HARNESSING OMIC APPROACHES TO UNDERSTAND HOW CARBON, NITROGEN, AND SULFUR CYCLING ARE PARTITIONED IN DEEP SEA AND SEDIMENT MICROBIAL COMMUNITIES

by

Brett Joseph Baker

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Geology)
in the University of Michigan
2014

Doctoral Committee:

Associate Professor Gregory James Dick, Chair
Research Professor Jeffrey C. Alt
Professor Joel D. Blum
Assistant Professor Vincent J. Denef
Professor Donald R. Zak

To my family: My parents, Bruce and Diana Baker, my son, Silas Baker, and my wife, Angela Baker.

ACKNOWLEDGEMENTS

This dissertation is a product of four years of fun I had learning about marine microbiology, which will likely continue the rest of my life. Firstly, I would like to thank my advisor, Gregory Dick, for providing a rich research atmosphere and laboratory. I consider Greg to be one of my greatest allies and friends. I am especially appreciative to him for teaching me so much about microbial oceanography. I have also been privileged to have an extremely helpful committee: Joel Blum, Jeffrey Alt, Vincent Denef and Donald Zak; whom I can honestly say I have learned a great deal from each of them. You never stop learning and they have proven that to be true.

I would like to thank my co-authors and collaborators: Andreas Teske, Cody Sheik, Brandy Toner, Cassandre Lazar, Chris Tayler, James Cavalcoli, and Jill Banfield. Andreas has believed in my research from the beginning and has allowed me to invade his marine sediment research domain. Cody has become a friend and has provided endless insights about microbial physiology. Cassandre has been integral in initiating sediment project chapter V. Brandy has become one of my closest colleagues and will continue to be in the future. Thanks to Chris and James for all the lively discussions and contributions to chapter IV. Thanks to Jill for the support for the past 15 years and giving me the push I needed to go back and get my PhD. I have learned more from Jill than I can possibly say in a few sentences.

I would like to thank all the members of the Greg Dick lab at Michigan, past and present for their assistance including; Ryan Lesniewski, Alex Voorhies, Meng Li, Daniel Reed, Paul Den Uyl, Cody Sheik, Sunit Jain, Karthik Antharman and Daniel Marcus. I could tell you how wonderful it has been to be in laboratory where everyone gets along so well and helps each other.

I am grateful to funding agencies for supporting my dissertation research. The Guaymas Basin project was supported by Gordon and Betty Moore Foundation Grant GBMF2609, the National Science Foundation Grant OCE 1029242 and the University of Michigan Rackham Graduate School Predoctoral Fellowship Program. National Science Foundation Grant OCE1038006. Thanks to the Office Staff and Support Staff at the Department of Earth and Environmental Sciences.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

Microorganisms have mediated the cycling of elements on Earth for billions of years. However, the majority of microbes present in nature are uncultured and we know little about their physiologies or how geochemical niches are partitioned in nature. Innovations in DNA sequencing technologies and computational analyses now allow us to reconstruct genomes of individual community members from environmental samples. This process, referred to as "metagenomics", enables elucidation of metabolic pathways of microbes without having to culture them. Furthermore, *in situ* gene activity can be measured by sequencing community RNA, referred to as "metatranscriptomics". This dissertation uses these revolutionary approaches to investigate two biogeochemical hot spots in the oceans: deep-sea hydrothermal vent plumes and estuarine sediments.

The deep oceans are rich in bioavailable nutrients and contain the largest pool of inorganic carbon near the Earth's surface. Very little was known about the diversity of microbes or metabolic pathways involved in carbon and nitrogen cycling in the deep oceans. Therefore, I have applied metagenomic and transcriptomic approaches to the Guaymas Basin (Gulf of California) deep sea hydrothermal plume and surrounding waters. Nitrification is the main source of nitrate to primary producers in the oceans. This process is mediated by two distinct groups of microbes, ammonia oxidizers (NH_3 to NO_2^-) and nitrite oxidizers (NO_2 to NO_3). We found that ammonia oxidation in the plume and background waters are driven by Thaumarchaea capable of CO_2 fixation. Surprisingly, these Archaea are closely related to those found in surface ocean waters. However, these deep sea populations are physiologically distinct, for example they are able to utilize urea as a source of nitrogen. Interestingly, nitrite oxidation is mediate by a novel group of bacteria. Also among the most active genes in the plume and background waters are those involved in the amino acid transporters, suggesting the ability to recycle carbon and nitrogen from amino acids. These genes belong to a group of bacteria, similar to the genus *Alteromonas*, and to novel Archaea generically referred to as "marine group II". This suggests that these uncultured Archaea, which are ubiquitous in the deep sea, are involved in the remineralization of organic carbon to CO_2.

I have also used metagenomics to reconstruct the genomes of 82 bacteria from estuary (in the White Oak River, North Carolina) sediments. The genomes obtained are common to sediments worldwide, and have provided insights into the metabolisms of these uncultured bacteria. Many of the genomes belong to branches of life that had not been sequenced before and are involved in important processes including organic carbon degradation and fermentation. Among the most abundant organisms in the sulfate-rich layer are novel, uncultured members of the *Gammaproteobacteria* that are capable of anaerobic sulfur oxidation coupled to nitrate and nitrite reduction. This high resolution dataset enabled us to construct the first realistic map of the flow of carbon and energy through estuary sediment microbial communities.

CHAPTER I

INTRODUCTION

Microorganisms derive energy by transferring electrons from various sources (e.g., organic carbon, hydrogen, ammonia, sulfide) to various electron acceptors (e.g., oxygen, nitrate, sulfate), depending on their availability and energy yield. Microbial enzymes catalyze these non-equilibrium redox reactions, and thus drive nearly all the elemental cycles on the planet. For over a hundred years, since Sergei Winogradsky began culturing microorganisms from the environment (Dworkin, 2011), microbial ecology relied on growing community members in the laboratory to understand their physiologies. In the late 1980's, the development of PCR-amplification and DNA sequencing of single genes from the environment (16S rRNA gene) by Pace and others (Pace et al. 1986, Pace et al. 1991) resulted in a paradigm shift in microbiology. This led to realization that the vast majority of microbes (>99.9%) in nature are novel and have not been cultured (Pace, 1996, Hugenholtz et al. 1998). Newly discovered divisions of life are as different to what had been cultured as humans are to mushrooms. As revolutionary as this was, the sequencing of individual genes yields little information about the metabolisms (or geochemical roles) of these uncultured groups.

The fields of microbial ecology, geomicrobiology, and oceanography are being rapidly transformed by innovations in DNA sequencing technologies and computational analyses. This transformation is being fueled by the decreasing cost and increase in the yield of DNA sequencing. It is now possible to reconstruct hundreds of genomes of microbes directly, without cultivation, from the environment. This approach, referred to as "metagenomics", enables us to piece together genes into complete metabolic pathways that are present in uncultured microbes. The reconstruction of genomes has begun to inform about the biogeochemical roles of several uncultured taxa including; OD1, OP11, OP9 bacteria, and Marine Group II Archaea to name a few (Wrighton et al. 2011, Dodsworth et al. 2013, Iverson et al. 2012). However, the presence of genes does not always equate to activity in nature. Therefore, recently a new approach to quantify the activity of genes in natural communities has been developed, called transcriptomics

(Frias-Lopez et al. 2008). Transcriptomics is the sequencing of RNA (transcripts), which is what is produced in the cell from DNA (genes) when active.

In this dissertation I use metagenomic and transcriptomic (omic) approaches to better understand the microbes and metabolic pathways involved in geochemical cycling of carbon and nitrogen in Guaymas Basin (GB) deep sea hydrothermal plumes (Gulf of California), and carbon, nitrogen, and sulfur cycling in White Oak River (WOR) estuary sediments (North Carolina). In the process of doing this I have also been involved in enhancing the computational analyses used to reconstruction both genomes and the quantification of transcript data from these microbial communities.

1.1 Nitrogen cycling in the Guaymas Basin deep sea and hydrothermal plumes

The Guaymas Basin contains a deep-sea hydrothermal system that is located at a water depth of about 2000 m and that forms the northern-most segment of the East Pacific Rise. It is unique compared to other hydrothermal systems in that it is located in a semi-enclosed basin, close to the coast, and sits underneath highly productive surface waters. This results in relatively high sedimentation rates, which has produced a thick layer of organic-rich sediments in the basin. As a result, the hydrothermal fluids that circulate through these sediments are enriched in ammonia, hydrocarbons, and methane (Bazylinki et al. 1988, Von Damm et al. 1985). The Guaymas Basin vent fluids have a temperature of 315^{o}C, a pH of 5.9, and concentrations of H_2 and H_2S of 3.4 and 6 mM, respectively.

Nitrogen is essential to the structure of all life, as it is a large component of proteins. Nitrate, a key nutrient for primary production in the oceans, is primarily generated by microbial nitrification. Nitrification is a two-step process mediated by two distinct groups of microbes referred to as ammonia oxidizers (NH_3 to NO_2) and nitrite oxidizers (NO_2 to NO_3) (Francis et al. 2007). Ammonia oxidizers were thought to be driven by a few groups of Proteobacteria (Kowalchuk and Stephen, 2001), but our understanding of nitrogen cycling has change dramatically in recent years.

Among the most abundant and widespread microbes on the planet are the so-called “cold water Crenarchaea” or “Marine Group I”, now referred to as Thaumarchaea (Fuhrman, 2010, Brocheri-Armanet et al. 2008). These Thaumarchaea have been shown to comprise over 20% of

the picoplankton in the oceans, an estimated total of 10^{28} cells, and up to 40% in the deep sea (Karner et al. 2001). One of the first metagenomic studies was conducted on surface waters in Bermuda and found putative ammonia monooxygenase (*amoA*) gene present in Thaumarchaea (Venter et al. 2004). Soon after, a member of this group was isolated from an aquarium and shown to be capable of ammonia oxidation to nitrite with biocarbonate as the carbon source (Könneke et al. 2005). This isolate, named *Nitrosopumilus maritimus*, was also shown to have a high affinity for ammonia, it is thought that this results in Thaumarchaea's ability to outcompete ammonia oxidizing bacteria in low ammonia waters in the oceans (Konneke et al. 2005).

It is now known that Thaumarchaea are the key drivers of ammonia oxidation in the oceans, however little is known about the next step of nitrification, nitrite oxidation. Nitrite rarely accumulates in the oceans. Given the large numbers of ammonia oxidizers present there, nitrite oxidizers must be capable of keeping pace (Ward et al. 2007). Members of the genus *Nitrospina* have been shown to be dominant in Monterey Bay, and the deep North Pacific Gyre (Delong et al. 2006, Mincer et al. 2007). Two other bacteria, *Nitrospira* and *Nitrococcus*, have been implicated to be dominant in oxygen minimum zones (Fussel et al. 2012). However, these bacteria are commonly seen at much lower abundances than Thaumarchaea. This is has led to speculation that there are unknown, yet-to-be-discovered organism that are driving nitrite oxidation in nature (Ward et al. 2007).

The major source of nitrate to the surface ocean is through diffusion and upwelling of nitrate-rich deep ocean water (Zehr and Ward, 2002). Nitrate in the deep oceans is produced entirely by microbial nitrification. Studies of ammonia oxidizers (Thaumarchaea) have mostly focused on surface waters (Tully et al. 2012), and cultures from soils (Lehtovirta-Morley et al. 2011, Jung et al. 2014) and aquariums (Konneke et al. 2005). Therefore, we were interested in using metagenomic and transcriptomics to interrogate both ammonia and nitrite oxidizing populations in the Guaymas Basin deep sea hydrothermal plume, which has high concentrations of ammonium and abundant Thaumarchaea.

1.2 Microbial geochemical cycling in sediments

Marine and estuary sediments contain massive pools of organic carbon (Raven and Falkowski, 2002). Microbes degrade organic carbon and produce electron donors, namely acetate and H_2, for respiratory processes via anaerobic metabolisms and fermentation. Electron acceptors are consumed in order of the energy yield of the redox reaction, with the most energetically favorable electron acceptors consumed first (i.e., $O_2 > NO_3 > SO_4 > CO_2$). This results in vertical zonation of microbial redox niches in the sediment profiles. Organic carbon is remineralized, the energy derived from reduction of the electron acceptors is used for microbial cell growth, the products of reduction reactions (e.g., HS^-, CH_4, N_2O) serve as electron donors for other microbial metabolisms or in some cases accumulate or escape to the water column. It has been shown that up to 95% of the sulfide and thiosulfate produced by sulfate reduction is reoxidized in coastal sediments (Jørgensen, 1990). Although sulfate reduction has been shown to be mediated by *Deltaproteobacteria*, the microorganisms involved in sulfur oxidation, and the energetics of this process, are mostly unknown. Broadly speaking, sediment microbes assimilate and mobilize buried organic carbon produced in the overlying water column. In the ocean this results in the sequestration of carbon. While in estuaries it can represent a net sink of carbon or the release of methane and CO_2 to the atmosphere (Bauer et al. 2013, Cai, 2011).

Unfortunately, our understanding of sediment microbiology is largely limited to surveys of the diversity of single genes (Teske and Sorensen, 2008) and cultivation of sulfate reducers and methanogens. These communities are considerably complex and we now know that most of what is out there has not been cultured (Whitman et al. 1998). This is a fundamental gap in our understanding of one of the largest habitats on the planet. It has been shown that there is considerable overlap in the membership of Archaea in marine sediments and estuaries (Kubo et al. 2012), suggesting that key geochemical processes are mediated by microbial groups that are ubiquitous. A recent study looking at the genomes of a few members of these widespread marine sediment Archaea, Misc. Crenarchaea Group (MCG) and Marine Benthic Group (MBG-D), revealed that they are capable of protein degradation (Lloyd et al. 2013). However, these are just a small portion of total diversity present in sediments.

To better understand the links between sediment microbial metabolism and biogeochemistry, we conducted metagenomic sequencing of sediments in the White Oak River Estuary (WOR) in North Carolina (34°44.49N; 77°07.44W). The White Oak River is an approximately 40 mile-long coastal blackwater stream flowing through coastal plain hardwood

forests before it enters a 10-mile-long tidal estuary; early diagenetic pathways (sulfate reduction, methanogenesis) in sediments of the freshwater lower river and the seawater-influenced central estuary (the sampling site) have been studied extensively (Martens and Goldhaber 1978; Kelley et al. 1990, 1995; Martens et al. 1998; Avery and Martens 1998; Avery et al. 2003). Locally dominant plants (*Spartina* spp., cord grass, Juncus black needle rush, Red Cedar) contribute significantly to the buried organic material of the estuarine sediments (3-6% TOC) and its $\delta^{13}C$ isotopic signature (Kelley et al. 1990). Additionally, phytoplankton biomass (especially phytoflagellates, cryptomonads in cool seasons, and cyanobacteria in summer blooms; Malin 1994) contributes seasonally to the organic carbon input into the river. The abundance and diversity of organic biomass could be a key factor that selects for the benthic microbial communities. This location has been shown to be rich in microbial groups that are common to deep sea and coastal sediments throughout the world, making it a model site to study sediment biogeochemical cycling (Kubo et al. 2012).

1.3 Structure of the dissertation

The research I present in this dissertation is motivated by fundamental questions such as: Which microbes are present in the deep sea hydrothermal plumes and estuary sediments? What are their metabolic and ecological roles? How are geochemical niches partitioned within the communities? While addressing these questions, in every chapter, I have also been involved in the development and enhancement of novel metagenomic and transcriptomic computational approaches. **Chapter II** is a feature article on the status of analyzing omic data (metagenomics and metatranscritpomics) in microbial ecology that I was asked to write for Microbe magazine. In this chapter I detail how DNA and RNA (cDNA) sequence assembly is important for piecing together individual genomes and pathways from natural microbial communities. In **Chapter III**, I ask the questions: Who are the dominant ammonia oxidizers in deep sea hydrothermal plumes and how do their genome sequences and predicted physiologies compare to those from other environments? I reconstructed genomes of Thaumarchaea populations involved in ammonia oxidation in the deep Guaymas Basin and use them as a basis to compare gene expression in hydrothermal plumes to the background waters. In **Chapter IV**, I ask the question: Which community members are mediating important geochemical transformations of carbon and

nitrogen in the deep sea? I again focus on the Guaymas Basin, but conduct the first *de novo* assemblies of metatranscriptomic data to assess gene expression activity in the plume and background microbial communities. In **Chapter V**, I ask the question: What are the metabolic capabilities of estuary sediment bacteria and how are carbon and energy cycled through the community? I reconstructed dozen of bacterial genomes, many belonging to new branches of life, which enabled us to construct a detailed map of cycling of carbon, nitrogen, iron, and sulfur in an estuary sediments. In **Chapter VI**, I provide overall conclusions and a brief discussion of future directions.

1.4 References

Avery BG, and Martens CS (1999) Controls on the stable carbon isotopic composition of biogenic methane produced in a tidal freshwater estuarine sediment. *Geochim Cosmochim Acta* 63, 1075-1082.

Avery BG, et al. (2003) Controls on methane production in a tidal freshwater estuary and a peatland: methane production via acetate fermentation and CO2 reduction. *Biogeochemistry* 62, 19-37.

Bauer JE, Cai W-J, Raymond PA, Bianchi TS, Hopkinson CS, Regnier PAG (2013) The changing carbon cycle of the coastal ocean. *Nature* 504, 61-70.

Brochier-Armanet C, et al. (2008) Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6, 245-252.

Cai, W-J. (2011) Estuarine and coastal ocean carbon paradox: CO2 sinks or sites of terrestrial carbon incineration? *Annu Rev Mar Sci* 3, 123–45.

DeLong EF, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311, 496-503.

Dodsworth JA, et al. (2013) Single-cell and metagenomic analyses indicate a fermentative saccharolytic lifestyle for members of the OP9 lineage. *Nat Comm* 4, 1854.

Dworkin, M. (2011) Sergei Winogradsky: a founder of modern microbiology and the first microbial ecologist. *FEMS Micrbiol. Ecol*. 36, 364-379.

Frias-Lopez J, et al. (2008) Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci* 105, 3805-3810.

Fussel J, et al. (2012) Nitrite oxidation in the Namibian oxygen minimum zone. *ISME J* 6, 1200-1209.

Hugenholz P, Goebel BM, and Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J Bacteriol* 180, 4765-4774.

Iverson V, et al. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335, 587-590.

Jung M-Y, et al. (2014) A mesophilic autotrophic ammonia-oxidizing archaeon of the thaumarchael group I.1a cultivate from a deep oligotrophic soid horizon. *Appl Environ Microbiol* in press.

Jørgensen, BB (1990) A thiosulfate shunt in the sulfur cycle of marine sediments. *Science* 249:152-154.

Kelley CA, et al. (1990) Variations in sedimentary carbon mineralization rates in the White Oak River estuary, North Carolina. *Limnol Oceanogr* 35, 372-383.

Kelley CA, et al. (1995) Methane dynamics across a tidally flooded riverbank margin. *Limnol Oceanogr* 40, 1112-1129.

Kubo K, et al. (2012) Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J*. 6, 1949-1965.

Könneke M, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437, 543-546.

Lazar C, et al. Environmental controls and intragroup diversity of the MCG lineage naturally enriched in anoxic sediments of the White Oak River Estuary (North Carolina, USA). in review.

Lehtovirta-Morley LE, et al. (2011) Cultivation of an obligate acidophilic ammonia oxidizer from a nitrifying acid soil. *Proc Natl Acad Sci* 108, 15892-15897.

Martens CS, and Goldhaber MB. (1978) Early diagenesis in transitional sedimentary environments of the White Oak River estuary, North Carolina. *Limnol Oceanog*r 23, 428-441.

Mincer TJ, et al. (2007) Quantitative distribution of presumptive archaeal and bacterial nitrifiers in the Monterey Bay and the North Pacific Subtropical Gyre. *Environ. Microbiol* 9, 1162-1175.

Pace NR (1996) New perspective on the natural microbial world: molecular microbial ecology. *ASM News* 62, 463-470.

Pace NR, Stahl DA, Lane DJ, Olsen GJ (1986) The analyses of natural microbial populations by ribosomal RNA sequences. *Adv Microbial Ecol* 9, 1-55.

Pace NR, Delong EF (1991) Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing". *J Bacteriol* 173, 4371–4378.

Raven JA, Falkowski PG (2002) Oceanic sinks for atmospheric CO2. *Plant, Cell, Environ* 22, 741-755.

Teske A, and Sorensen KB (2008) Uncultured archaeal in deep marine subsurface sediments: have we caught them all? *ISME J* 2, 3-18.

Tully BJ, et al. (2012) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* 14, 254-267.

Walker CB, et al. (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci* 107, 8818-8823.

Ward BB, et al. (2007) What's new in the nitrogen cycle? *Oceanography* 101-109.

Whitman WB, et al. (1998) Prokaryotes: The unseen majority. *Proc Natl Acad Sci* 95, 6578-6583.

Wrighton KC, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 28, 1661-1665.

CHAPTER II

OMIC APPROACHES IN MICROBIAL ECOLOGY: CHARTING THE UNKNOWN

Brett J. Baker[a], and Gregory J. Dick[a,b]

a. Department of Earth and Environmental Sciences , University of Michigan, Ann Arbor, MI 48109
b. Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109

Originall published in 2013 in *Microbe* 8: 353-359.

2.1 Summary

- Investigators are deducing genomes, transcriptomes, and proteomes of novel microbes, based on analysis of DNA sequence data from complex environmental samples.
- Because genome sequence data fall short of covering the microbial diversity in nature, database-dependent approaches to analyzing that diversity have important limitations and biases.
- Recent glimpses, using omics approaches, into the uncultured microbial biosphere have started to reveal fascinating organisms that expand our spectrum of knowledge of biology.
- Using whole-community omics to track microbial communities in nature can resolve the roles of novel uncultured groups and shed light on fundamental links between ecological and evolutionary processes.

2.2 Introduction

"*Somewhere, something incredible is waiting to be known.*" Carl Sagan

Much like astronomy, the microbial sciences are confronted with staggering unknowns and are reliant on tools and technology to probe the frontiers of knowledge. The first realizations of just how extraordinarily diverse and unexplored microorganisms are occurred by analyzing microbial ribosomal RNA (rRNA) gene sequences directly from environmental samples. Norm Pace of the University of Colorado, Boulder, and his collaborators 20 years ago helped to pry open the extraordinarily diverse but largely unexplored array of microorganisms from hot springs and comparable environments.

However, the novel microorganisms that they uncovered are not restricted to extreme environments. In fact, we need to look no farther than the dirt outside our homes and workplaces, the lakes and streams that provide us with water to drink, and the surfaces on and within our bodies. Pace's approach, applied to many environments, led many other microbiologists to realize that many phyla in the natural world are waiting to be discovered. Indeed, these ongoing efforts to learn more about microorganisms from so many different environments continue to alter our views of the tree of life (Fig. 2.1 and 2.2).

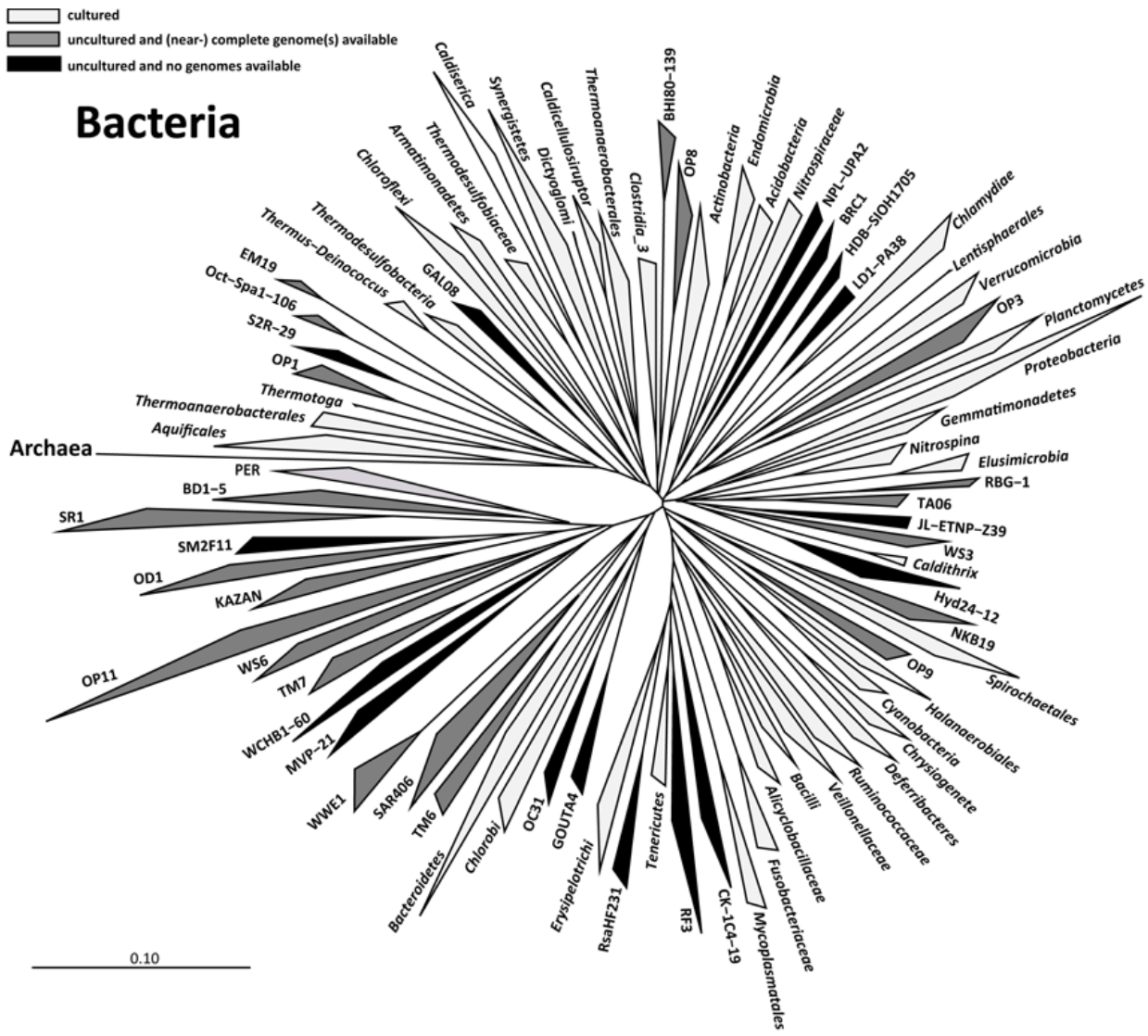

Figure 2.1 Phylogenetic tree reflecting our knowledge of the diversity of Bacteria. Sequences are clustered into bacterial phyla and archaeal divisions based on the Silva classification with manual curations and additions.

Advances in DNA sequencing techniques made it possible to move from analyzing rRNA genes to entire genomes. Although this approach, called metagenomics, is based on random shotgun DNA sequencing of microbial communities and has proved powerful for interrogating microbial communities, it does not provide information about which genes are being expressed at any specific moment. However, whole-community transcriptomics and proteomics are helping to address questions about gene expression. Here we highlight how such whole community omic

analyses provide unique perspectives of microbial diversity and discuss the importance of using *de novo* assembly to analyze this type of data.

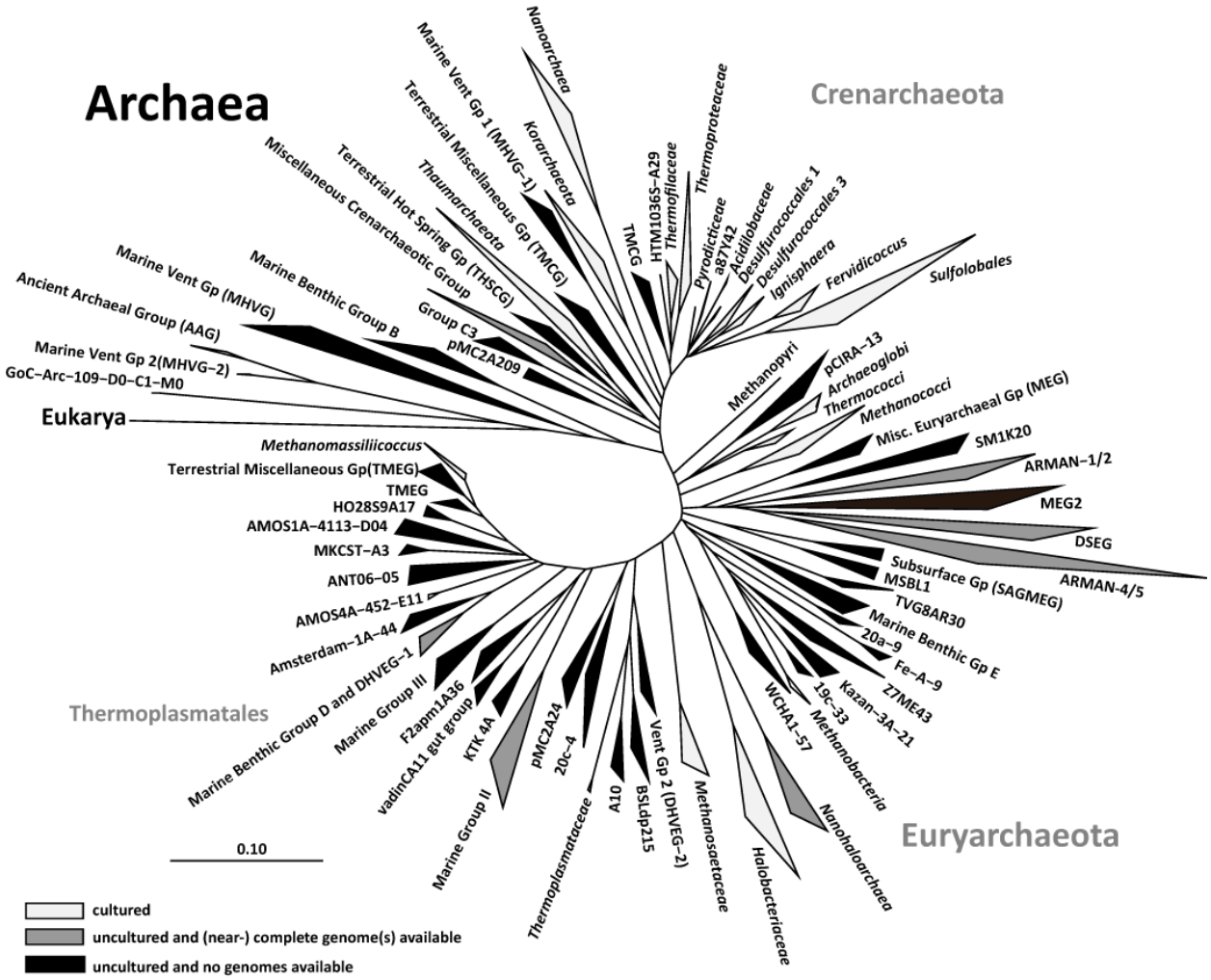

Figure 2.2 Phylogenetic tree reflecting our knowledge of the diversity of Archaea. Sequences are clustered into bacterial phyla and archaeal divisions based on the Silva classification with manual curations and additions.

2.3 Understanding Microbial Communities in the Wild

Nearly a decade ago, Jill Banfield of the University of California, Berkeley, and her collaborators applied random shotgun DNA sequencing to microbial biofilms from extremely acidic waters in the Iron Mountain mine in California, assembling genomes for the most abundant of those microbes. Although they relied on Sanger sequencing, whose yields are

modest compared to other DNA-sequencing techniques that are now available, their efforts showed that genome sequences could be reconstructed directly from environmental samples, bypassing cultivation while providing insights into the uncultured members of that microbial community.

Subsequent studies illustrate that metagenomics contributes on at least three distinct levels to our understanding of microbial diversity. Taken together, these three levels provide insights into diversity that are inherent to microbial communities in the wild and that are readily tracked by whole-community genomic approaches.

First, metagenomic analysis reveals entirely new microbial groups that might otherwise be overlooked if one were relying on rRNA or other traditional analytic techniques. For example, Banfield and her collaborators in 2006 detected groups of archaea called ARMAN (archaeal Richmond Mine acidophilic nanoorganisms), that were not previously recognized because their 16S rRNA genes are mismatched with standard archaeal PCR primers.

Second, metagenomics highlights how even well-known microbial taxa harbor novel genes, which sometimes are harbored within hypervariable regions called genomic islands. Such novel genes are sometimes expressed at high levels in particular environments, suggesting that they are important even if their functions are not known, according to Edward DeLong of the Massachusetts Institute of Technology in Cambridge and his collaborators.

