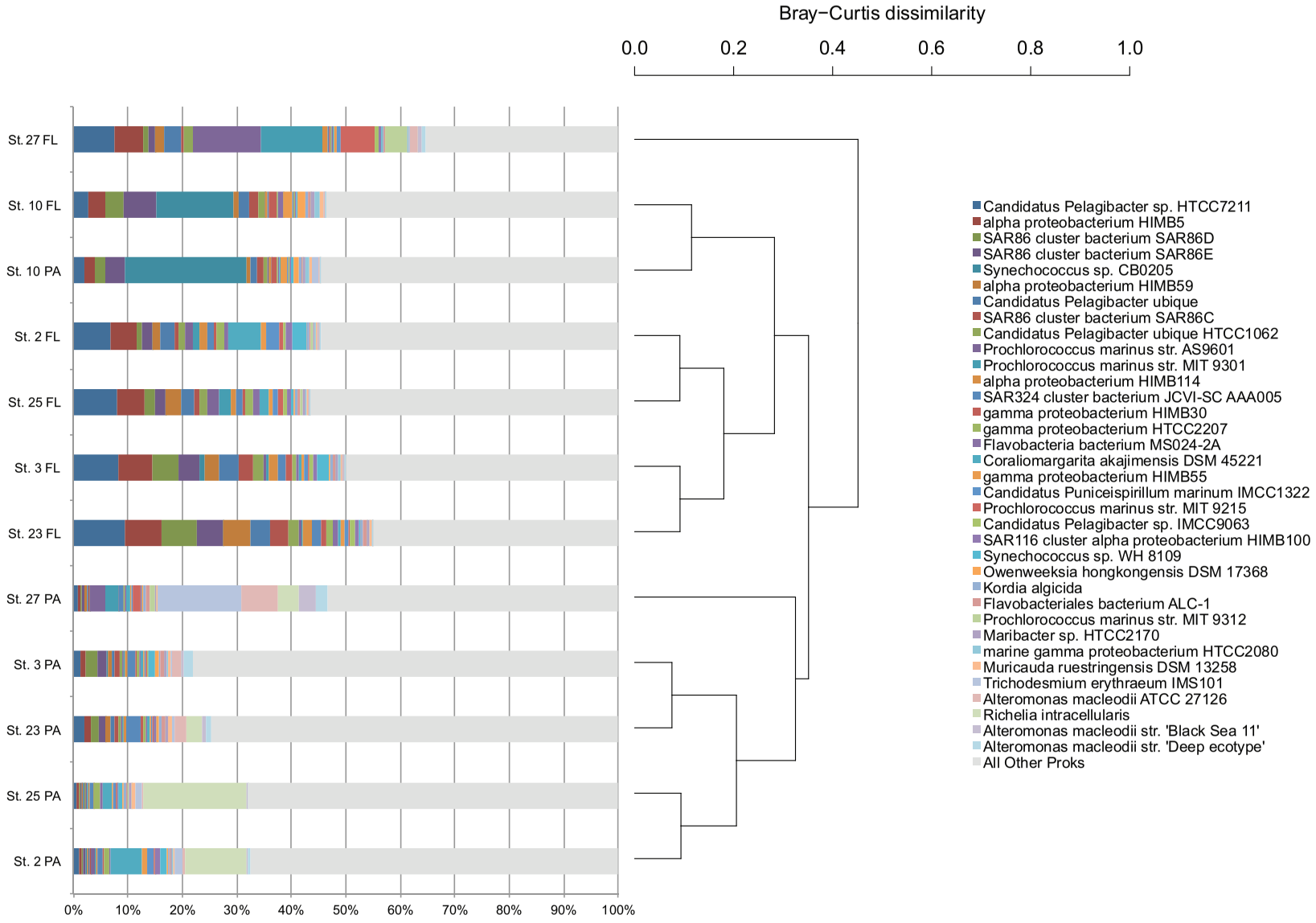

FL
PA
Whole Taxon
GAPDH
pgi
metF
pepM
pepN
amtB
glnA
pstC
pstS
Transcripts per Gene Copy
Transcripts per Liter
Station
10
23
3
2
25
27
Coraliomargarita akajimensis
marine Gammaproteobacteria HIMB30
SAR324
Transcripts per Gene Copy
Transcripts per Liter

Figure 4.S2 Relative abundance patterns in the metagenomes of 30 highly abundant taxonomic bins for each station and size fraction clustered hierarchically.


Bray−Curtis dissimilarity
0.0
0.2
0.4
0.6
0.8
1.0
St. 27 FL
St. 10 FL
St. 10 PA
St. 2 FL
St. 25 FL
St. 3 FL
St. 23 FL
St. 27 PA
St. 3 PA
St. 23 PA
St. 25 PA
St. 2 PA
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Candidatus Pelagibacter sp. HTCC7211
alpha proteobacterium HIMB5
SAR86 cluster bacterium SAR86D
SAR86 cluster bacterium SAR86E
Synechococcus sp. CB0205
alpha proteobacterium HIMB59
Candidatus Pelagibacter ubique
SAR86 cluster bacterium SAR86C
Candidatus Pelagibacter ubique HTCC1062
Prochlorococcus marinus str. AS9601
Prochlorococcus marinus str. MIT 9301
alpha proteobacterium HIMB114
SAR324 cluster bacterium JCVI-SC AAA005
gamma proteobacterium HIMB30
gamma proteobacterium HTCC2207
Flavobacteria bacterium MS024-2A
Coraliomargarita akajimensis DSM 45221
gamma proteobacterium HIMB55
Candidatus Puniceispirillum marinum IMCC1322
Prochlorococcus marinus str. MIT 9215
Candidatus Pelagibacter sp. IMCC9063
SAR116 cluster alpha proteobacterium HIMB100
Synechococcus sp. WH 8109
Owenweeksia hongkongensis DSM 17368
Kordia algicida
Flavobacteriales bacterium ALC-1
Prochlorococcus marinus str. MIT 9312
Maribacter sp. HTCC2170
marine gamma proteobacterium HTCC2080
Muricauda ruestringensis DSM 13258
Trichodesmium erythraeum IMS101
Alteromonas macleodii ATCC 27126
Richelia intracellularis
Alteromonas macleodii str. 'Black Sea 11'
Alteromonas macleodii str. 'Deep ecotype'
All Other Proks

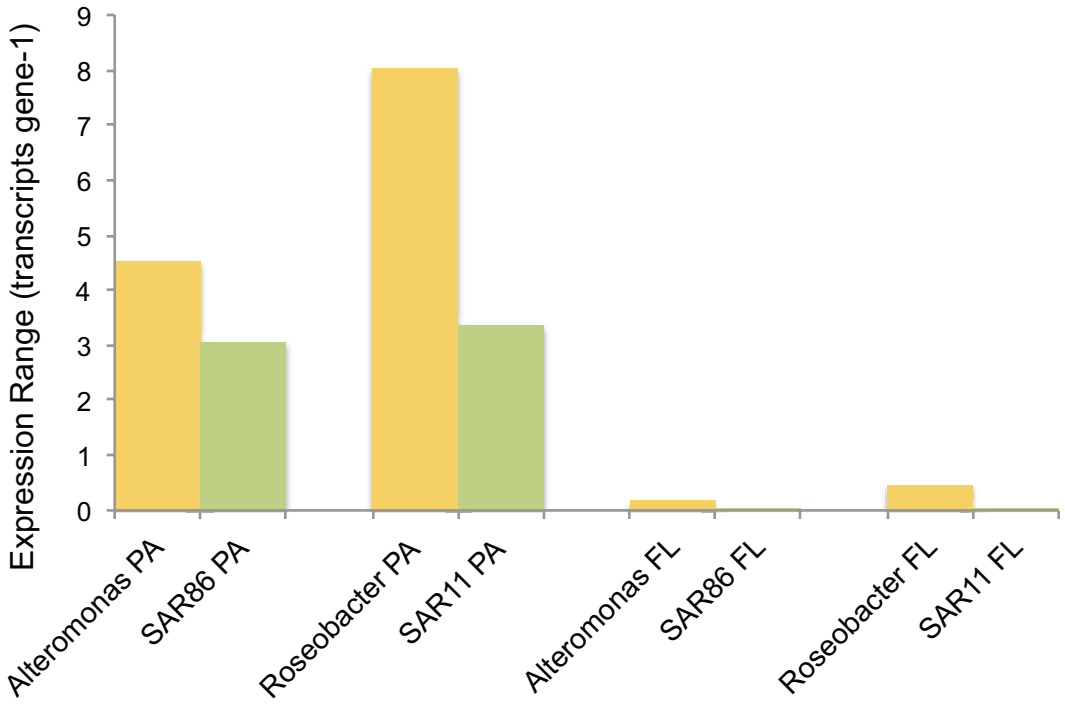


Figure 4.7 Difference between the highest and lowest expression levels for two pairs of streamlined (green) and non-streamlined (yellow) taxonomic groups representing Alphaproteobacteia (pooled SAR11 genomes versus pooled Roseobacter genomes) and Gammaproteobacteria (pooled SAR86 genomes versus pooled Alteromonadales genomes).

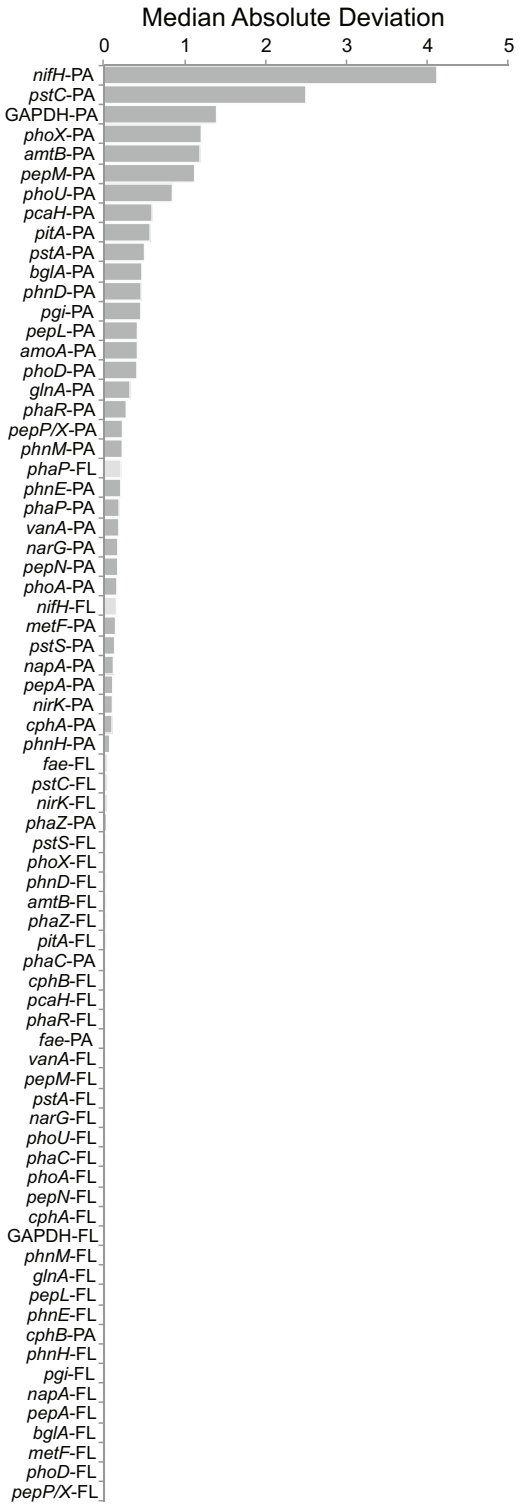


Figure 4.8 Rank order of the median absolute difference for each gene and size fraction (FL, light gray; PA, dark gray).

Table 4.S2. Metadata for 6 Amazon River plume stations in June 2010.

	Station 10	Station 23	Station 3	Station 2	Station 25	Station 27
Temperature (C°)	29.36 ± 0.06	29.22 ± 0.08	28.85 ± 0.13	28.63 ± 0.02	29.21 ± 0.09	28.4 ± 0.02
Salinity (PSU)	22.55 ± 1.36	26.49 ± 0.04	30.8 ± 0.61	31.8 ± 0.02	31.87 ± 0.03	36.03 ± 0.02
Depth (m)	4.26 ± 0.89	3.64 ± 0.48	3.76 ± 0.42	4.47 ± 0.49	3.93 ± 0.19	3.89 ± 0.34
Pressure (Db)	4.28 ± 0.9	3.66 ± 0.49	3.78 ± 0.42	4.5 ± 0.49	3.95 ± 0.19	3.92 ± 0.34
Density (kgm^{-2})	1012.65 ± 1.02	1015.64 ± 0.02	1018.98 ± 0.42	1019.8 ± 0.02	1019.66 ± 0.01	1023.05 ± 0.02
Sigma-theta	12.63 ± 1.02	15.62 ± 0.02	18.97 ± 0.42	19.78 ± 0.02	19.65 ± 0.01	23.04 ± 0.02
Oxygen (µmol/kg)	251.56 ± 22.28	194.25 ± 0.36	188.26 ± 0.6	196.25 ± 0.53	213.84 ± 0.76	188.39 ± 0.39
Oxygen_sat (pct)	120.81 ± 11.15	95.39 ± 0.26	94.43 ± 0.28	98.7 ± 0.29	108.6 ± 0.52	96.95 ± 0.22
Fluorescence (mgm^{-3})	7.25 ± 0.54	0.16 ± 0.01	0.28 ± 0.03	0.61 ± 0.05	0.99 ± 0.32	0.01 ± 0.01
Turbidity (NTU)	3.02 ± 0.13	1.1 ± 0.02	1.12 ± 0.01	1.17 ± 0.01	1.23 ± 0.02	1.07 ± 0.03
Surface Irradiance (µEm-2sec-1)	1913.71 ± 1043.71	1449.22 ± 1045.6	2690.35 ± 68.39	796.66 ± 258.27	2057.58 ± 925.74	1621.32 ± 970.75
PAR (µEm-2sec-1)	22.08 ± 33.17	27.94 ± 21.55	65.38 ± 33.93	16.71 ± 5.95	73.71 ± 47.37	80.36 ± 70.11
Bacterial Production (pmol leu/L/hr)	327.96	79.80	139.68	112.48	78.48	16.11
Dissolved Organic Carbon (micromoles/L)	97.00	116.00	101.00	98.00	108.00	88.00
Dissolved Inorganic Carbon (micromoles/L)	1373.00	1576.00	1846.00	1802.00	1775.00	2005.00
Chlorophyll A (mg/L)	22.40	0.15	0.58	2.66	5.25	0.13
PO_4 (micromolar)	0.39	0.29	0.19	0.11	0.00	0.52
NO_3 (micromoles/L)	0.14	0.00	0.00	0.00	0.00	0.00
NO_2 (micromoles/L)	0.05	0.00	0.00	0.00	0.00	0.00
NH_4 (micromoles/L)	0.08	0.11	n.d.	0.00	0.04	0.00

CHAPTER 5

QUANTITATIVE MICROBIAL GENE EXPRESSION PATTERNS OF THE AMAZON RIVER[1]

[1] Satinsky, B.M., Crump, B.C., Smith, C.B., Sharma, S., Yager. P.L., and Moran, M.A. To be submitted to *International Society of Microbial Ecology Journal.*

Abstract

The Amazon River is the world's largest river system and plays a central role in global nutrient cycling, with fundamental aspects of energy production and consumption within the river under the control of prokaryotes. Here, we use metagenomics and metatranscriptomics benchmarked with internal standards to generate the first quantitative transcript inventory of freshwater prokaryotes and explore expression patterns of elemental cycling genes along the lower Amazon River. On average, prokaryotic transcript pools in the lower Amazon River harbored more than 1×10^{11} transcripts L^{-1}. The transcriptomes were dominated by taxa related to Actinobacteria, Thaumarchaea, Betaproteobacteria, and at stations with reduced turbidity, cyanobacteria related to *Synechococcus* and *Microcystis*. Expression levels of genes involved in ammonia oxidation, ammonium transport, and denitrification were dynamic spatially along the river as well as between microenvironments, and were frequently among the most highly expressed elemental cycling genes. Despite inputs of vascular plant-derived organic material from surrounding watersheds, signals for aromatic carbon metabolism were often times lower in the river than in the plume waters offshore from the Amazon mouth. These patterns in prokaryotic gene expression shed light on microbial activities in the world's largest river system, and generate new questions about the roles of prokaryotes in river biogeochemistry.

Introduction

The Amazon River is the world's largest riverine system based on both volume and watershed area (Coles *et al.,* 2013), and runs nearly 6,500 km across the South American continent before emptying into the Western Tropical North Atlantic Ocean. The rainforest of the Amazon Basin is responsible for nearly 10% of global primary production (Field *et al.,* 1998),

and much of the 8.5 Pg C fixed per year (Malhi *et al.,* 2008) ultimately ends up in the river. The main channel of the Amazon is well mixed and turbid, and high levels of CO_2 as well as low light penetration into the water suggest an environment dominated by heterotrophic bacteria. These microbes remove, transform, and stabilize riverine organic matter during transit, and their activities lead to river outgassing of 500 Tg C per year to the atmosphere (Richey *et al.,* 2002).