Third, metagenomics can reveal genomic variability at the strain level within natural populations of Bacteria, Archaea, and viruses. Such variability between slightly different strains sometimes contributes in important ways to community functioning.

2.4 Data Assembly Is Critical when Analyzing Microbial Communities

Investigators now have two main options for analyzing whole-community genomic and transcriptomic datasets. One approach involves comparing individual DNA sequences, or reads, to those in available databases. This approach helps them to determine the function of specific genes and the organism in which those genes are functioning (grey arrows, Fig. 3.2).

One ongoing problem with taking this “read mapping” or “recruitment” approach is that many or, in some cases, most of the reads may fail to match sequences in public databases such as GenBank and MG-RAST. Although these and other genomic databases are growing rapidly,

they vastly underrepresent microbial diversity in nature. Genomes from cells grown in culture or other sources that populate such databases are mere snapshots of the microbial world, and fall short of the full genomic repertoire. Put another way, much of the DNA and RNA sequence data from environmental samples is novel and thus unlike what is available in major databases.

There are at least two critical challenges facing anyone interpreting reads that do have matches in databases. First, comparisons of DNA sequences from the environment to those in databases typically use low thresholds of sequence similarity, or BLAST scores, to define positive matches. This approach can result in finding considerable diversity within reads mapping to single database sequences (Fig. 2.3). For example, we found that many bacterial reads readily map to an archaeal genome with commonly used parameters. Second, inferring functions of genes based on environmental reads can be difficult because read lengths are considerably shorter than full-length genes.

The other major approach to analyzing whole microbial communities involves assembling individual reads into ever-larger fragments on the basis of sequence overlaps (Fig. 2.3). While this approach may seem daunting, especially when working with diverse microbial communities in which sequence coverage of individual taxa may be limited, it can prove fruitful, and it is now commonly used to reconstruct near-complete genomes for microorganisms within samples from environments such as seawater and groundwater.

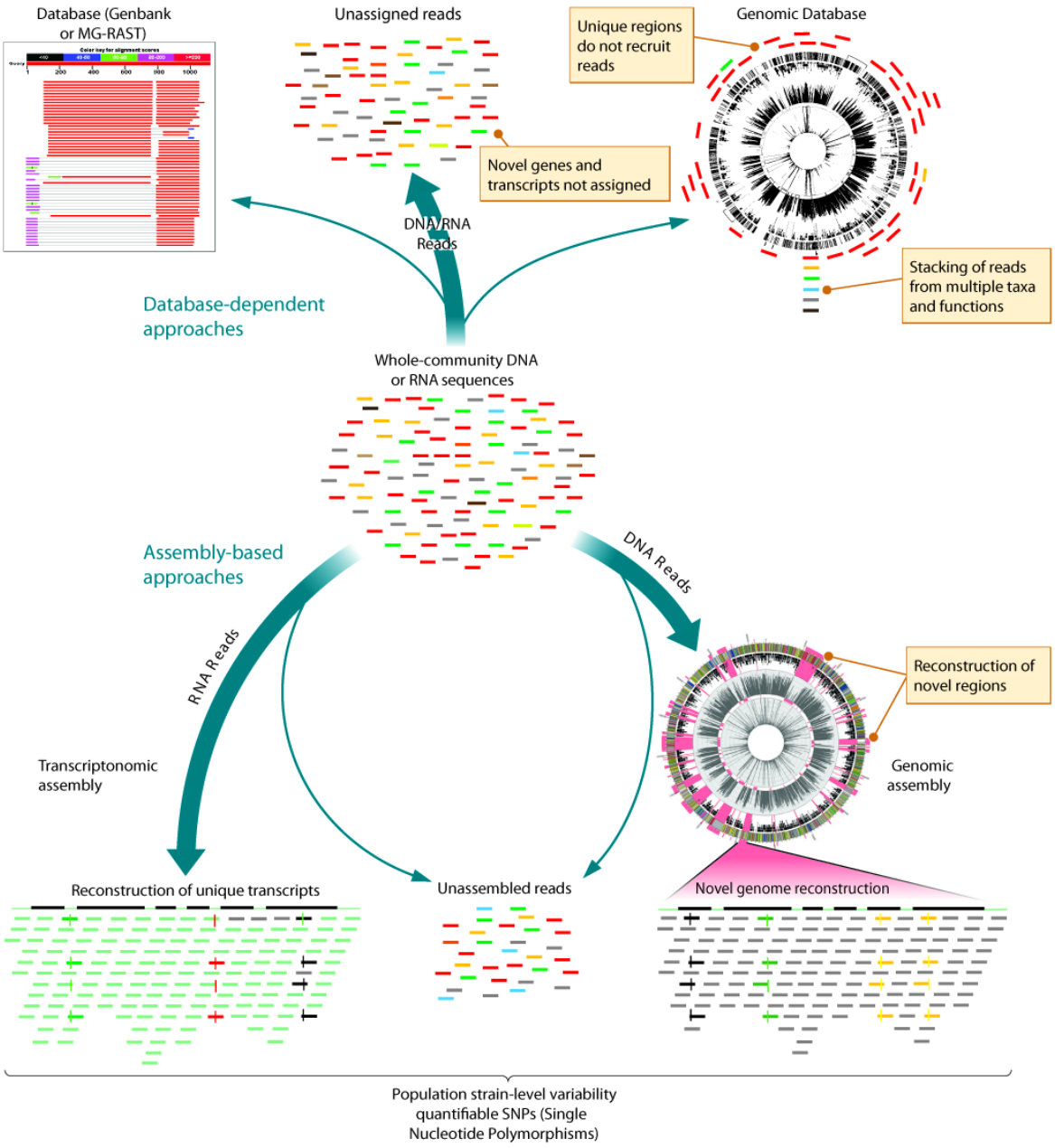

Figure 2.3 Schematic of approaches for analyzing whole-community genomic and transcriptomic sequence data, database-dependent (on top) and assembly-based (on the bottom). Multi-colored lines represent DNA sequences ("reads") from different microbial genomes. The thickness of arrows roughly represents number of reads. Database mapping and searching of both DNA (genomic) and RNA (transcriptomic) reads is biased by the completeness of the database being used to map and assign function to the reads. Both transcriptomic and genomic assemblies enable the reconstruction of natural community members and genes that are commonly overlooked by database-dependent approaches.

This analytic approach can lead to useful insights. For instance, genes that are linked via the genome can also reveal how genes associate on finer scales, such as operons, providing potential insights into how those genes function. Moreover, this approach enables investigators to predict what metabolic pathways may be functioning within uncultured organisms. Perhaps most importantly, metagenomic assembly can help investigators to piece together novel genomic regions as well as full genomes for microorganisms that are absent from databases. Even sets of fragments that do not belong to other near-complete genomes can be linked, or "binned," with one another on the basis of signatures such as tetranuclueotide frequencies. However, binning is not currently possible when relying on individual reads.

Metagenomic assemblies can also enhance analyses of parallel metatranscriptomic data. We took this approach to look at ammonia-oxidizing archaeal populations within deep-sea hydrothermal plume communities from the Gulf of California. From these samples, we assembled several similar genotypes of archaea and could differentiate the transcriptional activity of each of these types under different regimes of ammonium concentration. Thus, sequences from different organisms can be assembled and used to estimate their abundance and transcriptional activity in the community (Fig. 2.4). Without assembly to resolve these different types, all of those reads might be assigned mistakenly to only one or a few entries in public databases, thus collapsing queries into database-dependent compartments that do not fully reflect the full diversity within the natural environment.

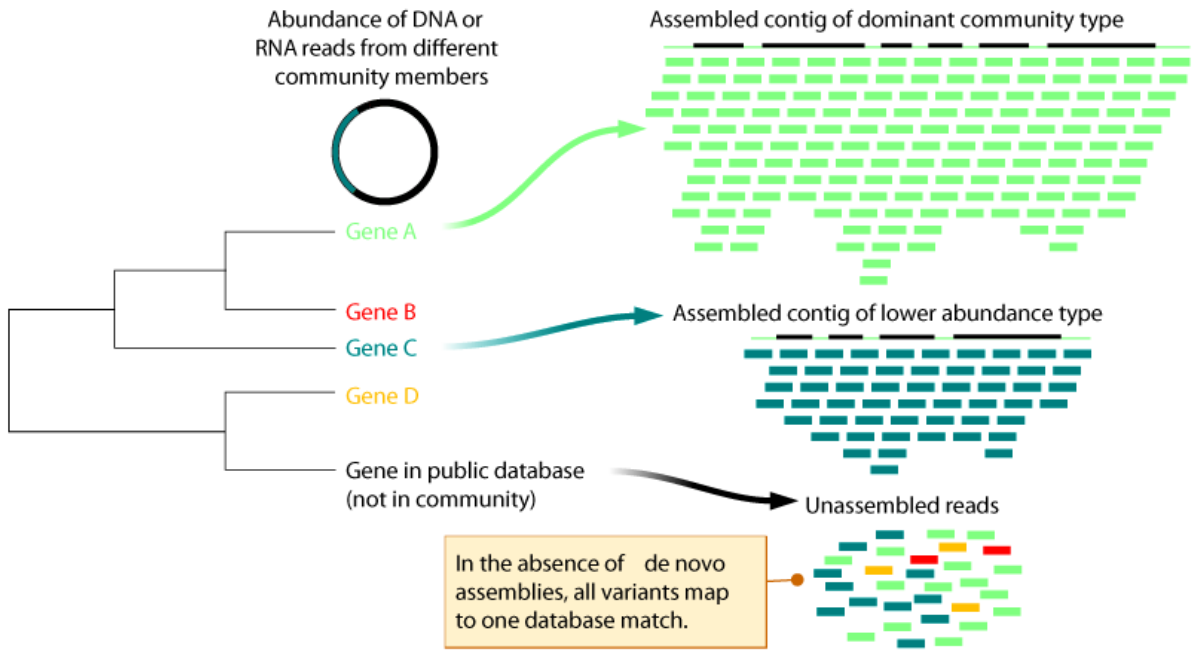

Figure 2.4 Community assembly enables the reconstruction of fine-scale genomic and transcriptomic variation that is inherent to natural microbial populations. Thus, we can determine which genotypes are the most abundant and/or transcriptionally active in the community.

Complementary metagenomic and metatranscriptomic datasets can also be used for assessing the relative abundance and activity of different microbial populations. We find extreme differences in DNA and RNA compositions from the same community. Some of the most abundant mRNA sequences belong to rare members of the community.

Relying on metagenomic data when analyzing metatranscriptomics has pitfalls, however. For example, novel, highly expressed genes from low-abundance organisms may not be captured in metagenomic datasets or public databases. To better characterize these novel mRNA sequences, we assembled transcript libraries from the deep sea to reconstruct whole operons from uncultured microbial populations. Some of the most abundant transcripts are those from rare—that is, accounting for less than 1% of the community—bacteria that are involved in oxidizing nitrite. There is value in doing transcriptomic and genomic assemblies in parallel.

2.5 Filling in the Tree of Life

During the past several decades, rRNA gene surveys have uncovered a substantial number of new microbial phyla or divisions. Few members of these phyla have been cultured, thus, our knowledge of them is limited to abundances of single marker genes in the environment and, in some instances, visualization of cells with rRNA-targeted fluorescent probes.

How do we begin to understand the physiology and ecology of these mysterious lineages, which some experts call biological dark matter? In lieu of traditional culture-based analyses, genome sequence analysis of environmental samples is providing insights into their lifestyles and evolutionary history.

A few years ago, for example, we determined the genomic sequences of the members of two uncultured archaeal phyla, now referred to as Parvarchaeota (two genomes, ARMAN-4 and 5) and Micrarchaeota (ARMAN-2). The latter sequence was obtained directly from whole-community assemblies, while the two Parvarchaeota sequences depended in part on enriching our samples for ultrasmall (less than 500 nm in diameter) cells. Recognizing that their genomes indicated that these cells interact with others, we determined that other nearby archaeal species mysteriously penetrate their cell walls.

This degree of genomic information remains a rarity for Archaea, most of whose phyla are poorly characterized (Fig. 2.2). However, massive sequencing of natural microbial communities is bearing fruit more quickly in terms of furnishing genomes for candidate Bacteria phyla. For example, Banfield and her collaborators recently assembled genomes for members of several such phyla, including OD1, BD1-5, OP11, and a new group PER (Fig. 2.1).

Single-cell genomics (SCG) is another approach to filling gaps in the tree of life. SCG depends on isolating individual cells from microbial communities, and then amplifying individual whole genomes before determining their DNA sequences. Of the known Bacteria phyla, many uncultured phyla are published and others are in the works. Some of the novel bacterial genomes, including OP1, OP9, and OP3, as well as several archaea such as Misc. Crenarchaeota Group and Marine Benthic group D, were determined by SCG.

Although powerful, SCG provides mere snapshots of the genomic content of single cells. High-throughput methods make it possible to take many snapshots, but only community-wide approaches provide a direct measure of their full diversity. This diversity can be tracked further using community transcriptomic and proteomic approaches. Ultimately, community genomics and SCG are complementary, with the latter providing references for fragment recruitment and

nucleotide compositional binning, while metagenomic data can provide quantitative information and a wider lens for addressing questions concerning whole-communities and their dynamics.

2.5 Outlook

Microbiologists now have the analytic tools to reconstruct genomes, transcriptomes, and proteomes of entire microbial communities. Because so much of what is out there is uncharacterized, we are facing many challenges before we more fully understand how novel genes and proteins from members within such communities function and how they determine the ecological roles of novel phyla.

However, now that we can also quantify gene expression in nature and how it varies in different niches, we can begin to assess how these novel microorganisms within communities behave and interact at the system level. Such data may yield clues about how to culture cells within novel microbial groups, leading to experiments that will link gene sequences to their functions. However, even if we could culture all members from a particular microbial community, removing them from that environment will change how each of them behaves. Moreover, because ecology and evolution are linked, our efforts to understand the latter will very much depend on our ability to track microbial genomes and their expression patterns in complex assemblages in nature.

Acknowledgements. We thank Jillian Banfield for insights and sequences for the phylogenetic trees. This article is funded by the Alfred P. Sloan Foundation BR2013-027, Gordon and Betty Moore Foundation GBMF2609 to Dr. Gregory Dick, and by a University of Michigan Rackham Pre-doctoral fellowship to Brett Baker.

2.7 References

Baker BJ, Comolli LR, Dick GJ, Hauser L, Land M, Hyatt D, VerBerkmoes N, Hettich R, and Banfield JF (2010) Enigmatic, ultra-small uncultivated Archaea. *Proc. Natl. Acad. Sci.* USA 107, 8806-8811.

Baker BJ, Lesniewski RA, and Dick GJ (2012) Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J* 6, 2269-2279.

Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, and Banfield JF (2006) Lineages of acidophilic Archaea revealed by community genomic analysis. *Science* 314, 1933-1935.

Baker BJ, Taylor C, Jain S, Shiek CS, Bhasi A, Cavalcoli JD, and Dick GJ (2012) Community transcriptomic assembly reveals novel and low abundance microbes contribute to carbon and nitrogen cycling in the deep sea. *ISME J* 7, 1962-1973.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, and DeLong EF (2008) Microbial community gene expression in ocean surface waters. *Proc Nat Acad Sci* USA 105, 3805-3810.

Ram RJ, VerBerkmoes N, Thelen MP, Tyson GW, Baker BJ, Shah M, Blake II RC, Hettich R, and Banfield JF (2005) Community proteomics of a natural microbial biofilm. *Science* 308, 1915-1920.

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

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NV, Wilkins MJ, Hettich RL, Lipton MS, Williams KH, Long PE, and Banfield JF (2012) Fermentation, hydrogen, and sulfur metabolisms in multiple uncultivated bacterial phyla. *Science* 337, 1661-1665.

CHAPTER III

GENOME-ENABLED TRANSCRIPTOMICS REVEALS ARCHAEAL POPULATIONS THAT DRIVE NITRIFICATION IN A DEEP-SEA HYDROTHERMAL PLUME

Brett J. Baker[1], Ryan A. Lesniewski[1] and Gregory J. Dick[1,2,3]

[1]Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA

[2]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[3]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

Originally published in 2012 in *ISME Journal* 8: 353-359.

3.1 Abstract

Ammonia-oxidizing Archaea (AOA) are among the most abundant microorganisms in the oceans and play crucial roles in biogeochemical cycling of nitrogen and carbon. To better understand AOA inhabiting the deep sea, we obtained community genomic and transcriptomic data from ammonium-rich hydrothermal plumes in the Guaymas Basin (GB) and from surrounding deep waters of the Gulf of California. Among the most abundant and active lineages in the sequence data were marine group I (MGI) Archaea related to the cultured autotrophic ammonia-oxidizer, *Nitrosopumilus maritimus*. Assembly of MGI genomic fragments yielded 2.9 Mb of sequence containing seven 16S rRNA genes (95.4 to 98.4% similar to *N. maritimus*), including two near-complete genomes and several lower-abundance variants. Equal copy numbers of MGI 16S rRNA genes and ammonia monooxygenase genes and transcription of ammonia oxidation genes indicates that all of these genotypes actively oxidize ammonia. *De novo* genomic assembly

revealed the functional potential of MGI populations and enhanced interpretation of metatranscriptomic data. Physiological distinction from *N. maritimus* is evident in the transcription of novel genes, including genes for urea utilization, suggesting an alternative source of ammonia. We were also able to determine which genotypes are most active in the plume. Transcripts involved in nitrification were more prominent in the plume and were among the most abundant transcripts in the community. These unique datasets reveal populations of deep-sea AOA thriving in the ammonium-rich GB that are related to surface types but with key genomic and physiological differences.

3.2 Introduction

Marine Group I (MGI) Archaea are a diverse group of Archaea that are ubiquitous in marine environments and are thought to play a significant role in global nitrification (Delong 1992, Delong *et al.*, 1994, Francis *et al.*, 2005, Fuhrman *et al.*, 1992, Kalanetra *et al.*, 2009, Wuchter *et al.*, 2006). Originally classified as Crenarchaeota, recent phylogenetic analysis suggests that the MGI are part of the distinct and deeply-branching phylum Thaumarchaeota (Brochier-Armanet *et al.* 2008; Pester *et al.* 2011). These Archaea are particularly abundant in the deep, dark ocean (Church *et al.* 2010), where they account for up to 40% of microbial communities (Karner *et al.*, 2001). Despite their abundance and biogeochemical importance, fundamental questions remain regarding the physiology and metabolism of MGI. Several studies of MGI have provided evidence for both autotrophic ammonia oxidation (Ingalls *et al.*, 2006, Konneke *et al.*, 2005) and heterotrophy (Tourna *et al.*, 2011, Ouverney and Fuhrman 2000, Agogué *et al.*, 2008, Muβmann *et al.*, 2011). Autotrophic ammonia oxidation has now been confirmed in a few cultured representatives (Tourna *et al.* 2011, De La Torre *et al.*, 2008), including *Nitrosopumilus maritimus* (Konneke *et al.*, 2005). Physiological characterization of *N. maritimus* showed that it has a high affinity for ammonia, providing a mechanism of niche differentiation with ammonia-oxidizing bacteria (AOB) (Martens-Habbena *et al.*, 2009), which are active in soils and other environments with higher ammonium concentration (Verhamme *et al.*, 2011).

Given the difficulty in culturing MGI, only two genomes have been fully sequenced. Both come from shallow waters, the sponge symbiont *Cenarchaeum symbiosum* (Hallam *et al.*, 2006) and the aquarium isolate *N. maritimus* (Walker *et al.*, 2010). Recently, draft genome

sequence has been obtained from single cells and San Francisco Bay sediment enrichments of *Nitrosoarchaeum limnia,* recovered from an estuary in San Francisco bay (Blainey *et al*., 2011), and from a soil isolate, *Nitrososphaera viennensis* (Tourna *et al*., 2011). Characterization of these genomes suggests they use a modified 3-hydroxypropionate/4-hydroxybutryrate pathway for carbon fixation, and have a copper-dependent system for ammonia oxidation and electron-transfer that is distinct from ammonia-oxidizing bacteria. Additionally, comparison of these genomes with marine metagenomic datasets revealed widespread conservation of gene content, highlighting the ubiquity of these oligophiles throughout the world. A recent genomic characterization of communities of MGI Archaea from surface waters in the Gulf of Maine revealed that *N. maritimus* has several genomic islands that are not present in marine populations (Tully *et al.*, 2012).

Although MGI are particularly abundant in the deep oceans (Karner *et al*., 2001), these deep populations are not well studied compared to those from shallower depths. A recent PCR-based study found that deep waters (>1000m depth) of the North Atlantic have lower ratios of MGI *amoA* to 16S rRNA gene copies than subsurface waters, suggesting that most deep-sea MGI are heterotrophic (Agogué *et al*., 2008). However, metagenomic sequencing of North Pacific waters at 4,000 m depth revealed an equal ratio of MGI *amoA* to 16S rRNA genes (Konstantidis *et al*., 2009). Furthermore, it has recently been shown that expression of ammonia monoxygenase does not always signify autotrophy (Muβmann *et al*., 2011). In this study, we utilize deep-sea hydrothermal vent plumes in Guaymas Basin of the Gulf of California as natural laboratories in which to study ecological and physiological responses of deep-sea MGI to ammonium inputs. Sedimented hydrothermal systems such as Guaymas Basin are enriched in ammonium due to interactions of hydrothermal fluids with organic-rich sediments as they ascend en route to the water column (Von Damm et al., 1985). As a result, ammonium concentrations in Guaymas Basin end-member fluids (10.3-15.6 mM) (Von Damm *et al*., 1985) are considerably higher than unsedimented ridge discharge fluids (<0.01 mM) (Lilley *et al*., 1993). These hydrothermal inputs contribute to ammonium concentrations of up to 3 μM in GB deep waters (1800 to 2000 m depth) (Lam, 2004). Gene-based surveys have shown that the MGI dominate the Guaymas Basin plume archaeal community (Dick and Tebo, 2010, Lesniewski *et al*., 2012), and that MGI are more abundant in plumes than background seawater in the deep Indian Ocean and Okinawa Trough (Takai *et al*., 2004).

Here we use community genomics and transcriptomics to survey the genomic diversity and activity of MGI populations in ammonium-enriched Guaymas Basin plumes compared to surrounding background waters. Community genomics and transcriptomics have proved to be valuable in understanding ecology of microbial communities (Hallam *et al*., 2006, Baker *et al*., 2010, Frias-Lopez *et al*., 2008, Shi *et al*., 2009). To date metatranscriptomic studies have relied almost entirely on comparisons to public genomic databases (Frias-Lopez *et al*., 2008, Shi *et al*., 2009, Stewart *et al*., 2011), isolate genomes (Hollibaugh *et al.*, 2011), and unassembled DNA sequence (Shi *et al.*, 2011). Instead, we utilized *de novo* genomic assembly of community DNA to evaluate the genomic diversity of MGI and provide a framework for recruitment of transcripts to closely-relate gene variants from plume and background waters. These analyses provide a unique glimpse into deep-sea MGI genomic diversity, and suggest that a cluster of closely-related Archaea dominate nitrification in the deep waters of the Gulf of California.

3.3 Materials and Methods

Sample Collection and processing. Samples were obtained by CTD Rosette from Guaymas Basin and Carmen Basin on three cruises aboard the *R/V New Horizon* in 2004 and 2005. Once on deck, plume and background waters were immediately filtered by N_2 gas pressure onto 0.2-µm pore size, 142-mm diameter polycarbonate filters and fixed and frozen in RNAlater as previously described (Dick *et al.* 2009b, Dick and Tebo 2010). Further details of sample processing, locations, and environmental conditions are provided in Table S1 and in Lesniewski et al., (2012). Plume-1 and Plume-2 were used for genomics while Plume-3 and Plume-4 were transcriptomics samples from the plume. Two background samples were each used for both metagenomics and metatranscriptomics. Because it is not possible to obtain true background samples from sub-sill depths of Guaymas Basin, Background-1 was taken from just above the Guaymas Basin plume and Background-2 was from the next basin south of Guaymas, Carmen Basin (Lesniewski *et al.*, 2012).

RNA was isolated using a modification of the mirVana miRNA Isolation kit (Ambion) as described previously (Stewart *et al.*, 2011, Hollibaugh *et al.*, 2011). The RNA was then purified and concentrated using the RNeasy MinElute Cleanup kit (Qiagen). cDNA synthesis was conducted as described previously (Hollibaugh *et al.*, 2011). Genomic DNA and cDNA libraries

were prepared for sequencing using standard protocols (454 Life Sciences, Roche) and randomly shotgun sequenced by 454 Titanium pyrosequencing. All of the cDNA reads presented here are available in the NCBI Sequence Read Archive under accession number SRA045655. The entire metagenomic assembly has been deposited NCBI Bioproject number PRJNA77837.

Genomic analyses. Genomic reads were assembled using MIRA 3 (http://chevreux.org/projects_mira.html), and resulting contigs were manually checked with consed (Gordon *et al.*, 1998) and annotated using the JGI IMG/MER system (Markowitz et al., 2009). Initial binning of the assembled fragments was done using tetra-nucleotide frequencies signatures and ESOM mapping as detailed in Dick *et al.* (2009a). Since this binning only obtained fragments larger than 2.5 kb, we also searched the entire assembly for additional fragments using reciprocal BLAST searches with the *N. maritimus* genome. GB fragments were then checked for synteny by manually comparing gene order to that of *N. maritimus*. Fragments were chosen to be added to the bin manually based on synteny and sequence match qualities. All phylogenetic trees were generated using maximum likelihood method within ARB software package (Ludwig *et al.* 2004).

cDNA analyses. Transcript reads were mapped to predicted proteins using BLASTX (cutoff of bit score >45 and >70% similarity). Previous transcriptome studies have relied on publically available databases to recruit cDNA reads using bit score of >40-45 (Frias-Lopez *et al.* 2008, Gifford *et al.* 2011, Hollibaugh *et al.* 2011, Shi *et al.* 2009). We used a bit score cutoff of >40 to assign mRNA reads to the *de novo* assembled MGI genomes, resulting in recruitment of a total of 10,747 reads (Figure S4). Of these, 4,382 hits have less than 70% sequence similarity to the MGIC proteins. Comparison of these transcripts with Genbank revealed that most are not archaeal. In fact, only 13% had top hits to Archaea and just 10% of those matched MGIC (bit score >40). Furthermore, recruitment of reads originally identified as archaeal (bit score >40, %ID <70) to the entire GB genomic assembly revealed that the majority of them (72%) are more accurately assigned to other members of the community (Figure S5). In contrast, 93% of the cDNA reads recruited at >70% sequence similarity had top hits to Archaea in NCBI, and 88% of those had top hits to MGI. Comparisons of recruitment patterns between the samples within the

plume and the two backgrounds did not reveal significant differences in transcription profiles (data not shown), therefore we pooled the two plume and two background samples for analyses.

To confirm the absence of AOB, we searched the transcript libraries using *amoA* from *Nitrosomonas marinus* and *Nitrosococcus oceani* with e-value cutoffs of $1x10^{-10}$. Some genes like ammonia monooxygenases are highly conserved, for example there is 97-99% similarity (at the protein level) between *amoA* genes in the community. Therefore, in order to accurately differentiate expression among variants we found it necessary to recruit reads at the DNA level. For normalized comparison analyses the total number of mapped transcripts was divided by the length of the gene and the total number of transcripts from each sample for comparison between them. A total of 1,651,287 (696,718 from plume-1 and 954,569 from plume-2) and 1,117,284 (570,580 from background-1 and 546,704 from background-2) reads in the plume and background libraries respectively. The circular diagram for comparative genomics and transcriptomics was generated using Circos (Krywinski *et al.* 2009). Suspected replicate reads due to artifacts of 454 pyrosequencing were manually removed from ammonia monooxygenase gene analyses and from all DNA read coverage-based analyses.

3.4 Results and disccusion

Community genomics and transcriptomics reveals multiple populations of Marine Group I Archaea in the deep Gulf of California. Plume samples yielded more RNA as well as more DNA and cDNA reads than the background samples (Table S1). *De novo* genomic assembly and binning by tetranucleotide frequency and emergent self-organizing maps (Dick *et al.*, 2009a) revealed a well-defined MGI bin that contained 449 DNA sequence fragments with length greater than 2.5 kb, totaling 1.79 Mb of consensus sequence (Figure S1). Based on BLAST searches of the whole community versus the *N. maritimus* genome, we identified 790 additional fragments belonging to MGI, bringing the total length of assembled fragments identified as MGI to 2.9 Mb. The average GC content of the bin was 31%, similar to that of *N. maritimus* (34%). The average fragment size was 2.3 kb and there were 80 sequences (of 1,239 total) longer than 5 kb.

Seven different MGI 16S rRNA genes were identified. Phylogenetic analyses placed them all in group I.1a (Figure S2). 16S rRNA sequence similarity to *N. maritimus* ranged from

95.4 to 98.4% (pair-wise gene aligned), indicating that these GB populations likely represent distinct species of the *Nitrosopumilus* genus. To estimate the abundance and overall metabolic activity of each phylotype across samples, metagenomic and metatranscriptomic reads were mapped to the MGI 16S rRNA genes (Figure S3). MGI 16S rRNA genes consistently recruited more cDNA reads from plumes than background. The balance of DNA reads in background versus plume was more variable, and in some cases much higher in background than plume (Figure S3). In terms of the whole community, these MGI SSU rRNA genes were the most abundantly represented Archaea in both plume and background metatranscriptomic datasets (Lesniewski *et al.*, 2012).

The GB genomic assembly also contained seven different sequence types of MGI *amoA* genes. The average coverage of *amoA* genes was 6X (294 total genomic reads), comparable to the 4X coverage of MGI 16S rRNA genes (390 total genomic reads). Normalization of read numbers by gene length resulted in a roughly equivalent copy number of *amoA* and 16S genes (*amoA*:16S ratio of ~1.4), indicating that the majority of GB MGI cells have *amoA* and are thus capable of ammonia oxidation. These findings are consistent with those of Konstantinidis *et al.* (2009) from 4,000 m depth at station ALOHA and distinct from those of Agogué *et al.*, 2008, which found smaller *amoA*:16S ratios in deep Atlantic waters.

Comparison of the metagenome and metatranscriptome to N. maritimus. Comparison of the Guaymas Basin MGI metagenome to the genome of *N. maritimus* (Walker et al., 2010) revealed both similarities and differences. 85% of the ORFs in the Guaymas Basin MGI bin (4,875 of 5,744) are homologous to proteins from *N. maritimus*. These homologs average 78% protein sequence similarity, and appear to stem primarily from four different genotypes, two of which are well-covered in the Guaymas Basin metagenome (Figure 3.1). The remaining 15% of putative proteins in the MGI bin do not have homology to proteins in *N. maritimus* (e-value cutoff $1x10^{-10}$). Most of these Guaymas Basin-unique proteins (76%) could not be assigned any putative function. There were also predicted proteins from the *N. maritimus* genome that could not be identified in the Guaymas Basin genomic data (205 of the 1,799), the majority of which were annotated as hypothetical proteins. Many of these *N. maritimus*-unique genes clustered in certain regions of the *N. maritimus* genome (Figure 3.1), which correspond to recently identified genomic islands that are also absent in MGI populations from surface waters of the Gulf of

Maine (Tully *et al*., 2012). This suggests that these regions are unique to the *N. maritimus* genome and that our metagenomic assembly contains near-complete genomes of MGI populations.

Figure 3.1. Mapping of Guaymas Basin metagenomic fragments to the Nitrsospumilus maritimus genome. The outer-most ring is the complete genome *N. maritimus* with ORFs colored based on COG categories. The black tiles inside are assembled genomic fragments from the Guaymas assembly that map by BLASTn to regions of the *N. maritimus* genome. The outer grey shaded circle is a histogram showing percent sequence identity of top GB proteins that match *N. maritimus* proteins (scaled from 50% inside to 100% outside). The inner-most grey circle is the raw number of transcripts that map to those homologous proteins (range 0-30) for visualization of region with no recruitment. Genomic islands (> 1 kb) missing in archaeal

metagenomic data from Gulf of Maine surface waters (Tully *et al.*, 2012) are highlighted with light red wedges. Note that nearly all the gaps in the Guaymas Basin genomic data occur in these genomic islands.