Within the mainstem of the Amazon, heterotrophic bacteria rely on the allochthonous input of carbon and nutrients from the surrounding rainforest and drainage basins. Humic and fulvic acids, likely derived from lignin and other terrestrial plant components, account for ~60% of riverine dissolved organic carbon (DOC) (Ertel *et al.,* 1986), and recent work suggests that degradation of terrestrially-derived organic compounds by bacteria contributes significantly to the bulk river respiration and outgassing (Ward *et al.,* 2013). In contrast, many of the tributaries along the river have much lower turbidity levels and may provide conditions suitable for the growth of photosynthetic microorganisms. As the prokaryotic community plays a major role in carbon cycling in the river, it also plays a critical role in the processing and cycling of nitrogen, phosphorus, iron, and other elements.

There are an estimated 1.2×10^{24} prokaryotic cells in rivers globally, with average cell concentrations typically twice those in the upper ocean (Whitman *et al.,* 1998). Yet we know comparatively little about the diversity and activity of the freshwater microbial communities. To date most research efforts attempting to resolve microbial community structure in freshwater environments have used 16S rRNA gene based methodology (Newton *et al.,* 2011), including in the Amazon River (Peixoto *et al.,* 2011). This single-gene taxonomic marker, however, cannot address microbial function and activity. The recent development of metagenomics enables deeper understanding of microbial activity in natural environments through the assembly of a

community gene inventory. The methodology has recently been used to study microbes of lakes (Debroas *et al.,* 2009, Oh *et al.,* 2011, Pope and Patel 2008, Rusch *et al.,* 2007) and the upper course of the Amazon River (Ghai *et al.,* 2011). Metatranscriptomics enables the assembly of a community transcriptome (Poretsky *et al.,* 2005), focusing on the genes being actively expressed (transcripts) and reflecting microbial community responses to their environment. Metatranscriptomics-based methodologies have been applied sparsely in freshwater systems, and thus far only in lakes (Tsementzi *et al.,* 2014, Vila-Costa *et al.,* 2013). When used in tandem with internal standards (Gifford *et al.,* 2011, Satinsky *et al.,* 2013) metagenomics and metatranscriptomics can generate detailed inventories of microbial gene abundances, transcript abundances, and levels of gene expression (transcripts gene copy^{-1}), allowing for direct comparison of spatially and temporally segregated datasets (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b).

Here we detail the first fully-quantitative assessment of prokaryotic gene expression in the world's largest river ecosystem. Five stations within the lower reaches of the Amazon River (Óbidos to Macapá and Belém) during a period of high discharge in May 2011 were investigated using both metagenomics and metatranscriptomics methodologies enhanced by high-throughput sequencing technologies and benchmarked with internal standards.

Methods

Sample Collection

In May 2011, microbial cells were collected by filtration for both free-living (0.2 – 2.0 μm) and particle-associated (>2.0 μm) size fractions, with duplicate samples collected for each sample type and size fraction, at 5 stations in the lower reaches of the Amazon River system

(Table 5.1). All samples were collected from 50% of river depth by pumping water to the boat's deck using a Shurflo submersible pump fitted with a 297 μm stainless steel screen. Water collected into 20 L carboys was then sequentially filtered through a 2.0 μm pore-size, 142 mm diameter polycarbonate (PCTE) membrane filter (Sterlitech) and a 0.22 μm pore-size, 142 mm diameter Supor membrane filter (Pall), and the volume of filtrate was recorded. Filtration was completed within 30 minutes of collection for the metatranscriptomic samples. Following filtration all membrane filters were immediately submerged in RNAlater (Applied Biosystems) in sterile 15 mL conical tubes and refrigerated until returning to shore, at which point the samples were frozen until processing.

Sample Processing

For all metatranscriptomic samples, filters were thawed, removed from the RNAlater preservation solution, placed in Whirl-Pak® bags (Nasco), and flash-frozen in liquid nitrogen. Frozen filters in bags were crushed into small pieces using a rubber mallet, and each crushed sample was transferred to a prepared 50 mL lysis tube containing 10 mL of Denaturation Solution (Ambion), 500 μL of Plant RNA Isolation Aid (Ambion), 2 mL of sterilized zirconium beads (OPS Diagnostics), and internal standards (Satinsky *et al.,* 2013). Tubes were vortexed for 10 min to lyse cells, after which time the tubes were centrifuged for 1 min at 5,000 rpm. The lysates were transferred to a sterile 15 mL conical tube and then centrifuged for 5 min at 5,000 rpm. The clarified lysates were transferred to sterile 50 mL conical tubes and 3.5 mL of saturated phenol (pH 4.3) was added to each lysate and vortexed thoroughly. The tubes were centrifuged for 8 min at 12,000 x g, after which the non-viscous phase in each tube was transferred to a fresh 50 mL conical tube, and 3.5 mL of a phenol:chloroform solution (1:1, pH 5) was added and tube

contents were mixed well. Following another 8 min centrifugation at 12,000 x g, the aqueous phase in each tube was transferred to a sterile 50 mL conical tube and 5 mL of chloroform:isoamyl alcohol solution was added. Tubes were vortexed and contents were centrifuged for 5 min at 12,000 x g. The final aqueous phase in each tube was transferred to a fresh 50 mL conical tube prior to the addition of an equal volume of 100% ethanol. Each mixture was homogenized by passage through a syringe several times. RNA purification was completed for each sample using the Direct-Zol RNA Kit (Zymo Research) according to manufacturer's protocol. Following RNA extraction, residual DNA was removed, rRNA was depleted, and mRNA was linearly amplified before the synthesis of double-stranded cDNA as previously described (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b).

For metagenomic samples, filters were processed as previously described (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b) with the inclusion of an internal standard consisting of *Thermus thermophilus* HB8 genomic DNA (American Type Culture Collection) (Satinsky *et al.,* 2013). In short, each filter was thawed and removed from RNAlater, rinsed with 0.1% phosphate-buffered saline, sliced into pieces, and added to a prepared lysis tube containing the internal standard and extraction buffer. After treatments with proteinase-K, lysozyme, and sodium dodecyl sulfate the DNA was purified using a phenol:chloroform extraction and isopropanol precipitation.

Resulting cDNA and DNA preparations were sheared ultrasonically to ~200-250 bp fragments and libraries were constructed for paired-end sequencing (150 x 150) using the HiSeq 2500 platform (Illumina Inc.). Following sequencing, reads were paired using PandaSeq (Masella *et al.,* 2012) and filtered with FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) with a minimum score cutoff of 20 over 80% of a read. From the paired, quality-controlled reads, internal standard sequences were quantified and removed, and any rRNA sequences were

removed from the metatranscriptomes. Remaining reads were annotated against the RefSeq Protein database and against a smaller, custom functional gene database using RAPsearch2 (Zhao *et al.,* 2012) and blastx respectively. Transcript and gene abundances as well as expression ratios were calculated as previously described (Table 5.S1) (Satinsky *et al.,* 2013).

Results

In May 2011, samples were collected in the lower reaches of the Amazon River at five stations (Tables 5.1 and 5.S2) from 50% of river depth at that location, which ranged from 10 to 33 m depth. Samples were collected for metagenomic (gDNA) and metatranscriptomic (mRNA) analyses from two operationally defined size fractions (free-living; 0.2 – 2.0 μm, particle-associated; 2.0 – 297 μm). Among the 5 stations (Fig. 5.1), the Óbidos station (OB) was the furthest upriver, above which ~90% of the Amazon's water enters the mainstem. Downriver of Óbidos, near Santarém, the Tapajós station (TAP) captured the major input of the Tapajós river into the mainstem of the Amazon, and this station is characterized by much clearer water than found in the river proper. Further downriver the Macapá north channel (MCPN) and Macapá south channel (MCPS) stations represent two of the major outflows of the Amazon mainstem into the ocean, while the Belém (BLM) station to the south represents the third outflow and also captured the input of water from the Tocantins River. Duplicate samples were collected for each station and analysis type, generating a total of 20 metagenomes and 20 metatranscriptomes, and more than 500 million paired, quality controlled reads were obtained giving 10-30 million potential protein encoding reads per library (Table 5.1). Functional and taxonomic binning were carried out for the reads, and here we detail the molecular inventories and gene expression

patterns of prokaryotic (Bacteria and Archaea) communities in this lower region of the Amazon River system.

Prokaryotic genes were present in higher numbers than transcripts at the Amazon stations, and on average there were more genes and transcripts present in the particle-associated fraction than in the free-living (Table 5.1). Across the five stations, prokaryotic transcript inventories averaged of 1.1 x 10^{11} transcripts L^{-1} (Table 5.1), and about 23% (~2.5 x 10^{10} L^{-1}) were contributed by the ten most dominant taxonomic bins at each station. The taxonomic composition and rank order in terms of contribution to the transcript pool of these top 10 bins varied among stations, but several were among the most dominant taxa at many stations (Fig. 5.1). The Actinobacterium SCGC AAA027-L06 reference genome, for example, was the highest recruiting bin, accounting for between 2.2 and ~11% of the prokaryotic transcript pool across all five stations. Four of the five stations (OB, MCPN, MCPS, BLM) had multiple highly-recruiting Thaumarchaeota bins (reference genomes *Nitrosopumilus maritimus* SCM1, *Candidatus* Nitrosoarchaeum limnia, *Candidatus* Nitrosopumilus *sp.* AR2, *Candidatus* Nitrosoarchaeum koreensis, *Candidatus* Nitrosopumilus salaria, and *Candidatus* Nitrosopumilus koreensis AR1). The largest contributions from these bins were at the three stations nearest the ocean, and they accounted for as much as ~21% (1.7 x 10^{10} transcripts L^{-1}) of the total transcriptome at MCPS (Fig. 5.1). In addition, heterotrophic bacterial bins related to Betaproteobacteria, Planctomycetia, Verrucomicriobia, and Sphingobacteria were often dominant contributors to the transcript pool at more than one station. Autotrophic taxon bins were also among the most dominant at some stations, including *Synechococcus* sp. CB0101 at the clearwater TAP station and *Microcystis aeruginosa* at BLM (Fig. 5.1). Typically, the dominant taxa contributed more transcripts per liter to the particle-attached size fraction than to the free-living fraction; the extreme case was at

MCPN where roughly 11 times more transcripts were present in the PA community compared to the FL community. TAP was the only station where more transcripts were present in the free-living bacterial and archaeal cells; here, the FL fraction harbored 4 times more transcripts L^{-1} compared to the PA fraction (Fig. 5.1).

Transcript Inventories - Transcript abundances were determined for 90 prokaryotic genes involved in a range of biogeochemically-relevant processes, including transport and metabolism of carbon, nitrogen, phosphorus, and iron. Among the genes related to heterotrophy, more transcripts were identified for GAPDH than any other gene when averaged across the five stations, followed by carbon storage genes (*phaP*, *phaC*) and aminopeptidases genes (*pepM*, *pepN*) (Fig. 5.2). Abundances ranged over more than 2 orders of magnitude for these heterotrophy genes (~1 x 10^8 transcripts L^{-1} for GAPDH to ~8 x 10^5 transcripts L^{-1} for *fsdD*) when averaged across the 5 stations. For genes involved in light utilization and carbon fixation, transcript abundances ranged nearly four orders of magnitude with the highest abundance attributed to the photosystem II gene *psbA* (~3 x 10^8 transcripts L^{-1} when averaged across the 5 stations) followed by proteorhodopsin and other photosystem II genes (*psbB, psbC*) (Fig. 5.2). RuBisCO transcripts were dominated by *rbcL* (ID) and *rbcL* (IA).

Of the 90 genes surveyed, 3 of the 4 most abundant in the transcriptomes were involved in nitrogen transport and metabolism (*amoA, amtB, nirK*), and each accounted for an average of >10^8 transcripts L^{-1} at each station. Among phosphorus-related genes, more transcripts were produced from genes encoding proteins involved in high-affinity phosphate transport (*pstS, pstC, pstA, phoU*) than the other phosphorus-related genes surveyed, and when combined the four high-affinity phosphate transport related genes averaged ~2 x 10^8 transcripts L^{-1}. The most abundant transcripts related to sulfur cycling were those for sulfur assimilation (*cysK*, *cysI*),

while the most abundant related to iron and vitamins were those for the thiamine biosynthesis gene *thiC* (Fig. 5.2).

Looking at the individual stations, *amoA* contributed more transcripts than for any other gene surveyed except all stations except TAP (Fig. 5.2). Phosphate transport transcripts reached their highest abundances at the BLM and TAP stations, and these were also the stations with highest abundance of transcripts mediating photosynthesis (photosystem genes *psbA*, *psbB*, and *psbC*, as well subtypes of the carbon fixation gene *rbcL*) (Fig. 5.2). In addition, both BLM and TAP had higher abundances of transcripts encoding proteorhodopsin than the other stations.

Gene Expression - For each of the 90 genes we also calculated expression levels by normalizing transcript numbers from the metatranscriptomes to gene numbers from the metagenomes (transcripts gene copy^{-1}). Expression ratios were consistently higher in the FL fraction relative to the PA fraction at both OB (84 of 90 genes) and MCPS (77 of 90 genes), while ratios were consistently higher in the PA fraction at MCPN (87 of 90) and BLM (89 of 90); TAP expression was balanced, with 44 genes higher in each of the two fractions (Fig. 5.2). Heterotrophic metabolism gene *phaP*, which had one of the highest inventories (see above), also had one of the highest levels of expression (Fig. 5.2), ranging up to 1.8 transcripts gene copy^{-1} in the PA fraction at BLM (Fig 2). Methane monooxygenase (*mmoB*) and a tetrahydromethanopterin-linked C1 metabolism protein (*fae*, formaldehyde activating enzyme) also had very high transcript gene^{-1} ratios, and both reached peak expression in the PA fraction at BLM. Photosystem II genes *psbA*, *psbB*, and *psbC*, also among the genes with the highest inventories, were consistently among those with the highest transcript gene^{-1} ratios, and all peaked in expression in the PA fraction at BLM (Fig. 5.2). Consistent with transcript inventories, the two most highly expressed forms of RuBisCO were *rbcL* (IA) and *rbcL* (IB), both of which

had quite variable expression ratios across the stations (*rbcL* (IA), ~35 fold; *rbcL* (IB), ~50 fold) (Fig. 5.2). The ammonia oxidase gene (*amoA)* had high transcript gene^{-1} ratios at all stations and in both fractions (Fig. 5.2). Nitrogenase gene *nifH*, nitric oxide reductase gene *norB*, and nitrous oxide reductase gene *nosZ* had some of the lowest transcripts gene copy^{-1}, while genes encoding high-affinity phosphate transport (*pstS, pstC, pstA,* and *phoU*) typically had among the highest (Fig. 5.2).

Single Taxon Analyses - We asked whether there was evidence that bacterial and archaeal taxa were regulating biogeochemically-relevant genes differently at various river stations. To do this, analysis was restricted to taxonomic bins with consistent population structure at multiple stations. For each of eight reference genome bins with high transcript coverage (3 heterotrophic bacteria and 5 thaumarchaeota), transcripts mapping to selected ribosomal protein genes were clustered into operational taxonomic units (OTUs) based on 95% nucleotide identity. Reads from each station and size fraction were then assigned to the OTUs and the resulting distributions were clustered hierarchically for each gene. Because reads have an average length (~215 nt) that is shorter than the full-length ribosomal protein genes, even a single population of identical genomes will produce multiple OTUs that map to different gene regions. However, as long as the same population(s) are present in the same relative abundances, OTU patterns will be similar between samples. Three of the reference bins (Actinobacterium SCGC AAA027-L06, the planctomycete *Schlesneria paludicola*, and the verrucomicrobium *Pedosphaera parvula*) generated fingerprints with little similarity between stations, indicating that the population structure of the reads binning to the reference genes was shifting spatially along the river (Fig. 5.3). However, in the case of the Actinobacterium SCGC AAA027-L06 bin, the same OTU pattern emerged from both size-fractions for three of the stations (OB, BLM, TAP). Further, each

of the five archaeal bins produced congruent OTU fingerprints at the OB, MCPN, MCPS, and BLM stations, suggesting that similar populations were binning to these reference genomes at multiple locations in the river (Figs. 5.3 and 5.S1).

Using the reference genome bins that passed the test for population homogeneity, gene regulation was assessed in individual taxa at different river locations. For the thaumarchaeal genomes, expression patterns were fairly similar at all stations within a bin, particularly in the case of three highly-expressed genes related to chemoautotrophic nitrogen metabolism (*amoA*, *nirK*, and *amtB*) (Fig. 5.4). However, the same population varied its transcript gene^{-1} ratios substantially between stations. For example, *nirK* expression varied 18-fold within the *Candidatus* Nitrosopumilus salaria genome bin, ranging from a high of ~12 transcripts gene copy^{-1} (PA fraction at OB) to a low of 0.65 transcripts gene copy^{-1} (FL fraction at MCPN). The expression ratios were similar across the five Thaumarcheota reference bins (Fig. 5.4), although clusters of samples with similar transcriptomes could be differentiated, suggesting both taxonomic and geographic influences on expression patterns (Fig. 5.5).