The advent of transcriptomic sequencing of microbial communities is advancing knowledge of the transcriptional activity organisms in the environment (Frias-Lopez *et al.*, 2008, Stewart *et al.*, 2011). However, accurate assignment and phylogenetic placement of transcripts from natural populations of uncultivated microorganisms is hindered by a lack of coverage of genomes present in the environment. We applied a stringent threshold (>70% sequence identity and bit score >45) to recruit 8,520 high similarity reads to the Guaymas Basin MGI metagenome (Figure 3.2). 6,363 of these transcripts came from the plume and 2,157 from the background samples, with an average of 94% amino acid similarity. Using the same parameters only 6,849 transcripts were mapped to *N. maritimus* genes (Figure S4), highlighting the value of genomes assembled directly from the same environment where metatranscriptomic data was collected. Given the considerable diversity of MGI in deep Gulf of California waters (Figure S2) and the modest quantity of mRNA transcripts recovered, the metatranscriptomic data presented here likely represents only the most abundantly transcribed genes of MGI populations.

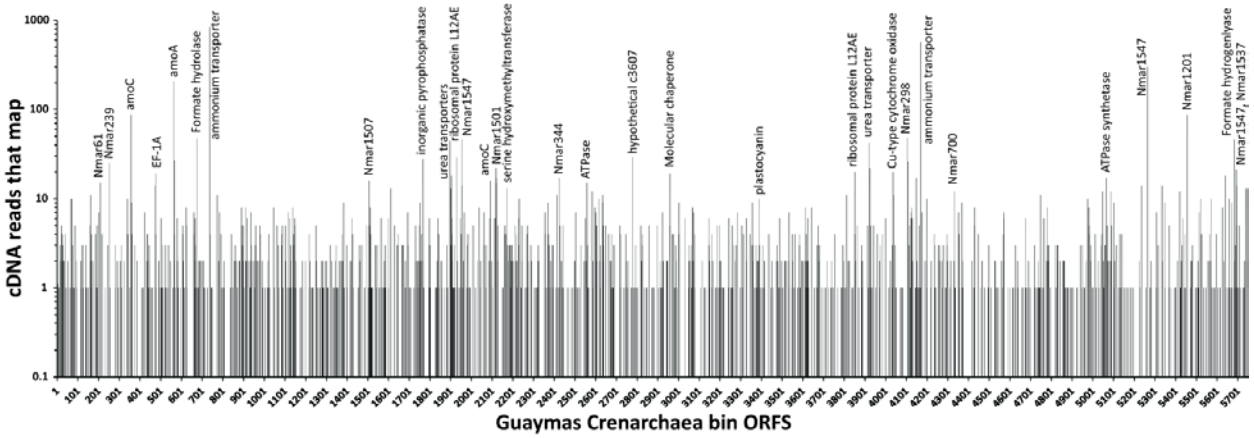

Figure 3.2. Abundance of raw (not normalized) transcripts mapped to genes in the Guaymas Basin MGI metagenomic bin (5,744 total genes). Predicted hypothetical proteins that have matches to *N. maritimus* genes are labeled "Nmar".

Enhancement of AOA in plumes and dominance over AOB. Several recent studies have investigated how the balance of AOA and AOB varies as a function of ammonium concentration

(Martens-Habbena *et al.* 2009, Verhamme *et al.* 2011). Hydrothermal inputs into the deep Guaymas Basin lead to ammonium concentrations of 0.2-3 µM in plumes (Lam, 2004), which spans the range proposed to delineate niches of AOA and AOB (Martens-Habbena *et al.*, 2009). We found that transcripts of MGI genes encoding ammonia monooxygenase (*amoA*) and an ammonium transporter were among the most abundant protein-coding transcripts in the deep Guaymas Basin microbial community (total of 405 and 1713 transcripts, respectively) and were more abundant in plume samples compared to background (Figure 3.3). In contrast, no bacterial ammonia monooxygenase genes were identified in any of the Guaymas Basin metagenomic or metatranscriptomic datasets (plume or background). This suggests that ammonia oxidation in the deep Gulf of California, including ammonium-enriched hydrothermal plumes, is dominated by AOA.

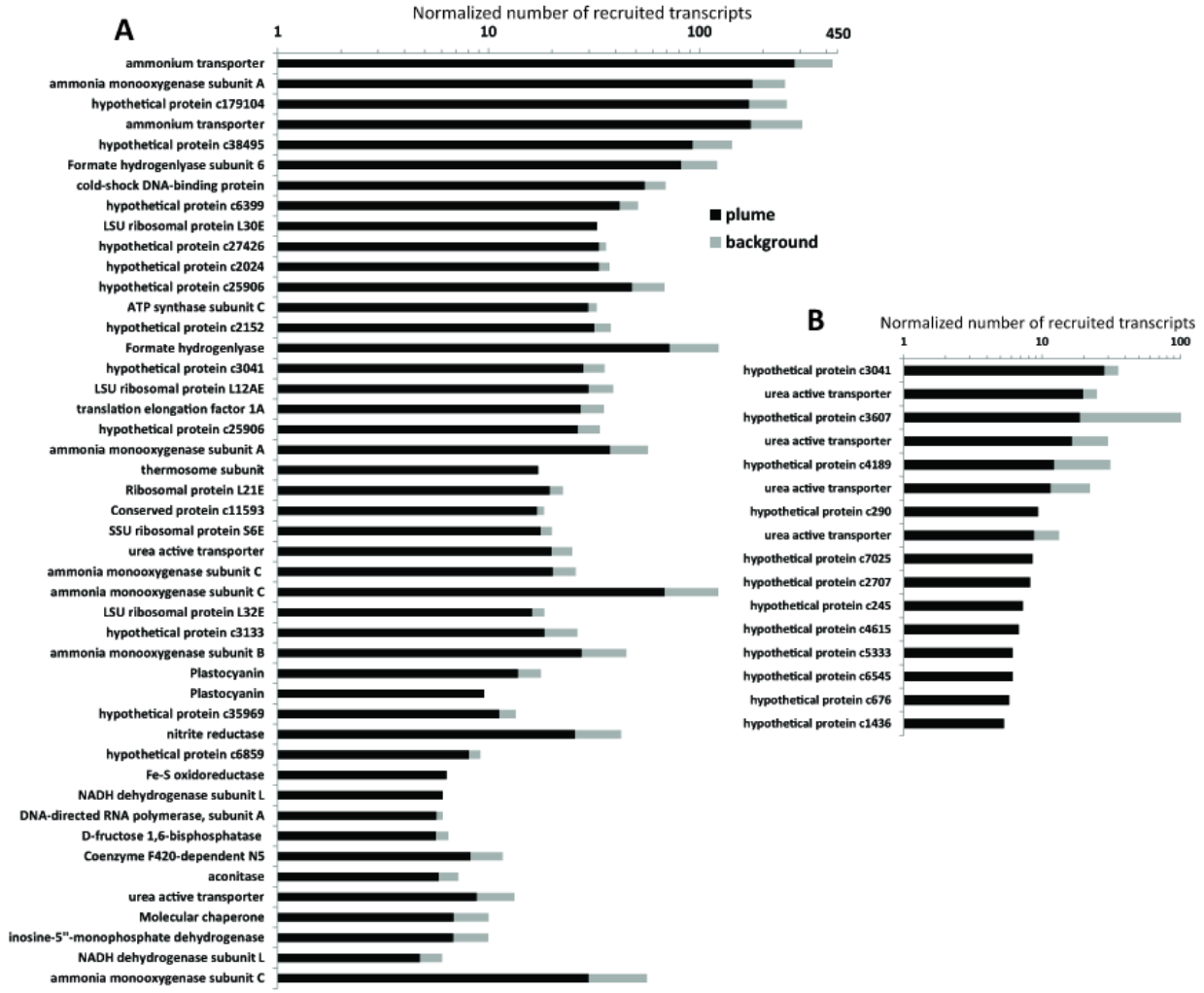

Figure 3.3. Stacked bar graph showing the number of transcripts recruited to MGI Archaea genes in the plume and background samples and sorted by difference between the plume and background recruited, with the greatest being at the top. Numbers are normalized to length of the genes as well as the total number of transcripts per sample (raw number of recruited divided by gene length and library size, then multiplied by a million to it comparable to the raw number of reads). (A) shows transcripts that are most abundant in the plume. (B) shows transcripts of genes not present in *Nitrosompumilus maritimus* that are most up-regulated in the plume.

Species-resolved transcriptomics of ammonia oxidation genes. Detailed analysis of *amoA* transcripts revealed dynamic transcription patterns of particular AOA populations. The Guaymas Basin metagenome contains 27 contigs that have amo genes from at least seven different genotypes (Figure 3.4). These well-assembled amo loci represent the dominant AOA genotypes present in the genomic data. To assess the ammonia-oxidizing transcriptional activity of each of these genotypes in ammonia-rich and ammonia-poor settings, we compared transcript recruitment from plume and background samples to all ammonia monooxygenase genes (*amoA*, *amoB*, *amoC*, and the *amo*-associated hypothetical) from all genotypes. Transcription of *amoA* genes from three of the abundant GB genotypes (c1374, c45409 and c51705) is dramatically higher in plume compared to background (Figure 3.4). Interestingly, there are several low-abundance variants that are highly active in the plume (*amoA* and hypothetical from c51705, *amoC* from c113214, and hypothetical from c225589).

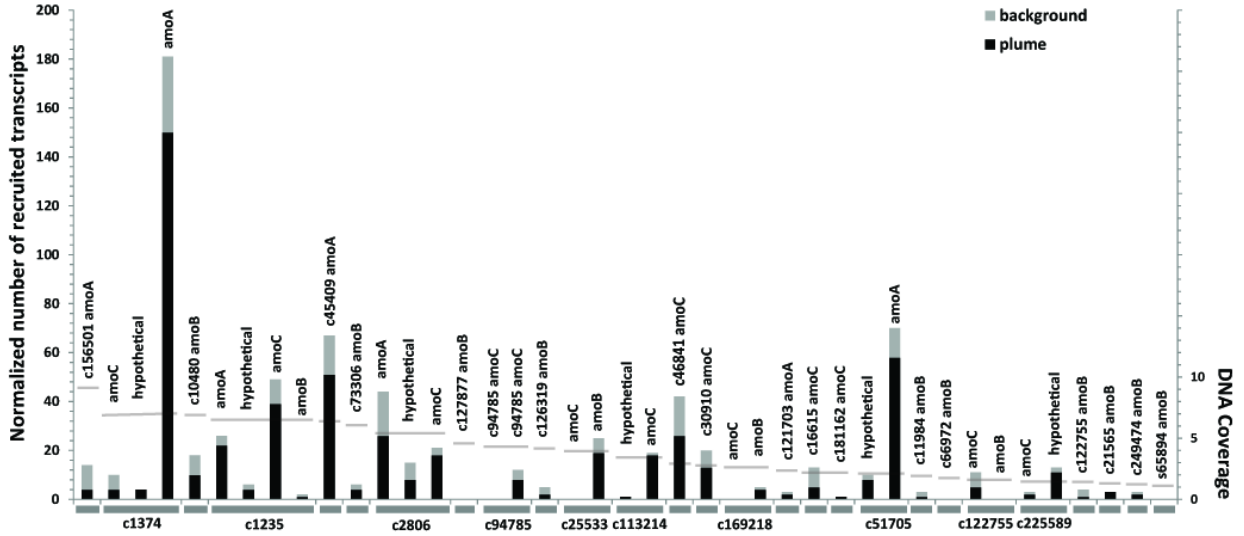

Figure 3.4. Transcript levels of sequence variants of ammonia monooxygenase genes in plume and background samples. Transcript numbers are normalized to gene length and library size. DNA fragments (contigs) with more than one gene are designated with grey bars on the x-axis. Individual genes are labeled on top. Thin grey horizontal lines indicate contig coverage in the genomic libraries (see scale on the right).

The four Guaymas Basin *amoA* variants that are most active in the plume (c1235, c1374, c45409, and c51705) fall within a tight phylogenetic group (Figure 3.5). Interestingly, the most abundant transcript type (c1374) is most closely related to a clone recovered from deep waters (2956 m) of the Japan Sea (Nakagawa *et al.* 2007). Furthermore, these deep-sea genotypes types are distinct from those that have been recovered from the upper 650 m of the water column at Guaymas and Carmen Basins (Beman et al. 2008). These Guaymas Basin gene sequences are >97% similar to one another and 92.3 to 93.4% similar to *N. maritimus*. Thus the genotypes that dominate *amoA* transcription in the deep Guaymas Basin likely represent strains of a novel species of *Nitrosopumilus*, a notion that is supported by sequence similarity and phylogeny of the dominant 16S rRNA genes (Figure S2). Our data suggests that expression of *amoA* genes from this deep GB group is enhanced in ammonium-rich hydrothermal plumes of the Guaymas Basin. Several other *amoA* sequences in this phylogenetic cluster were recovered from a site in the Arctic Ocean that has high ammonium concentrations (Kalanetra *et al.*, 2009). Taken together, this evidence reveals a cluster of MGI that thrives in geographically widespread ammonium-rich marine environments.

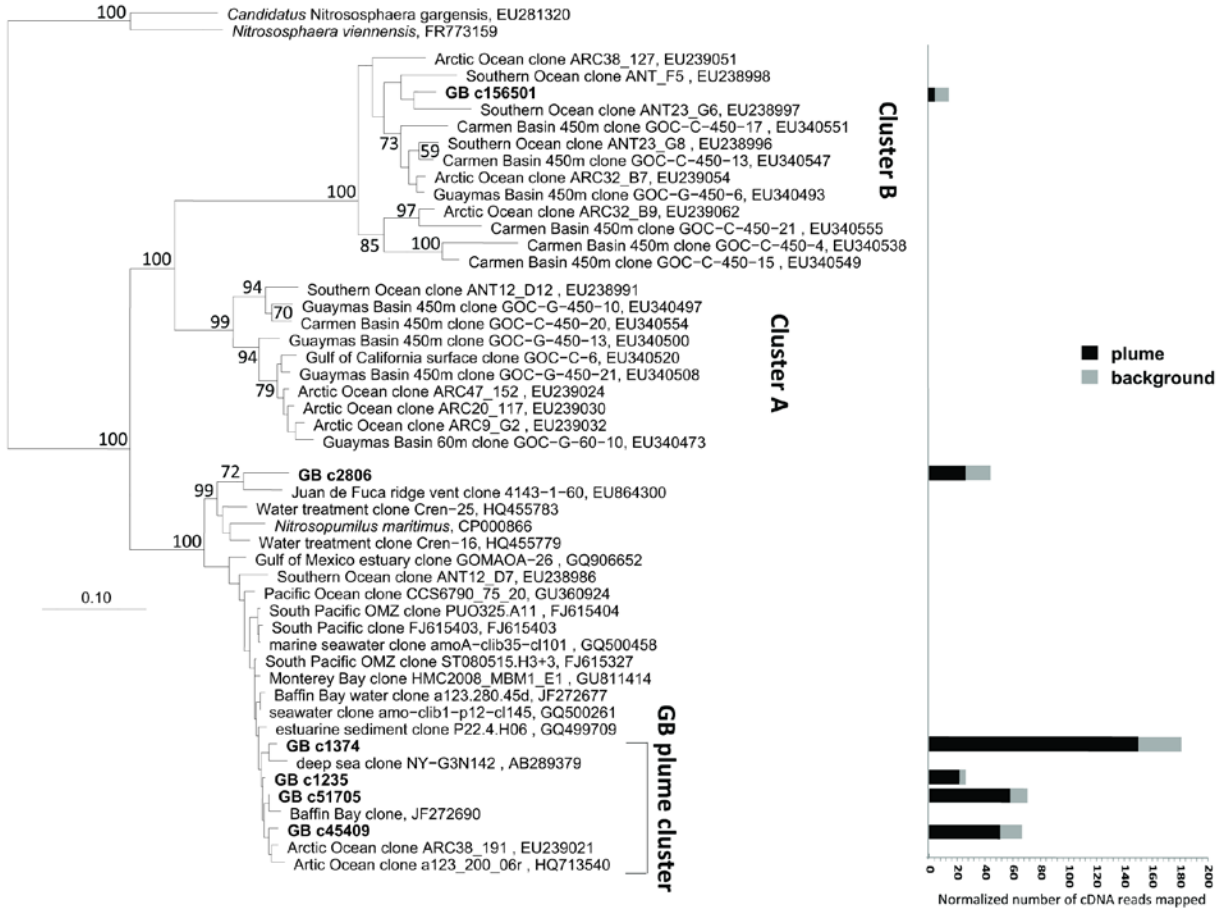

Figure 3.5. Phylogenetic tree of ammonia monooxygenase (amoA) genes and abundance of their transcripts in plume and background datasets. Recruitment numbers were normalized to gene lengths and the library sizes. Notice the Gulf of California sequences recovered from the surface waters (Beman *et al.* 2008) are not related to the types found in this study.

Genomic insights into the carbon metabolism of GB MGI. Given their abundance in the oceans and potential role in the carbon cycle, defining the carbon metabolism of MGI is an important yet unfinished task. Conflicting results leave open the question of whether individual MGI are capable of both heterotrophy and autotrophy or there are sub-groups that specialize in each. Whereas studies of cultures and surface waters indicate autotrophy, observations of a lower ratio of MGI *amoA* to 16S rRNA gene copies (Agogué *et al.*, 2008) and decreasing MGI carbon fixation with depth (Varela *et al.*, 2011) in the Atlantic suggest that deep-sea MGI are predominantly organoheterotrophic. Recent studies also show that the presence (and expression) of *amoA* genes does not necessarily indicate CO_2 fixation (Tourna *et al.*, 2011, Muβmann *et al.*,

2011). The Guaymas Basin metagenome contains genes homologous to *N. maritimus* genes encoding the 3-hydroxypropionate/4-hydroxybutyrate pathway for CO_2 fixation (Berg *et al.*, 2007), including 4-hydroxybutyryl-CoA dehydratase (HCD), methylmalonyl-CoA epimerase and mutase, and acetyl-CoA carboxylase (ACCase). Genes for ACCase recruited 25 transcripts (16 from plume and 9 from background) as did genes for methylmalonyl-CoA epimerase and mutase (19 plume and 6 background). Representation of these autotrophy genes in the metatranscriptomic data supports the idea that MGI fix CO_2 in the deep Gulf of California.

The Guaymas Basin MGI genomes contain 57 predicted ABC-type transporters for uptake of amino acids, which might be are an important source of carbon, nitrogen and energy for marine heterotrophs including MGI (Ouverney and Fuhrman, 2000, Fuhrman, 1987, Suttle *et al.*, 1991). Although this indicates genomic potential for heterotrophy in the Guaymas Basin MGI, these transporters recruited few or no transcripts ($\leq$2), suggesting that transcription of genes encoding MGI amino acid transporters was lower than those for carbon fixation.

Nitrogen and energy metabolism of GB MGI. Although MGI show high affinity for ammonium (Martens-Habbena *et al.*, 2009), low ammonium concentration still presents a potential bottleneck for energy metabolism of MGI. The Guaymas Basin MGI show evidence of several strategies for ammonia acquisition. First, genes encoding ammonium transporters are the most abundant protein-coding transcripts in the MGI metatranscriptome (Figure 2). Such high transcription of MGI ammonium transporters is consistent with prior observations from surface waters (Hollibaugh *et al*. 2011, Stewart *et al.*, 2011) and likely reflects the much higher concentration of ammonium than ammonia at seawater pH. The fact that ammonium transporters are the most highly expressed protein coding gene of deep GB MGI suggests that ammonium must be first transported into the cell for oxidation to occur. Regardless, this gene is clearly critical to the MGI's success in the community, and may account for their high N affinity (Martens-Habbena *et al.*, 2009).

Second, the deep-sea Guaymas Basin MGI metagenome contains three operons of *ure* genes for urea utilization. One genomic fragment (c229) has *ureE*, *ureF*, *ureG*, and *ureH* genes and another (c464) has *ureB*, *ureG*, and *ureE*, and urease-associated metallopeptidase genes. Two additional fragments contain urea active transporters and one of these has a second urease-associated metallopeptidase gene. It has been recently shown that the soil AOA isolate, *N.*

viennensis, is capable of growth on urea (Tourna *et al.*, 2011). Both *C. symbiosum* and *N. viennensis* contain urease genes (Tourna *et al.*, 2011, Hallum *et al.*, 2006), however *N. maritimus* lacks any recognizable genes for urea utilization. Thus our results and other recent environmental studies (*Konstantinidis et al.*, 2009; Tully *et al.*, 2012; Yakimov *et al.*, 2011) highlight an important difference in N acquisition between natural populations of MGI and *N. maritimus*.

All genes for the proposed AOA respiratory pathway (Walker *et al.*, 2010) are present in the GB genomic data except the plastocyanin-like subunit of complex III. Genes present include those encoding NADH dehydrogenase (NuoABCDHIJKMLN), ATP F_0F_1-type synthetase, complex III, multicopper oxidases (MCO), and the terminal oxidase (complex IV). Many of the respiratory pathway genes have multiple variants (up to seven) in the Guaymas Basin, but in nearly every instance one specific genotype recruited the majority of transcripts (see Table S2 for complete list).

Nitrite, the product of ammonia oxidation, inhibits growth of AOA (Tourna *et al.*, 2011). However, a recent study suggests that AOA reduce nitrite through a pathway known as "nitrifier-denitrification", resulting in globally significant production of nitrous oxide (N_2O), an important greenhouse gas (Santoro *et al.*, 2011). Although culture-based studies of MGI physiology have not demonstrated nitrite reduction, genes with homology to nitrite reductase (*nirK*) and several cupredoxin domain-containing MCOs thought to be involved in nitrite reduction were identified in the *N. maritimus* genome (Walker *et al.*, 2010). We identified single copies of *nirK*-like genes Nmar_1259 and Nmar_1667 in the Guaymas Basin genomic data. The Nmar_1259 *nirK* homolog (c632) is well-represented in both plumes (61 transcripts) and background (35 transcripts), whereas only a few transcripts of the Nmar_1667 homolog were detected. Nearly all the *nirK*-associated MCOs are also present in the GB MGI bin (Nmar_1354 was not found) but they are not expressed at significant levels. A potential source of electrons for nitrite reduction is formate (Ruiz-Herrera and DeMoss, 1969), which is likely present in Guaymas Basin plume. Some of the most abundant MGI transcripts that are highly enriched in the plume come from two variants of formate dehydrogenase (c1456 and c85331) that are highly similar (100% and 97%) to this protein from *N. maritimus* (Figure 3.3). Taken together, the evolutionary conservation and abundant transcriptional activity of this formate hydrogenase suggests that it serves a critical role in Guaymas Basin MGI. The overall magnitude and extensive enrichment of transcripts of

formate dehydrogenase and nitrite reductase genes that we observe in the Guaymas Basin plume implies that AOA actively reduce nitrite in these deep waters.

3.6 Conclusions

It is becoming increasingly apparent that MGI are widespread and globally significant players in the nitrogen and carbon cycles, yet the extent and implications of their influence are unclear due to questions surrounding their physiology and ecology. This is especially true for deep-sea MGI, which are numerically dominant but not well studied. In this study, *de novo* assembly of community genomic sequence provided a framework for investigating the activity of naturally occurring populations of MGI in the Gulf of California. This approach proved to be especially useful for differentiating transcriptional activity among closely-related genotypes. Additionally, it provided a catalog of genes not present in reference genomes, including those for urea utilization and many hypothetical genes.

Our findings show that the dominant Archaea in the deep Gulf of California are ammonia oxidizers. Archaeal genes for ammonia oxidation are among the most highly transcribed protein-coding genes in microbial communities inhabiting ammonium-enriched Guaymas Basin deep-sea hydrothermal plumes, suggesting vigorous MGI-mediated nitrification. This is surprising in light of the prevailing view that Bacteria tend to dominate at higher ammonium concentrations. Instead, we found a dominant clade of deep-sea AOA that thrive under ammonium-rich conditions, perhaps indicating that the marine AOA niche has a broader range of ammonia concentration than previously recognized. This group is closely related to *N. maritimus*, sharing with it the ability to oxidize ammonia and fix carbon, but is also characterized by genomic novelty reflecting important physiological differences such as acquisition of nitrogen via urea. These insights highlight populations of MGI Archaea in the deep Gulf of California that are distinct from those in surface waters and deep Atlantic waters, and that respond to geochemical perturbation in the plume environment.

Acknowledgements This project is funded in part by the Gordon and Betty Moore Foundation and the National Science Foundation (OCE 1029242). We also thank the University of Michigan Rackham Graduate School Faculty Research Fellowship Program for their support and Kathryn

Iverson and Sunit Jain for assistance with metagenomic analyses. DNA and cDNA sequencing was conducted in the laboratory of Stephan Schuster at Penn State University and we are grateful for the assistance of Lynn Tomsho. We also thank Dr. Cody Sheik and Karthik Anantharaman for their helpful comments and revisions to the manuscript.

3.7 Appendix A

CHAPTER III Supplementary Information

Contents

Table 3.1. Summary of samples characteristics and sequencing results.

Cast	Sample	Depth m	O_2 μM	T °C	Total RNA conc. ng/L[d]	No. sequencing reads[e]	
						MetaG.	MetaT.
11-2 #14	GD-1	1996	27.38	3.0	N.D.	640,069	-
11-1 #8	GD-2	1775	27.04	3.0	N.D.	616,464	-
21-6 #2	GD-5	1963	26.05	2.9	167/251	-	758,433
12-27a #1	GD-6	1950	27.7	3.0	203/319	-	954,569
12-8 #12	GD-7/9	1600	28.5	3.0	64/69	421,004	570,580

34-2 #7	GD-8/10	1900	46.46	2.6	57/82	462,850	546,704

[d]total RNA concentrations shown are results of two independent extractions for each sample; N.D. = not determined

[e]Number of pyrosequencing reads produced

**more information on the location and times of samples in Dick et al. 2009 and Dick & Tebo 2010.

Table 3.2. List of genes thought to be involved in *Nitrospumilus maritimus* ammonia oxidation electron transfer and their representation in the GB genomic assembly.

Function	homolog in N. maritimus	variants in GB assembly	fragment with dominant variant	number of plume cDNA reads	number of background cDNA reads
Multi-Copper Oxidases (MCO) with plastocyanin domains	Nmar_1226	4	c1026	5	0
	Nmar_1102	4	c4782	11	2
	Nmar_1665	1	c1407	0	0
	Nmar_1307	1	c197	0	0
	Nmar_1443	2	none	0	0
	Nmar_1142	4	c6328	2	1
NADH dehydrogenase					
NuoA	Nmar_0276	4	c586	1	1
NuoB	Nmar_0277	6	c586	12	10
NuoC	Nmar_0278	3	c2015	6	0
NuoD	Nmar_0279	4	c586	5	0
NuoH	Nmar_0280	7	c1640	4	5
NuoI	Nmar_0281	4	c1640	2	0
NuoJ	Nmar_0282	3	c1640	2	0
NuoK	Nmar_0283	3	c17106	2	0
NuoM	Nmar_0284	4	c1640	3	1
NuoL	Nmar_0285	3	c341	17	3
NuoN	Nmar_0286	3	c341	1	0
Complex III					
Plastocyanin-type subunit	Nmar_1542	0			
Transmembrane subunit	Nmar_1543	4	c49089	9	6
Rieske-type subunit	Nmar_1544	4	c7149	5	1
Complex IV terminal oxidase	Nmar_0182	4	c2483	5	2
	Nmar_0183	3	c78268	7	7
	Nmar_0184	3	c6181	15	11
	Nmar_0185	1	c6181	6	5
ATP F_0F_1 synthetase					
ATP sythetase C	Nmar_1688	3	c964	16	1
ATP sythetase D	Nmar_1689	4	c1542	2	0
ATP sythetase B	Nmar_1690	4	c336	4	0
ATP sythetase A	Nmar_1691	2	c336	13	5
ATP sythetase E	Nmar_1692	3	c1842	2	2
ATP sythetase I	Nmar_1693	6	c288	4	1
total	30	101		161	64

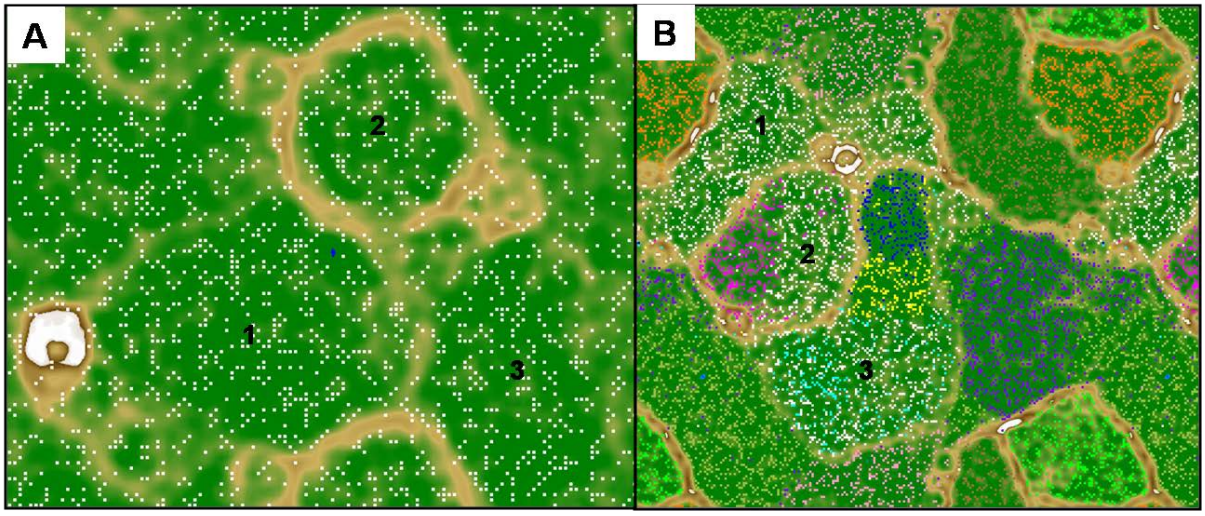

Figure 3.6. Assignment of metagenomic contigs to populations by binning with tetra-ESOM as described in Dick et al., 2009. Maps are continuous from top to bottom and side to side. White points represent metagenomic sequence fragments that have been clustered according to tetranucleotide frequency patterns. Background color indicates tetranucleotide frequency distance between data points; thus brown ridges represent large distances between fragments, delineating borders between genomes, whereas green represents short distances between fragments within genomes. Numbers indicate population based on presence of 16S rRNA genes on member contigs: (1) SAR324 *Deltaproteobacteria*; (2) *MGI*; (3) SUP05 *Gammaproteobacteria*. **A:** *de novo* clustering of GB plume and background contigs. **B:** GB contigs as in (A) with the addition of reference genomes to evaluate binning accuracy: *turquoise* – SUP05 contigs from (WALSH et al., 2009); *light green – Cenarchaem symbiosum*; *bright pink – Nitrosopumilis maritimus*; *blue – Ruthia magnifica*; *yellow – Calyptogena okutanii*; *orange – Methylococcus capsulatus*; *purple – Alteromonas macleodii*; *brown – Marinobacter algicola* DG893; *pale pink – Methylophaga thiooxidans* DMS010.

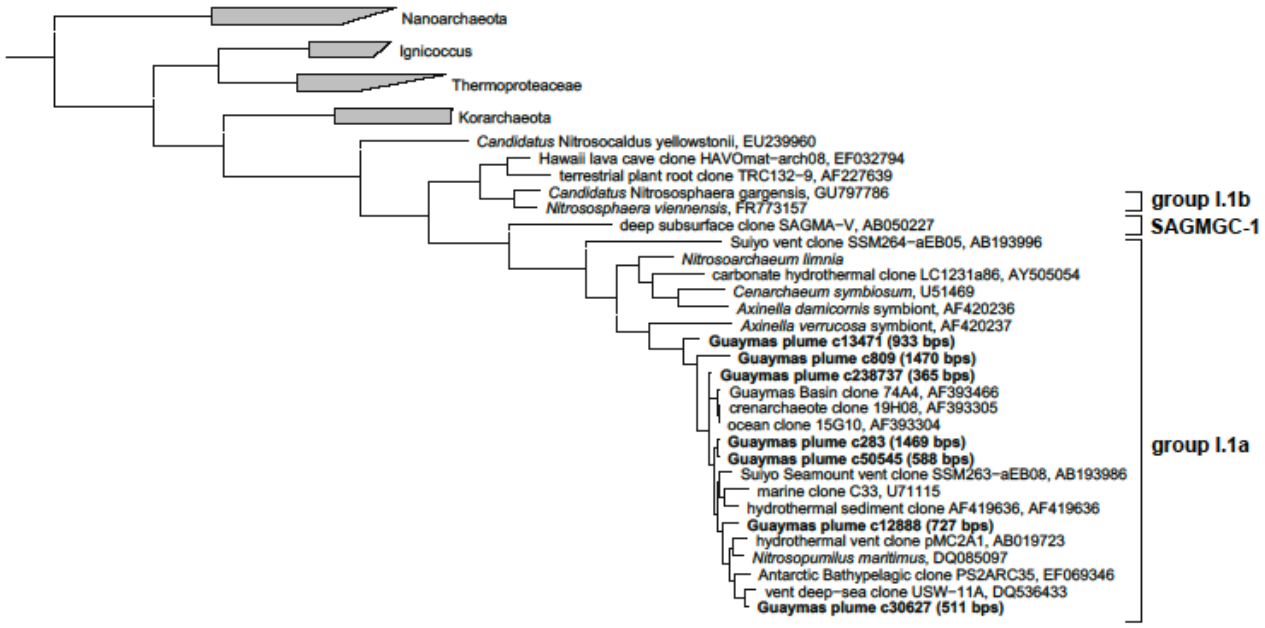

Figure 3.7. Phylogeny of 16S rRNA genes assembled from the Guaymas genomic library. The tree was generated using maximum likelihood method with *Aquifex pyrophilus* used as the outgroup. Note that all GB sequences cluster with *N. martimus*. The groups that have been tentatively namely by Pester et al (2011) are labeled on the right.