Discussion

Because relatively few genomes have been sequenced for freshwater prokaryotes (Ghai *et al.,* 2011), binning reads from the Amazon River metagenomes and metatranscriptomes to existing reference genomes in some cases resulted in amalgamated taxonomic groups containing reads from multiple species. Further, the composition of these groups often varied spatially. Analysis of OTU patterns for ribosomal protein transcripts (selected because they are single-copy housekeeping genes with high coverage in the metatranscriptomes) highlighted this issue, showing that the same reference genome recruited different populations of Actinobacterium

SCGC AAA027-L06 at each station (Fig. 5.3), and different populations of planctomycete *Schlesneria paludicola* and verrucomicrobium *Pedosphaera parvula* even for the two microenvironments at the same station. The exception to this was the five Thaumarchaea genome bins that had consistent population structure across most stations and size fractions.

In cases in which population structure was homogeneous within a reference bin, expression levels of biogeochemically-relevant genes were compared. The addition of internal standards to metagenomic and metatranscriptomic samples allowed direct calculation of transcript abundance per gene copy for each taxon and station, a calculation that is not possible is relative data been collected, an approach that characterizes nearly all meta-omics analyses to this point (that is, metagenomic and metatranscriptomic data is typically analyzed as percent of metagenome and percent of metatranscriptome). With the advantage of standard-enabled absolute gene and transcript counts, it was evident that individual Thaumarchaea taxa varied expression levels of genes for nitrogen and phosphorus acquisition and metabolism by over an order of magnitude at different locations in the river (Fig. 5.4). Patterns in expression ratios were sometimes more consistent within a Thaumarchaeal taxon than between (Fig. 5.5). This was the case for *Candidatus* Nitrosoarchaeum koreensis, which was differentiated by consistently higher *amoA* transcript gene^{-1} ratios regardless of station or microenvironment. Most of the *Candidatus* Nitrosoarchaeum limnia genome bins from different stations and size fractions also clustered together, characterized by high per-gene expression levels of *amoA* and *pstS* (Fig. 5.5). These data provide evidence that co-occurring Thaumarcheota occupy niches that are based in part on the dominance of ammonia oxidation relative to other cellular processes. However, in other cases, patterns in expression ratios were better explained more by location in the river than by taxonomic binning. This was the case for a cluster composed of bins at the Óbidos station with

extremely high ammonia transporter transcript gene^{-1} ratios but consisting of several different thaumarchaeal reference genomes. Despite observations of clustering by taxonomy and location, we did not see any clear clustering associated with the FL versus PA microenvironments. There were, however, differences in the relative importance of FL versus PA cells to the overall gene and transcript inventories at each station. The two Macapá stations had substantially more transcripts and genes in the PA fraction (more than 70% of all genes and more than 80% of transcripts), while the Belém station had substantially more genes and transcripts in the FL fraction. The fact that particle-associated cells can sometimes contribute more to the community genome and transcriptome than free-living cells is different from what has been observed in marine ecosystems (Satinsky *et al.,* 2014a).

To date there remains little information on the composition and activities of the microbial communities in this world's largest river system. In earlier work, a metagenomic approach was used to detail the microbial community structure at an Amazon mainstem site several hundred kilometers upriver from the Óbidos station (Ghai *et al.,* 2011). This previous metagenomic study found a functional gene pool dominated by species related to *Polynucleobacter* sp. QLW-P1DMWA-1 and *Polaromonas* sp. JS666 (Betaproteobacteria), *Acidothermus cellulolyticus* 11B and *Streptomyces* (Actinobacteria), and *Nitrosopumilus maritimus* SCM1 (Thaumarchaeota). In our study of the lower reaches of the Amazon River, Actinobacteria related to Actinobacterium SCGC AAA027-L06 contributed more genes across the five stations than any other genome (~5% of the functional genes at each site). Another highly recruiting bin was the betaproteobacterial genome *Polynucleobacter* sp. QLW-P1DMWA-1, a bin that was also highly recruiting in the upper Amazon (Ghai *et al.,* 2011); this reference genome recruited an average of ~1.5% of the protein-encoding genes and ~1% of the transcripts at each station. Cyanobacteria

were significant contributors to the transcript pools at both TAP (*Synechoccus* sp. CB0101; ~2.5% of genes and ~1.5% of transcripts) and BLM (*Microcystis aeruginosa*; ~0.4% of genes and ~5% of transcripts), but cyanobacteria were not dominant contributors in the upper Amazon. One of the most pronounced signals in the lower reaches was the transcriptional dominance of thaumarchaeal genomic bins, combining to account for >9% of the transcripts at each of the four stations, and up to ~21% of the transcripts at MCPS. However, these thaumarchaeota were less important in the metagenomes, accounting for a maximum of ~3% of genes (at MCPS), a number similar to that found by Ghai et al. (2011) further upriver. Noting that many of the highly-recruiting reference genomes in the lower reaches of the Amazon were not available at the time of the Ghai et al. (2011) study, there were nonetheless consistencies at broad taxonomic levels between the two studies. Actinobacteria, Betaproteobacteria, and Thaumarchaea were significant members of the microbial communities in both the upper and lower reaches of the Amazon. Cyanobacteria, verrucomicrobia, and planctomycetes were more important in the lower mainstem, however.

The tremendous influence of the Amazon River and the microbes living within it continues beyond the river itself. Discharge from the Amazon accounts for 18% of the world's river input to the oceans (Richey *et al.,* 1989, Subramaniam *et al.,* 2008), and the dissolved and particulate nitrogen, phosphate, silica, and iron delivered to the ocean stimulates marine microbial activity and affects both primary productivity and carbon sequestration at a global scale (Subramaniam *et al.,* 2008). A recent inventory of genes and transcripts of Amazon Plume using identical analytical methods and internal standard methodology (Satinsky *et al.,* 2014a, Satinsky *et al.,* 2014b) provides an opportunity to compare expression of elemental cycling genes by bacteria and archaea inhabiting the lower reaches of the river compared to those present

in the oceanic plume. When the relative transcript abundance of 90 biogeochemically-relevant genes are hierarchically-clustered across the river and plume stations, a strong difference related to nitrogen cycling emerged. River transcript pools were strongly dominated by Thaumarchaeota populations generating *amoA*, *nirK* and *amtB*; these transcripts were present but not nearly as abundant in the plume (Fig. 5.6). Denitrification genes, including *norB* (nitric oxide reductase) and *nosZ* (nitrous oxide reductase), were also more important in the river transcriptomes, with only a small denitrification signal at the two low salinity marine stations (Fig. 5.6) (Stations 10 and 23; (Satinsky *et al.,* 2014a)). Nitrogen storage in the form of cyanophycin, binning primarily to heterotrophic bacteria, was highly expressed in river transcriptomes, accounting for up to 4% of the 90 gene transcriptome at Óbidos (Fig. 5.6), but this was not the case in the marine environment; it may be that stronger limitation in marine systems reduces the opportunity to store excess nitrogen. In contrast to the river, nitrogen fixation transcripts were an important feature of the marine transcriptomes, particularly at two stations with populations of diatoms carrying an endosymbiotic N-fixing cyanobacterium (Hilton *et al.,* 2014, Satinsky *et al.,* 2014a).

Differences between river and plume transcriptomes also emerged in relation to phosphorus cycling. The alkaline phosphatase gene *phoX* is proportionally higher in marine transcriptomes compared to riverine, as expected from previous work investigating *pho* gene composition in marine metagenomes (Sebastian and Ammerman 2009). Transcripts for the regulatory gene *phoU* were present in very high abundance at the TAP station compared to the other riverine stations and to marine samples, possibly indicating a strong phosphorus limitation at this clear-water river station. Transcripts from the high-affinity phosphate transporter (*pstA,C,S*) were more important at the marine sites and TAP, while the low-affinity phosphate

transporter (*pitA*) was more important in the freshwater transcriptomes and coastal Stations 10 and 3 in the plume.

Patterns observed for transcripts involved in light utilization were reflective of turbidity levels in the river, with the marine stations, TAP and BLM all displaying high relative abundance of transcripts encoding photosystem II subunits (*psbA, psbB,* and *psbC*). Proteorhodopsin transcripts were higher at the marine stations and at TAP, with an actinobacterial proteorhodopsin gene largely responsible for the TAP signal. There was a clear increase in the relative investment in sulfur cycling genes in the plume samples relative to the river, especially with regards to N-acetyltaurine and DMSP metabolism, and there was a bias toward the marine system with regard to expression of vitamin B_6 transcripts.

The extent to which riverine prokaryotic transcriptomes were dominated by ammonia oxidizers was surprising. For four of the five river stations, *amoA*, *nirK*, and *amtB* transcripts accounted for between 1 and 4% of the river transcriptomes except at TAP; these are high values considering they represent just three genes out of many thousands that were present. Heterotrophic processing of terrestrially-derived carbon is typically considered the major biogeochemical role of prokaryotes in the Amazon and other large river systems (Battin *et al.,* 2009, Benner *et al.,* 1995). Although transcript numbers cannot be equated with a rates of transformation for a number of reasons (Moran *et al.,* 2013), the numbers of ammonia oxidation (*amoA*) transcripts in the lower Amazon are quite high compared to surface marine waters (~1 x 10^9 transcripts L^{-1} at the lower Amazon stations compared to ~5 x 10^7 transcripts L^{-1} at the plume station closest to the river mouth; (Satinsky *et al.,* 2014a). Equally surprising was that despite higher concentrations of aromatic carbon compounds derived from vascular plant material in the river (Ward *et al.,* 2013), expression ratios of genes for degradation of aromatic compounds

(*pcaH* and vanA)*,* tannin (*tanA)*, and cellulose (cellobiase *bglA*) were lower in the river, with average transcript gene^{-1} ratios 18-155 times lower compared to the plume. However, poor understanding of the enzymes mediating lignin degradation in aquatic systems, combined with the heterogeneous nature of lignin molecules, the likelihood that multiple enzymes are involved in its oxidation, and a possible role for photochemical oxidation (Ward *et al.,* 2013), may have influenced these results. Overall, metagenomic and metatranscriptomic analyses of the lower Amazon River provide new insights into the microbial activities of the world's largest river system, and suggest important questions regarding the roles of chemoautotrophic archaea and heterotrophic bacterioplankton in river biogeochemistry.

References

Battin TJ, Kaplan LA, Findlay S, Hopkinson CS, Marti E, Packman AI *et al.* (2009). Biophysical controls on organic carbon fluxes in fluvial networks. *Nat Geosci* **2:** 595-595.

Benner R, Opsahl S, ChinLeo G, Richey JE, Forsberg BR (1995). Bacterial carbon metabolism in the Amazon River system. *Limnol Oceanogr* **40:** 1262-1270.

Coles VJ, Brooks MT, Hopkins J, Stukel MR, Yager PL, Hood RR (2013). The pathways and properties of the Amazon River Plume in the tropical North Atlantic Ocean. *J Geophys Res-Oceans* **118:** 6894-6913.

Debroas D, Humbert JF, Enault F, Bronner G, Faubladier M, Cornillot E (2009). Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget--France). *Environ Microbiol* **11:** 2412-2424.

Ertel JR, Hedges JI, Devol AH, Richey JE, Ribeiro MDG (1986). Dissolved Humic Substances of the Amazon River System. *Limnol Oceanogr* **31:** 739-754.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281:** 237-240.

Ghai R, Rodriguez-Valera F, McMahon KD, Toyama D, Rinke R, Cristina Souza de Oliveira T *et al.* (2011). Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS One* **6:** e23785.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* **5:** 461-472.

Hilton JA, Satinsky BM, Doherty M, Zielinski BL, Zehr JP (2014). Metatranscriptomics of N_2-fixing cyanobacteria in the Amazon River plume. *ISME J* **In Revision**.

Malhi Y, Roberts JT, Betts RA, Killeen TJ, Li W, Nobre CA (2008). Climate change, deforestation, and the fate of the Amazon. *Science* **319:** 169-172.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC bioinformatics* **13:** 31.

Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan LK *et al.* (2013). Sizing up metatranscriptomics. *ISME J* **7:** 237-243.

Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* **75:** 14-49.

Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R *et al.* (2011). Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* **77:** 6000-6011.

Peixoto JC, Leomil L, Souza JV, Peixoto FB, Astolfi-Filho S (2011). Comparison of bacterial communities in the Solimoes and Negro River tributaries of the Amazon River based on small subunit rRNA gene sequences. *Genet Mol Res* **10:** 3783-3793.

Pope PB, Patel BK (2008). Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *FEMS Microbiol Ecol* **64:** 9-27.

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.

Richey JE, Nobre C, Deser C (1989). Amazon river discharge and climate variability: 1903 to 1985. *Science* **246:** 101-103.

Richey JE, Melack JM, Aufdenkampe AK, Ballester VM, Hess LL (2002). Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO2. *Nature* **416:** 617-620.

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 Biol* **5:** e77.

Satinsky BM, Gifford SM, Crump BC, Moran MA (2013). Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. In: DeLong EF (ed). *Methods Enzymol*. Academic Press. pp 237-250.

Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski BL, Doherty M *et al.* (2014a). Microspatial gene expression patterns in the Amazon River Plume. *Proc Natl Acad Sci U S A* **111:** 11085-11090.

Satinsky BM, Zielinski BL, Doherty M, Smith CB, Sharma S, Paul JH *et al.* (2014b). The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. *Microbiome* **2:** 17.

Sebastian M, Ammerman JW (2009). The alkaline phosphatase PhoX is more widely distributed in marine bacteria than the classical PhoA. *ISME J* **3:** 563-572.

Subramaniam A, Yager PL, Carpenter EJ, Mahaffey C, Bjorkman K, Cooley S *et al.* (2008). Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci U S A* **105:** 10460-10465.

Tsementzi D, Poretsky R, Rodriguez-R LM, Luo C, Konstantinidis KT (2014). Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environ Microbiol Rep*.

Vila-Costa M, Sharma S, Moran MA, Casamayor EO (2013). Diel gene expression profiles of a phosphorus limited mountain lake using metatranscriptomics. *Environ Microbiol* **15:** 1190-1203.

Ward ND, Keil RG, Medeiros PM, Brito DC, Cunha AC, Dittmar T *et al.* (2013). Degradation of terrestrially derived macromolecules in the Amazon River. *Nat Geosci* **6:** 530-533.

Whitman WB, Coleman DC, Wiebe WJ (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95:** 6578-6583.

Zhao Y, Tang H, Ye Y (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28:** 125-126.

Table 5.1. Metagenome and metatranscriptome data summaries for 5 Amazon River Stations in May 2011. Duplicate samples were collected for each of two size fractions for each data type at all six stations. Per liter calculations are based on recovery of internal standards (Table 5.S1).

	Óbidos	Tapajós	Macapá north	Macapá south	Belém
Location (Lat, Long)	1° 55.141'S, 55° 31.543'W	2° 29.063'S, 55° 0.450'W	0° 5.033'S, 51° 3.085'W	0° 9.415'S, 50° 37.353'W	1° 31.162'S, 48° 55.077'W
Depth (m)	33	15	14	10	19
Raw reads					
Metagenomic	1.76×10^{8}	1.06×10^{8}	1.39×10^{8}	1.95×10^{8}	8.97×10^{7}
Metatranscriptomic	2.89×10^{8}	2.65×10^{8}	2.75×10^{8}	2.94×10^{8}	3.32×10^{8}
Joined reads post QC					
Metagenomic	7.17×10^{7}	3.45×10^{7}	5.45×10^{7}	8.04×10^{7}	3.70×10^{7}
Metatranscriptomic	1.15×10^{8}	7.48×10^{7}	1.09×10^{8}	8.98×10^{7}	1.22×10^{8}
Mean read length (bp)	206	216	216	200	220
Protein Encoding Reads					
Metagenomic	2.85×10^{7}	1.05×10^{7}	1.40×10^{7}	2.77×10^{7}	1.37×10^{7}
Metatranscriptomic	1.96×10^{7}	1.29×10^{7}	2.30×10^{7}	1.47×10^{7}	3.27×10^{7}
Prokaryotic Gene Copies L^{-1}	1.04×10^{13}	1.03×10^{13}	9.79×10^{12}	1.04×10^{13}	1.01×10^{13}
% FL	24.19%	65.93%	19.38%	16.48%	61.09%
% PA	75.81%	34.07%	80.62%	83.52%	38.91%
Prokaryotic Transcripts L^{-1}	5.90×10^{10}	8.74×10^{10}	1.42×10^{11}	8.15×10^{10}	1.64×10^{11}
% FL	45.94%	72.77%	7.61%	33.90%	34.26%
% PA	54.07%	27.23%	92.39%	66.10%	65.74%