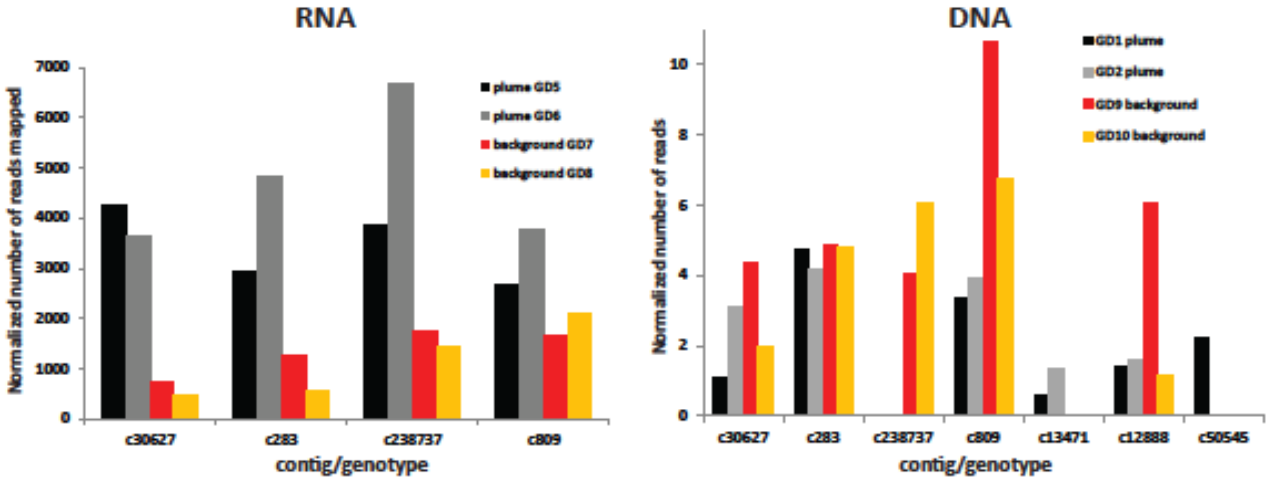

Figure 3.8. Comparison of the number of metatranscriptomic and metagenomic reads mapped from plume and background samples to the MGIC 16S rRNA genes. The RNA reads (left) were normalized to length 1166-1472, *E. coli* numbering, and the total number of transcripts. The DNA read counts (right) are normalized to the length of each fragment and the total number of reads from each library. Only four of the 16S rRNA genes have overlapping sequence used in this analysis, which is why there are fewer contigs for RNA recruitment (left). Replicates reads resulting from 454 artifacts were manually identified and removed from these analyses.

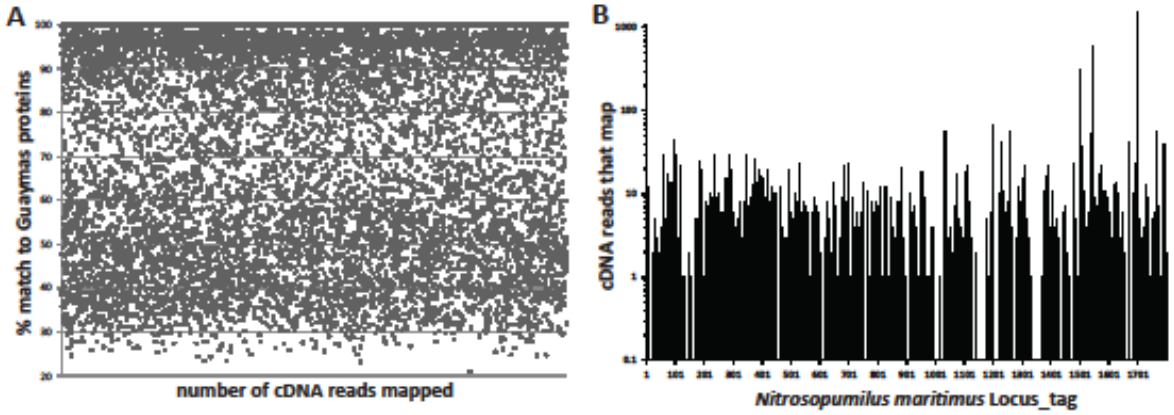

Figure 3.9. Mapping of GB plume and background transcript reads. (A) shows the percent match distribution of all transcripts that map to GB MGI proteins with bit scores >40 using BLASTx. (B) shows the number of transcripts (4067 total) that map to *Nitrosopumilus maritimus* proteins at more stringent parameters (bit score >45 and >70% similarity).

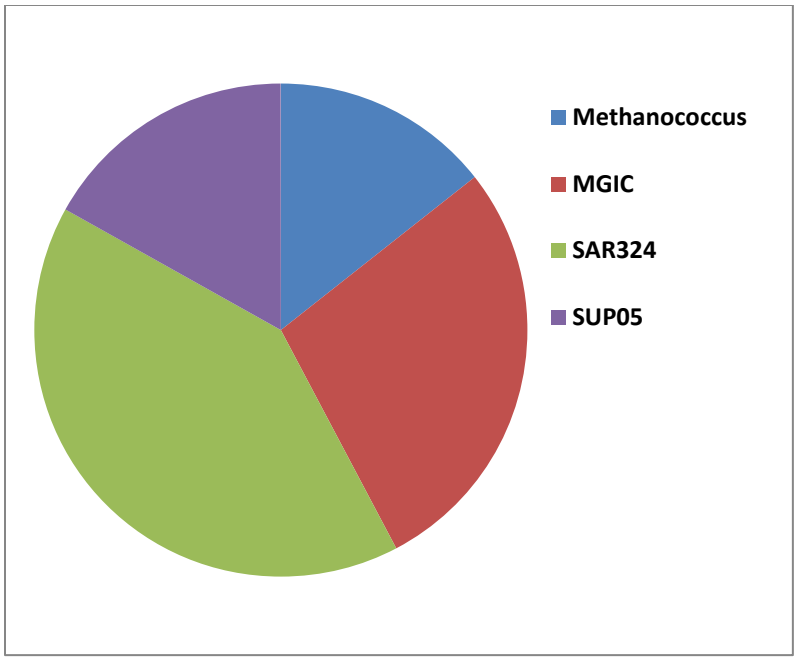

Figure 3.10. Recruitment of cDNA reads originally identified as MGIC (> 40 bit score but <70% ID) to binned fragments in the entire genomic Guaymas Basin assembly.

3.8 References

Agogué H, Brink M, Dinasquet J, Herndl GJ (2008) Major gradients in putatively nitrifying and non-nitrifying Archaea in the deep North Atlantic. *Nature* 456, 788-792.

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, et al. (2010) Enigmatic, ultra-small uncultivated Archaea. *Proc Natl Acad Sci* 107, 8806-8811

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

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

Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR (2011) Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analyses. *PLoS One* e16626 6, 1-12.

Brochier-Armanet C, Boussau B, Bribaldo S, Forterre P (2008) Mesophilic crenarchaeota, proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Rev. Microbiol*. 6, 245-252.

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.

De La Torre J, Walker CB, Ingalls AE, Konneke M, Stahl DA (2008) Cultivation of a thermophilic ammonia oxidizing archaeon synthesizing crenarchaeol. *Environ Microbiol* 10,810-818.

DeLong DF (1992) Archaea in costal marine environments. *Proc Natl Acad Sci* USA 89, 5685-5689.

DeLong EF, Wu KY, Prézelin BB, Jovine RVM (1994) High abundance of Archaea in Antarctic marine picoplankton. *Nature* 371, 695-697.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, *et al.* (2009a) Community-wide analysis of microbial genome sequence signatures. *Genome Biology*. 10, R85.

Dick GJ, Clement BG, Webb SM, Fodrie FJ, Bargar JR, Tebo BM (2009b) Enzymatic microbial Mn(II) oxidation and Mn biooxide production in the Guymas Basin deep-sea hydrothermal plume. *Geochim et Cosmo Acta* 73, 6517-6530.

Dick GJ, Tebo BM (2010) Microbial diversity and biogeochemistry of the Guaymas Basin deep-sea hydrothermal plume. *Environ Microbiol* 12, 1334-1347.

Francis CA, Roberts KF, Beman JM, Santoro AE, Oakley BB (2005) Ubiquity and diversity of ammonia-oxidizing Archaea in water columns and sediments of the ocean. *Proc. Natl. Acad. Sci USA* 102, 14683-14688.

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

Fuhrman JA (1987) Close coupling between release and uptake of dissolved free amino acids in seawater studied by an isotope dilution approach. *Mar Ecol* 37, 45-52.

Fuhrman JA, McCallum K, Davis AA (1992) Novel major archaebacterial group from marine plankton. *Nature* 356, 148-149.

Gordon D, Abajian C, Green P (1998) Consed, a graphical tool for sequence finishing. Genome Research 8, 186-194.

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

Hallam SJ, Konstantinidis KT, Putnam N, Schleper C, Watanabe Y, Sugahara J, *et al*. (2006) Genomic analysis of the uncultivated marine crenarchaeote *Cenarchaeum symbiosum*. *Proc Natl Acad Sci* 103, 18296-18301.

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

Ingalls AE, Shah SR, Hansman RL, Aluwihare, Santos GM, Drufel ERM, Pearson A (2006) Quantifying archaeal community autotrophy in the mesopelagic using natural radiocarbon. *Proc Natl Acad Sci* 103, 6442-6447.

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

Kryzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ and Marra MA (2009) Circos, An information aesthetic for comparative genomics. *Genomic Res* 19, 1639-1645.

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

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

Lesniewski R, Jain S, Anantharaman K, Schloss PD, Dick GJ (2011) The metatranscriptome of the Guaymas Basin deep-sea hydrothermal plume is dominated by water column methanotrophs and chemolithoautotrophs. *ISME J* 6, 2257-2268.

Martens-Habbena W, Berube PM, Urakawa H, de la Torre JR, Stahl DA (2009) Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* 461, 976-981.

Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009) Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific Subtropical Gyre. *Appl Environ Microbiol* 75, 5345-5355.

Lam P (2004) Microbial ammonia oxidation in deep-sea hydrothermal plumes. PhD Thesis. Hawaii, USA, University of Hawaii.

Lilley MD, Betterfield DA, Olson EJ, Lupton JE, Macko SA, McDuff RE (1993) Anomalous CH_4 and NH_4^+ concentrations at an unsedimented mid-ocean-ridge hydrothermal system. *Nature* 364, 45-47.

Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC (2009) IMG ER, a system for microbial genome annotation expert review and curation. Bioinformatics 25, 2271-2278.

Molina V, Belmar L, Ulloa O (2010) High diversity of ammonia-oxidizing archaea in permanent and seasonal oxygen-deficient waters of the eastern South Pacific. *Environ Microbiol* 12, 2450-2465.

Muβmann M, Brito I, Pitcher A, Sinninghe Damste JS, Hatzenpichler R, Richter A, *et al.* (2011) Thaumarchaeotes abundant in refinery nitrifying sludges express *amoA* but are not obligate autotrophic ammonia oxidizers. *Proc Natl Acad Sci.* in press.

Nakagawa T, Mori K, Kato C, Takahashi R, Tokuyama T (2007) Distribution of cold-adapted ammonia-oxidizing microorganisms in the deep-ocean of the northestern Japan Sea. *Microbes Environ*. 4, 365-372.

Ouverney CC, Fuhrman JA (2000) Marine planktonic archaea take up amino acids. *Appl Environ Microbiol* 66, 4829-4833.

Pester M, Schleper, C, Wagner M (2011) The Thaumarchaeota, an emerging view of their phylogeny and ecophysiology. *Curr Opin Microbiol* 14, 300-306.

Ruiz-Herrera J, DeMoss JA (1969) Nitrate Reductase Complex of Escherichia coli K-12, Participation of specific formate dehydrogenase and cytochrome b1 components in nitrate reduction. *J Bacteriol* 99, 720-729.

Santoro AE, Buchwald C, McIlvin MR, Cascitti KL (2011) Isotopic signature of N_2O produced by marine ammonia-oxidizing Archaea. Science 333, 1282-1285.

Suttle CA, Chan AM, Fuhrman JA (1991) Dissolved free amino acids in the Sargasso Sea, uptake and repiration rates, turnover times, and concentrations. *Mar Ecol Prog Ser* 70, 189-199.

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

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

Stewart FJ, Ulloa O, DeLong EF (2011) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* 14, 23-40.

Takai K, Oida H, Suzuki Y, Hirayama H, Nakagawa S, Nunoura T, *et al.* (2004) Spatial distribution of marine Crenarchaeota group I in the vicinity of deep-sea hydrothermal systems. *Appl. Environ. Microbiol.* 70, 2404-2413.

Tourna M, Stieglmeier M, Spang A, Konneke M, Schintlmeister A, Urich T, *et al.* (2011) *Nitrosophaera viennensis*, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci USA* 108, 8420-8425.

Tully BJ, Nelson WC, Heidelberg JF (2012) Metagenomic analysis of a complex marine planktonic thaumarchaea community from the Gulf of Maine. *Environ. Microbiol.* Online advance of publication.

Varela MM, van Aken HM, Sintes E, Reinthaler T, Herndl GJ (2011) Contribution of Crenarchaeota and Bacteria to autotrophy in the North Atlantic interior. *Environ Microbiol* 13, 1524-1533.

Verhamme DT, Prosser JI, Nicol GW (2011) Ammonia concentration determines differential growth of ammonia-oxidising archaea and bacteria in soil microcosms. *ISME J* 5, 1067-1071.

Von Damm KL, Edmond JM, Measures CI, Grant B (1985) Chemistry of submarine hydrothermal solutions at Guaymas Basin, Gulf of California. *Geochim et Cosmochim Acta* 49, 2221-2237.

Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, *et al.* (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci* 107, 8818-8823.

Wuchter C, Abbas B, Coolen MJL, Herfort L, van Bleijswijk J, Timmers P, *et al.* (2006) Archaeal nitrification in the ocean. *Proc Natl Acad Sci* 103, 12317-12322.

Yakimov MM, La Cono V, Smedile F, DeLuca TH, Juarez S, Ciordia S, *et al.* (2011) Contribution of crenarchaeal autotrophic ammonia oxidizers to the dark primary production of Tyrrhanian deep waters (Central Mediterranean Sea). *ISME J.* 5, 945-961.

CHAPTER IV

COMMUNITY TRANSCRIPTOMIC ASSEMBLY REVEALS MICROBES THAT CONTRIBUTE TO DEEP-SEA CARBON AND NITROGEN CYCLING

Brett J. Baker[a], Cody Sheik[a], Chris Taylor[b], Sunit Jain[a], Ashwini Bhasi[b], James D. Cavalcoli[b], and Gregory J. Dick[a,b]

a. Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA
b. Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

Originally published in 2013 in *ISME Journal* 7:1962-1973

4.1 Abstract

The deep ocean is an important component of global biogeochemical cycles because it contains one of the largest pools of reactive carbon and nitrogen on Earth. However, the microbial communities that drive deep-sea geochemistry are vastly unexplored. Metatranscriptomics offers new windows into these communities, but it has been hampered by reliance on genome databases for interpretation. We reconstructed the transcriptomes of microbial populations from Guaymas Basin, in the deep Gulf of California, through shotgun sequencing and *de novo* assembly of total community RNA. Many of the resulting mRNA contigs contain multiple genes, reflecting co-transcription of operons, including those from dominant members. Also prevalent were transcripts with only limited representation (2.8X coverage) in a corresponding metagenome, including a considerable portion (1.2Mb total assembled mRNA sequence) with similarity (96%) to a marine heterotroph, *Alteromonas macleodii*. This *Alteromonas* and euryarchaeal marine

group II (MGII) populations displayed abundant transcripts from amino acid transporters, suggesting recycling of organic carbon and nitrogen from amino acids. Also among the most abundant mRNAs were catalytic subunits of the nitrite oxidoreductase (NXR) complex and electron transfer components involved in nitrite oxidation. These and other novel genes are related to novel *Nitrospirae* and have limited representation in accompanying metagenomic data. High throughput sequencing of 16S rRNA genes and rRNA read counts confirmed that *Nitrospirae* are low abundance yet widespread members of deep-sea communities. These results implicate a novel bacterial group in deep sea nitrite oxidation, the second step of nitrification. This study highlights metatranscriptomic assembly as a valuable approach to study microbial communities.

4.2 Introduction

Microorganisms mediate the marine carbon and nitrogen cycles and thus control nutrient bioavailability, primary productivity, and production and consumption of greenhouse gases such as N_2O and CO_2 in the oceans (Füssel *et al*., 2012; Karl *et al*., 2012; Ward *et al*., 2007). The deep ocean represents the largest active reservoir of carbon on the planet, containing ~50 times more inorganic carbon than the atmosphere (Raven and Falkowski, 1999). Thus, understanding the primary agents of carbon cycling in the deep sea is of considerable interest. The “biological pump” has been considered a driving force for sequestration of carbon to the ocean interior (Raven and Falkowski, 1999) and the “microbial carbon pump”, in which heterotrophic bacteria generate recalcitrant dissolved organic carbon, represents a more recently recognized form of carbon sequestration (Jiao and Zheng, 2011). Therefore, elucidating the key microbial players that mediate interconversions between dissolved inorganic carbon, particulate organic carbon, and dissolved organic carbon is crucial to understand the carbon cycle as it pertains to global change.

Because nitrogen is often the co-limiting nutrient for productivity in the oceans, the carbon cycle is intimately linked to biogeochemical transformations of nitrogen (Zehr and Kudela, 2011). Recent transformative advances in environmental DNA sequencing have revealed the pathways, organisms, and genes involved in the nitrogen cycle including anaerobic ammonia oxidation (anammox) (Strous *et al*., 2006), denitrification (Ward *et al*., 2007), N_2O

production (Santoro *et al*., 2011), and ammonia oxidation (Könneke *et al*., 2005). Ammonia-oxidizing Archaea (AOA) are now recognized as major contributors to oceanic nitrification (Wuchter *et al*., 2006) by catalyzing the first step, oxidation of ammonia to nitrite (Könneke *et al*., 2005). These AOA are numerically abundant, especially in the deep sea, where they account for up to 40% of total cells (Karner *et al*., 2001), thus they have been estimated to be among the most abundant Archaea on Earth (Pester *et al*., 2011). Despite these new insights, fundamental questions about the marine nitrogen cycle remain open. For example, because nitrite produced by AOA typically does not accumulate in the environment (Dore and Karl, 1996), nitrite oxidation must be equally prevalent as AOA in nitrification (Ward *et al*., 2007). Correlation between nitrite-oxidizing bacteria (NOB) and ammonia oxidizing Archaea populations suggests metabolic coupling between these groups (Mincer *et al*., 2007; Santoro *et al*., 2010), yet NOB are observed at much lower abundance than their ammonia-oxidizing counterparts (Mincer *et al*., 2007; Santoro *et al*., 2010; Koops and Pommerening-Roser, 2001). This high AOA:NOB ratio is unexplained even when the greater free energy available from ammonia oxidation is taken into account, thus the mechanisms and organisms responsible for nitrite removal remain unresolved (Ward *et al*. 2007, Zehr and Kudela, 2011). Previously unrecognized nitrite reduction by AOA has recently been highlighted as another potential sink for nitrite (Santoro *et al*., 2010; Baker *et al*., 2012).

Metatranscriptomics is emerging as a valuable tool for tracking the metabolic activity of microbial communities as they occur in nature. Although the relationship between the abundance of RNA and protein is not simple, thus complicating efforts to use transcript abundance as a direct proxy for metabolic activity, metatranscriptomics still provides highly informative views of the interactions between microbes and their environments (Moran *et al*., 2012). This approach offers the ability to sequence and quantify mRNA of specific genes and populations within an entire community, potentially including those that have not been previously identified. To date, analysis of metatranscriptomic sequence data has primarily relied on mapping of cDNA reads to genomic datasets derived from either public databases (Frias-Lopez *et al*., 2007; Stewart *et al*., 2011; Shi *et al*., 2009) or from accompanying metagenomic sequencing (Shi *et al*., 2011; Lesniewski *et al*., 2012). These approaches are limited by reference datasets that lack the full diversity inherent to natural communities and by public databases that are biased towards readily cultured representatives. Thus a large fraction of metatranscriptomic

data is typically unclassified (Frias-Lopez *et al*., 2007). Here we attempt to resolve the metabolic activity of novel and minor community members through *de novo* assembly of metatranscriptomic sequence reads from a hydrothermal plume in Guaymas Basin, Gulf of California, where enhanced primary production is fueled by ammonia oxidation, methanotrophy, and sulfur oxidation (Lesniewski *et al*., 2012). Reconstruction of transcriptomes of deep-sea community members enabled identification of abundant transcripts involved in nitrite oxidation and carbon cycling from organisms with limited representation in metagenomic datasets.

4.3 Materials and Methods

Sample collection and processing. Samples were collected in 10L bottles by CTD-Rosette aboard the R/V New Horizon (Table 1) as described previously (Dick and Tebo, 2010). Briefly, samples were collected by "tow-yo" of the CTD-rosette then immediately filtered onto 0.2 um polycarbonate membranes with N_2 gas once on deck and preserved in RNAlater (Ambion). Although potential changes in the RNA pool during collection are a concern, as discussed previously (Lesniewski *et al*., 2012), these changes are minimized by the fact that samples are kept under *in situ* conditions (cold, dark) throughout collection and immediately filtered and preserved once onboard. RNA was extracted from filters using the MirVana miRNA Isolation kit (Ambion) and treated with DNAase I, and concentrated and re-purified using RNeasy MinElut Kit. RNA amplification by random priming and cDNA synthesis was performed as described previously (Stewart *et al*., 2011; Shi *et al*., 2009). Sequencing performed on an Illumina HiSeq2000 instrument at the University of Michigan Sequencing Core.

rRNA-based taxonomy abundance assessment. In order to assess the diversity of organisms present in the RNA dataset we mapped cDNA reads from the plume community to the SILVA small subunit (SSU) rRNA gene database (Pruesse *et al*., 2007) using riboPicker software (Schmieder *et al*., 2012).

cDNA sequencing and analyses. cDNA reads were dereplicated by removing duplicated reads (100% match, identical length), then quality trimmed using Sickle (github.com/najoshi/sickle). For all read mapping we used trimmed and dereplicated datasets. Dereplication reduced the

number of reads from 206 to 45 million in the plume and 245 to 130 million in the background sample. We also mapped reads prior to dereplication to confirm that the general trends seen were not artifacts of preprocessing of the sequences. Reads were assembled with Velvet (1.2.01) and subsequently processed using the transcriptomic assembler Oases (0.2.04) (Schulz *et al*., 2012). Abundance of cDNA reads was determined by mapping all of the cDNA reads to the assembled transcripts fragments. Mapping was done using BWA (Li and Durbin, 2009) with default settings (maximum mismatch = 4%). We manually checked the mRNA transcripts discussed in depth here for chimeras by viewing the read mapping in Integrated Genome Viewer (IGV). The trends reported for the NOB and other low abundance members were observed in that analysis as well. Assembled transcript contigs were searched for functions using DOE JGI IMG/MER annotation pipeline (Markowitz *et al*., 2012). The cDNA reads are available at NCBI SRA under accession numbers SRX134769 (plume) and SRX134768 (background). The assembled and annotation plume transcript library is available via IMG under taxon object ID 236347000. All comparisons of cDNA assemblies with metagenomic data was done with previously described data Lesniewski *et al*. (2012), which was a co-assembly of reads from the same sample (Bkgrd-1), and additional ones (Plume-1 and -2, and Bkgrd-2).

Phylogenetic analyses. All phylogenetic trees were generated using maximum likelihood (RaxML) with ARB software (Ludwig *et al*., 2004). rRNA-containing transcripts were identified using riboPicker package (Schmieder *et al*., 2012). 16S rRNA sequences were aligned in GreenGenes (DeSantis *et al*., 2006). Alignments of mRNA sequences were done using CLUSTALW with manual refinement. In order to identify all of the 16S and 23S rRNA sequences in the transcript assembly we first searched the plume assembly with *Ca*. Nitrospira defluvii 16S and 23S rRNA genes. Matches were then imported and aligned to the Greengenes 16S rRNA and the Silva 23S rRNA databases (DeSantis *et al*. 2006). We then generated large neighbor joining trees with thousands of reference sequences. Only contigs larger than 350 bp were used in the 16S tree and only those larger than 500 were used in the 23S rRNA tree. Only those sequences that fell within the *Nitrospirae* were kept. The 23S rRNA phylogeny was generated using 1909 characters. Group names in the 16S rRNA tree are based on those characterized by Lebedeva *et al*. (2011).

Identification of NOB and anammox transcripts. We searched all annotated genes on the Guaymas mRNA transcripts using all of the *Ca*. N. defluvii genes. We then compared these hits to the non-redundant NCBI protein database. Only those that had top hits to *Ca*. N. defluvii and *Leptospirillum* spp. were then considered to belong to the *Nitrospirae*. We searched the metatranscriptomic assembly for transcripts of key anammox genes, hydrazine oxidoreductase (*hzo*) from the genome of *Ca.* Kuenenia stuttgartiensis and hydrazine hydrolase from *Candidatus* Scalindua sp. (FM163627). Supplementary Figure S7 is based on the number of reads that match with e-value $<1E^{-5}$ (BLASTx). One gene transcript, 2236391221, had a match of 57% (bitscore of 56.6, e-value $7E^{-10}$). However, comparison of this transcript to Genbank revealed that it is most similar to several sequences obtained from microbes not thought to be capable of anammox, including *Shewanella woodyi* (79% similarity), and SUP05 (68% similarity).

Analyses of transcript sequence variants. cDNA reads were mapped to assembled contigs using BWA mapping software (Li and Durbin, 2009). SNPs were identified by visually comparing reads mapped using IGV.

16S rRNA gene pyrosequencing. DNA was extracted from ¼ of a filter with the MoBio PowerSoil DNA isolation kit (Carlsbad, CA, USA). In addition to bead beating, filters were incubated at 65^{o}C for 20 min to facilitate cellular lysis. Bead beating was performed using the MP-Bio FastPrep-24 (Santa Ana, CA, USA) for 45 seconds at setting 6.5. 16S rRNA genes were amplified in triplicate 25 µL reactions containing the following (final concentration): 12.5 µL 5 Prime HotMasterMix (Gaithersburg, MD, USA), 2 µL (15 µM) each forward and reverse primers, 1 µL community DNA. PCR thermocycler conditions were as follows: initial denaturation 95^{o}C -4 min followed by 30 rounds of 95^{o}C for 30 sec, 50^{o}C for 1 min, 72^{o}C for 1 min and final elongation 72^{o}C for 10 min. Triplicate PCRs were combined and cleaned using a MoBio UltraClean PCR Clean-up kit. DNA was quantified using PicoGreen (Invitrogen, Carlsbad, CA, USA). Previously described 16S rRNA gene primers targeting the V4 region (515F/806R) (Bates *et al*., 2010) were used with reverse primers containing a 12-base barcode. Individual barcoded samples were combined into a single sample at equivalent concentrations then sent to Engencore (http://engencore.sc.edu) for pyrosequencing using Titanium chemistry. Sequences were error corrected with Pyronoise (Quince *et al*., 2009) implemented in Mothur (v

1.25.0) (Schloss *et al*., 2009). Species level OTUs were binned at 97% similarity and chimera checked using the OTUpipe (http://drive5.com/otupipe) command within QIIME (Caporaso *et al.*, 2010). Default parameters were used with the exception of low abundance OTUs being kept for downstream analysis. OTUs were taxonomically classified with BLASTn (Altschul *et al*., 1990) (ver 2.2.22, e-values cutoff 10^{-8}) using Greengenes taxonomy (available at http://qiime.wordpress.com) and customized to include NOB 16S rRNA sequences recovered from Guaymas Basin transcriptomic libraries.

4.4 Results and Discussion

De novo assembly of transcripts. Random shotgun metatranscriptomic sequencing was conducted on a sample from the Guaymas Basin hydrothermal plume (1950 m water depth) and from a location just above the plume (1600 m), referred to as "background", for comparison. *De novo* assembly of metatranscriptomic reads yielded 78,250 assembled contiguous sequences (contigs) containing 81,452 predicted genes. 18,501 (23%) of these were putative protein coding genes (non-tRNA or rRNA) of which 12,605 (68%) were assigned putative functions. The large number of non-protein coding transcripts can be explained by a high level of fragmentation (due to fine-scale variability in highly similar sequences) of rRNA genes that is commonly seen in short-read sequencing data (Miller *et al*. 2011). Several of the mRNA contigs have homology to multiple genes of related function, reflecting assembly of co-transcribed genes from operons (Supplementary Figure S1). Among the most abundant were transcripts involved in oxidation of sulfur (*sox*), ammonia (*amo*), and methane (*pmo*) from dominant community members. Also highly expressed were genes encoding ribosomal proteins (Figure S2) from dominant groups, including ammonia-oxidizing Archaea (Baker *et al*., 2012), sulfur-oxidizing SUP05 *Gammaproteobacteria*, and methanotrophs (Lesniewski *et al*., 2012). These results are consistent with previous analyses of the same samples based on genome databases (Lesniewski *et al*. 2012). The majority (53%) of the 16S rRNA containing reads belong to members of the *Gammaproteobacteria* (including methanotrophs and the sulfur-oxidizer, SUP05) (Figure 4.1). This is consistent with previous findings (Lesniewski *et al*. 2012; Dick and Tebo, 2010) and their high coverage in genomic assemblies.

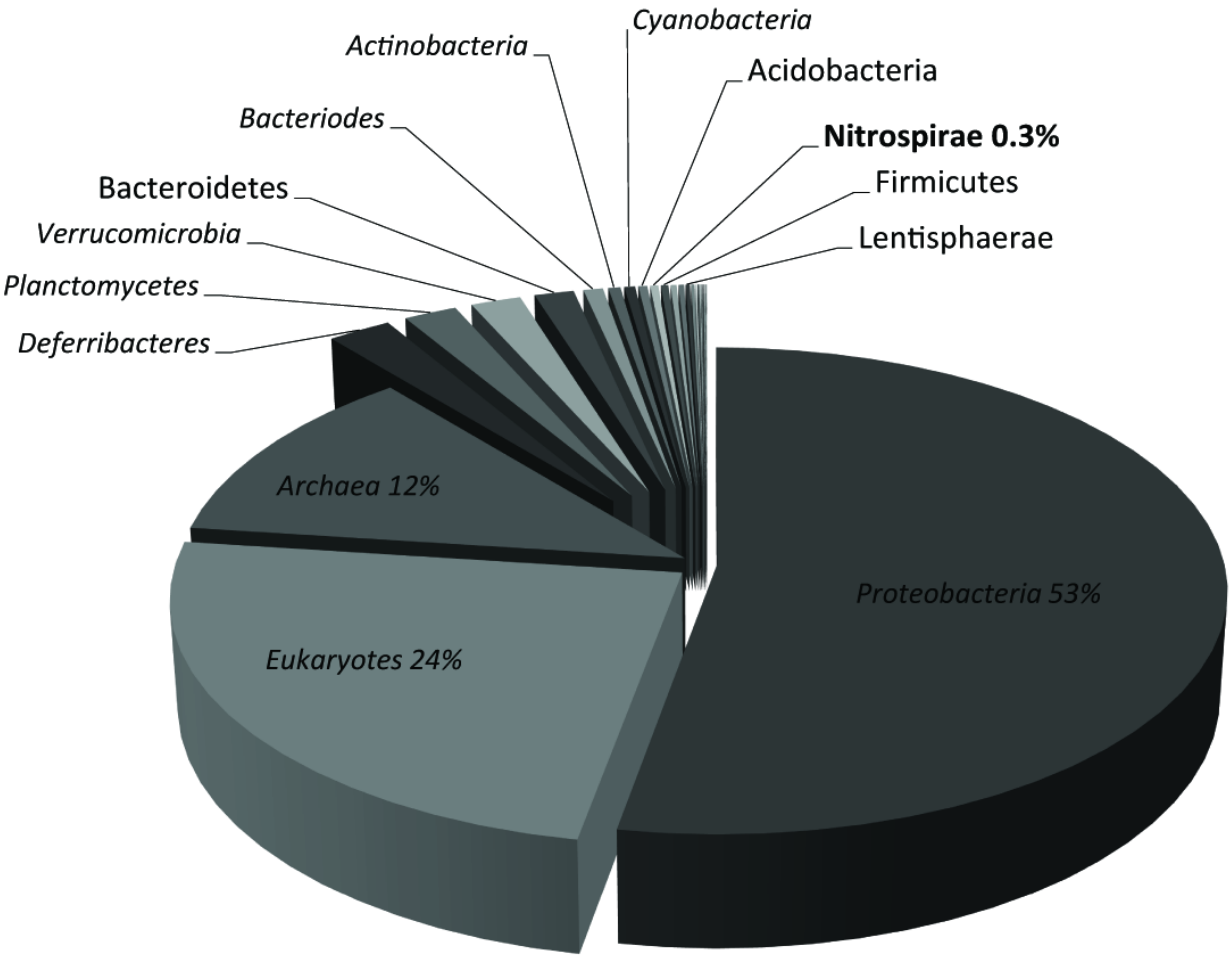

Figure 4.1. Abundance of major phyla based on classification of rRNA transcript reads. All 16S rRNA reads (total of 14,571,562) were mapped (>75% over half the read length cutoff) to a 16S rRNA gene database (SILVA).

Heterotrophy. Metatranscriptomic assembly revealed abundant transcripts from community members that were not well-represented in corresponding metagenomic datasets (Lesniewski *et al*. 2012). The most abundant ribosomal protein-coding transcripts were highly similar (up to 99% DNA similarity) to a deep-sea heterotroph, *Alteromonas macleodii* (Ivars-Martinez *et al*., 2008) (Supplementary Figure S2), that has limited representation in corresponding metagenomic data (averaging 2.8X coverage) and 16S rRNA gene surveys (Dick and Tebo, 2010). The longest of these assembled transcripts is nearly 5 kb and contains an operon of 11 co-transcribed ribosomal protein genes. Overall, 1968 mRNA contigs were identified totaling just over 1.2Mb of consensus sequence (Figure 4.2), with an average similarity of 96% to *A. macleodii*. These

transcripts are generally less abundant in the background compared to plume (Figure 4.2). The four most abundant *Alteromonas*-like transcripts are for TonB, an amino acid transporter, ribosomal protein S2, and a hypothetical protein (Figure 4.2). TonB is a membrane-bound receptor that is commonly involved in iron uptake systems in a variety of bacteria. However, it has been shown that this protein family is also involved in transport of other metals and various carbohydrates (Schauer *et al*., 2008).

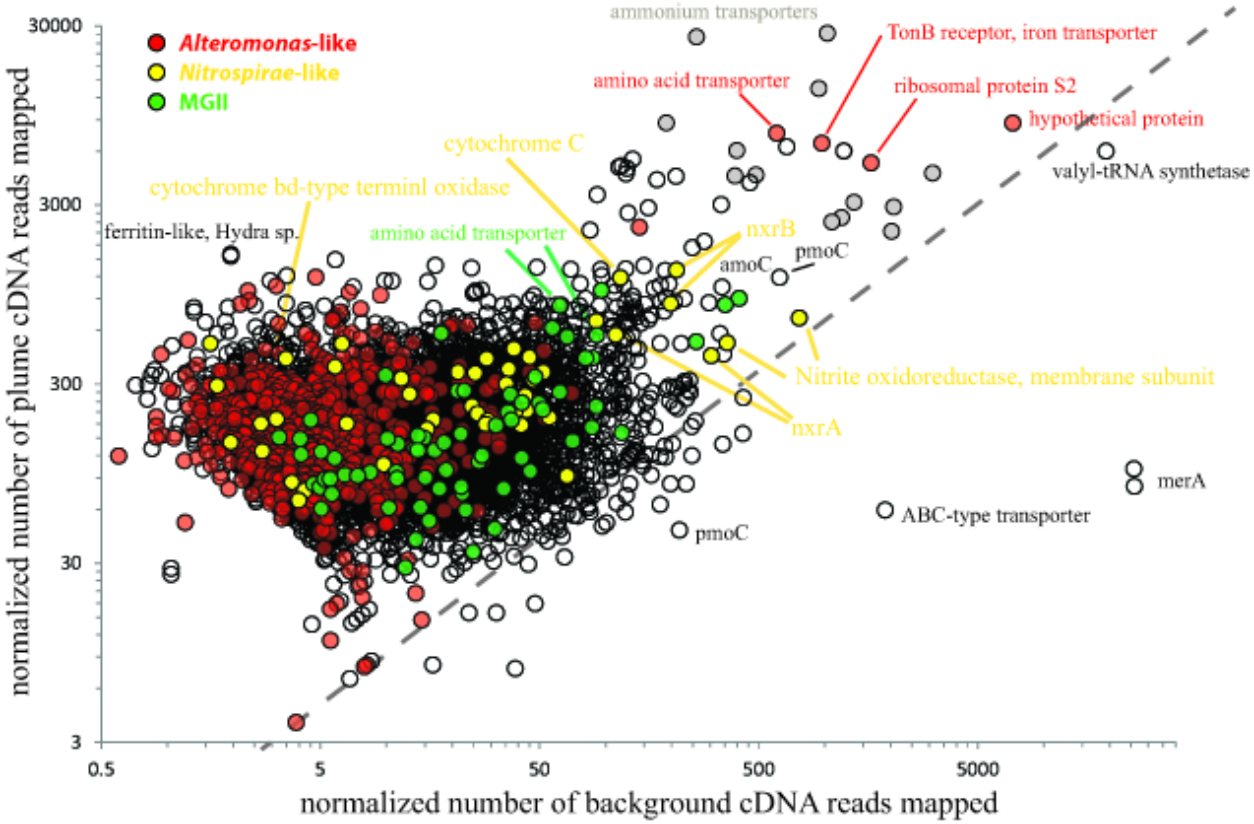

Figure 4.2. Abundance of gene transcripts in plume and background based on mapping transcripts to the plume de novo metatranscriptomic assembly. Red filled circles are mRNAs that have high similarity to *Altermonoas* spp., yellow are those related to *Nitrospirae*, and green are MGII. Grey filled circles are highly transcribed ammonium transporters, most of these belong to AOA, consistent with previous findings (Baker *et al*., 2012). The dotted line indicates equal representation of transcripts in plume and background.

Given that ATP binding cassette (ABC) transporters are an essential component of hetrotrophy and uptake of dissolved organic carbon (DOC) in the oceans (Jiao and Zheng, 2011), we compared transcriptional activity among putative ABC amino acid transporters present in the metatranscriptome. Interestingly, 5 of the 10 most abundantly represented amino acid

transporters in the plume metatranscriptome have high similarity to Euryarchaea Marine Group II (MGII), suggesting this group utilizes exogenous amino acids as a carbon and/or nitrogen source. Because of the low coverage of MGII in the metagenome (Figure 4.3 and Supplementary Figure S3), we searched the transcript assembly using a recently obtained MGII genome (Iverson *et al.*, 2012). A total of 112 transcript contigs (nearly 72kb total) were identified with an average similarity of 91% to the MGII genome. Putative functions could be assigned to only 37 of these assembled contigs; the vast majority was annotated as "hypothetical proteins", underscoring the lack of knowledge of this group. MGII are ubiquitous in marine environments (Martin-Cuadrado *et al.*, 2008) yet their physiology and function has remained enigmatic until their recent implication in heterotrophy (Iverson *et al.*, 2012).

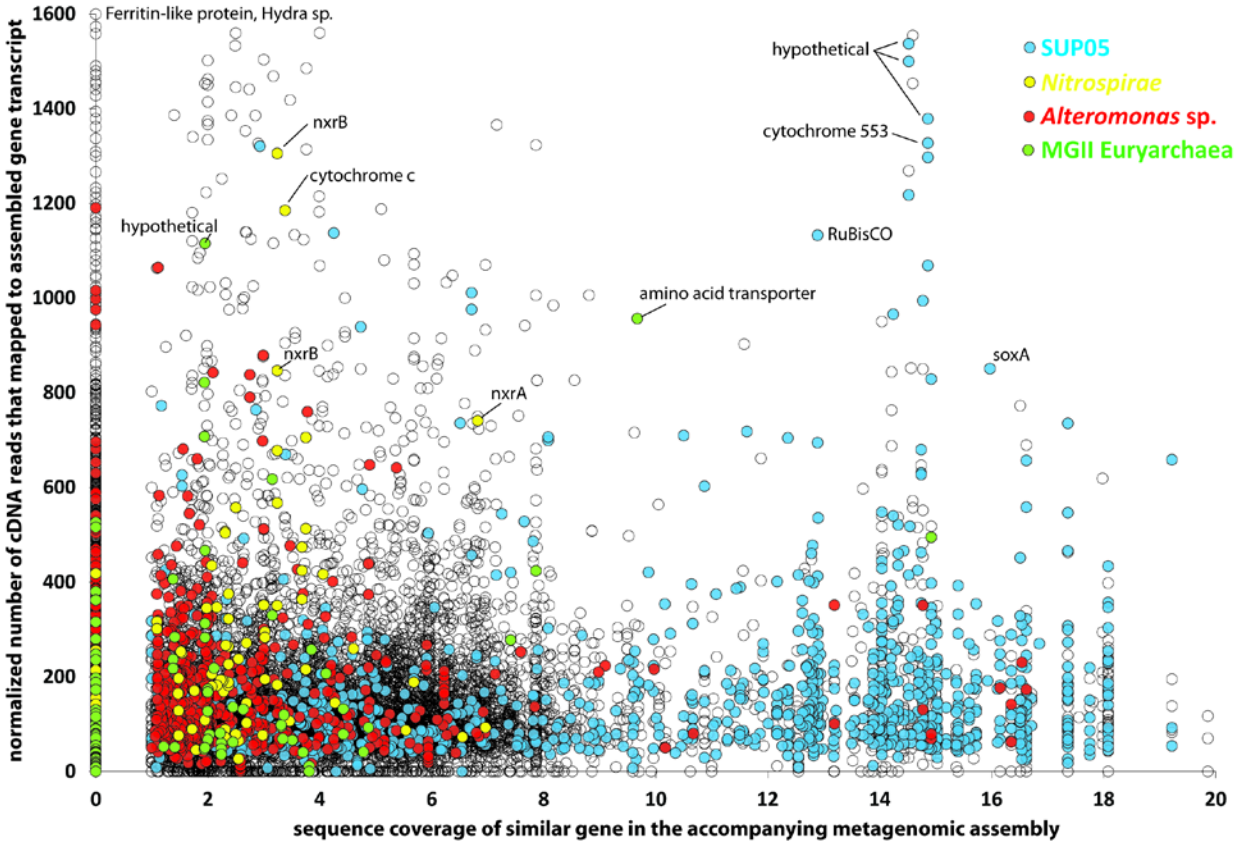

Figure 4.3. Plot of gene transcript abundance vs. coverage in the metagenomic assembly from Lesniewski et al., 2012. Abundance is the number of cDNA reads mapped to the transcript, normalized to the length of the gene. Top matches in the gDNA library assembly are greater than e-value of $1E^{-10}$.

MGII have proteorhodopsin genes for energy generation in the photic zone (Frigaard *et al*., 2006). Upon searching for proteorhodopsin genes in the deep Guaymas metatranscriptome none were idnentified, as expected for a dark environment and consistent with previous 454-based results (Lesniewski *et al*., 2012). We did however identify expression of a V-type H^+-translocating inorganic pyrophosphatase (H^+-PPases) gene, which are implicated in energy generation in symbionts (Kleiner *et al*., 2012). These results hint that deep-sea MGII Archaea utilize H^+-PPases as an important mechanism of energy conservation and ATP generation. Also prominent among MGII transcripts were several RNA processing genes including multiple RNA-binding Rrp4 and RNase PH genes (Supplementary Figure S4).

Transcripts absent from metagenomic data. To further assess the extent of sequences present in the metatranscriptome but absent from the metagenome, we compared the transcripts to a prior metagenomic assembly derived from the same samples (Lesniewski *et al*., 2012). 8360 metatranscriptome-specific mRNAs were found, totaling over 3.4 Mb of consensus sequence. We were unable to assign potential function to 41% (3419) of these genes, and 2447 did not have confident matches to sequences in public databases. Many of the most active genes present in this category are of unknown function (Supplementary Figure S5). Overall, 16% (1378 of 8360) of the metatranscriptome-specific genes are closely related to genes from *Alteromonas*, including the abundant TonB receptor and ribosomal proteins (Figure 4.4).

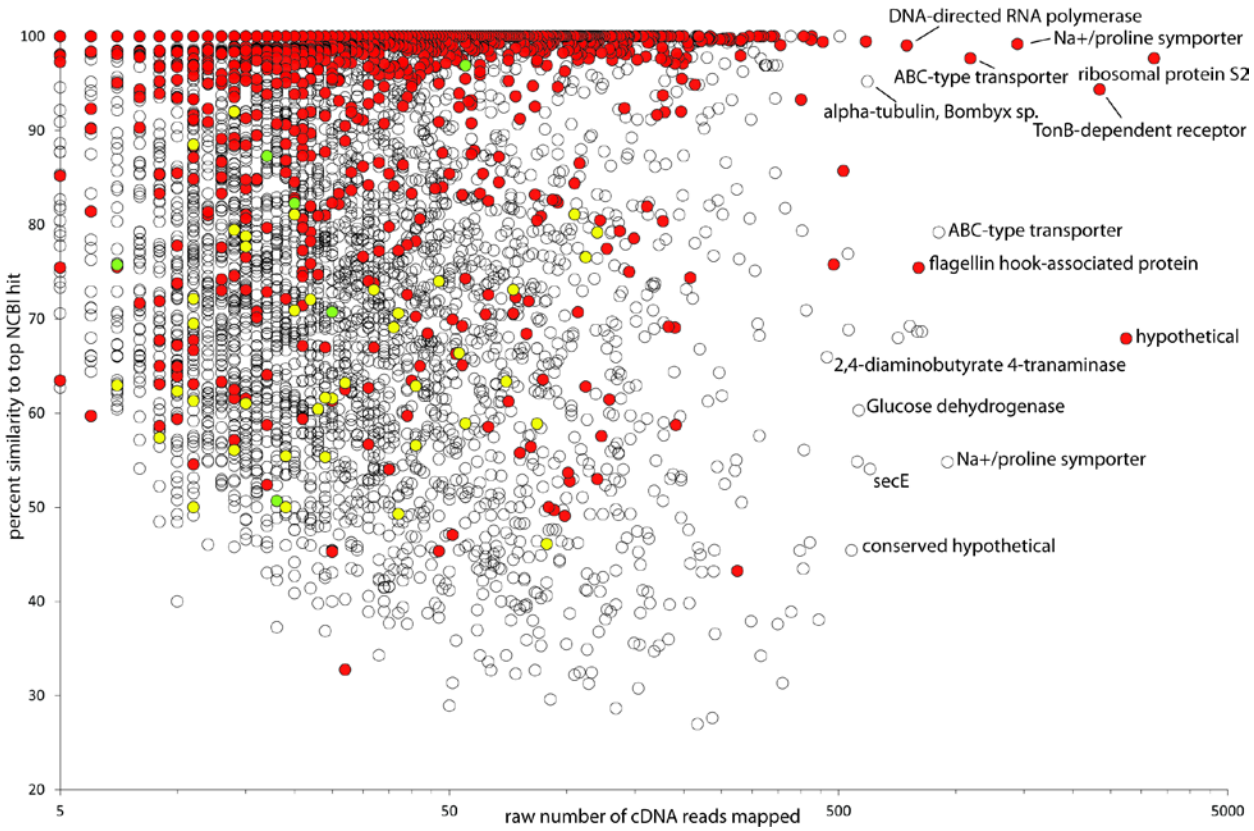

Figure 4.4. Transcripts not present in accompanying metagenomic data but with similarity to sequences in public databases. Each circle represents an assembled mRNA contig. Plotted is percent similarity to NCBI sequence versus the number of plume cDNA reads recruited. Coloring is consistent with Figure 2 and 3; red are *Alteromonas*, yellow are *Nitrospirae*, and green are MGII.

Identification of nitrite oxidation transcripts. Some of the most abundant transcripts in the community are from genes for nitrite oxidation and associated energy metabolism (Figure 4.2). These highly transcribed genes encode the key enzyme for nitrite oxidation, nitrite oxidoreductase (NxrA, NxrB, and the membrane subunit), as well a *c*-type cytochrome and cytochrome *bd*-type terminal oxidase for reduction of O_2 (Lücker *et al*., 2010). Except for *nxrA*, all of these genes are most similar to *Candidatus* Nitrospira defluvii (Supplementary Figure S6 and Table S1), a nitrite-oxidizing member of the phylum *Nitrospirae* (Lücker *et al*., 2010). These *nxr* genes are phylogenetically distinct from those recently discovered in *Chloroflexi* sp. (Sorokin *et al*., 2012). Many of the components proposed to oxidize nitrite and reduce O_2 in *Ca*. N. defluvii are present and highly transcribed in the Guaymas Basin metatranscriptome (Figure 4.5).

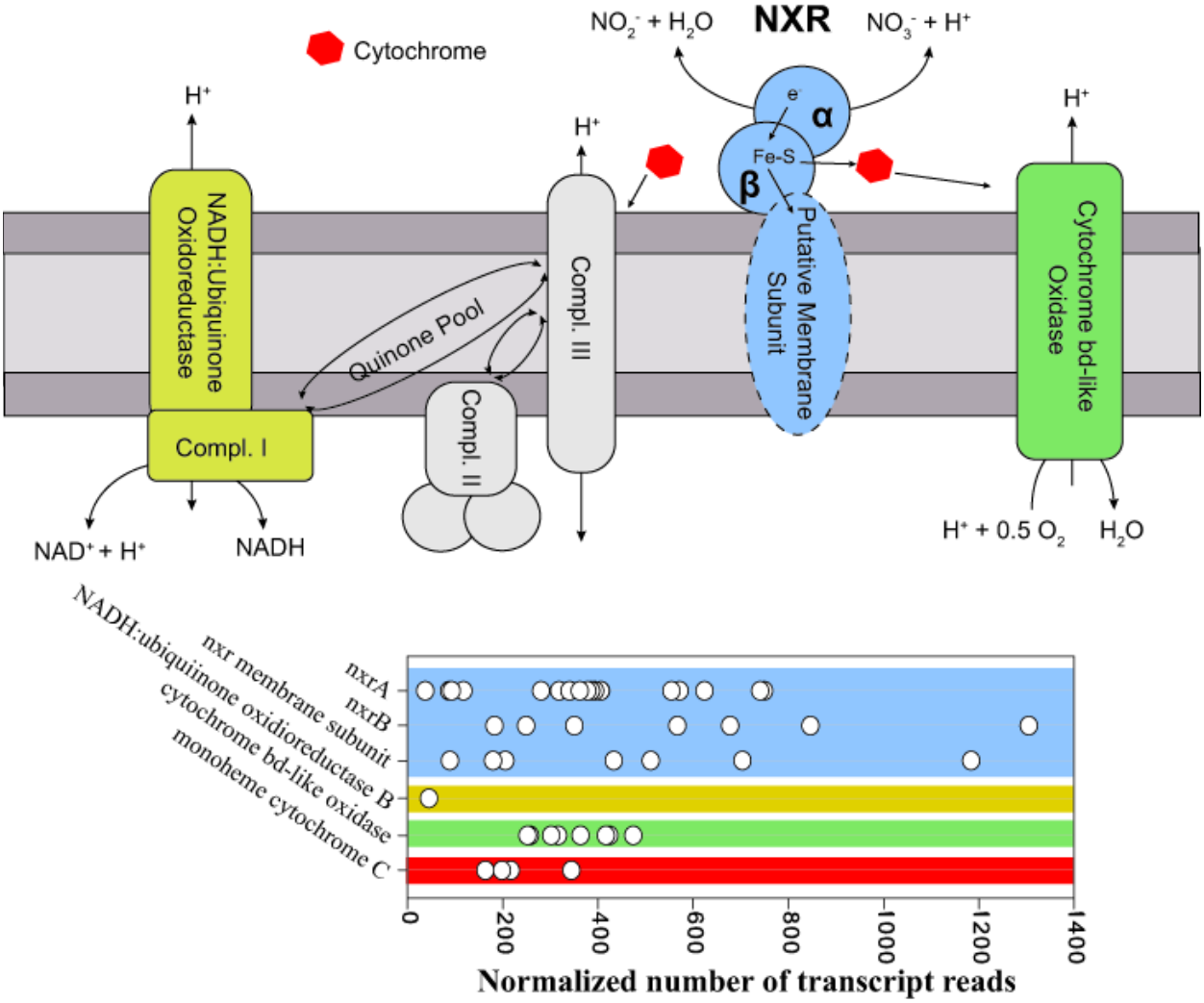

Figure 4.5. Schematic model and abundance of transcripts in the plume for proteins involved in nitrite oxidation and associated electron transfer. Colored proteins were detected in the plume cDNA libraries. Complexes in grey were not identified but are included in the model of electron transport for reference. Arrows show movement of electrons and protons. For transcript abundance, multiple circles for each gene represent multiple closely related gene sequence variants. Normalization is calculated as the number of cDNA reads mapped divided by lengths of the genes and multiplied by 1000.

Assignment of *nxr* genes to aerobic nitrite oxidation by *Nitrospirae* is complicated because at least two members of the phylum *Planctomycetes*, *Candidatus* Kuenenia stuttgartiensis and *Candidatus* Scalindua profundus, also contain *nxrA* and *nxrB*-like genes, which are thought to be involved in nitrite oxidation during anammox (Strous *et al.*, 2006). Although the prevalence of anammox seems unlikely in oxic waters of the deep Guaymas Basin

(~28 µM O_2), it could take place in particle-associated, anoxic microenvironments (Wright *et al.*, 2012). Thus to evaluate the possibility that the *nxr*-like transcripts we observed are from anammox microorganisms, attempts to identify additional anammox-related gene transcripts were made. No appreciable abundance of transcripts of key anammox genes, hydroxylamine oxidoreductase (*hao*) or hydrazine hydrolase (*hzh*), were identified in the metatranscriptomic assembly (Supplementary Figure S7). Further, the concurrent abundant expression of the cytochrome bd-terminal oxidase most closely related to a *Ca*. N. defluvii homolog indicates aerobic metabolism, ruling out a role for NXR in anammox or H_2-linked denitrification, which has been suggested for *Nitrospira moscoviensis* (Ehrlich *et al*., 1995). Therefore, we conclude that the abundant transcripts encoding a novel NXR and associated electron transport chain are involved in aerobic nitrite oxidation.

Guaymas Basin plumes are enriched in ammonium and hydrocarbons (Bazylinski *et al*., 1989) thus may be more representative of areas of intense nitrogen and carbon cycling (e.g. oxygen minimum zones (Wright *et al.*, 2012)) than the typical deep ocean. However, the high abundance of transcripts from *nxr* and associated electron transport genes in the non-plume background sample shows that their prominence is not restricted to ammonium-rich hydrothermal plumes (Figure 4.2 and Supplementary Figure S8).

Recovery and Characterization of Nitrospirae 16S rRNA and rRNA genes. Because the novel *nxr* transcripts are not directly linked to conserved phylogenetic markers (i.e. do not co-occur on a single assembled contig), it is not possible to definitively assign these genes to taxa present in our dataset. To probe this question further, we searched the transcript library for known NOB rRNA gene sequences. No 16S rRNA genes from common NOB genera (*e.g. Nitrospina*, *Nitrobacter*, *Nitrococcus, Nitrospira*) were present, however two phylotypes that fall within phylum *Nitrospirae* were identified. Phylogenetic analyses of the *Nitrospirae* 16S rRNA and 23S rRNA genes indicated two distinct clusters (81% and 82% 16S rRNA gene similarity to *Ca*. N. defluvii) that represent novel members of the *Nitrospirae* phylum (Figure 4.6). One of these phylotypes (referred to hereafter as Deep Ocean *Nitrospirae* Nitrifier, "DONN") recruited four times more rRNA transcripts than the other in the plume and was the only transcriptionally active *Nitrospirae*-like phylotype in the background sample. The closest match to DONN in public databases shares only 88% sequence identity to uncultured *Nitrospirae* clones (Figure

4.6), highlighting the novelty of this group. Comparison of all cDNA reads to a comprehensive 16S rRNA gene database revealed that only 0.3% of the rRNA reads matched most closely to *Nitrospirae* (Figure 4.1). In addition, <0.001% of all rRNA gene-containing reads from the genomic library were identified as *Nitrospirae*.

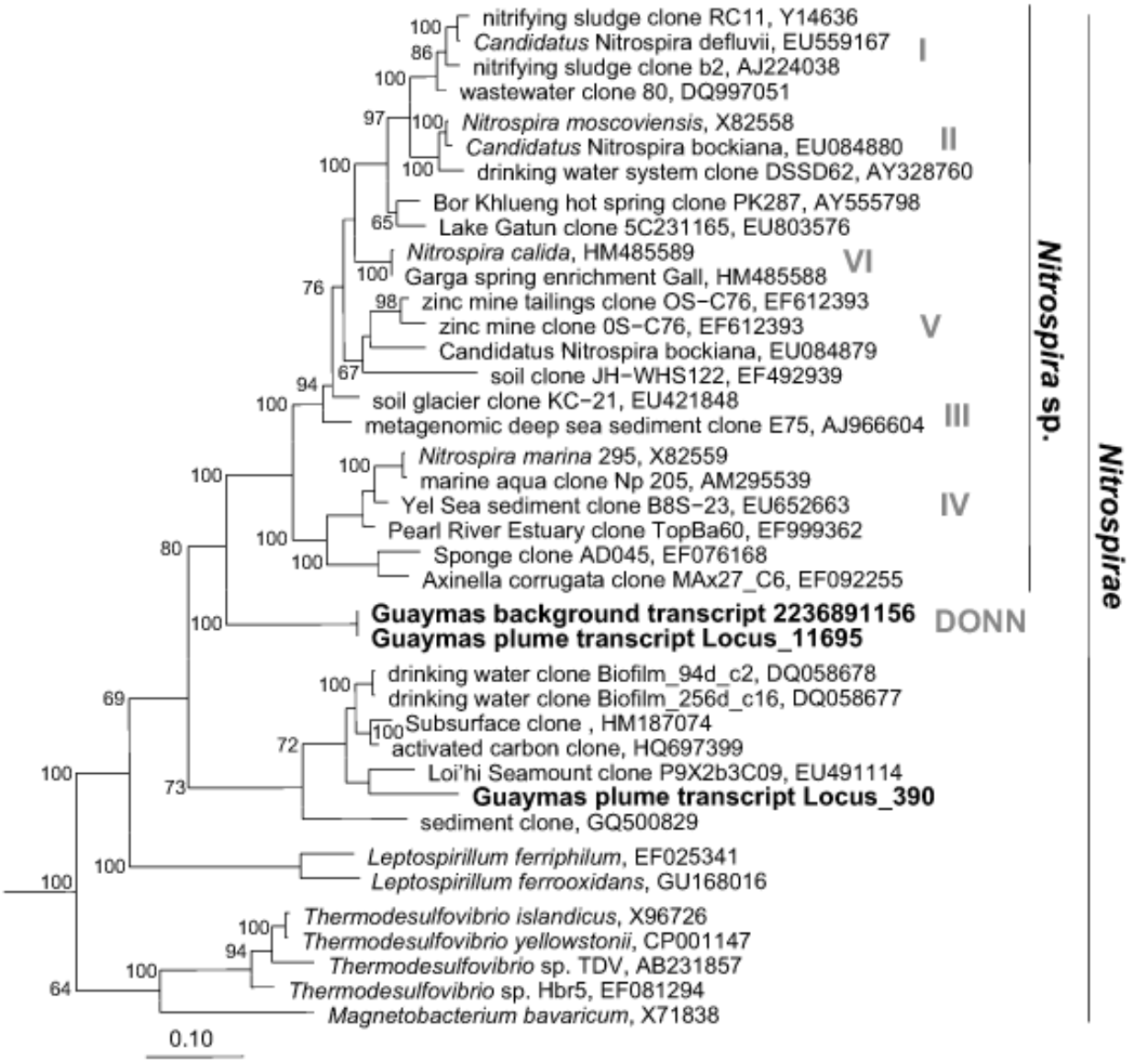

Figure 4.6. Phylogeny of Nitrospira-like 16S rRNA genes from assembled transcripts. Trees were generated using the maximum likelihood method and *Planctomycetes brasiliensis* as the outgroup.

Prevalence of Nitrospirae metabolic gene transcripts. Further support for the assignment of *nxr* genes to *Nitrospirae* comes from the prevalence of additional abundant mRNA transcripts with high similarity to *Nitrospirae*. In total, we identified 160 *Nitrospirae*-like genes (including several species/strain variants) on 142 assembled mRNA fragments (Figure 4.7). Interestingly, 115 of these have similarity to contigs in the accompanying metagenomic dataset (Figure 4.3) but are present at low coverage (2.4X). To confirm this we searched all the previously published 454 datasets (Lesniewski et al. 2012) and found the same trend of a high (5:1) cDNA:DNA ratio in total community nxr genes. Taken together, these results reveal the low-abundance yet high transcriptional activity of *Nitrospirae* in the deep Guaymas Basin (Table S1).

Figure 4.7. Abundance of assembled transcripts most closely related to Nitrospirae from the plume transcript assembly. Each circle represents a distinct gene sequence, with assigned functions listed on the left. Thus, multiple data points for each gene represent sequence variants present in the community. Abundance is based on the number of reads that mapped to the assembled transcript. Normalization is calculated as the number of cDNA reads mapped divided by lengths of the genes and multiplied by 1000.

Nitrite-oxidizing bacteria are thought to be primarily autotrophic, but there is evidence for enhanced growth of *Nitrospira* spp. when supplied with simple organic carbon sources such as pyruvate (Lücker *et al*., 2010, Ehrlich *et al*., 1995, Boon and Laudelout, 1962). All described *Nitrospirae* utilize the reductive tricarboxylic acid (rTCA) cycle. Among the abundant *Nitrospirae*-like transcripts in the Guaymas Basin metatranscriptome were those from genes integral to carbon metabolism via the TCA cycle. Transcripts of several strain variants of all three subunits (α, B, and γ) of 2-oxoglutarate:ferredoxin oxidoreductase and pyruvate:ferredoxin oxidoreductase genes were identified (Figure 4.7). However, genes encoding the ATP-citrate lyase (indicative of CO_2 fixation via reductive TCA) were not recovered, thus we are unable to verify if the reductive TCA cycle operates for CO_2 fixation in these DONN populations as it does in *Ca*. N. defluvii (Lücker *et al*., 2010).

Abundance and distribution of low-abundance yet transcriptionally active microbial groups. Given that a large proportion of transcripts originate from minor community members (MGII, *Alteromonas*, and DONN groups), we sought to further assess the abundance and distribution of these groups in Guaymas Basin as well as in hydrothermal plumes of the Eastern Lau Spreading Center, which is located in the southwestern Pacific and hosts geochemically diverse hydrothermal vents. Analysis of high-throughput 16S rRNA gene pyrosequencing libraries taken from various depths of the water column yielded operational taxonomic units (OTUs) corresponding to all three groups, and confirmed their presence across geographically disparate microbial communities. At Guaymas, three dominant 16S rRNA gene phylotypes of *Alteromonas* sp. were present, which collectively represent 1.01-4.04% of the total Guaymas Basin community at depths of 1300-1900 m. These phylotypes were not detected in near surface samples (12.5m) but increased to 1.0 and 1.76% near the oxygen minimum zone (356 and 554 m respectively). At Lau Basin, only two of the three Guaymas Basin *Alteromonas* phylotypes were detected. Two dominant MGII phylotypes were present as minor community members at Guaymas, comprising only 0-0.67% of the total community. MGII were not detected in the two near surface Guaymas Basin samples (12.5 and 356 m). At Lau only one of the two Guaymas MGII phylotypes were detected, and ranged in abundance from 0-1.3%.