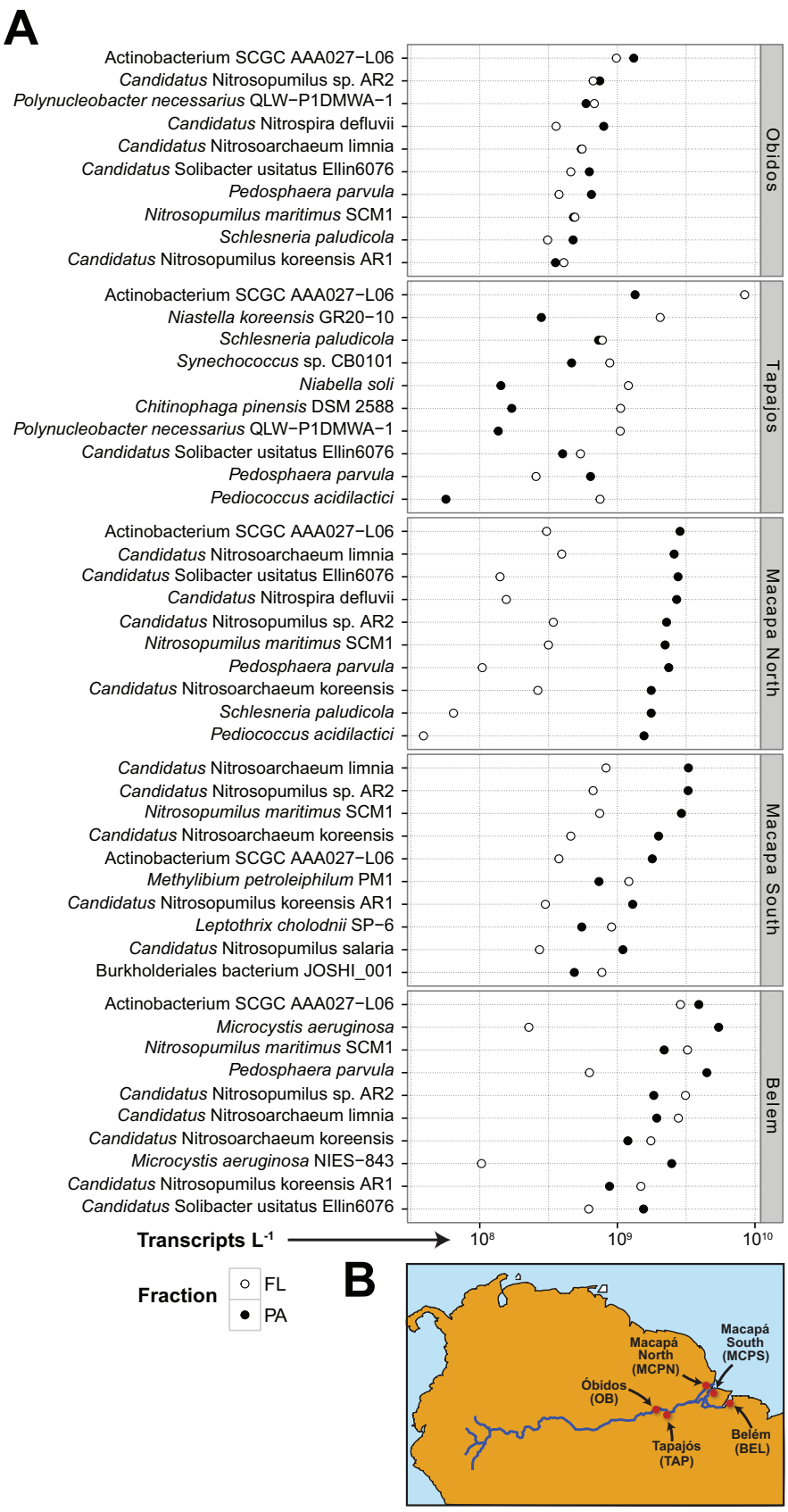
Table 5.S1. May 2011 Amazon River metagenome and metatranscriptome dataset descriptions.

Sample	Site	Sample Type	Size Fraction	Collection Time	Volume Filtered (L)	Paired Reads	Avg. Read Length (nt)	Normalization Factor
RCM5	TAP	RNA	2.0 - 297 µm	5/20/11 9:30	1.72	18,049,826	175.7	39,960
RCM6	TAP	RNA	0.2 - 2.0 µm	5/20/11 9:30	1.72	12,415,459	201.9	15,541
RCM7	TAP	RNA	2.0 - 297 µm	5/20/11 9:30	1.60	24,102,523	182.1	42,446
RCM8	TAP	RNA	2.0 - 297 µm	5/20/11 9:30	1.60	20,228,483	224.1	10,146
RCM9	BLM	RNA	2.0 - 297 µm	5/15/11 15:00	1.80	21,675,520	224.6	13,195
RCM10	BLM	RNA	2.0 - 297 µm	5/15/11 15:00	1.80	19,836,121	247.3	37,938
RCM11	BLM	RNA	0.2 - 2.0 µm	5/15/11 15:00	1.34	43,365,572	240.6	5,418
RCM12	BLM	RNA	2.0 - 297 µm	5/15/11 15:00	1.34	37,075,161	194.8	19,783
RCM13	OB	RNA	2.0 - 297 µm	5/19/11 13:00	1.06	40,158,920	216	5,713
RCM14	OB	RNA	0.2 - 2.0 µm	5/19/11 13:00	1.06	27,643,448	209.6	4,217
RCM15	OB	RNA	2.0 - 297 µm	5/19/11 13:00	0.73	27,060,530	192.8	5,000
RCM16	OB	RNA	0.2 - 2.0 µm	5/19/11 13:00	0.73	20,228,361	213	17,109
RCM17	MCPS	RNA	2.0 - 297 µm	5/8/11 8:45	0.70	13,740,236	181.5	21,777
RCM18	MCPS	RNA	0.2 - 2.0 µm	5/8/11 8:45	0.70	24,487,314	168.3	43,832
RCM19	MCPS	RNA	2.0 - 297 µm	5/8/11 8:45	0.70	25,766,486	182.5	14,073
RCM20	MCPS	RNA	0.2 - 2.0 µm	5/8/11 8:45	0.70	25,763,156	207.2	23,472
RCM21	MCPN	RNA	0.2 - 2.0 µm	5/7/11 9:00	0.65	26,806,801	219.2	10,521
RCM22	MCPN	RNA	2.0 - 297 µm	5/7/11 9:00	0.65	27,147,770	222.1	15,029
RCM23	MCPN	RNA	0.2 - 2.0 µm	5/7/11 9:00	0.53	24,034,777	209.9	3,667
RCM24	MCPN	RNA	2.0 - 297 µm	5/7/11 9:00	0.53	31,083,593	217.9	114,936
RCM25	MCPN	DNA	2.0 - 297 µm	5/7/11 9:00	0.46	15,003,791	222	158,034
RCM26	MCPN	DNA	2.0 - 297 µm	5/7/11 9:00	0.66	24,431,236	198.1	758,599
RCM27	MCPN	DNA	2.0 - 297 µm	5/7/11 9:00	0.50	7,070,719	220.8	934,437
RCM28	MCPN	DNA	0.2 - 2.0 µm	5/7/11 9:00	0.50	8,037,703	218.1	2,382,985
RCM29	MCPS	DNA	2.0 - 297 µm	5/8/11 8:45	0.67	20,632,031	237.2	107,128
RCM30	MCPS	DNA	0.2 - 2.0 µm	5/8/11 8:45	0.60	10,753,118	217.4	2,155,856
RCM31	MCPS	DNA	2.0 - 297 µm	5/8/11 8:45	0.81	35,211,964	197.5	179,176
RCM32	MCPS	DNA	0.2 - 2.0 µm	5/8/11 8:45	0.81	13,756,496	211.5	1,343,560
RCM33	BLM	DNA	0.2 - 2.0 µm	5/15/11 15:00	1.31	8,962,213	200.7	2,546,761
RCM34	BLM	DNA	2.0 - 297 µm	5/15/11 15:00	1.31	14,947,161	199.4	546,133
RCM35	BLM	DNA	0.2 - 2.0 µm	5/15/11 15:00	1.30	10,160,926	226.7	1,215,522
RCM36	BLM	DNA	2.0 - 297 µm	5/15/11 15:00	1.30	2,955,867	225.2	5,706,669
RCM37	OB	DNA	0.2 - 2.0 µm	5/19/11 13:00	1.03	30,234,753	197.8	170,520
RCM38	OB	DNA	2.0 - 297 µm	5/19/11 13:00	1.03	20,954,148	201.2	903,094
RCM39	OB	DNA	0.2 - 2.0 µm	5/19/11 13:00	0.61	10,918,268	195.4	629,099
RCM40	OB	DNA	2.0 - 297 µm	5/19/11 13:00	0.61	9,639,824	219.9	2,183,485
RCM41	TAP	DNA	2.0 - 297 µm	5/20/11 9:30	1.40	8,422,780	229.8	2,393,401
RCM42	TAP	DNA	2.0 - 297 µm	5/20/11 9:30	0.73	11,357,451	218.6	943,179
RCM43	TAP	DNA	2.0 - 297 µm	5/20/11 9:30	1.57	8,248,989	244.6	1,842,684
RCM44	TAP	DNA	2.0 - 297 µm	5/20/11 9:30	1.57	6,477,633	252.7	1,379,903

Table 5.S2. Metadata for 5 Amazon River plume stations in May 2011.

	Óbidos	Tapajós	Macapá north	Macapá south	Belém
Temp. (C°)	28.4	29.5	29.0	28.8	29.3
Depth (m)	33	15	14	10	19
Conductivity (µS /cm)	56.40	16.80	61.50	55.30	37.10
Bacterial Abundance (counts/ml)	3.76E06 ± 0.64%	3.66E06 ± 0.88%	3.96E06 ± 3.57%	3.94E06 ± 3.07%	3.64E+06
Dissolved Inorganic Carbon (µmol C / kg)	551 ± 4	156	507 ± 0.5	459 ± 2.7	308 ± 0.4
Chlorophyll A (µg/L)	3.23 ± 0.23	3.99 ± 0.10	1.67 ± 0.08	1.99 ± 0.31	2.03 ± 0.88

Figure 5.1 (a) Transcript inventories (transcripts L^{-1}) for the ten most transcriptionally abundant prokaryotic taxonomic bins at each river station. Bubbles are colored by microenvironment (free-living, white; particle-associated, black). (b) Map of the lower Amazon River showing the locations of the stations sampled in May 2011.


A
Actinobacterium SCGC AAA027−L06
Candidatus Nitrosopumilus sp. AR2
Polynucleobacter necessarius QLW−P1DMWA−1
Candidatus Nitrospira defluvii
Candidatus Nitrosoarchaeum limnia
Candidatus Solibacter usitatus Ellin6076
Pedosphaera parvula
Nitrosopumilus maritimus SCM1
Schlesneria paludicola
Candidatus Nitrosopumilus koreensis AR1
Obidos
Actinobacterium SCGC AAA027−L06
Niastella koreensis GR20−10
Schlesneria paludicola
Synechococcus sp. CB0101
Niabella soli
Chitinophaga pinensis DSM 2588
Polynucleobacter necessarius QLW−P1DMWA−1
Candidatus Solibacter usitatus Ellin6076
Pedosphaera parvula
Pediococcus acidilactici
Tapajos
Actinobacterium SCGC AAA027−L06
Candidatus Nitrosoarchaeum limnia
Candidatus Solibacter usitatus Ellin6076
Candidatus Nitrospira defluvii
Candidatus Nitrosopumilus sp. AR2
Nitrosopumilus maritimus SCM1
Pedosphaera parvula
Candidatus Nitrosoarchaeum koreensis
Schlesneria paludicola
Pediococcus acidilactici
Macapa North
Candidatus Nitrosoarchaeum limnia
Candidatus Nitrosopumilus sp. AR2
Nitrosopumilus maritimus SCM1
Candidatus Nitrosoarchaeum koreensis
Actinobacterium SCGC AAA027−L06
Methylibium petroleiphilum PM1
Candidatus Nitrosopumilus koreensis AR1
Leptothrix cholodnii SP−6
Candidatus Nitrosopumilus salaria
Burkholderiales bacterium JOSHI_001
Macapa South
Actinobacterium SCGC AAA027−L06
Microcystis aeruginosa
Nitrosopumilus maritimus SCM1
Pedosphaera parvula
Candidatus Nitrosopumilus sp. AR2
Candidatus Nitrosoarchaeum limnia
Candidatus Nitrosoarchaeum koreensis
Microcystis aeruginosa NIES−843
Candidatus Nitrosopumilus koreensis AR1
Candidatus Solibacter usitatus Ellin6076
Belem
Transcripts L-1
10^8
10^9
10^10
Fraction
FL
PA
B
Macapá North (MCPN)
Macapá South (MCPS)
Óbidos (OB)
Tapajós (TAP)
Belém (BEL)

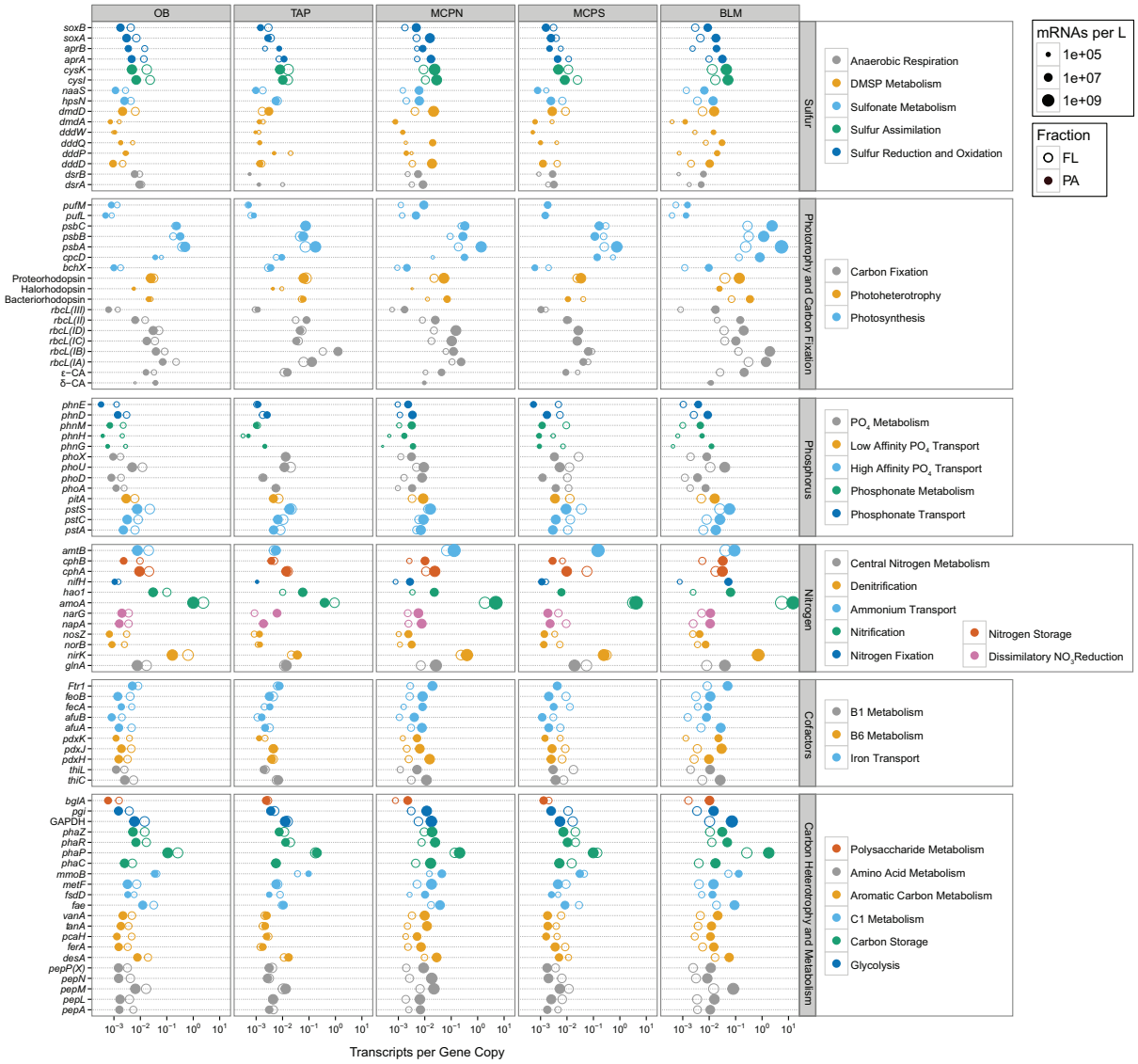


Figure 5.2 Expression ratios (transcripts gene copy^{-1}) of 90 biogeochemically-relevant genes across all five river stations. Genes are grouped according to broad functional category, and colored based on functional subcategories. Circle size is representative of transcript abundance (transcripts L^{-1}). Circles are filled according to microenvironment (free-living, no fill; particle-associated, fill).