We found that both *Nitrospirae* phylotypes are present as low abundance community members at Guaymas Basin, both in previously obtained clone libraries from Dick and Tebo (2010) and in new pyrosequencing data (Supplementary Figure S9). The DONN group is most abundant in the deep basin, but even there it only accounts for ~0.25% of the community (Supplementary Figure S10). Similar *Nitrospirae* phylotpes were also identified as minor members of Lau Basin communities (Supplementary Figure S11), further suggesting that the novel *Nitrospirae* phylotypes reported here are widespread and consistently low abundance members of deep–sea microbial communities. It is also important to note that commonly used probes for the *Nitrospira* (Füssel *et al*., 2012) have two nucleotide mismatches to the DONN group, so these organisms may have been missed by previous studies. Thus more work is needed to assess the distribution of DONN in diverse marine environments where nitrification is prevalent.

The stark contrast in NXR abundance between transcript (high-abundance) and metagenomic (low-abundance) libraries calls attention to the concept that keystone ecological functions can be performed by low abundance species of the biosphere. In the case of NOB, low abundance despite high metabolic activity may be inherent to their physiology. Cultured NOB grow slowly (Watson *et al*., 1986), presumably due to low free energy yield from nitrite oxidation (Boon and Laudelout, 1962), which likely constrains the abundance of *in situ* NOB populations. Further, the disparity in population size between NOB (low abundance) and AOA (high abundance) at Guaymas Basin implies that cell-specific nitrite oxidation rates must be large relative to those of ammonia oxidation (assuming quantitative conversion to nitrate by the NOB). In the common terrestrial NOB *Nitrobacter winogradskyi*, enzyme saturation is evident under μM concentrations of nitrite (Watson *et al*., 1986), and it is estimated that the nitrite oxidoreductase enzyme may comprise 10-30% of total cell protein (Bock *et al*., 1991). Increased transcription of *nxr* genes but not rRNA genes has also been observed in *Ca*. N. defluvii enrichments (Lücker *et al*., 2010) and actively fertilized soils where *Nitrobacter*-like *nxr* expression was elevated (Wertz *et al*., 2011). Our data suggests a similar scenario occurs in the deep-sea, where NOB highly transcribe *nxr* genes to maximize nitrite oxidation that provides only modest energy and growth yield. Another possible explanation is that the disparity between DNA and RNA abundance of the NOB reflects of a recent transcriptional response to nitrite in the environment, and the subsequent increase in DNA (cell division) would soon follow.

Intra-population variability. The high sequence coverage produced by *de novo* metatranscriptomic assembly provides opportunities to investigate gene sequence variation and ecological dynamics of strains within natural populations. In many cases, multiple sequence variants of each gene involved in nitrite oxidation were recovered, indicating the presence of several closely related strains or multiple gene copies within a genome (Figure 4.7). The most highly expressed transcripts tended to have the greatest number of variants within the dataset, likely as a result of greater coverage of those regions. The NXR variants cluster into two divergent groups (82-84% similar at the DNA level), likely representing the two different *Nitrospirae* groups, but it is impossible to rule out the alternative interpretation of gene duplicates within a single genome as in *Ca*. N. defluvii (Lücker *et al*., 2010). The most highly expressed *nxrAB* type in the plume is also the dominant type in the background, suggesting that the same strain is dominant in both communities. Many of the sequence substitutions between these transcripts are synonymous. For example, the two most abundant *nxrAB* operon variants (GBPt_c08738 and GBPt_c08738) have eight nucleotide polymorphisms within a 72 bp region in the *nxrB* transcripts yet they have identical amino acid sequences. However, we also identified minor variants that were only present in the plume and that have seven distinct nucleotide substitutions clustered solely within the metal-coordinating [Fe-S] center of NxrB (Supplementary Figures. S12 and S13). This site is homologous to a region of nitrate reductase of *E. coli* that mediates intramolecular electron transfer. The high frequency of nucleotide polymorphisms around this region suggests that selective pressures (perhaps substrate concentrations) maintain such variation.

4.5 Conclusions

Despite the recent explosion of microbial genome sequencing, environmental shotgun sequencing continues to reveal vast genetic novelty, which presents fundamental challenges to our ability to fully characterize natural microbial communities. Our findings demonstrate that *de novo* metatranscriptomic assembly offers the ability to assess transcriptionally active populations of diverse and novel microbial communities at high resolution (to the strain level). More importantly, it enables the

reconstruction and functional characterization of transcripts that would have otherwise been overlooked by mapping to reference genomic databases. In the deep Gulf of California, this approach revealed the functional importance of low-abundance populations of *Alteromonas* and archaeal MGII in heterotrophy and novel *Nitrospirae* in nitrite oxidation. The high RNA:DNA ratio and novelty of genes implicated in nitrite oxidation suggest explanations for why these *Nitrospirae* have eluded detection and are under-represented relative to their ammonia-oxidizing counterparts. These new insights into novel nitrite-oxidizing Bacteria indicate that the distribution and role of this functional group should be re-considered as we seek to understand the fate of nitrite in terms of nutrient cycling and production of greenhouse gases in the oceans.

Acknowledgements

This work was supported by the Gordon and Betty Moore Foundation and National Science Foundation (OCE 1029242). Thanks to Drs. Anders Andersson and Meng Li for their helpful discussions. We especially thank Donald Zak for revisions to the manuscript. We also thank Karthik Anantharaman for cDNA library preparations and Prashanna Balaji for assistance with read mapping.

4.6 Appendix B

CHAPTER IV Supplementary Information

Contents

14. Supplementary Figure 4.14

15. Supplementary Figure 4.15

16. Supplementary Figure 4.16

17. Supplementary Figure 4.17

18. Supplementary Figure 4.18

19. Supplementary Figure 4.19

20. Supplementary Figure 4.20

21. Supplementary Table 4.1

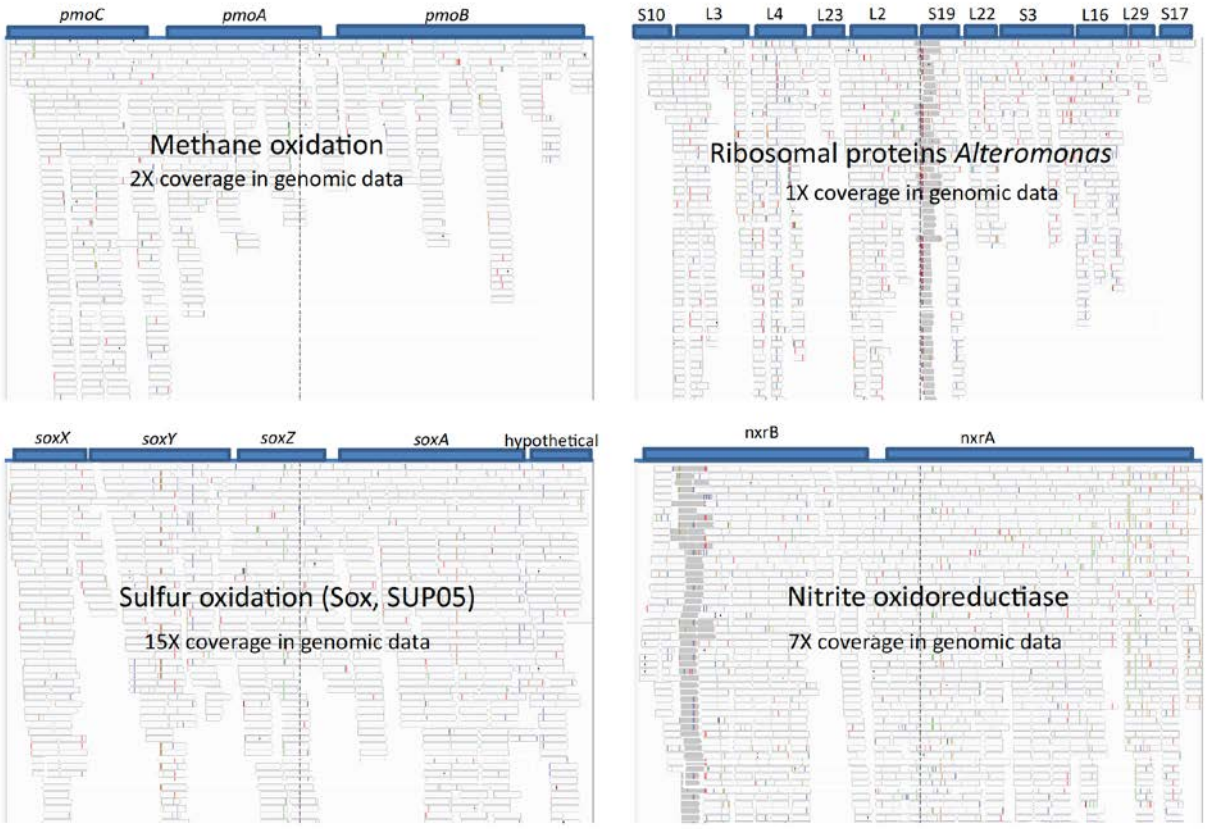

Figure 4.8. Examples of four operons assembled from the metatranscriptome. Genes are indicated by bars at top and individual cDNA reads are shown below. Coverage of these sequences in the assembled 454 metagenomic data (Lesniewski *et al*., 2012) is shown.

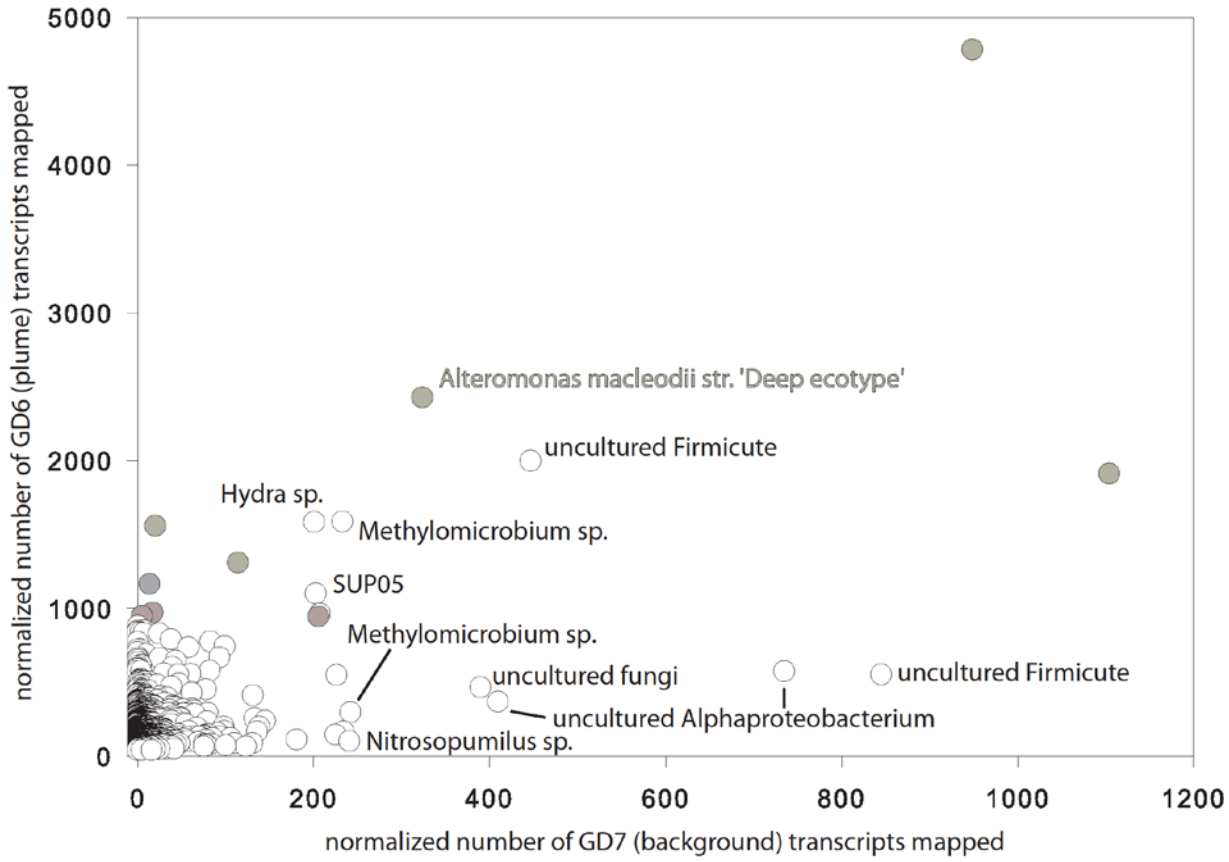

Figure 4.9. Comparison of abundance of transcripts containing ribosomal proteins (some have multiple genes present co-transcribed on operons) in plume versus background. Grey filled circles are those related to *Altermonas macleodii*.

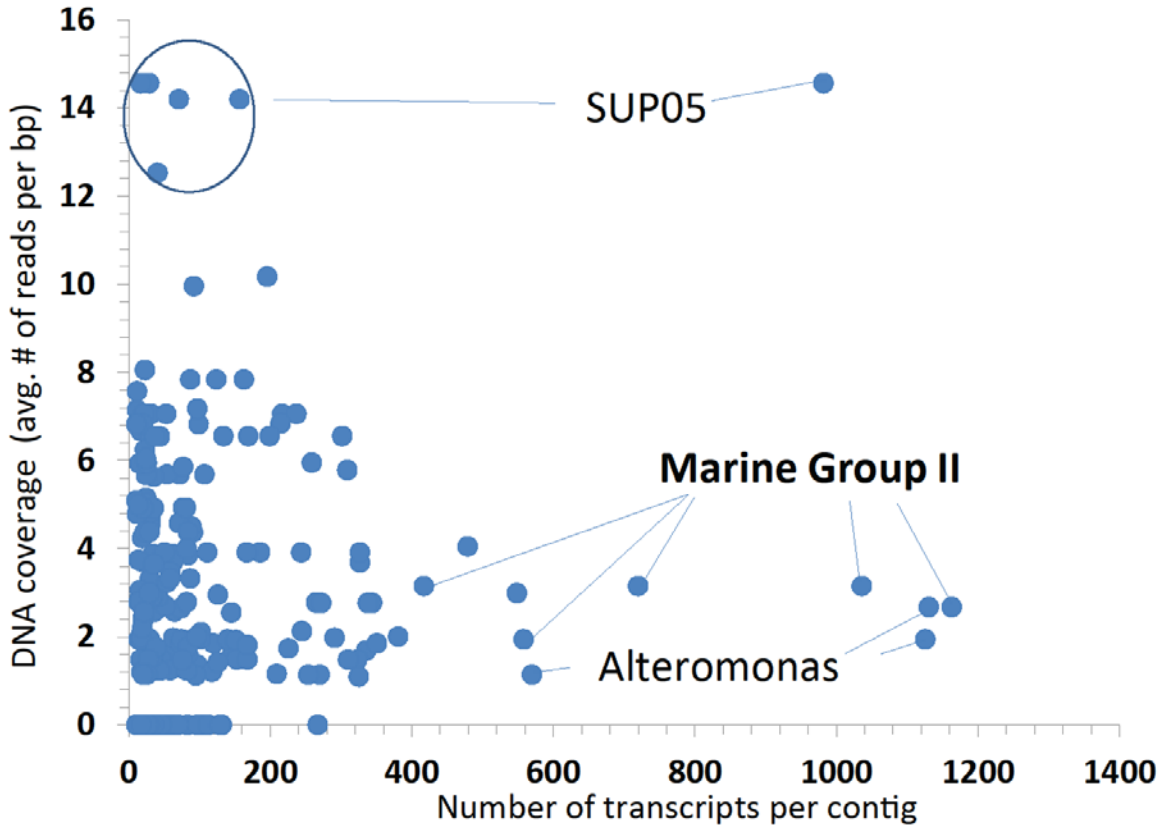

Figure 4.10. Plot of transcript abundance versus DNA coverage for genes identified as amino acid transporters in the transcript assembly. DNA coverage is based on previously published metagenomic data (Lesniewski *et al*., 2012) reconstructed from the same samples used in this study.

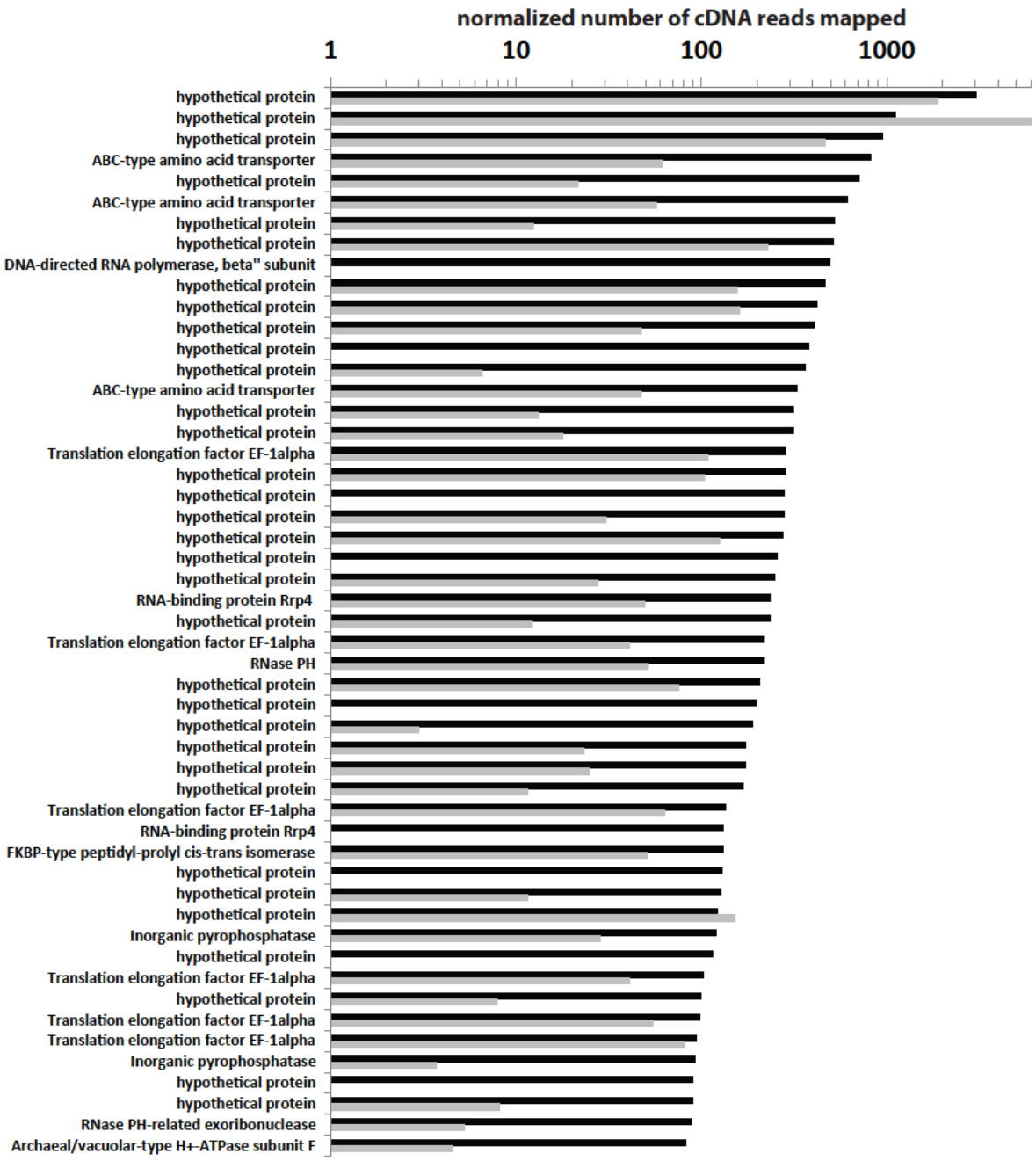

Figure 4.11. Abundance of the top 50 MGII transcripts in the plume (black bars) and background (grey bars). The number of reads mapped is normalized to length of gene transcript.

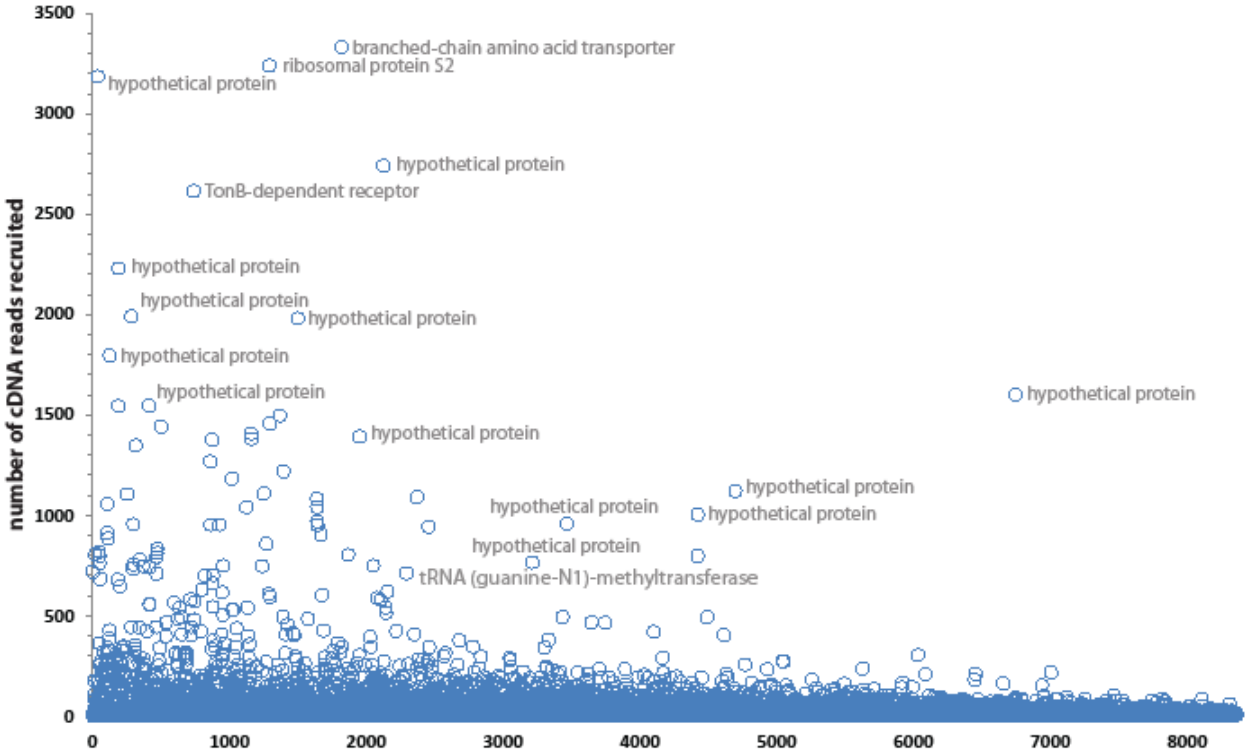

Figure 4.12. Transcriptional activity of novel genes on assembled transcripts. Each circle is a transcript ORF that is not present in the accompanying metagenomic dataset. The x-axis is valueless, the genes were arranged in order across the bottom and were sorted based on JGI ID numbers

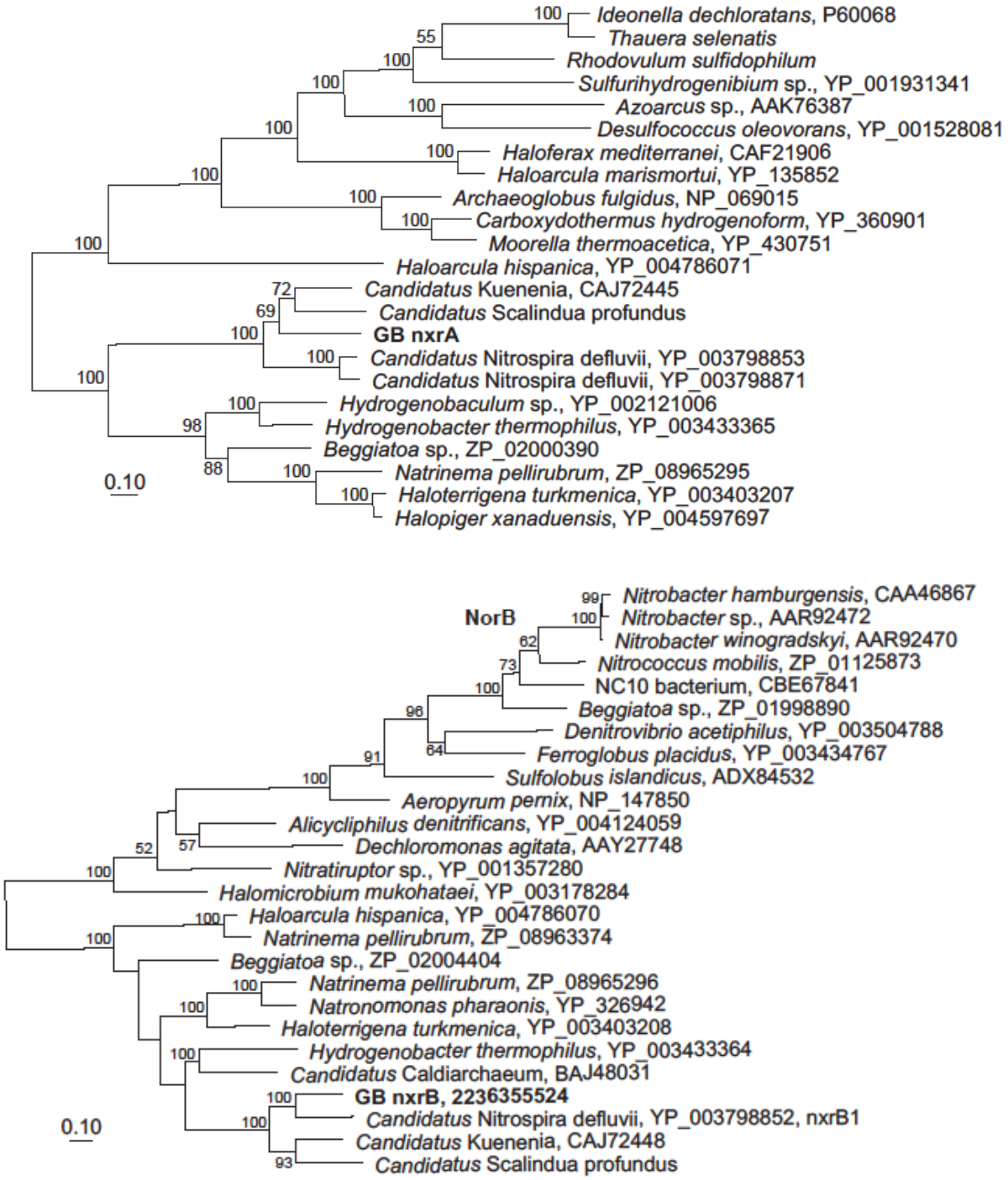

Figure 4.13. Phylogeny of nitrite oxidioreductase subunit A (top, NxrA) and B (bottom, NxrB) proteins from the transcript assembly and several other DMSO-like proteins. Both trees were generated using maximum likelihood (PHYML) methods in the ARB software package. Bootstrap values were estimated maximum parsimony using Phylip with a 1000 replicates.

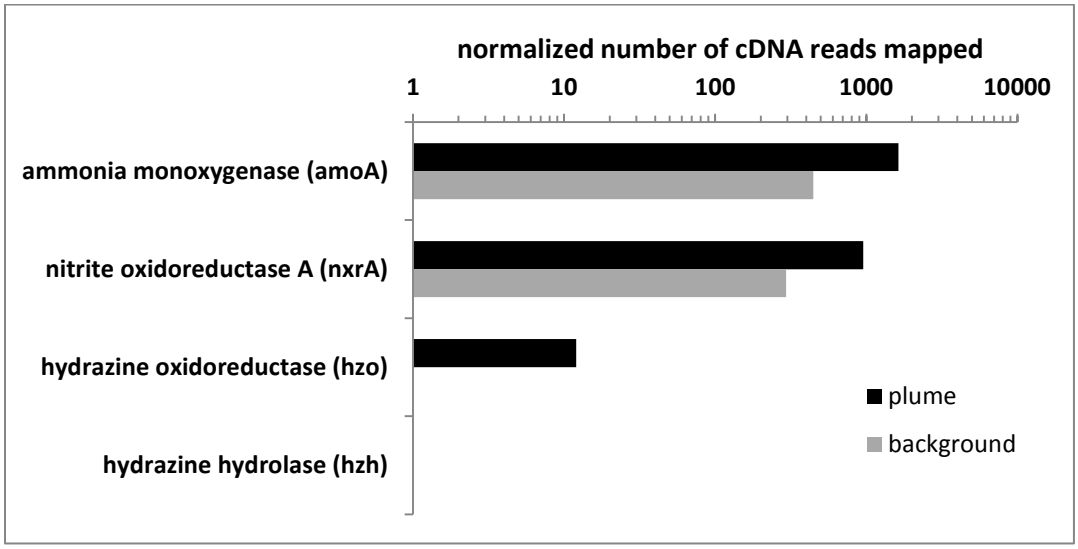

Figure 4.14. Comparison of abundance of cDNA reads matching key genes for ammonia oxidation, nitrite oxidation, and anaerobic ammonia oxidation in the plume and background samples. Normalized (to the number of total reads in each library) number of cDNA reads mapped to ammonia monooxygenase (*amoA)* of *Nitrosopumilus maritimus*, nitrite oxidoreductase *(nxrA1)* of *Ca*. N. defluvii, hydrazine oxidoreductase (*hzo)* of *Ca*. Kuenenia stuttgartiensis, and hydrazine hydrolase (*hzh) Candidatus* Scalindua profundus.

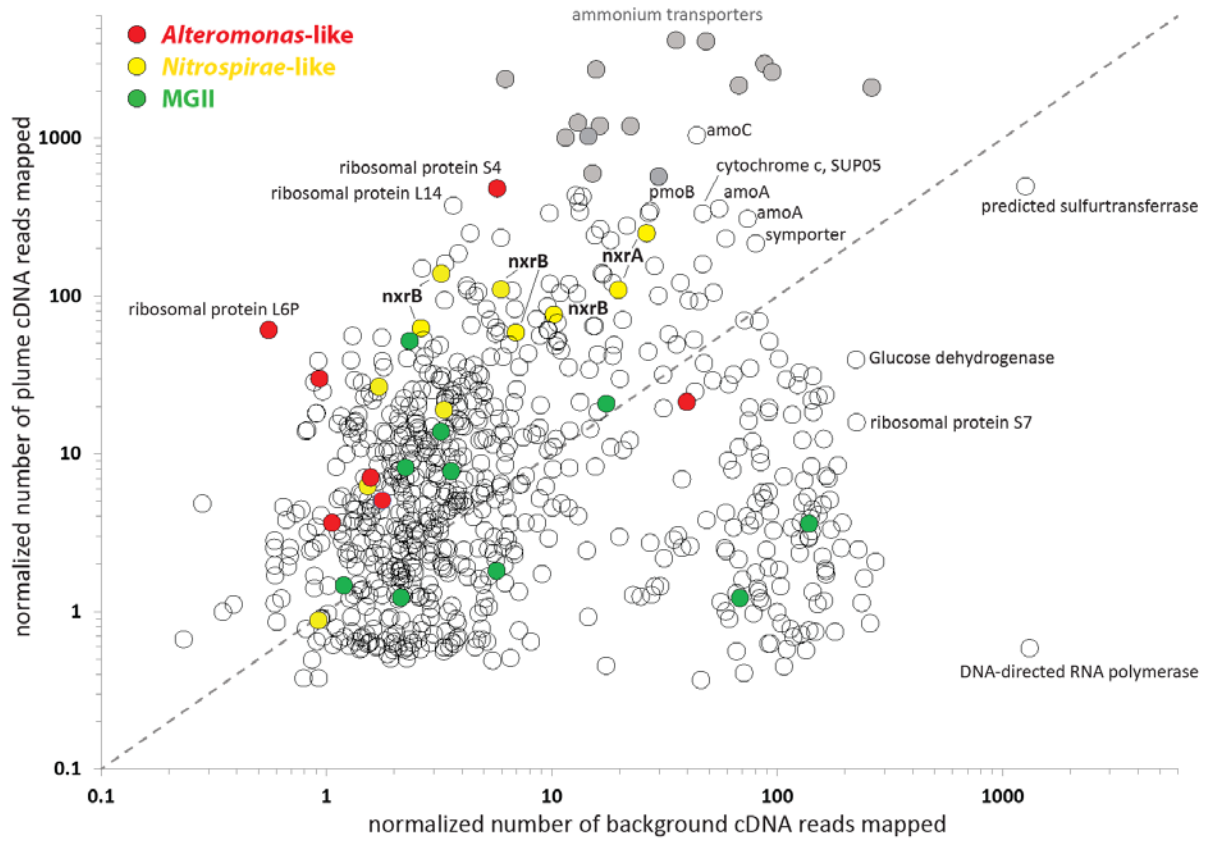

Figure 4.15. Abundance of gene transcripts in plume and background based on mapping transcripts to the background *de novo* metatranscriptomic assembly. Red filled circles are mRNAs that have high similarity to *Altermonoas* spp., yellow are those related to *Nitrospirae*, and green are MGII. Grey filled circles are highly transcribed ammonium transporters, most of these belong to AOA, consistent with previous findings (Baker *et al*., 2012). The dotted line indicates equal representation of transcripts in plume and background.