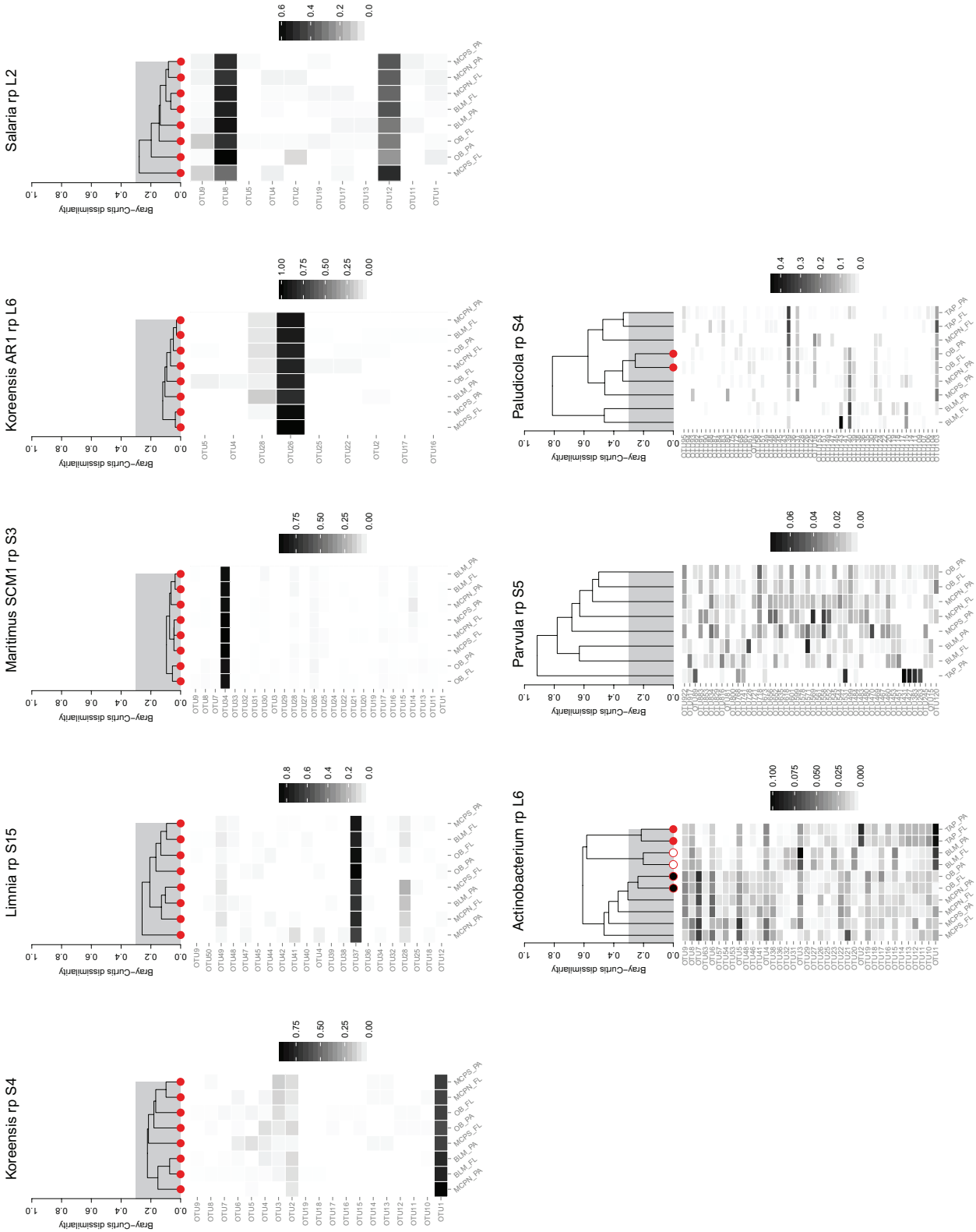


Figure 5.3 OTU fingerprints and hierarchical clustering of an individual ribosomal proteins for each of 8 reference genome bins. Individual stations and microenvironment identified as containing the same population structure are marked with circles.

Figure 5.S1 OTU fingerprints and hierarchical clustering of additional ribosomal proteins for each of five archaeal reference genome bins.

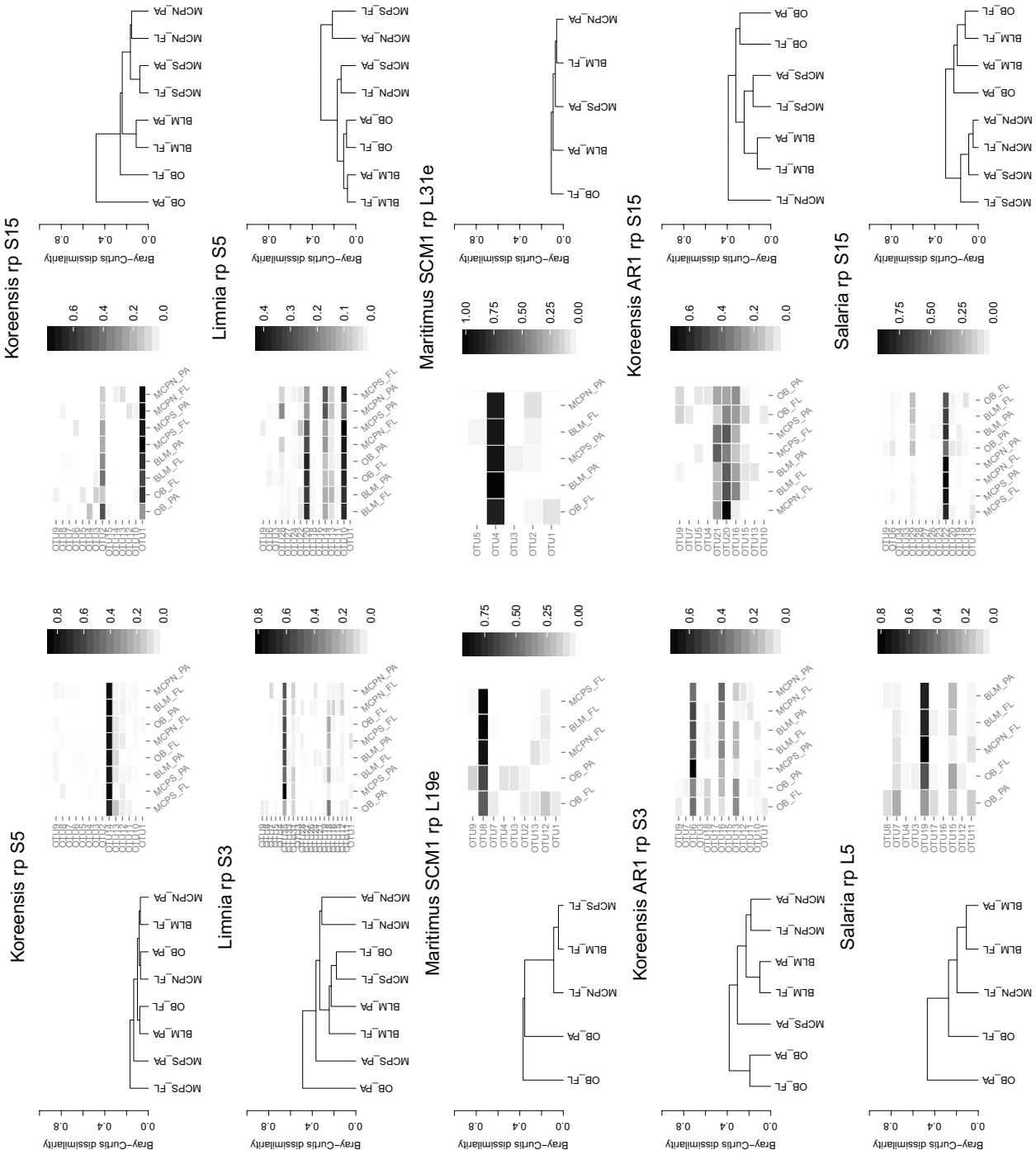
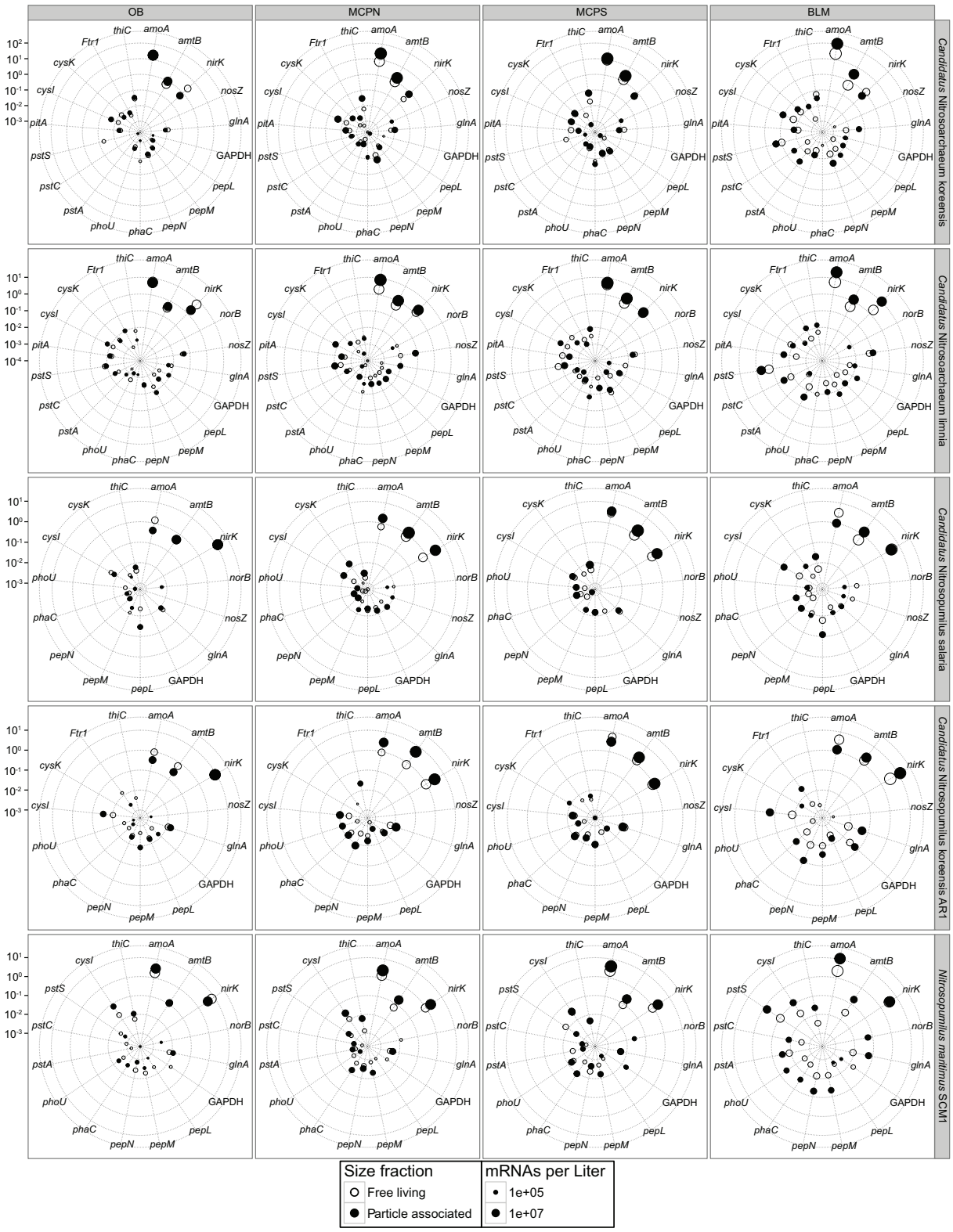

Koreensis rp S5
Limnia rp S3
Maritimus SCM1 rp L19e
Koreensis AR1 rp S3
Salaria rp L5
Koreensis rp S15
Limnia rp S5
Maritimus SCM1 rp L31e
Koreensis AR1 rp S15
Salaria rp S15
Bray−Curtis dissimilarity

Figure 5.4 Expression ratios (transcripts gene copy^{-1}) of biogeochemically-relevant genes in five individual archaeal reference genome bins (*Candidatus* Nitrosoarchaeum koreensis, *Candidatus* Nitrosoarchaeum limnia, *Candidatus* Nitrosopumilus salaria, *Candidatus* Nitrosopumilus koreensis AR1, and *Nitrosopumilus maritimus* SCM1) across river stations containing the same population structure within a bin. Circle size is representative of transcript abundance (transcripts L^{-1}). Circles are filled according to microenvironment (free-living, no fill; particle-associated, fill).


OB
MCPN
MCPS
BLM
Candidatus Nitrosoarchaeum koreensis
Candidatus Nitrosoarchaeum limnia
Candidatus Nitrosopumilus salaria
Candidatus Nitrosopumilus koreensis AR1
Nitrosopumilus maritimus SCM1
thiC
amoA
amtB
nirK
nosZ
glnA
GAPDH
pepL
pepM
pepN
phaC
phoU
pstA
pstC
pstS
pitA
cysI
cysK
Ftr1
norB
Size fraction
Free living
Particle associated
mRNAs per Liter
1e+05
1e+07

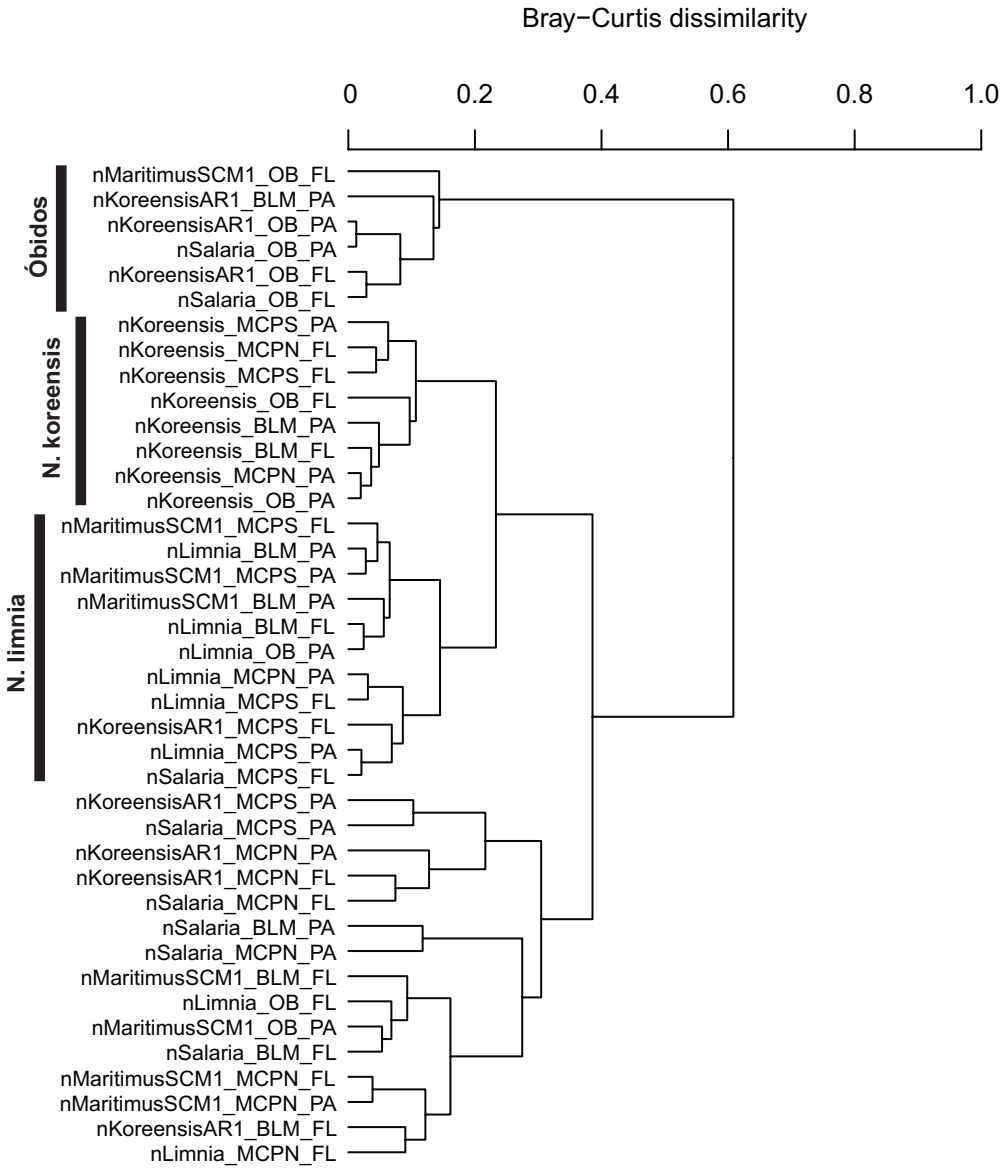
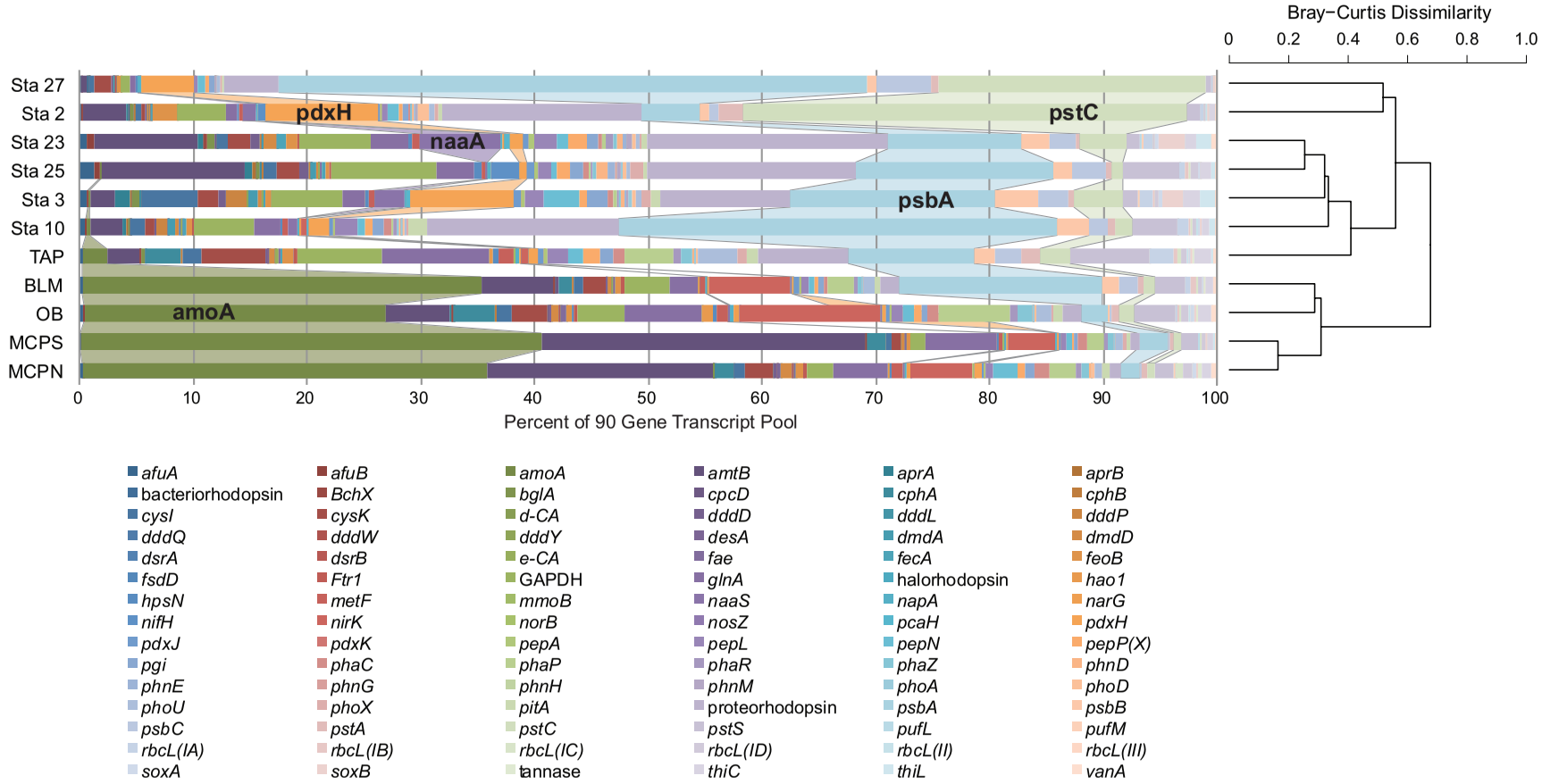


Figure 5.5 Hierarchical clustering of five Thaumarchaeota reference genome bins for each river station and microenvironment containing the same population structure within a bin (see Fig. 2) based upon the relative expression ratio patterns of 20 elemental cycling genes. Samples codes indicate the reference bin, station name, and microenvironment (FL or PA). The *Candidatus* N. koreensis cluster, *Candidatus* N. limina cluster, and Óbidos cluster mentioned in the text are indicated.