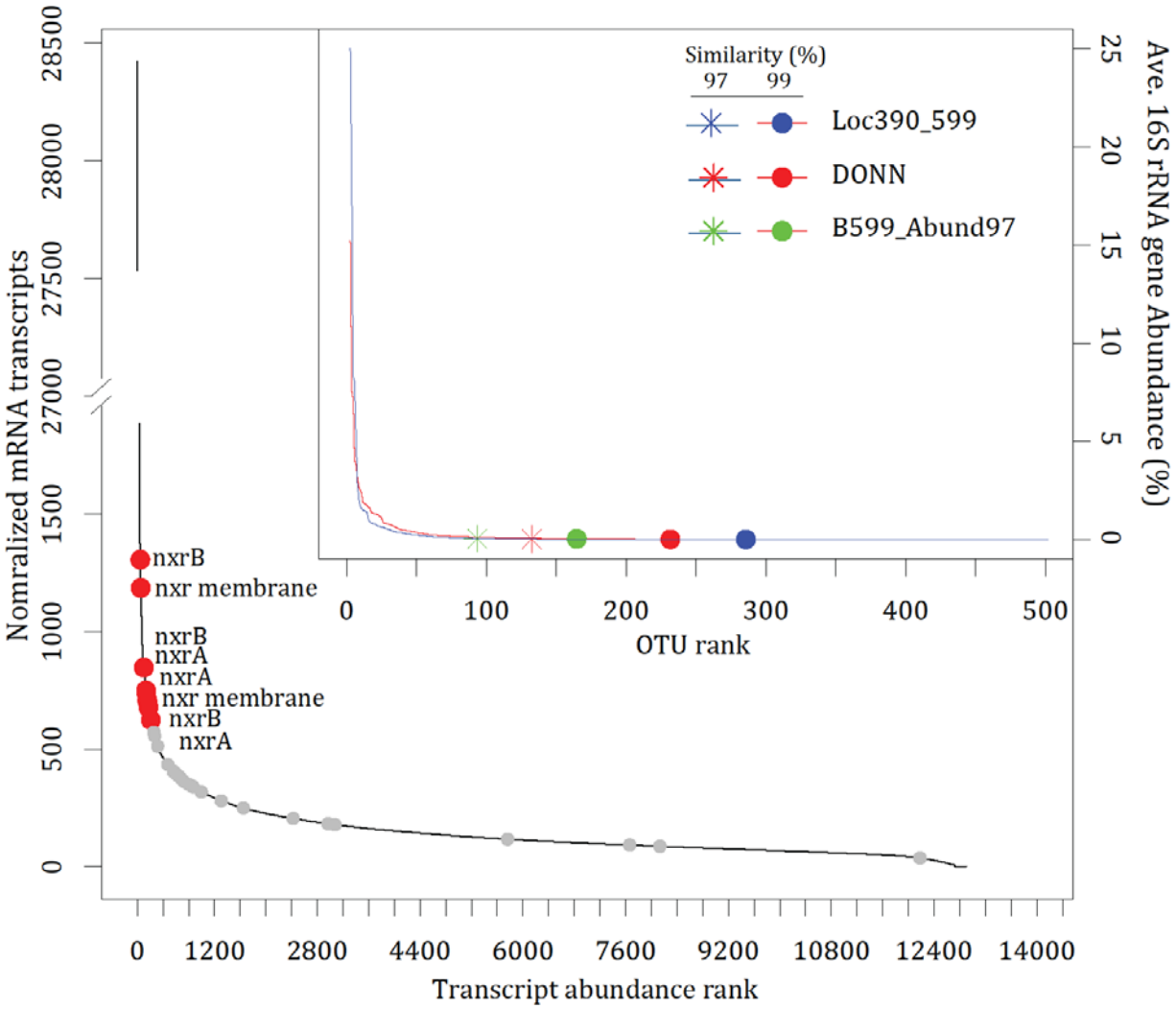

Figure 4.16. Rank abundance plots of genes on mRNA contigs and 16S rRNA genes (inset) of the whole plume community. Highlighted in red (top 8 are labeled) and grey are all the sequences that we identified as belonging to *Nitrospirae*. On the 16S rRNA gene plot (inset) we highlight OTUs that match the *Nitrospirae* groups at 97 and 99% similarity. Note that all of these fall within the rare portion of abundance in the community. The number of mRNA transcripts mapped was normalized by the lengths of the genes multiplied by 1000.

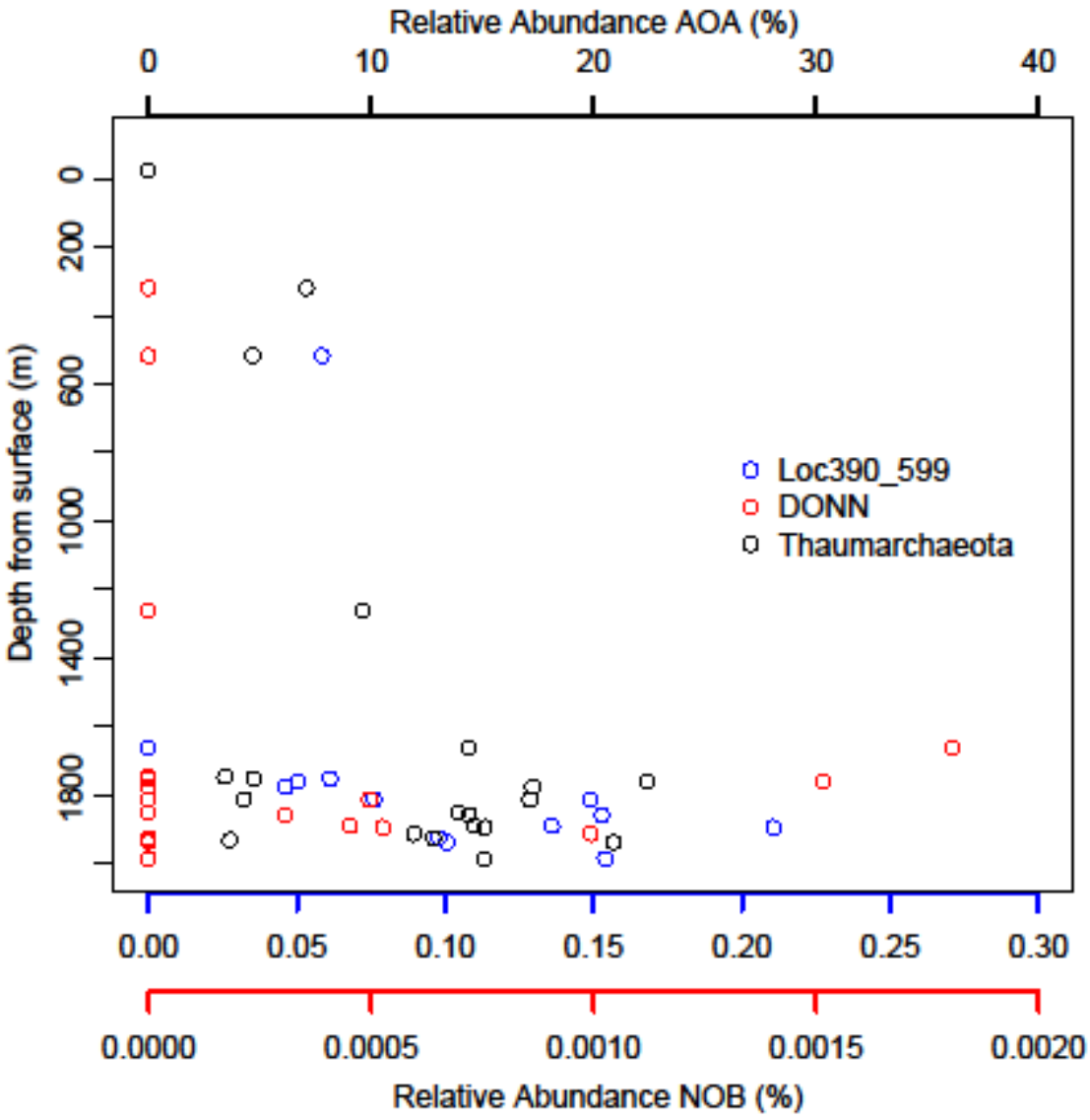

Figure 4.17. Distribution of NOB (Loc390_599 and DONN) and AOA (Thaumarchaeota) 16S rRNA genus-level OTUs (95% similarity) using pyrosequencing generated data taken from depth profiles at Guaymas Basin. Note the different scales for NOB and AOB.

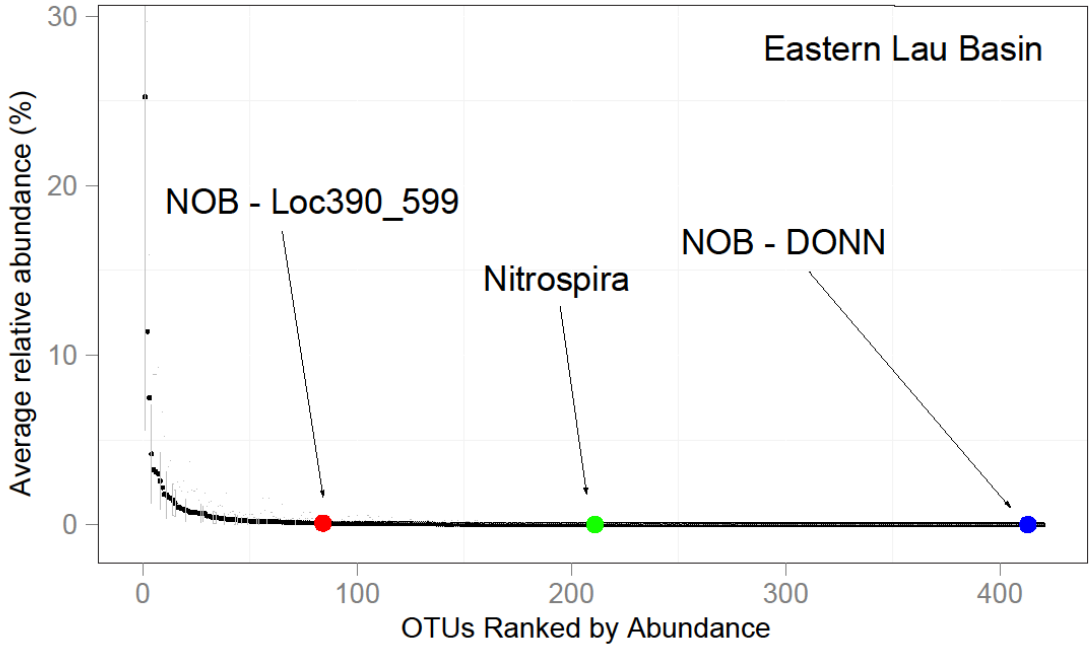

Figure 4.18. Ranked average abundance of putative nitrite oxidizing bacteria from rising plume and background waters at five vent sites (Tahi Moana, Tui Manila, ABE, Mariner and Kilo Moana) within the Eastern Lau hydrothermal spreading center. Operational Taxonomic Units (OTUs 97% Similarity) related to the known NOB group Nitrospira (green) and Nitrospira-like OTUs (blue and red) identified in Guaymas Basin metatranscriptomes are highlighted.

179--**GEIRPYCAICKESVRTTGRYMPKKFPCQEVC**--*210*

- Red letters are [Fe-S]-binding residues : 179,183,200,203
- Blue letters are residues that contain SNPs: 193,194,195, 197,201,202,205,207

Figure 4.19. Amino acid sequence of an iron-sulfur binding region in the metal-coordinating center in of the NxrB gene transcript variants, 2236355947 and 2236348949.

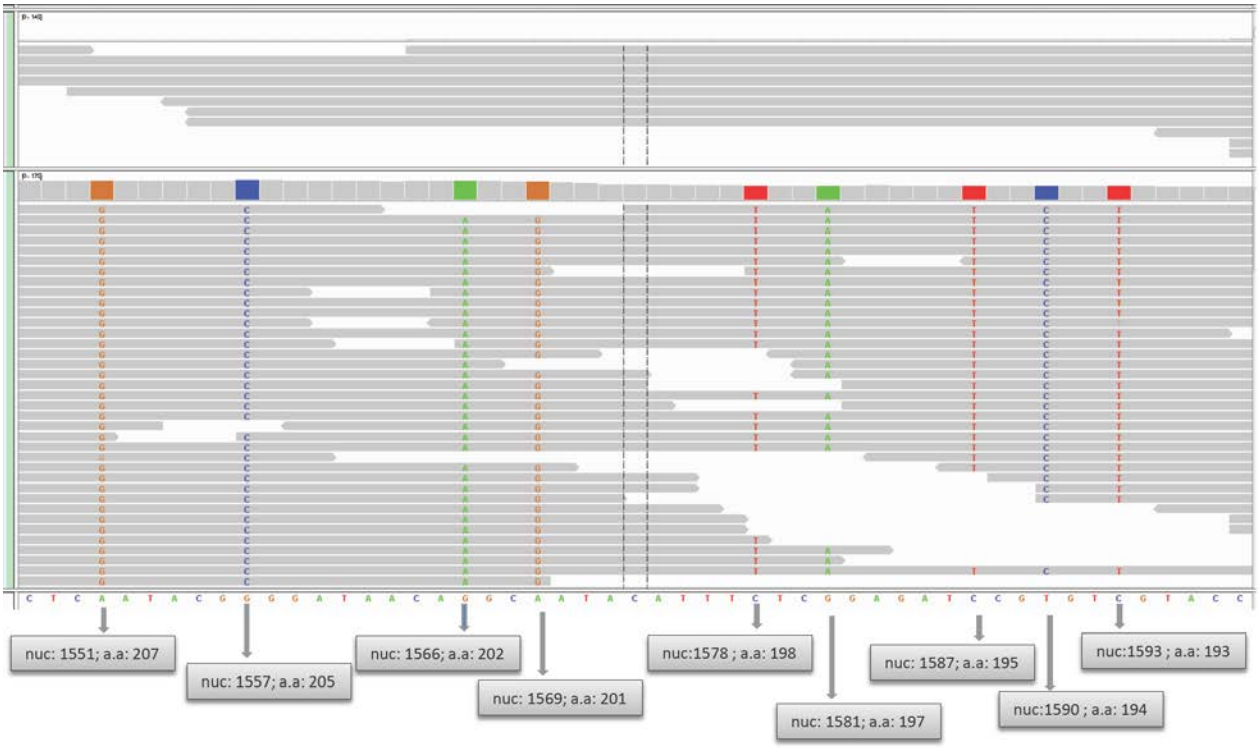

Figure 4.20. View of reads mapped to the [Fe-S]-binding region of the NxrB transcripts. The consensus sequence is shown at the bottom. Grey bars represent individual cDNA reads and the sequence variations that occur in 2236348949 (the minor variant) are the colored letters in the reads. The positions of the SNPs are labeled at the bottom.

Table 4.1. List of all gene transcripts binned as *Nitrospirae*. The following is an explanation of the columns starting on the left: IMG ID number, IMG annotation, top hit to the NCBI non-redundant database, accession number of ortholog in *Ca*. Nitrospira defluvii genome, % match to Ca. N. defluvii gene, *Ca*. Kuenenia stuttgartiensis, Leptospirillum spp., number of reads that map to plume (GD6) and background (GD7), and the coverage of that sequences in the genome (454 assembly) if found. Blank spaces in the columns of % similtarity to genomes represents no hits found (e-value <$1E^{-10}$). Colors of annotation boxes match colors found in Figure 4.

GB Plume protein	IMG annotation	top nr hit	N. defluvii gene	Candidatus Nitrospira defluvii	Candidatus Kuenenia stuttgartiensis	% Leptosprillum	gd6 N	gd7 N	DNA coverage
2236424278	2-oxoglutarate:ferredoxin oxidoreductase, alpha subunit	Leptospirillum ferrodiazotrophum	NIDE0827	61.82		61.5	121.7	17.8	
2236393701	2-oxoglutarate:ferredoxin oxidoreductase, alpha subunit	Leptospirillum rubarum	NIDE0827	68.82		62.7	196.5	13.49	
2236405518	2-oxoglutarate:ferredoxin oxidoreductase, alpha subunit	Leptospirillum rubarum	NIDE0827	75		78	125	0	3.46
2236385866	2-oxoglutarate:ferredoxin oxidoreductase, beta subunit	Leptospirillum ferrodiazotrophum	NIDE0823	71.59		75	85.19	0	2.9
2236379976	2-oxoglutarate:ferredoxin oxidoreductase, beta subunit	Leptospirillum rubarum	NIDE0823	70.17		73.48	159.5	17.92	
2236379977	2-oxoglutarate:ferredoxin oxidoreductase, beta subunit	Leptospirillum rubarum	NIDE0823	72.73		75.97	142.3	33.47	
2236368251	2-oxoglutarate:ferredoxin oxidoreductase, beta subunit	Leptospirillum rubarum	NIDE0823	81.68		83.85	229.8	22.73	3.16
2236350102	2-oxoglutarate:ferredoxin oxidoreductase, beta subunit	Leptospirillum rubarum	NIDE0823	64.71		69.12	248.8	48.78	
2236403362	2-oxoglutarate:ferredoxin oxidoreductase, gamma subunit	Leptospirillum ferrodiazotrophum	NIDE0824	70.37		79.63	132.4	0	
2236416995	50S ribosomal protein L11	Leptospirillum ferrodiazotrophum	NIDE1300	71.88		73.68	188.2	5.618	
2236427913	ABC-type molybdate transport permase	Candidatus Poribacteria sp. WGA-A3	NIDE3149				65.93	5.495	1.83
2236351417	ammonium transporter	Could not find annotation in database file					592.9	1.63	2.27
2236398111	ATP-dependent Zn proteases	Candidatus Nitrospira defluvii	NIDE2737	72.06		73.53	120.2	0	3.46
2236417421	CBS-domain-containing membrane protein	Candidatus Nitrospira defluvii	NIDE3802	50	40.45		78.07	0	1.72
2236418504	Chaperonin GroEL (HSP60 family)	Leptospirillum ferrodiazotrophum	NIDE1378	57.89		68.81	190.9	33.33	2.27
2236425622	Chaperonin GroEL (HSP60 family)	Leptospirillum sp. Group II '5-way CG'	NIDE1378	50.77	50	59.38	134	15.46	
2236379008	DNA-directed RNA polymerase, beta subunit/140 kD subunit	Candidatus Nitrospira defluvii	NIDE1303	74.16		74.16	573.2	41.4	
2236398779	DNA-directed RNA polymerase, beta subunit/140 kD subunit	Leptospirillum ferrodiazotrophum	NIDE1303	72.86	69.29	75	130.5	2.331	7.87
2236394039	DNA-directed RNA polymerase, beta'' subunit/160 kD subunit	Candidatus Nitrospira defluvii	NIDE1304	63.36	48.33	58.54	195.1	24.39	3.88
2236393487	DNA-directed RNA polymerase, beta'' subunit/160 kD subunit	Leptospirillum ferrodiazotrophum	NIDE1304	91.11		92.39	351.4	10.87	4.06
2236420746	DnaK suppressor protein	Candidatus Nitrospira defluvii	NIDE0281	60.44		54.44	96.77	3.584	
2236350912	Ferredoxin-nitrite reductase	Candidatus Nitrospira defluvii	NIDE1367	54.98			628	1.431	2.5
2236358570	Ferredoxin-nitrite reductase	Candidatus Nitrospira defluvii	NIDE1367	64.64			416.1	1.963	2.5
2236358574	Ferredoxin-nitrite reductase	Candidatus Nitrospira defluvii	NIDE1367	64.64			415.7	6.359	3.02
2236379802	Ferredoxin-nitrite reductase	Candidatus Nitrospira defluvii	NIDE1367	67.57			153.1	0	1.76
2236401838	Ferredoxin-nitrite reductase	Candidatus Nitrospira defluvii	NIDE1367	59.38			74.26	0	2.5
2236393164	Formate dehydrogenase, gamma subunit	Leptospirillum rubarum	NIDE1930	34.78	38.3	43.48	105.4	13.61	

2236397255	Formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase	Candidatus Nitrospira defluvii	NIDE4140	70.89		46.84	89.15	0	
2236351418	Glutamine synthetase	Candidatus Nitrospira defluvii	NIDE1363	65.22	37.78	66.67	592.9	1.63	
2236352223	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	55.17			230.3	12.01	2.37
2236352221	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	56.74			262.7	8.102	2.37
2236377463	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3281	36.78	32.02		310.3	50.82	4.6
2236385375	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	68.29			141	22.26	1.43
2236415742	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	53.12			283.3	31.25	2.69
2236396947	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	67.36			283.3	0	2.32
2236378967	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	70.92			163.3	31.32	2.69
2236382997	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	74.55			240.8	59.63	1.48
2236397567	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	63.21			113	0	
2236375749	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	59.22			279	3.361	2.37
2236391359	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	81			115	0	2.69
2236399526	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	56.57			140.1	0	
2236424645	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	74.23			81.08	0	2.55
2236396077	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	60.24			298.8	7.968	2.55
2236366660	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	81.08			100	0	1.52
2236370494	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	58.9			251.6	22.58	4.27
2236378664	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3308	55.1			206.6	20.66	2.23
2236383136	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	62.16			79.44	9.346	1.48
2236421079	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3519	64.91			168.3	9.615	3.46
2236406698	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3313	74.6			108.8	4.184	3.91
2236373673	hypothetical protein	Candidatus Nitrospira defluvii	NIDE3276	54.24			86.96	16.3	3.44
2236401687	hypothetical protein	Candidatus Nitrospira defluvii	NIDE1338	82.5		76.92	81.18	0	
2236351506	hypothetical protein	Candidatus Nitrospira defluvii		47.5			783.4	100.1	4.48
2236352220	hypothetical protein						262.7	8.102	2.37
2236353357	hypothetical protein						445.1	48.66	1.85
2236354822	hypothetical protein						401.3	73.3	3.13
2236354828	hypothetical protein						425.8	76.36	3.13
2236353359	hypothetical protein						458.8	55.19	1.85
2236352222	hypothetical protein						230.3	12.01	2.37
2236377488	Membrane protease subunits, stomatin/prohibitin homologs	Candidatus Nitrospira defluvii	NIDE3711	70.59			122.2	3.215	
2236403053	Membrane protease subunits, stomatin/prohibitin homologs	Candidatus Nitrospira defluvii	NIDE3711	61.54			142.9	0	
2236418054	Multicopper oxidase.	Candidatus Nitrospira defluvii	NIDE3313	69.49			97.74	0	
2236415243	NADH:ubiquinone oxidoreductase 20 kD subunit and related Fe-S oxidoreductases	Candidatus Nitrospira defluvii	NIDE0614	89.83		84.75	54.64	0	1.79
2236355525	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	55.28	64.02		1027	205.4	6.82
2236355522	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	55.28	64.02		856.4	181.2	6.82
2236354819	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	52.82	57.86		440.4	103.3	1.85

2236354817	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	52.82	57.86		436.2	78.19	3.13
2236354825	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	52.82	57.86		415.4	67.2	3.13
2236354823	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	53.01	57.86		401.3	73.3	3.13
2236354829	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	52.82	57.86		425.8	76.36	3.13
2236354827	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	53.12	59.59		389.4	64.13	3.13
2236354689	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	56.42	58.69		630.4	161.3	6.96
2236354690	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	56.17	58.69		554.1	77.6	6.96
2236354821	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	57.41	62.06		412.1	77.11	2.68
2236370746	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	49.86	61.02		328	81.06	6.82
2236370745	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	50.14	61.3		332.7	78.23	6.82
2236354826	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	49.72	55.97		396	55.08	3.13
2236355582	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	64.94	72.78		594.6	185.5	6.82
2236407001	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	63.06	73		100.9	17.29	2.68
2236355519	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	47.12	61.9		386.7	108.8	2.03
2236396531	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	52.38	68.27		160.7	136.9	3.13
2236389678	Nitrite oxidoreductase, alpha subunit	Candidatus Kuenenia stuttgartiensis	NIDE3237	68.25	72.58		72.92	52.08	3.5
2236366173	Nitrite oxidoreductase, alpha subunit	Candidatus Nitrospira defluvii	NIDE3237	64.62	58.46		121.8	152.3	6.96
2236355524	Nitrite oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE3236	68.78	63.89		1027	205.4	6.82
2236355521	Nitrite oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE3236	68.78	63.89		856.4	181.2	6.82
2236348871	Nitrite oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE3236	69.12	63.77		530.3	81.55	3.24
2236348868	Nitrite oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE3236	69.12	63.77		505.8	84.67	3.24
2236348869	Nitrite oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE3236	70.57	64.33		371.8	71.19	3.24
2236365965	Nitrite oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE3236	65.33	65.33		223.7	30.7	3.24
2236409832	Nitrite oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE3236	67.69	64.62		268.9	396.2	1.54
2236348870	Nitrite oxidoreductase, membrane subunit	Candidatus Nitrospira defluvii	NIDE3278	38.36			530.3	81.55	3.24
2236348867	Nitrite oxidoreductase, membrane subunit	Candidatus Nitrospira defluvii	NIDE3278	38.36			505.8	84.67	3.24
2236351505	Nitrite oxidoreductase, membrane subunit	Candidatus Nitrospira defluvii	NIDE3293	45.06	37.78		783.4	100.1	4.48
2236376574	Nitrite oxidoreductase, membrane subunit	Candidatus Nitrospira defluvii	NIDE3278	54.17	49.14		2440	1042	3.75
2236355673	Nitrite oxidoreductase, membrane subunit	Candidatus Nitrospira defluvii	NIDE3278	51.16			1109	477.3	3.75
2236355674	Nitrite oxidoreductase, membrane subunit	Candidatus Nitrospira defluvii	NIDE3278	51.16			650.7	285.3	3.75
2236364616	Nitrite oxidoreductase, membrane subunit	Candidatus Nitrospira defluvii	NIDE3293	66.07			159.9	29.74	1.72
2236358575	Nitrite transporter	Candidatus Nitrospira defluvii	NIDE1364	55.05	40.82		415.7	6.359	3.02
2236358573	Nitrite transporter	Candidatus Nitrospira defluvii	NIDE1364	55.05	40.82		318.5	0	3.02
2236358571	Nitrite transporter	Candidatus Nitrospira defluvii	NIDE1364	54.7	38.46		416.1	1.963	2.5
2236384125	Nitrite transporter	Candidatus Nitrospira defluvii	NIDE1364	55.32			89.35	0	
2236351914	Nitrite transporter	Candidatus Nitrospira defluvii	NIDE1364	55.17	47.62		213	0	3.02
2236413789	Nitrite transporter	Candidatus Nitrospira defluvii	NIDE1364	72.13			60	0	1.46
2236351419	Nitrite transporter	Candidatus Nitrospira defluvii	NIDE1364	66.67	50		592.9	1.63	2.27
2236376310	Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclophilin family	Candidatus Nitrospira defluvii	NIDE1683	72.38			224.2	3.03	2.27
2236416751	Polyribonucleotide nucleotidyltransferase	Leptospirillum ferrodiazotrophum	NIDE4127	72.46		76.09	148.9	16.55	
2236400377	Predicted transcriptional regulator	Candidatus Nitrospira defluvii	NIDE1347	60.38			110.6	0	1.95

2236424351	Proto-chlorophyllide reductase 57 kD subunit.	Candidatus Nitrospira defluvii	NIDE0413	62.96			54.35	0	3.44
2236374791	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	62.43	35.56		372	44.3	1.09
2236374792	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	62.61	35.56		314.9	51.58	1.09
2236353360	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	62.69	34.62		458.8	55.19	1.85
2236353358	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	62.69	34.81		445.1	48.66	1.85
2236353356	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	62.69	34.62		405.3	35.01	1.85
2236397810	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	58.88	37.04		294.8	0	
2236393944	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	79.17	60.42		483.7	3.268	
2236350389	putative cytochrome bd-like oxidase	Candidatus Nitrospira defluvii	NIDE3296	67.27			543.5	21.74	1.52
2236384231	putative monoheme cyctochrome C	Candidatus Nitrospira defluvii	NIDE3295	43.07			223.3	0	
2236363341	putative monoheme cyctochrome C	Candidatus Nitrospira defluvii	NIDE3350	54.55		46.55	344.3	0	1.98
2236363340	putative monoheme cyctochrome C	Candidatus Nitrospira defluvii	NIDE3350	62.86		57.63	344.3	0	1.98
2236405281	putative monoheme cyctochrome C	Candidatus Nitrospira defluvii	NIDE3294	78.43			172.4	0	1.47
2236347574	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	65.32			361.1	29.63	3
2236396890	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	50.93			352	38	2.16
2236382019	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	55.78			249.1	34.6	2.37
2236382513	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	47.66			198.1	7.075	2.16
2236373071	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	65.42			438.9	19.08	2.37
2236373070	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	66.36			370.2	11.45	2.37
2236422675	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	55.42			83.67	0	2.55
2236363366	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	67.57			215.1	11.95	2.55
2236423378	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	61.64			125.6	0	
2236425237	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	56.06			75.38	0	
2236417255	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	61.02			132.7	9.479	2.37
2236425626	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	57.38			91.4	0	5.74
2236375812	Putative multicopper oxidases	Candidatus Nitrospira defluvii	NIDE3937	69.09			207.3	10.36	
2236378910	Pyruvate:ferredoxin oxidoreductase, alpha subunit	Candidatus Nitrospira defluvii	NIDE0971	76.56		75.24	198.4	48.82	
2236389232	Pyruvate:ferredoxin oxidoreductase, alpha subunit	Candidatus Nitrospira defluvii	NIDE0971	73.1		74.12	148.9	0	
2236388614	Pyruvate:ferredoxin oxidoreductase, alpha subunit	Candidatus Nitrospira defluvii	NIDE0971	81.1		79.27	268.1	48.39	
2236414119	Pyruvate:ferredoxin oxidoreductase, alpha subunit	Candidatus Nitrospira defluvii	NIDE0971	78.79		74.24	83.74	4.926	1.47
2236410692	Pyruvate:ferredoxin oxidoreductase, alpha subunit	Leptospirillum sp. Group II '5-way CG'	NIDE0971	73	n.a.	80	200	38.1	
2236396345	Pyruvate:ferredoxin oxidoreductase, alpha subunit	Leptospirillum sp. Group II '5-way CG'	NIDE0971	79.37	n.a.	85.71	163.3	5.102	
2236389369	Pyruvate:ferredoxin oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE0970	79.76	n.a.	72.62	120.8	7.547	2.35
2236410434	Pyruvate:ferredoxin oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE0970	79.45	n.a.	76.71	94.22	3.04	
2236394921	Pyruvate:ferredoxin oxidoreductase, beta subunit	Candidatus Nitrospira defluvii	NIDE0970	88.52	n.a.	83.61	72.16	0	
2236347475	Pyruvate:ferredoxin oxidoreductase, beta subunit	Leptospirillum sp. Group II '5-way CG'	NIDE0970	76.07	n.a.	78.63	192.4	16.26	3.88
2236417779	Pyruvate:ferredoxin oxidoreductase, gamma subunit	Candidatus Nitrospira defluvii	NIDE0969	69.62	n.a.	64.1	90.53	8.23	2.98
2236366485	Response regulator containing CheY-like receiver, AAA-type ATPase,	Candidatus Nitrospira defluvii	NIDE3585	40	n.a.	n.a.	266.7	54.32	3.75

	and DNA-binding domains								
2236404975	Riboflavin biosynthesis, RibBA	Leptospirillum ferrodiazotrophum	NIDE1383	68.5	n.a.	72.8	162.3	0	
2236407700	ribosomal protein L14	Candidatus Nitrospira defluvii	NIDE1321	72.22	n.a.	68.52	186.6	29.85	
2236424068	Ribosomal protein L5	Candidatus Nitrospira defluvii	NIDE1323	73.98	n.a.	n.a.	134.9	2.646	
2236392527	Ribosomal protein S1	Candidatus Nitrospira defluvii	NIDE0426	73.08	n.a.	70.31	88.61	0	
2236401688	Ribosomal protein S11						81.18	0	
2236428182	ribosomal protein S12	Candidatus Nitrospira defluvii	NIDE1306	90.54	n.a.	83.78	80.36	0	2.3
2236419086	Ribosomal protein S12	Candidatus Nitrospira defluvii	NIDE1306	92	n.a.	88	85.71	0	
2236400350	Ribosomal protein S7	Candidatus Nitrospira defluvii	NIDE1307	77.66		63.74	63.83	0	
2236405975	RNA polymerase sigma factor, sigma-70 family	Candidatus Nitrospira defluvii	NIDE2309	53.4		53	233.2	46.65	
2236351507	RNA polymerase, sigma-24	Candidatus Poribacteria sp. WGA-A3		39.64	n.a.	n.a.	783.4	100.1	4.48
2236411333	Serine-pyruvate aminotransferase/archaeal aspartate aminotransferase	Candidatus Nitrospira defluvii	NIDE3444	61.29	48.39	54.84	62.83	0	
2236366352	Short-chain alcohol dehydrogenase	Candidatus Nitrospira defluvii	NIDE0797	66.36	n.a.	n.a.	165.2	0	
2236414421	Translation elongation factor Ts	Candidatus Poribacteria sp. WGA-A3	NIDE2758	50.79	52.38	58.73	140	0	1.46
2236405443	Transposase IS200 like.	Candidatus Nitrospira defluvii	NIDE2608	75		78.26	65.42	0	1.48
2236410178	Uncharacterized conserved protein	Candidatus Nitrospira defluvii	NIDE3451	49.3	42.67		156.8	0	
2236423820	Uncharacterized conserved protein	Candidatus Nitrospira defluvii	NIDE3299	50			63.06	0	1.12

4.6 References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Baker BJ, Lesniewski R, Dick GJ (2012) Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J* 6, 2269-2279.