Figure 5.6 Hierarchical clustering of relative abundance patterns of 90 biogeochemically-relevant genes in the metatranscriptomes from five Amazon River stations and six Amazon plume stations.


Bray−Curtis Dissimilarity
0
0.2
0.4
0.6
0.8
1.0
Sta 27
Sta 2
Sta 23
Sta 25
Sta 3
Sta 10
TAP
BLM
OB
MCPS
MCPN
pdxH
naaA
pstC
psbA
amoA
0
10
20
30
40
50
60
70
80
90
100
Percent of 90 Gene Transcript Pool
afuA
afuB
amoA
amtB
aprA
aprB
bacteriorhodopsin
BchX
bglA
cpcD
cphA
cphB
cysI
cysK
d-CA
dddD
dddL
dddP
dddQ
dddW
dddY
desA
dmdA
dmdD
dsrA
dsrB
e-CA
fae
fecA
feoB
fsdD
Ftr1
GAPDH
glnA
halorhodopsin
hao1
hpsN
metF
mmoB
naaS
napA
narG
nifH
nirK
norB
nosZ
pcaH
pdxH
pdxJ
pdxK
pepA
pepL
pepN
pepP(X)
pgi
phaC
phaP
phaR
phaZ
phnD
phnE
phnG
phnH
phnM
phoA
phoD
phoU
phoX
pitA
proteorhodopsin
psbA
psbB
psbC
pstA
pstC
pstS
pufL
pufM
rbcL(IA)
rbcL(IB)
rbcL(IC)
rbcL(ID)
rbcL(II)
rbcL(III)
soxA
soxB
tannase
thiC
thiL
vanA

CHAPTER 6

SUMMARY

Next generation sequencing-enabled metatranscriptomic and metagenomic methodologies were used in tandem to elucidate the functional diversity and expression patterns of microbial communities spatially within the Amazon River and Amazon Plume. The first chapter described the use of internal DNA and mRNA standards in meta–omics datasets, and highlighted how data collected in an absolute framework (per L or per cell) provides increased comparative power and insight into underlying causes of expression differences between samples. The remaining three chapters put this methodology into practice, using meta-omics datasets benchmarked with internal standards to detail patterns of gene expression across microspatial environments in the Amazon River and plume, and subsequently address the implications of those patterns ecologically.

In Chapter 2 developments in methodology for the use of internal standards in both metatranscriptomics and metagenomics were described. The benefits of using internal standards in meta-omics studies were demonstrated through two different examples highlighting how studies conducted in an absolute framework provide insights into spatial and temporal variations in gene regulation by bacteria in their natural environment. The chapter provided a detailed, step-by-step protocol for the synthesis and subsequent implementation of internal standards in metagenomics and metatranscriptomics, highlighting important considerations and alternatives in internal standard design, and recognized limitations as well.

In Chapter 3 the first fully quantitative inventory of genes and transcripts was detailed and the expression patterns of genes mediating elemental cycling at the microspatial scale at a low-salinity site within the ocean's largest river plume were described. It was determined that each liter of plume seawater contains 1 trillion genes and 50 billion transcripts from thousands of different bacterial, archaeal, and eukaryotic taxa. These inventories revealed that genomes from free-living cells were more abundant than those from particle-associated cells, and that more transcripts per liter for carbon fixation, heterotrophy, nitrogen and phosphorus uptake, and iron acquisition were generated in the free-living microhabitat, while genomes from particle-associated cells contributed more transcripts for sulfur cycling, aromatic compound degradation, and the synthesis of biologically essential vitamins. Contributions by individual taxa to the community transcriptome were shown to be determined not just by the abundance of each taxon's genome but also by the regulation of its genes, and quantitatively, gene regulation differences were more important than genome abundance differences in explaining microenvironment transcriptome differences. Taxa contributing genomes to both free-living and particle-associated communities had up to 65% of their expressed genes regulated differently between the two, and on average, particle-associated cells had twice as many transcripts as free-living cells. Expression of biogeochemically-relevant genes varied by up to 6-fold between microenvironments in response to patchiness in carbon, nutrients, and light at the micron scale. This patchiness within a single water mass signaled the partitioning of functions driving elemental cycling, and highlighted that the scale of spatial heterogeneity relevant to ecological processes is at the micron level in this ecosystem.

In Chapter 4 quantitative metatranscriptomics and metagenomics datasets generated from for free-living and particle-associated microbial communities at six stations were used to detail

transcriptional patterns of prokaryotic genes spatially through the Amazon River Plume. Transcript pools with average of 3.3 x 10^{11} transcripts L^{-1} were dominated by taxa related to Gammaproteobacteria, Flavobacteria, and the cyanobacterium *Synechococcus* at the lower salinity stations; Verrucomicrobia and the endosymbiotic diazotroph *Richelia intracellularis* dominated in the mesohaline regions; and *Prochlorococcus* and *Alteromonas* at the higher salinity oceanic station. Patterns in gene expression provided insights into the roles of bacteria and archaea along the plume, pointing to the environmental conditions most relevant to microbial cells at each station and microenvironment. Particle-associated prokaryotes had higher expression levels (transcripts $gene^{-1}$) for most genes mediating carbon, nitrogen, and phosphorus transport and metabolism when compared to prokaryotes in the free-living microenvironment. However, expression ratios of key biogeochemical genes in particle-associated cells were highly variable while those in free-living cells remain relatively constant as the plume waters mixed into the western tropical North Atlantic. Expression ratios of nitrogen fixation and phosphorus acquisition genes were shown to be the most spatially variable across the plume, and expression levels of nitrogen fixation, phosphorus acquisition, and heterotrophic carbon processing genes all peaked for particle-associated cells at the mesohaline stations,. Carbon storage was the most dynamic activity in the free-living microenvironment, peaking closest to the river mouth. In Chapter 4 it was also demonstrated that changes in both gene regulation within a taxon and shifts in taxonomy among stations drove the high variability in expression observed in particle-associated cells, suggesting that particle-associated bacteria and archaea experience more chemically dynamic conditions spatially in the Amazon Plume than do free-living cells.

In Chapter 5, meta-omics datasets benchmarked with internal standards was used to generate the first quantitative transcript inventory of freshwater prokaryotes, including both free-

living and particle-associated microenvironments at five sites along the lower Amazon River. Prokaryotic transcript pools dominated by in the lower Amazon River were harbored more than 1 x 10^{11} transcripts L^{-1} dominated by taxa related to Actinobacteria, Thaumarchaea, Betaproteobacteria, as well as cyanobacteria related to *Synechococcus* and *Microcystis* at stations with lower turbidity. Patterns in transcript abundance and expression ratios for genes mediating nutrient cycling varied along the river as well as between microenvironments, and led to the surprising discovery that riverine prokaryotic transcriptomes were frequently dominated by transcripts from ammonia oxidizers. Largely due to the influence of actively regulating Thaumarchaeal populations, transcripts from nitrogen related genes *amoA*, *nirK*, and *amtB* accounted for between 1 and 4% of the river transcriptomes at each site, with the exception of the clearwater station at Tapajós. Though heterotrophic processing of terrestrially-derived carbon is considered the major biogeochemical role of prokaryotes in the Amazon river, expression levels for genes diagnostic of aromatic carbon metabolism were often lower than those seen in the plume. The patterns of prokaryotic gene expression discussed in Chapter 5 suggest important roles for chemoautotrophic archaea and heterotrophic bacterioplankton in river biogeochemistry.

APPENDIX A

THE AMAZON CONTINUUM DATASET: QUANTITATIVE METAGENOMIC AND METATRANSCRIPTOMIC INVENTORIES OF THE AMAZON RIVER PLUME, JUNE 2010[1]

[1] Satinsky BM, Zielinski BL, Doherty M, Smith CB, Sharma S, Paul JH *et al.* (2014b). The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. *Microbiome* **2:** 17. Reprinted here with permission of publisher.

Abstract

Background: The Amazon River is by far the world's largest in terms of volume and area, generating a fluvial export that accounts for about a fifth of riverine input into the world's oceans. Marine microbial communities of the Western Tropical North Atlantic Ocean are strongly affected by the terrestrial materials carried by the Amazon plume, including dissolved (DOC) and particulate organic carbon (POC) and inorganic nutrients, with impacts on primary productivity and carbon sequestration.

Results: We inventoried genes and transcripts at six stations in the Amazon River plume during June 2010. At each station, internal standard-spiked metagenomes, non-selective metatranscriptomes, and poly(A)-selective metatranscriptomes were obtained in duplicate for two discrete size fractions (0.2 to 2.0 mm and 2.0 – 156 mm) using 150 x 150 paired-end Illumina sequencing. Following quality control, the data set contained 360 million reads of ~200 bp average size from Bacteria, Archaea, Eukarya, and viruses. Bacterial metagenomes and metatranscriptomes were dominated by *Synechococcus, Prochlorococcus*, SAR11, SAR116, and SAR86, with high contributions from SAR324 and Verrucomicrobia at some stations. Diatoms, green picophytoplankton, dinoflagellates, haptophytes, and copepods dominated the eukaryotic genes and transcripts. Gene expression ratios differed by station, size fraction, and microbial group, with transcription levels varying over three orders of magnitude across taxa and environments.

Conclusions: This first comprehensive inventory of microbial genes and transcripts, benchmarked with internal standards for full quantitation, is generating novel

insights into biogeochemical processes of the Amazon plume and improving prediction of climate change impacts on the marine biosphere.

Background:

The Amazon River runs nearly 6,500 km across the South American continent before emptying into the Western Tropical North Atlantic Ocean; in terms of both volume and watershed area it is the world's largest riverine system (Coles *et al.,* 2013). The river carries a significant load of terrestrially-derived nutrients to the ocean, and this has global consequences to marine primary productivity and carbon sequestration (Richey *et al.,* 1989, Subramaniam *et al.,* 2008). Productive phytoplankton blooms harboring cyanobacteria, coastal diatom species, and oceanic diatoms with endosymbiotic diazotrophs take advantage of the riverine nutrient supplements and enhance carbon export from the upper ocean to deeper waters via sinking particles (Goes *et al.,* 2014, Subramaniam *et al.,* 2008). Heterotrophic bacteria also remineralize organic nutrients in the plume, further fueling primary production and increasing the flux of organic material to deep water.

We inventoried the microbial genes and transcripts at six stations in the Amazon River plume aboard the R/V *Knorr* between May 22 and June 25, 2010 (Fig. A.1) using Illumina sequencing with 150 x 150 bp overlapping paired-end reads. Metagenomic and metatranscriptomic data have typically been analyzed within a relative framework (i.e., % of metagenome and % of metatranscriptome), but this approach is problematic for dynamic communities because a change in the abundance of one type of gene or transcript imposes a change in the percent contribution of the others. By incorporating

internal standards, we are able to assess meta-omics datasets within an absolute framework that facilitates comparisons of communities sampled at different times and places in the environment. In the Amazon plume sequence libraries, known copy numbers of internal standards were added at the initiation of sample processing and consisted of genomic DNA from an exotic bacterium for the metagenomes (*Thermus thermophilus* HB8) and artificial mRNAs and poly(A)-tailed mRNAs for the metatranscriptomes; these standards were identified, counted, and removed from the natural sequences during quality control steps.

For each station, metagenomes and non-selective metatranscriptomes were each obtained in duplicate for two discrete size fractions (0.2 to 2.0 mm and 2.0 to 156 mm), while poly(A)-selective metatranscriptomes were obtained in duplicate only for the 2.0 to 156 mm size fraction (to increase coverage of the eukaryotic community), resulting in a total of 60 datasets (6 stations x 5 data types x 2 replicates). The data collection consisted of 360 million reads following quality control (removal of poor quality reads, removal of rRNAs from metatranscriptomes, removal of internal standards, and joining of overlapping 150 bp paired ends) and provides an unprecedented view of the metabolic functions of the Bacteria, Archaea, and Eukarya mediating carbon and nutrient cycling in the Amazon River plume.

Methods

Detailed sample collection and processing methodology can be found in the supplemental methods. Sample sites in the Amazon River plume were chosen to represent a range of salinity, nutrient concentrations, and microbial communities. Microbial cells were collected by filtration and preserved in RNAlater (Applied

Biosystems, Austin, TX). During sample processing, internal standards were added to each sample prior to cell lysis. Samples collected for non-selective metatranscriptomics were processed by extracting total RNA, removing residual DNA, depleting rRNA, linearly amplifying the remaining transcripts, and making double-stranded cDNA for library preparation and sequencing. Poly(A)-selective metatranscriptome samples were processed similarly except that poly(A)-tailed mRNAs were selectively isolated, eliminating the need for rRNA depletion steps. Metagenomic samples were processed by extracting DNA and removing residual proteins and RNA. Following sample processing, cDNA or DNA was sheared and libraries were constructed for paired-end sequencing (150 x 150) using either the Illumina Genome Analyzer IIx, HiSeq 2000, MiSeq, or HiSeq 2500 platform.

From 60 samples, we obtained 8.21×10^8 raw sequences containing 1.23×10^{11} nt. Following sequence quality control, 3.59×10^8 reads with a mean length of 195 bp were obtained. Internal standards were quantified and removed, along with any remaining rRNA sequences. Remaining reads were annotated against the RefSeq Protein database or a custom marine database using RAPSearch2 (Zhao *et al.,* 2012), and abundance per liter was calculated based on internal standard recovery (Satinsky *et al.,* 2013).

Biological and chemical data measured concurrently with sample collection provides environmental context for sequence data. These metadata include temperature, salinity, oxygen concentration, irradiance, chlorophyll concentration, nutrient concentrations, and bacterial abundance and production. Datasets describing the phytoplankton communities and other features of the June 2010 plume ecosystem have

been previously published (Barada *et al.,* 2013, Chong *et al.,* 2014, Coles *et al.,* 2013, Goes *et al.,* 2014).

Quality Assurance

The She-ra program (Rodrigue *et al.,* 2010) was used to join the paired-end Illumina reads using the default parameters and a quality metric score of 0.5. Seqtrim (Falgueras *et al.,* 2010) was used to trim the joined reads using the default parameters. Ribosomal RNA and internal standard sequences were identified in the metatranscriptomes using a Blastn search against a custom database containing representative rRNA sequences and internal standard sequences; sequences with a bit score ≥ 50 were identified as either rRNA or internal standards and removed from the datasets. Internal standards were identified in metagenomes by first performing a Blastn search (bit score cutoff ≥ 50) against the *T. thermophilus* HB8 genome. Hits were subsequently queried against the Refseq protein database using Blastx (bit score cutoff ≥ 40) to identify and quantify all *T. thermophilus* HB8 protein encoding reads, and these reads were removed from the datasets.

Initial Findings

Metagenomic reads from surface waters of the six Amazon River plume stations were assigned to bacterial, archaeal, eukaryotic, and viral taxa based on best hits to reference genomes. Among autotrophic bacteria, *Synechococcus* was the largest contributor to the metagenomes at locations closest to the river mouth (Stations 10, 3; ~1.5 x 10^{12} genes L^{-1}) and was replaced by *Prochlorococcus* at more oceanic locations (Stations 25, 27) (Table A.2). Among heterotrophic bacteria, SAR86 had the largest gene abundance closest to the river mouth (Station 10; ~8.6 x 10^{11} genes L^{-1}). SAR11 clade

members (HTCC7211, HIMB5) were also abundant here, and became the dominant contributor of heterotrophic bacterial genes at more oceanic stations (up to 5.7 x 10^{12} genes L^{-1}) (Table A.2). Genes binning to SAR324 genomes were abundant at three stations (Station 2, 3, and 23; Table A.2), with the Amazon plume sequences aligning with heterotrophic members of this group (Chitsaz *et al.,* 2011). Station 2 had a distinctive bacterial community relative to the other plume stations, dominated by genes from Verrucomicrobia related to *Coraliomargarita akajimensis* DSM 45221 and strain DG1235 and with substantial contributions from SAR116 taxa (IMCC1322, HIMB100). *Coraliomargarita akajimensis* DSM 45221 was also among the most abundant genome bins at Station 25 (Table A.2).

Among eukaryotic taxa, diatoms and the green alga *Micromonas* contributed the greatest number of genes at lower salinities, while Haptophytes (binning to *Phaeocyctis antarctica*), dinoflagellates (binning to *Alexandrium tamarense* CCMP1771) and relatives of the green alga *Pyraminomonas obovata* CCMP722 increased in importance at more saline stations (Table A.2). Among Archaea, members of the ammonia-oxidizing genus *Nitrosopumilus* and related genera contributed the most genes at stations closest to the river mouth, although they were 100-fold lower in numbers compared to the most abundant bacterial taxa. There were very few archaeal genes at the outermost stations (Stations 25 and 27), and these binned largely to methanogen sequences. The viral sequences were dominated by cyanobacterial phages (Table A.2).

Patterns of gene and transcript abundance provided insights into transcriptional activity by taxon and habitat [i.e., cells that were free-living versus those that were larger (>2 mm diameter) or particle-associated] for the dominant bacterial groups. Particle-

associated Verrucomicrobia (Order Puniceicoccales) maintained cellular transcript inventories of up to 14 transcripts/gene for particle-associated cells and averaged 2 transcripts/gene overall (Fig. A.2). In contrast, members of the Flavobacteria class averaged <0.5 transcripts/gene. Particle-associated cells in each of these major taxa typically had more transcripts per gene copy than did free-living cells (averaging 2.0 versus 0.15 transcripts/gene) (Fig. A.2). Abundance of transcripts originating from particle-associated versus free-living bacteria varied along the plume, with mRNAs from free-living cells contributing only 30-60% of the metatranscriptome in landward stations, but >90% at outer plume stations. Environmental data indicate that Station 10 had the lowest salinity (22.6) and Station 27 the highest (36.0). Station 10 was the most strongly influenced by riverine inputs, particularly of inorganic nitrogen.

Future Directions

The Amazon River plume is immense in scale and sensitive to anthropogenic forcing. This multi-omics dataset is the first of four high-throughput metagenomic and metatranscriptomic sequence collections being produced for the Amazon River Continuum as part of the ANACONDAS and ROCA projects (http://amazoncontinuum.org). These projects aim to improve predictive capabilities for climate change impacts on the marine biosphere, focusing on the Amazon ecosystem, and to better our understanding of feedbacks on the carbon cycle. Processes in the river and ocean are tightly linked from physical, biological, and biogeochemical perspectives. Thus the complete data collection will include two datasets from the Amazon plume (June 2010 and July 2013) and two from the Amazon River (Óbidos to Macapá and Belém; June 2011 and July 2013). These high-coverage, size-discrete, and replicated datasets are

all benchmarked with internal genomic and mRNA standards for comparative quantitative metagenomics and metatranscriptomics. Insights from these meta-omics datasets are enhancing predictive capabilities regarding the interplay between marine microbial communities, biogeochemical cycling, and carbon sequestration in the ocean.

Abbreviations

DOC: dissolved organic carbon; POC: particulate organic carbon; bp: base pairs; mRNA: messenger RNA; rRNA: ribosomal RNA; cDNA: complementary DNA; nt: nucleotides.

Availability of Supporting Data

Sequences from June 2010 Amazon Continuum study are available from NCBI under accession numbers SRP039390 (metagenomes), SRP037995 (non-selective metatranscriptomes), and SRP039544 (poly(A)-selected metatranscriptomes). The NCBI sequences are fastq files from which internal standard sequences and rRNA sequences (metatranscriptomes only) have been removed prior to deposition. Sequences are also available at the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) database under project number CAM_P_xxx. The CAMERA sequences are QC'd fasta files of joined paired-end reads, also with internal standards and rRNA sequences (metatranscriptomes only) removed. Metadata accompanying the omics datasets are provided in. ANACONDAS and ROCA project data are also available at the BCO-DMO data repository (http://www.bco-dmo.org/project/2097).

Authors' Contributions

BMS: conception and design of protocols, sample processing, data analysis, writing and final approval of the manuscript. BLZ: sample collection, sample processing, critical revision and final approval of the manuscript. MD: sample processing, protocol design, critical revision and final approval of the manuscript. CBS: sample processing, protocol design, critical revision and final approval of the manuscript. SS: data analysis, critical revision and final approval of the manuscript. JHP: critical revision and final approval of the manuscript. BCC: design of protocols, data analysis, critical revision and final approval of the manuscript. MAM: conception and design of protocols, data analysis, writing and final approval of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We appreciate the assistance of Roger Nilsen, Camille English, and Shulei Sun, and we thank P. Yager and scientists of the ROCA and ANACONDAS projects for helpful discussions. This research was funded by the Gordon and Betty Moore Foundation and NSF grant OCE-0934095. Resources and technical expertise were provided by the University of Georgia's Georgia Advanced Computing Resource Center and CAMERA.

References

Barada LP, Cutter L, Montoya JP, Webb EA, Capone DG, Sanudo-Wilhelmy SA (2013). The distribution of thiamin and pyridoxine in the western tropical North Atlantic Amazon River plume. *Front Microbiol* **4:** 25.

Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, Badger JH *et al.* (2011). Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol* **29:** 915-921.

Chong LS, Berelson WM, McManus J, Hammond DE, Rollins NE, Yager PL (2014). Carbon and biogenic silica export influenced by the Amazon River Plume: Patterns of remineralization in deep-sea sediments. *Deep Sea Res Part 1 Oceanogr Res Pap* **85:** 124-137.

Coles VJ, Brooks MT, Hopkins J, Stukel MR, Yager PL, Hood RR (2013). The pathways and properties of the Amazon River Plume in the tropical North Atlantic Ocean. *J Geophys Res-Oceans* **118:** 6894-6913.

Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC bioinformatics* **11:** 38.

Goes JI, Gomes HdR, Chekalyuk AM, Carpenter EJ, Montoya JP, Coles VJ *et al.* (2014). Influence of the Amazon River discharge on the biogeography of phytoplankton communities in the western tropical north Atlantic. *Prog Oceanogr* **120:** 29-40.

Richey JE, Nobre C, Deser C (1989). Amazon river discharge and climate variability: 1903 to 1985. *Science* **246:** 101-103.

Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ *et al.* (2010). Unlocking short read sequencing for metagenomics. *PLoS One* **5:** e11840.

Satinsky BM, Gifford SM, Crump BC, Moran MA (2013). Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. In: DeLong EF (ed). *Methods Enzymol*. Academic Press. pp 237-250.

Subramaniam A, Yager PL, Carpenter EJ, Mahaffey C, Bjorkman K, Cooley S *et al.* (2008). Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci U S A* **105:** 10460-10465.

Zhao Y, Tang H, Ye Y (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28:** 125-126.

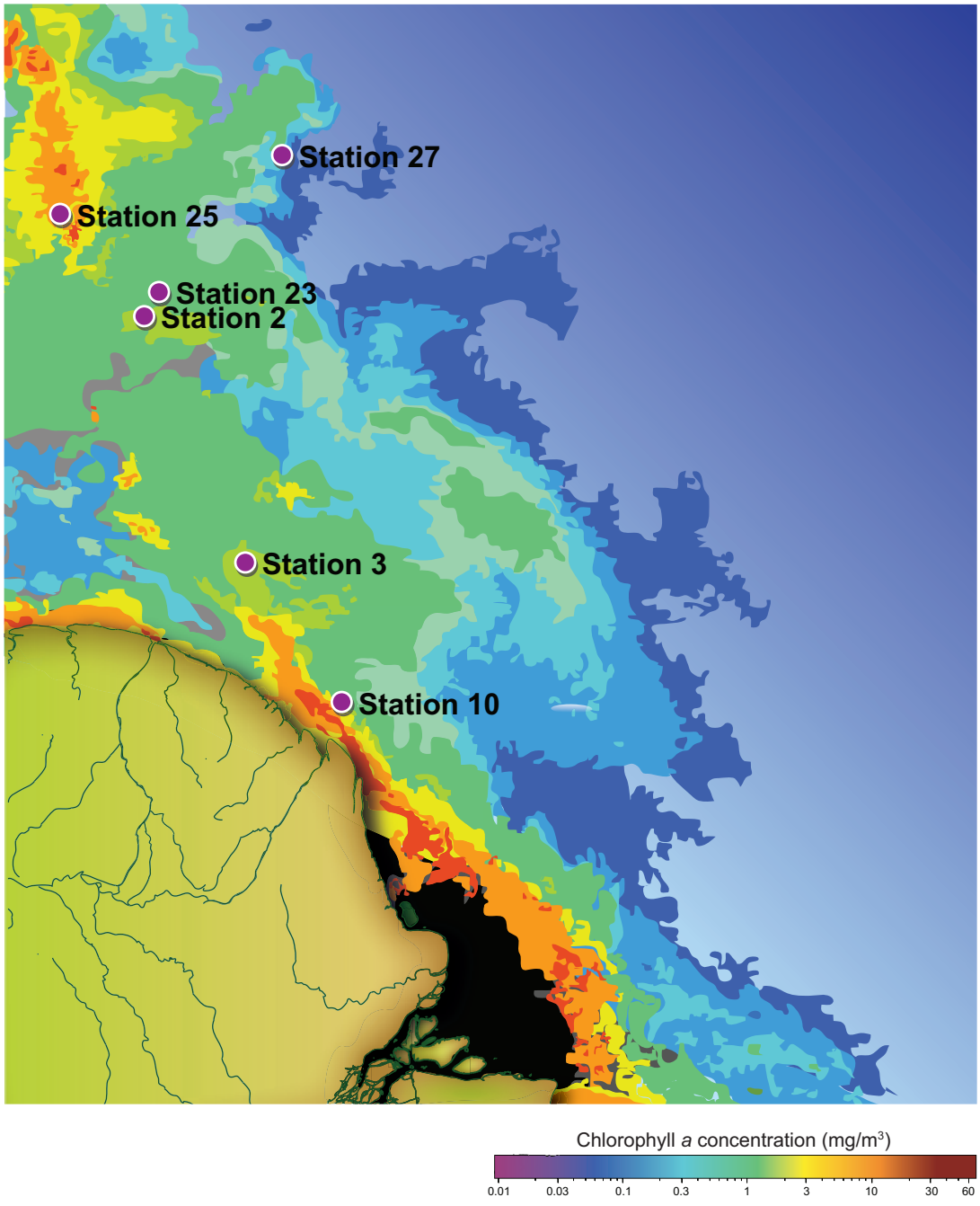


Figure A.1. Location of sampling sites in the Amazon River plume in June, 2010.

Table A.1. Number and types of libraries and reads obtained in the Amazon Continuum Project, June 2010, R/V *Knorr*.

	Metagenomes	**Non-selective Metatranscriptomes**	**Poly(A)-selective Metatranscriptomes**
Data Type	Total community DNA	Total community mRNA	Eukaryotic community mRNA[¥]
# Stations sampled	6	6	6
# Size fractions sampled	2	2	1
# Replicates	2	2	2
# Samples	24	24	12
# Raw reads	3.68×10^8	8.12×10^8	4.61×10^8
# Joined reads post QC	9.50×10^7	1.62×10^8	1.01×10^8
Average joined read length (bp)	205	190	185
# rRNA Reads	-	9.53×10^7	2.34×10^5
# Potential protein-encoding reads	9.44×10^7	6.52×10^7	9.86×10^7

[¥]The selective metatranscriptomes captured poly(A)-tailed transcripts and are therefore systematically biased against transcripts from eukaryotic organelles.

Table A.2. Reference genome bins garnering the most metagenomic reads, organized by station and Domain (top 10 Bacteria, 4 Eukarya, 2 Archaea, and 2 viruses). Bacterial, archaeal, and viral reads were annotated against the NCBI RefSeq database. Eukaryotic reads were annotated against a custom database containing marine eukaryotic genomes and transcriptomes from NCBI and 112 of the Marine Microbial Eukaryote Transcriptome Sequencing Project datasets that were public at the time of analysis (http://marinemicroeukaryotes.org).

Domain	Taxon	Genes L^{-1}	Domain	Taxon	Genes L^{-1}
Station 10					
Bacteria	*Synechococcus* sp. CB0205	1.46×10^{12}	Eukarya	*Thalassiosira oceanica* CCMP1005	1.26×10^{11}
Bacteria	SAR86 E	4.65×10^{11}	Eukarya	*Micromonas* sp. RCC299	8.27×10^{10}
Bacteria	SAR86 D	2.55×10^{11}	Eukarya	*Tetrahymena thermophila* SB210	2.41×10^{10}
Bacteria	Alphaproteobacterium HIMB5	2.32×10^{11}	Eukarya	*Strombidinopsis* sp. SopsisLIS2011	1.67×10^{10}
Bacteria	*Cand*. Pelagibacter sp. HTCC7211	2.22×10^{11}			
Bacteria	*Cand*. Pelagibacter ubique	1.54×10^{11}	Archaea	*Nitrosopumilus maritimus* SCM1	1.73×10^{10}
Bacteria	SAR86 C	1.39×10^{11}	Archaea	*Cand*. Nitrosopumilus koreensis AR1	1.02×10^{10}
Bacteria	Gammaproteobacterium HIMB55	1.33×10^{11}			
Bacteria	*Synechococcus* sp. CB0101	1.19×10^{11}	Virus	*Synechococcus* phage S-RSM4	1.74×10^{11}
Bacteria	Gammaproteobacterium HIMB30	1.15×10^{11}	Virus	*Synechococcus* phage S-SKS1	1.74×10^{11}
Station 3					
Bacteria	*Cand*. Pelagibacter sp. HTCC7211	2.79×10^{11}	Eukarya	*Micromonas* sp. RCC299	2.28×10^{10}
Bacteria	Alphaproteobacterium HIMB5	2.02×10^{11}	Eukarya	*Tetrahymena thermophila* SB210	4.93×10^{9}
Bacteria	SAR86 D	1.64×10^{11}	Eukarya	*Alexandrium tamarense* CCMP1771	3.71×10^{9}
Bacteria	SAR86 E	1.33×10^{11}	Eukarya	*Thalassiosira oceanica* CCMP1005	3.49×10^{9}
Bacteria	*Cand*. Pelagibacter ubique	1.17×10^{11}			
Bacteria	Alphaproteobacterium HIMB59	9.29×10^{10}	Archaea	*Nitrosopumilus maritimus* SCM1	3.01×10^{9}
Bacteria	SAR86 C	8.58×10^{10}	Archaea	*Cand*. Nitrosoarchaeum limnia	2.31×10^{9}
Bacteria	*Synechococcus* sp. WH 8109	7.33×10^{10}			
Bacteria	*Cand*. Pelagibacter ubique HTCC1062	6.37×10^{10}	Virus	*Synechococcus* phage S-RSM4	6.70×10^{10}
Bacteria	SAR324 JCVI-SC AAA005	5.06×10^{10}	Virus	*Synechococcus* phage S-SKS1	2.57×10^{10}
Station 2					
Bacteria	*Coraliomargarita akajimensis* DSM 45221	3.31×10^{12}	Eukarya	*Phaeocystis antarctica*	1.51×10^{12}
Bacteria	*Cand*. Puniceispirillum marinum IMCC1322	7.46×10^{11}	Eukarya	*Phytophthora sojae*	1.02×10^{12}
Bacteria	Gammaproteobacterium HIMB55	6.13×10^{11}	Eukarya	*Emiliania huxleyi*	9.44×10^{11}
Bacteria	*Synechococcus* sp. WH 8109	6.07×10^{11}	Eukarya	*Aplanochytrium kerguelense*	7.60×10^{11}
Bacteria	SAR116 HIMB100	5.95×10^{11}			
Bacteria	*Cand*. Pelagibacter sp. HTCC7211	5.09×10^{11}	Archaea	*Cand*. Nitrosopumilus salaria	1.54×10^{10}
Bacteria	SAR324 JCVI-SC AAA005	4.61×10^{11}	Archaea	*Methanomassiliicoccus* sp. M x 1-Issoire	5.89×10^{9}

Bacteria	Gammaproteobacterium HTCC2207	3.91×10^{11}			
Bacteria	Verrucomicrobiae DG1235	3.39×10^{11}	Virus	*Synechococcus* phage S-RIP1	9.55×10^{8}
Bacteria	*Prochlorococcus marinus* str. AS9601	3.16×10^{11}	Virus	*Phaeocystis globosa* virus	6.20×10^{8}
Station 23					
Bacteria	*Cand*. Pelagibacter sp. HTCC7211	1.36×10^{12}	Eukarya	*Tetrahymena thermophila* SB210	2.96×10^{10}
Bacteria	Alphaproteobacterium HIMB5	9.43×10^{11}	Eukarya	*Protocruzia adherens* Boccale	2.84×10^{10}
Bacteria	SAR86 D	9.31×10^{11}	Eukarya	*Strombidinopsis* sp. SopsisLIS2011	2.82×10^{10}
Bacteria	Alphaproteobacterium HIMB59	7.03×10^{11}	Eukarya	*Pseudo-nitzschia multiseries*	1.79×10^{10}
Bacteria	SAR86 E	6.95×10^{11}			
Bacteria	*Cand*. Pelagibacter ubique	5.17×10^{11}	Archaea	*Methanosarcina acetivorans* C2A	1.60×10^{9}
Bacteria	SAR86 C	4.69×10^{11}	Archaea	*Methanosarcina barkeri* str. Fusaro	1.37×10^{9}
Bacteria	*Cand*. Pelagibacter ubique HTCC1062	2.74×10^{11}			
Bacteria	SAR324 JCVI-SC AAA005	2.33×10^{11}	Virus	*Phaeocystis globosa* virus	1.01×10^{11}
Bacteria	Alphaproteobacterium HIMB114	2.31×10^{11}	Virus	*Synechococcus* phage S-SM2	4.62×10^{10}
Station 25					
Bacteria	*Cand*. Pelagibacter sp. HTCC7211	6.83×10^{11}	Eukarya	*Pyramimonas obovata* CCMP722	8.58×10^{9}
Bacteria	Alphaproteobacterium HIMB5	4.13×10^{11}	Eukarya	*Phaeocystis antarctica*	6.34×10^{9}
Bacteria	Alphaproteobacterium HIMB59	2.35×10^{11}	Eukarya	*Thalassiosira oceanica* CCMP1005	5.67×10^{9}
Bacteria	*Cand*. Pelagibacter ubique	2.07×10^{11}	Eukarya	*Volvox carteri* f. nagariensis	4.93×10^{9}
Bacteria	*Prochlorococcus marinus* str. AS9601	1.87×10^{11}			
Bacteria	*Prochlorococcus marinus* str. MIT 9301	1.70×10^{11}	Archaea	*Methanosarcina acetivorans* C2A	9.95×10^{8}
Bacteria	SAR86 E	1.67×10^{11}	Archaea	*Methanomassiliicoccus* sp. M x 1-Issoire	8.48×10^{8}
Bacteria	SAR86 D	1.61×10^{11}			
Bacteria	*Coraliomargarita akajimensis* DSM 45221	1.45×10^{11}	Virus	*Phaeocystis globosa* virus	4.57×10^{10}
Bacteria	Gammaproteobacterium HTCC2207	1.29×10^{11}	Virus	*Synechococcus* phage S-SM2	2.36×10^{10}
Station 27					
Bacteria	*Prochlorococcus marinus* str. AS9601	9.43×10^{12}	Eukarya	*Phaeocystis antarctica*	3.04×10^{10}
Bacteria	*Prochlorococcus marinus* str. MIT 9301	8.49×10^{12}	Eukarya	*Tetrahymena thermophila* SB210	2.25×10^{10}
Bacteria	*Cand*. Pelagibacter sp. HTCC7211	5.70×10^{12}	Eukarya	*Alexandrium tamarense* CCMP1771	1.56×10^{10}
Bacteria	*Prochlorococcus marinus* str. MIT 9215	4.46×10^{12}	Eukarya	*Monosiga brevicollis*	1.35×10^{10}
Bacteria	Alphaproteobacterium HIMB5	3.96×10^{12}			
Bacteria	*Prochlorococcus marinus* str. MIT 9312	3.13×10^{12}	Archaea	*Methanomassiliicoccus* sp. Mx1-Issoire	8.73×10^{9}
Bacteria	*Cand*. Pelagibacter ubique	2.22×10^{12}	Archaea	*Aciduliprofundum* sp. MAR08-339	6.10×10^{9}
Bacteria	*Prochlorococcus marinus*	1.75×10^{12}			
Bacteria	*Cand*. Pelagibacter ubique HTCC1062	1.31×10^{12}	Virus	*Prochlorococcus* phage P-SSM2	6.08×10^{11}
Bacteria	Alphaproteobacterium HIMB59	1.21×10^{12}	Virus	*Synechococcus* phage S-SM2	3.20×10^{11}

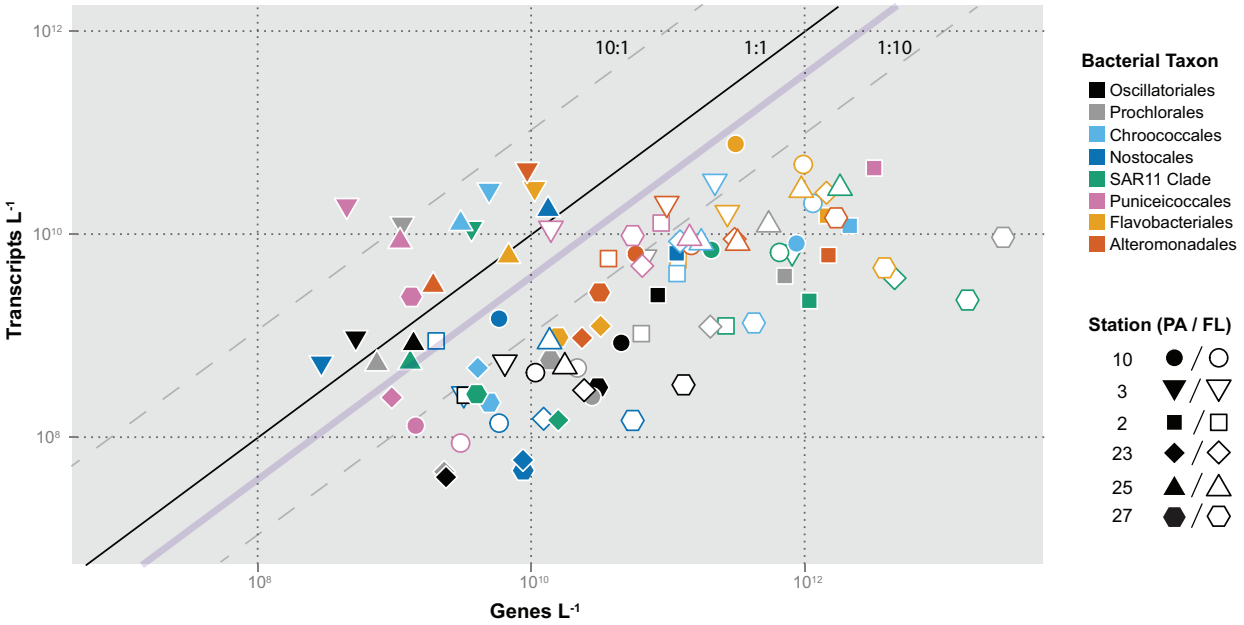


Figure A.2. Inventories of genes and transcripts for eight bacterial taxa in surface waters of the Amazon plume. Symbols represent the mean of duplicate analyses at six stations, color-coded by taxon and size fraction (particle-associated or free-living). Lines indicate a 1:1 ratio of transcripts:genes (black) or 10:1 and 1:10 ratios (gray). The purple line indicates the ratio of transcripts:genes for exponentially growing laboratory cultures of *Escherichia coli* [12,13]. Dominant bacterial groups are as follows: Oscillatoriales = *Trichodesmium*; Prochlorales = *Prochlorococcus*; Chroococcales = *Synechococcus*; Nostocales = *Richelia*; Puniceicoccales = Verrucomicrobia.

Supplemental Methods

Sample Collection

At each of the six stations selected for analysis, surface water was collected by gentle impeller pumping (modified Rule 1800 submersible sump pump) through 10 m of tygon tubing (3 cm) to the ship's deck where the water then flowed through a 156 μm mesh prefilter and was collected in 20 L carboys. The water was sequentially filtered (using a Masterflex peristaltic pump) through a 2.0 μm pore-size, 142 mm diameter polycarbonate (PCTE) membrane filter (Sterlitech Corporation, Kent, CWA) and a 0.22 μm pore-size, 142 mm diameter Supor membrane filter (Pall, Port Washington, NY). Duplicate samples were collected for each membrane size/type for total community metagenomics and metatranscriptomics. Duplicate 2.0 μm pore-size, 142 mm diameter polycarbonate (PCTE) membrane filters were collected for poly(A)-selective metatranscriptomics. After filtration, membranes were immediately submerged in RNAlater (Applied Biosystems, Austin, TX) in sterile 50 ml conical tubes, incubated at room temperature overnight, and then stored at -80^{o}C until extraction. All filtration and stabilization was completed within 30 min of water collection, and the volume of filtrate passed through each membrane was recorded.

RNA Processing for Total Community Metatranscriptomes

Prior to RNA extraction, the filters were thawed, removed from the preservative solution, placed in Whirl-Pak® bags (Nasco, Fort Artkinson, WI), and flash-frozen in liquid nitrogen. RNA extraction and DNA removal were carried out as previously described (Gifford *et al.,* 2011, Poretsky *et al.,* 2009a, Poretsky *et al.,* 2009b). For 0.2 – 2.0 μm samples, a lysis tube was

prepared for each sample consisting of a sterile 50 ml conical tube containing 8 ml of RLT Lysis Solution (Qiagen, Valencia, CA), 3 g RNA PowerSoil beads (Mo-Bio, Carlsbad, CA), and internal standards (described below). For ≥ 2.0 μm samples, a lysis tube was prepared for each sample consisting of a sterile 50 ml conical tube containing 10 ml of RLT Lysis Solution, 1.5 ml of 100 μm zirconium beads (OPS Diagnostics, Lebanon, NJ, USA), and internal standards. Filters inside the bags were broken into small pieces using a rubber mallet and transferred to the lysis tubes. Tubes were vortexed for 10 min to lyse cells, and RNA was purified from cell lysate using an RNeasy Kit (Qiagen, Valencia, CA) followed by two successive treatments with the Turbo DNA-free kit (Invitrogen, Carlsbad, CA) to remove residual DNA. Ribosomal RNA (rRNA) was selectively removed using community-specific biotinylated-rRNA probes prepared from DNA collected simultaneously (Stewart *et al.,* 2010). To maximize the removal of rRNA, probes were created for Bacterial and Archaeal 16S and 23S rRNA and Eukaryotic 18S and 28S rRNA. Probe-bound rRNA was removed via hybridization to streptavidin-coated magnetic beads (New England Biolabs, Ipswich, MA), and successful removal of rRNA from the samples was confirmed using either an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA) or a Bioanalyzer (Agilent Technologies, Santa Clara, CA). rRNA-depleted samples were linearly amplified using the MessageAmp II-Bacteria Kit (Applied Biosystems, Austin, TX), and amplified mRNA was converted into cDNA using the Superscript III First Strand synthesis system (Invitrogen, Carlsbad, CA) with random primers, followed by the NEBnext mRNA second strand synthesis module (New England Biolabs, Ipswich, MA), both according to manufacturer protocols. Synthesized cDNA was purified using the QIAquick PCR purification kit (Qiagen, Valencia, CA) followed by EtOH precipitation, resuspension in 100 μL of TE buffer, and storage at -80^{o} C until library preparation for sequencing.

RNA Processing for Poly(A)-tail Selected Metatranscriptomes

To ensure sufficient coverage of eukaryotic transcriptomes, a second metatranscriptome protocol was used that selectively sequenced messages with poly(A) tails; this was carried out for the >2.0 um pore-size filter only. Samples were prepared as described above with the following exceptions. An internal poly(A)-tailed mRNA standard was added to each lysis tube (see below). Following lysis, poly(A)-tailed mRNA was isolated from total RNA using an Oligotex mRNA kit (Qiagen, Valencia, CA), and mRNA was linearly amplified with a MessageAmp II-aRNA Amplification Kit (Applied Biosystems, Austin, TX). Double stranded cDNA was prepared as described above except cDNA was purified using the DNA Clean and Concentrator -25 Kit (Zymo, Irvine, CA) with five volumes of DNA binding buffer.

DNA Processing for Metagenomes

DNA was extracted and purified as previously described (Crump *et al.,* 1999, Crump *et al.,* 2003, Zhou *et al.,* 1996) with some modification. Briefly, each filter was thawed, removed from the preservative solution, and rinsed three times in autoclaved, filter-sterilized, 0.1% phosphate-buffered saline (PBS) to remove any residual RNAlater. Each filter was shattered as described above and placed in a tube containing DNA extraction buffer [DEB: 0.1 M Tris-HCl (pH 8), 0.1 M Na-EDTA (pH 8), 0.1 M Na2H2PO4 (pH 8), 1.5 M NaCl, 5% CTAB]. All liquid from the rinses as well as the original RNAlater was pushed through a Sterivex-GP filter capsule (EMD Millipore, Billerica, MA), which was subsequently rinsed 3 times to salvage any lost cells. The capsule was opened and the filter sliced into pieces and added to the tube with the original membrane filter and an internal genomic DNA standard (described below). Following

treatments with proteinase-K, lysozyme, and sodium dodecyl sulfate, DNA was purified via phenol:chloroform extraction and isopropanol precipitation.

Internal Standards

Omics processing included the addition of internal standards to allow for calculation of volume-based absolute copy numbers for each gene or transcript type, rather than just relative quantification (i.e., counts L^{-1} in addition to % of library) (Gifford *et al.,* 2011, Satinsky *et al.,* 2013). Two mRNA standards without poly(A) tails (to mimic prokaryotic and organelle mRNAs) were synthesized by in vitro transcription using a method modified from (Gifford *et al.,* 2011). The standards were constructed by linearizing a pTXB1 vector (New England Biolabs, Ipswich, MA) with NcoI restriction enzyme (New England Biolabs, Ipswich, MA) or pFN18A Halotag T7 Flexi Vector (Promega, Madison, WI) with BamHI restriction enzyme (New England Biolabs, Ipswich, MA). Each was purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. The 5' nucleotide overhangs were removed using Mung Bean Nuclease (New England Biolabs, Ipswich, MA), followed by purification via phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. Complete digestion of the vector was confirmed on a 1% agarose gel. The DNA fragment was then transcribed in vitro using the Riboprobe in vitro Transcription System (Promega, Madison, WI) according to the manufacturer's protocol using a T7 RNA polymerase to create 916 nt (pTXB1 standard) or 970 nt (pFN18A) artificial transcripts. Residual DNA was removed using RQ1 RNase-Free DNase and the RNA was purified by phenol:chloroform:isoamyl alcohol extraction and ethanol precipitation. The RNA standards were quantified using the Quant-iT Ribogreen RNA Reagent and Kit (Invitrogen, Carlsbad, CA), and RNA nucleotide length was confirmed with an Experion automated electrophoresis system

(Bio-Rad Laboratories, Hercules, CA). A known number of each standard (pTXB1 = 2.104 x 10^{10} copies; pFN18A = 1.172 x 10^{10} copies) was added independently to each lysis tube immediately prior to the addition of the sample filter.

An mRNA standard with a poly(A) tail (to mimic eukaryotic nuclear mRNA) was created from an HAP-1 Protolomerase viral gene. To create the standard, a 544 bp amplicon containing a poly(A) tail and a T7 promoter was produced from the template DNA through PCR. The PCR amplicons were then used as the template DNA for an in vitro transcription reaction to produce the resulting 499 nucleotide poly(A)-tailed mRNA. A known number of each standard (2.0 x 10^9 copies) was added to each tube immediately prior to lysis.

The genomic internal standard consisted of *Thermus thermophilus* DSM7039 [HB8] genomic DNA (American Type Culture Collection, Manassas, VA) added immediately prior to cell lysis. The amount of DNA standard added was estimated to be ~ 1% (8.4 ng per liter filtered) of sample DNA.

Sequencing

cDNA and DNA was sheared ultrasonically to ~200-250 bp fragments and TruSeq libraries (Illumina Inc., San Diego, CA) were constructed for paired-end (150 x 150) sequencing using the Illumina Genome Analyzer IIx, HiSeq2000, MiSeq, or HiSeq2500 platforms (Illumina Inc., San Diego, CA).

References

Crump BC, Armbrust EV, Baross JA (1999). Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia river, its estuary, and the adjacent coastal ocean. *Appl Environ Microbiol* **65:** 3192-3204.

Crump BC, Kling GW, Bahr M, Hobbie JE (2003). Bacterioplankton community shifts in an arctic lake correlate with seasonal changes in organic matter source. *Appl Environ Microbiol* **69:** 2253-2268.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011). Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* **5:** 461-472.

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

Poretsky RS, Hewson I, Sun S, 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.

Satinsky BM, Gifford SM, Crump BC, Moran MA (2013). Use of Internal Standards for Quantitative Metatranscriptome and Metagenome Analysis. In: DeLong EF (ed). *Methods Enzymol*. Academic Press. pp 237-250.

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.

Zhou J, Bruns MA, Tiedje JM (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62:** 316-322.