Bates ST, Berg-Lyons D, Caporaso, JG, Walters WA, Knight R, Fierer N. (2010) Examining the global distribution of dominant archaeal populations in soil. *ISME J*. 5, 908-917.

Bazylinski DA, Wirsen CO, Jannasch HW (1989) Microbial utilization of naturally occurring hydrocarbons at the Guaymas Basin hydrothermal vent site. *Appl. Environ. Microbiol.* 55, 2832-2836.

Bock E, Koops HP, Harms H, Ahlers B In, Shively, JM, Barton, LL. (Eds.), (1991) Variations in Autotrophic Life. Academic Press, London, pp. 171–200.

Boon B, Laudelout H. (1962) Kinetics of nitrite oxidation by *Nitrobacter winogradskyi*. *Biochem. J.* 85, 440-447.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Costello EK, Fierer N, *et al.* (2010) QIIME allows analyses of high-throughput community sequencing data. *Nat Meth* 7, 335-336.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72,5069-5072.

Dick GJ, Tebo BM (2010) Microbial diversity and biogeochemistry of the Guaymas Basin hydrothermal plume. *Environ. Microbiol.* 12, 1334-1347.

Dore JE, Karl DM (1996) Nitrite distributions and dynamics at station ALOHA. *Deep-sea Res. II* 43, 385-402.

Ehrlich S, Behrens D, Lebedeva E, Ludwig W, Bock E (1995) A new obligately chemolithoautotrophic, nitrite-oxidizing bacterium, *Nitrospira moscoviensis* sp. nov. and its phylogenetic relationship. *Arch. Microbiol.* 164, 16-23.

Frias-Lopez, J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, DeLong EF (2007) Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci*. 105, 3805-3810.

Frigaard NU, Martinez A, Mincer TJ, DeLong EF (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* 439, 847-850.

Füssel J, Lam P, Lavik G, Jensen MM, Holtappels M, Gunter M, Kuypers MM. (2012) Nitrite oxidation in the Namibian oxygen minimum zone. *ISME J.* 6, 1200-1209.

Ivars-Martínez E, Martin-Cuadrado AB, D'Auria G, Mira A, Ferriera S, Johnson J, *et al.* (2008) Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME J*. 2, 1194-1212.

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV (2012) Untangling genomes from metagenomes, revealing an uncultured class of marine Euryarchaeota. *Science* 33, 587-590.

Jiao N, Zheng Q (2011) The microbial carbon pump, from genes to ecosystems. *Appl. Environ. Microbiol.* 77, 7439-7444.

Karl DM, Church MJ, Dore JE, Letelier, Mahaffey C (2012) Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc. Natl. Acad. Sci*. 109, 1842-1849.

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

Kleiner M, Wentrup C, Lott C, Teeling H, Wetzel S, Young J, *et al*. (2012) Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proc. Natl. Acad. Sci.* early edition.

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

Koops H-P, Pommerening-Roser A (2001) Distribution and ecophysiology of the nitrifying bacteria emphasizing cultured species. *FEMS Microbiol. Ecol.* 37,1-9.

Lebedeva EV, Off S, Zumbrägel S, Kruse M, Shagzhina A, Lücker S, *et al.* (2011) Isolation and characterization of a moderately thermophilic nitrite-oxidizing bacterium from a geothermal spring. *FEMS Microbiol. Ecol*. 75, 195-204.

Lesniewski R, Jain S, Anantharaman K, Schloss PD, Dick GJ (2012) The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J.* in press

Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B, *et al*. (2010) A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc. Natl. Acad. Sci.* 107, 13479-13484.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, *et al.* (2004) ARB, a software environment for sequence data. *Nuc. Acids Res.* 32, 1363-1371.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrow-Wheeler Transform. *Bioinformatics* 25, 1754-1760.

Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Grechkin Y, *et al*. (2012) IMG/M, the integrated metagenome data management and comparative analysis system. *Nucl Acids Res*. 40, D123-D129.

Martin-Cuadrado A-B, Rodriquez-Valera F, Moreira D, Alba D, Ivars-Martínez E, Henn MR, Talla E, López-García P (2008) Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J.* 2, 865-886.

Miller CS, Baker BJ, Thomas BC, Singer S, Banfield JF (2011) EMRIGE, reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Gen. Biol.* 12, R44.

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

Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan L-K, *et al*. (2012) Sizing up metatranscriptomics. *ISME J* 7, 237-243.

Pester M, Schleper C, Wagner M (2011) The Thaumarchaeota, an emerging view of their phylogeny and ecophysiology. *Curr. Opin. Microbiol*. 14, 300-306.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner RO (2007) SILVA, a comprehensive online resourse for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nuc. Acids Res.* 35, 7188-7196.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, *et al.* (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat. Meth.* 6, 639-641.

Raven JA, Falkowski PG (1999) Oceanic sinks for atmospheric CO_2. *Plant, Cell & Eviron*. 22, 741-755.

Santoro AE, Buchwald C, McIlvin, Casciotti KL (2011) Isotopic signature of N2O produced by marine ammonia-oxidizing Archaea. *Science* 333, 1282-1285.

Santoro AE, Casciotti, KL, Francis CA (2010) Activity, abundance and diversity of nitrifying archaea and bacteria in the central California current. *Environ. Microbiol.* 12, 1989-2006.

Stewart FJ, Ulloa O, DeLong EF (2011) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ. Microbiol.* 14, 23-40.

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

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.

Schmieder R, Lim YW, Edwards R (2012) Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* 28, 433-435.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) *Oases*, Robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28, 1086-1092.

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

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

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

Strous M, Pelletier E, Mangenot S, Rattei T, Lehner A, Taylor MW, *et al*. (2006) Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature* 440, 790-794.

Sorokin DY, Lucker S, Vejmelkova D, Kostrikina NA, Kleerebezem R, Rijpstr WI, *et al*. (2012) Nitrification expanded, discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum Chloroflexi. *ISME J.* in press.

Ward BB, Capone, DG, Zehr JP (2007) What's new in the nitrogen cycle? *Oceanography* 20, 101-109.

Watson SW, Bock E, Valois FW, Waterbury JB, Schlosser U (1986) *Nitrospira marina* gen. nov., sp. nov., a chemolithotrophic nitrite-oxidizing bacterium. *Arch. Microbiol*. 144, 1-7.

Wertz S, Leigh AKK, Grayston SJ (2011) Effects of long-term fertilization of forest soils on potential nitrification and on the abundance and community structure of ammonia oxidizers and nitrite oxidizers. *FEMS Microbiol. Ecol*. 79, 142-154.

Wright J, Konwar KM, Hallam SJ (2012) Microbial ecology of expanding oxygen minimum zones. *Nature* 10, 381-394.

Wuchter C, Abbas B, Coolen MJL, Herfort L, Bleijswijk, Timmers P, *et al*. (2006) Archaeal nitrification in the ocean. *Proc. Natl. Acad. Sci*. 103, 12317-12322.

Zehr JP, Kudela RM (2011) Nitrogen cycle of the open ocean, from genes to ecosystems. *Ann. Rev. Mar. Sci*. 3, 197-225.

CHAPTER V

GENOMIC RESOLUTION OF LINKAGES IN CARBON, NITROGEN, AND SULFUR CYCLING AMONG WIDESPREAD ESTUARY SEDIMENT BACTERIA

Brett J. Baker[a], Cassandre Lazar[b,c], Andrease Teske[b], and Gregory J. Dick[a,d]

a. Department of Earth and Environmental Sciences, University of Michigan, Ann Arbor, MI, USA
b. Marine Sciences, University of North Carolina, Chapel Hill, NC
c. Organic Geochemistry Group, MARUM Center for Marine Environmental Sciences, Department of Geosciences, University of Bremen, Germany
d. Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

Publication in review at *Proc Natl Acad Sci.USA*

5.1 Abstract

Microbes associated with sediments in high productivity estuaries are important drivers of global cycling of carbon, nitrogen, and sulfur. However, little is known about how ecological and metabolic processes are partitioned among members of these communities. We reconstructed 82 near-complete bacterial genomes from three redox regimes (sulfate-rich, sulfate-methane transition (SMTZ), and methane-rich zones) in White Oak River (North Carolina) estuary sediments. These genomes belong to 21 bacterial groups that are widespread in estuary and marine sediments, including uncultured candidate phyla (BRC1, WS3, OD1, Zixibacteria, TA06, KSB1, and KD3-62), and two new candidate phyla (WOR-1 and BRC2). The candidate phyla are generally most abundant in the SMTZ and methane-rich zones, and mediate essential biogeochemical pathways of the estuarine environment, including organic carbon degradation and fermentation. Genes encoding production and utilization of acetate and hydrogen are widespread in the bacterial community; these processes likely represent important links between fermentation and respiratory metabolisms. Among the most abundant organisms in the sulfate-rich layer are novel, uncultured members of the *Gammaproteobacteria* that have the genetic

potential for anaerobic sulfur oxidation coupled to nitrate and nitrite reduction. These genome sequences of widespread yet uncultured bacteria provide unprecedented insights and suggest specific hypotheses about the physiologies and geochemical interactions between bacterial community members in estuarine sediments.

5.2 Introduction

Estuaries are a dynamic biogeochemical environment and represent a key crossroads in the global carbon cycle. Processes within estuaries mediate the transfer of carbon from land to sea, release a considerable amount of CO_2 to the atmosphere (Cai et al. 2011), and sequester carbon in sediments (Bauer et al. 2013). Microorganisms in estuary sediments mediate central geochemical processes including degradation of organic carbon and cycling of sulfur, nitrogen, and iron. However, our ability to partition these processes and the underlying metabolic pathways among specific microbial groups is limited by the complexity and abundance of uncultured groups present in sediment communities. Metagenomic studies have characterized the genetic potential of marine sediment microbial communities, highlighting their novelty and abundance of unknown genes (for example, see Biddle et al. 2008). However, the analysis of individual sequencing reads typically does not link individual community members and their metabolic pathways (Baker and Dick, 2013). Recently, single-cell genomics revealed metabolic pathways for protein degradation in two widespread lineages of marine benthic Archaea (Lloyd et al. 2013). Further culture-independent genomic reconstructions of estuary sediment microbial communities are lacking; thus we know little about uncultured communities and candidate phyla in this complex microbial ecosystem.

To better understand the metabolic capabilities of uncultured bacteria we obtained high-throughput genomic libraries from sediment profiles from the White Oak River estuary, North Carolina. This genomic dataset was assembled and binned to obtain near-complete genomes, and to reconstruct metabolic pathways of numerous community members. Placing the genome-encoded metabolic capabilities of specific bacterial groups into community context enabled us to identify several microbial interactions involving carbon, iron, nitrogen, and sulfur cycling in the sediments. This dataset provides a comprehensive genomic road map of geochemical and ecological functions of aquatic sediment microbial communities.

5.3 Results and Discussion

Genomic reconstruction and identification. Sediment samples were collected with push cores at three adjacent mid-estuary locations. Since the distinct redox layers shared similar bacterial and archaeal communities across the three sites, we combined genomic libraries from the three sites and conducted one genomic assembly for each zone: the sulfate-rich zone (8-12 cm), sulfate-methane transition zone (SMTZ) (24-32 cm), and methane-rich zone (52-54 cm). Subsequent binning by tetra-nucleotide frequency coupled with genomic coverage resulted in over 120 genomic bins of Bacteria (Fig. S1). Based on completeness, taxon coverage and genomic novelty, we chose 82 bins for detailed characterization of genome-encoded metabolic pathways. 26, 35, and 21 of the bins are from the sulfate-rich, SMTZ, and methane-rich zones, respectively. 88% and 67% of these genomes are estimated to be >70% and >90% complete, correspondingly (Table S1). Just 8 of the bins have more than one genome, based on numbers of single-copy genes (Raes et al. 2007). Only 30 of these genomes contain 16S rRNA genes (>300 bp) due to fragmentation commonly seen in short-read assemblies (Miller et al. 2012). Therefore, we also used ribosomal protein S3 phylogeny to determine the taxonomic identities of the remaining bins (Fig. 1) (Hug et al. 2013, Castelle et al. 2013). Ribosomal proteins belonging to novel phyla for which reference sequences were not available were identified by the presence of at least one 16S rRNA gene in the clade (Fig. S2).

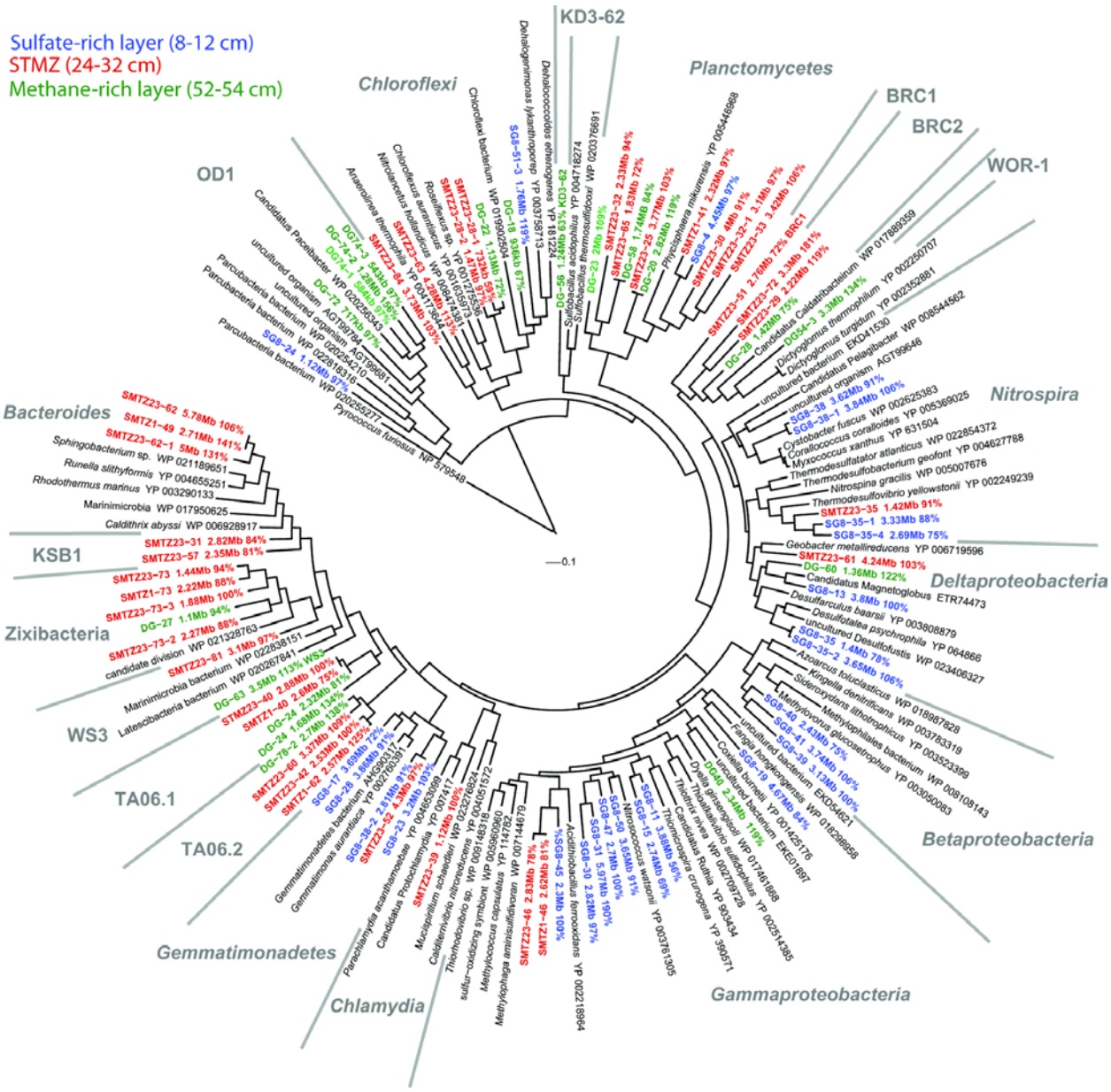

Fig. 5.1. Diversity of organisms from which genomes were reconstructed from the White Oak River sediments. Phylogenetic tree inferred from ribosomal protein S3 genes present within bacterial genomes from the sediment metagenomic assemblies. Each sequence in bold is from one genomic bin. Some clade designations are based on 16S rRNA gene phylogenetic analyses (Fig. S2); for example WOR-1 is a distinct novel phylum. Bin names and genome size and completeness are also shown. The total size (Mb) and estimated completeness (%) of the bins is shown.

Several of the genomic bins belong to groups that are commonly identified in rRNA gene surveys of marine and estuarine sediments, including *Betaproteobacteria*,

Gammaproteobacteria, *Deltaproteobacteria*, *Chloroflexi*, *Planctomycetes*, *Bacteroidetes*, *Gemmatimonadetes*, *Nitrospira*, *Chlamydiae*, and *Spirochetes* (Beazley et al. 2012, Bowen et al. 2012). Genomes of candidate phyla that are commonly identified in a variety of anoxic environments and marine sediments comprise OD1, WS3, TA06, Zixibacteria, and BRC1 (Table S2). Candidate phylum TA06 contains two distinct groups (hereby named TA06.1 and TA06.2) based on 16S rRNA and ribosomal protein gene phylogeny (Fig. 1 and S2), GC content (average 58% and 44%), and predicted physiology (detailed below). Three genomes forming a sister lineage to BRC 1 in the ribosomal protein S3 phylogenetic tree are specified as the new lineage BRC2, based on distinct GC content (42% vs 60%) (Fig. 1). One genomic bin belongs to a previously unnamed phylum-level lineage of 16S rRNA genes, mostly recovered from marine sediments, that we refer to as "WOR-1". This bin (DG-54-3) clusters with BRC1 and BRC2 phyla on the ribosomal protein tree, but it is deeply divergent from all currently recognized phyla based on 16S rRNA gene phylogeny (Fig. S2).

Genomic abundance of community members in the sediment profile. To quantify the genomic abundance of community members in each zone we mapped all the reads to all of the genes for ribosomal protein S3 in the assemblies (Fig. S3). The assembly from shallow, sulfate-rich sediments was dominated by *Beta-*, *Gamma-* and *Deltaproteobacteria* (including *Myxococcales*), *Bacteroidetes* and *Nitrospira*, and also contained more *Gemmatimonadetes* and *Planctomycetes* relative to the deeper sediment samples. The SMTZ and the deeper methane-rich sediments were dominated by Archaea and *Chloroflexi* (Fig. S3), consistent with previous qPCR and rRNA slot blot results showing that bacterial dominance in surficial White Oak River estuarine sediments is considerably reduced and even reversed downcore (Kubo et al. 2012). The Candidate bacterial phyla WS3, OD1, TA06.1, TA06.2, and Zixibacteria were sufficiently abundant for genome reconstruction only in the deeper sediment layers.

Organic Carbon degradation and fermentation. Sedimentary microbial communities process the input of photosynthetic organic matter from the overlying water column, and thus play a key role in the degradation of complex carbon substrates (Arnosti et al. 2011). We searched all the bacterial genomes present here for 26 known genes involved in degradation of chitin, protein, algal- and plant-derived material, and lipids (Berlemont and Martiny, 2013, Wrighton et al.

2014). Generally speaking, the genomes belonging to the *Bacteroidetes*, *Gemmatimonadetes*, *Planctomycetes*, WOR-1, BRC1, BRC2, and KD3-62 phyla contain the broadest array of glycoside hydrolases genes. These organisms have a variety of cellulose, hemicellulose, and polysaccharide degradation genes absent in the other community members (Table S3), suggesting they drive the initial degradation and hydrolysis of complex organic carbon compounds. Multiple genes are involved in the degradation of chitin, the long-chain polymer of N-acetylglucosamine and a major structural component of fungal and algal cell wells and of arthropod exoskeletons. Endo-acting chitinase genes were identified in two of the Gammaproteobacteria (SG8-31 and SG8-15) and the KSB1 bin SM23-31, and *N*-acetyl-glucosaminidase genes were found in the *Gemmatimonadetes*, *Bacteroidetes* and the three TA06.2 genomes (Table S3).

Proteins account for a large proportion of bioavailable carbon and nitrogen for sediment communities (Wakeham et al. 1997). In multiple estuary phyla, these resources are accessed via extracellular peptidases. We found the greatest numbers of peptidases in the candidate phyla TA06, WOR-1, and WS3, and in the *Bacteroidetes*, *Gemmatimonadetes*, *Chloroflexi*, and *Planctomycetes*, suggesting that these community members are involved in protein degradation along with sedimentary benthic Archaea (Lloyd *et al*. 2013).

Several of the *Chloroflexi* genomes (SG8-19, DG-18, SM23-63 and 84) contain the *β*-oxidation pathway to generate acetyl-CoA from fatty acids and organic acids, as found previously (Hug et al. 2013, Wasmund et al. 2014). This capability is also present in genomes from both of the *Myxococcales*, *Gemmatimonadetes* (SG8-23 and 28), several *Gammaproteobacteria* and *Deltaproteobacteria*, and the shallow sediment-dwelling *Betaproteobacteria*. Among the candidate phyla that contain the complete *β*-oxidation pathway are the BRC-1, and one of the Zixibacteria (SM23-57) genomic bins. Interestingly, genomes that have the *β*-oxidation pathway were primarily found in the shallow samples, where 9 of the 15 complete pathways are located.

Several of the bacteria groups capable of hydrolyzing complex organic carbon also have pathways for glycolytic fermentation of glucose to acetate, including BRC2, WOR-1 (Fig. 3), WS3, *Bacteroidetes*, *Nitrospira* (SG8-3), and *Spirochetes* (bin DG-61). All these groups have the reductive acetyl-CoA (Wood-Ljungdahl) and phosphate acetyltransferase-acetate kinase pathways for carbon fixation and acetate production; we note that the WS3 and WOR-1 genomes

appear to lack acetyl-CoA synthetase, essential to the Wood-Ljungdahl pathway. The *Spirochetes* and BRC2 genomes contain lactate dehydrogenase genes suggesting they are also capable of lactate fermentation. The *Bacteroidetes*, *Spirochetes* and *Thermodesulfovibrio*-like bins have genes that encode aldehyde dehydrogenase and alcohol dehydrogenase, suggesting they are capable of full fermentation to ethanol. Fermentation has been demonstrated in *Thermodesulfovibrio* spp. cultures (Sekiguchi *et al*. 2008). The genome bins did not yield an identifiable complete pathway for butyrate formation. The end products of these fermentation pathways fuel terminal respiration in the sediment community; the key electron donors acetate and hydrogen are among the principal drivers of sulfate reduction (Oremland and Polcin, 1982) (Fig. 5.2).

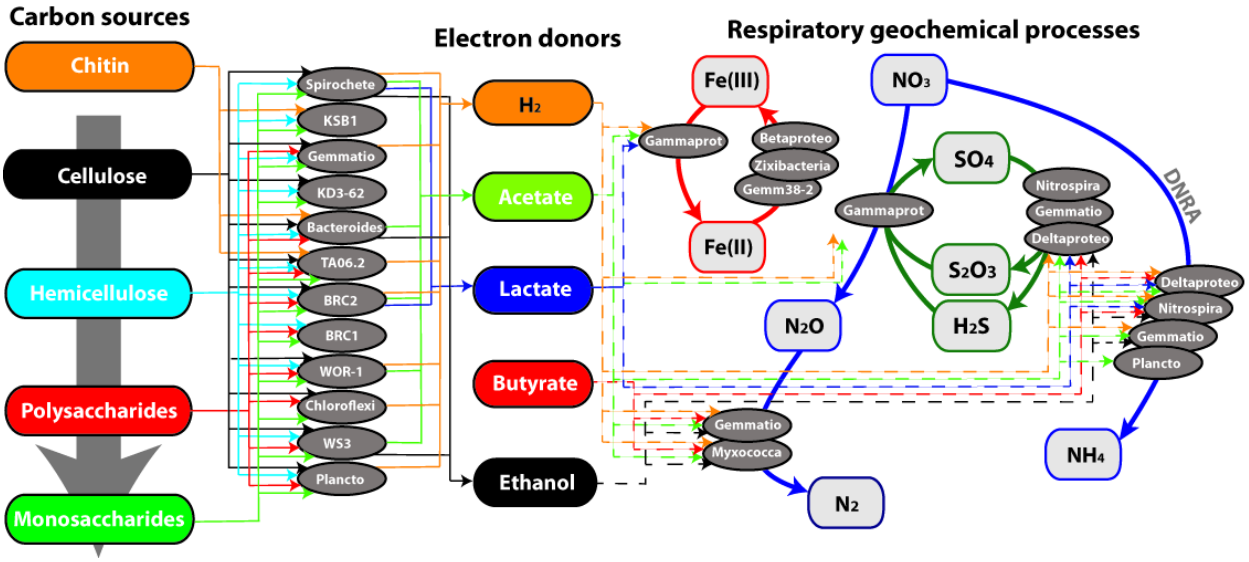

Figure 5.2. Flow diagram of the interactions between (left to right) organic carbon utilization, fermentation, and respiration identified in the bacterial genomes reconstructed in this study. Arrows represent metabolic capabilities that were identified in the metagenomic reconstruction from the White Oak River estuary. The dashed lines on the right represent potential electron donors for the anaerobic respiration processes. Note that the and *Gammaproteobacteria* are capable of coupling nitrate reduction to either thiosulfate or sulfide oxidation. DNRA = dissimilatory nitrate reduction to ammonia.

We did not identify any genes for Fe,Fe-hydrogenases, which are thought to primarily produce H_2 (Seiber et al. 2012). However, several of the bacteria possess genes for Ni,Fe-hydrogenases, which can be involved in H_2 production or consumption. These genes are ubiquitous in *Gemmatiomonas*, *Myxococcales*, *Delta-*, and Gammaproteobacteria (SG-11, 13,

15, 30, and 31) and are likely used for consumption of H_2 by respiratory processes (eg. sulfate reduction and denitrification). The *Nitrospira* bin (SG8-35-4, related to *Thermodesulfovibrio* spp., Fig. 1) contain genes for Ni,Fe-hydrogenases that are likely to participate in both sulfate reduction with hydrogen as the electron donor and the fermentative production of H_2, which have been demonstrated in *Thermodesulfovibrio* spp. (Sekiguchi et al. 2008). Several bacterial groups that are capable of organic carbon degradation and fermentation have Ni,Fe-hydrogenase genes, including *Planctomycetes*, *Spirochetes*, *Chloroflexi*, BRC2, TA06.2 (not TA06.1), and WOR-1 (Fig. 3). The extensive distribution of these hydrogenases among both fermenting and respiring bacteria indicates that H_2 is a highly dynamic electron carrier produced and consumed by a wide range of sediments microbes, as it is in other anoxic environments (Wrighton et al. 2014, Schmidt et al. 2010).

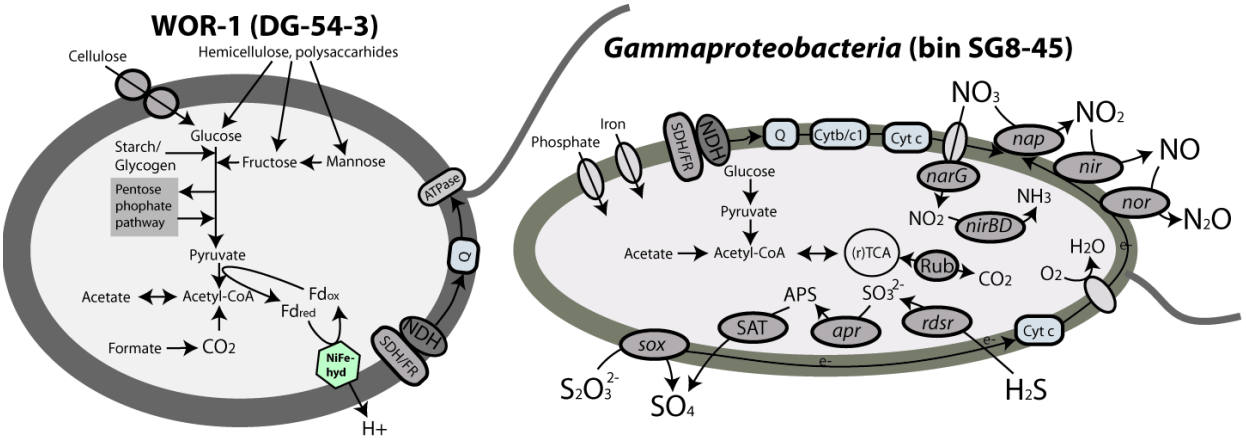

Figure 5.3. Diagrams of metabolic potential and electron transport of WOR-1 (bin DG-54-3) and Gammaproteobacteria (bin SG8-45), based on gene content. ATPase, ATP synthetase; FDH, formate dehydrogenase; NiFe-hyd, Ni,Fe-hydrogenase; Cytb/c1, quinone cytochrome oxidoreductase; Cyt c, cytochrome c; nap/nar, nitrate reductase; nir, nitrite reductase; nor, nitric oxide reductase; nir, nitrous oxide reductase; SAT, sulfate transferase; apr, APS reductase; rdsr, reverse dissimilatory sulfite reductase; Rub, RuBisCO; Q, quinine; SDH/FR, succinate dehydrogenase/fumarate reductase; NDH, NADH dehydrogenase; SOX, sulfur oxidation multienzyme complex.

Dissimilatory sulfur and nitrogen cycling. Aquatic sediments are characterized by redox gradients, as oxidized compounds (e.g., the electron acceptors O_2, nitrate and sulfate) are gradually reduced by anaerobic respiration. To determine the respiratory repertoire of microbial

community members, we surveyed the estuary genomes for key genes of respiration pathways. Several of the genomic bins contain dissimilatory sulfite reductase (*dsr*) genes, indicative of microbial sulfate and sulfite reduction. To account for the possibility that a related but phylogenetically distinct group of genes (rdsr) mediate the reverse reaction (Loy et al. 2009), we generated a phylogenetic tree of all the *dsr*-like genes recovered (Fig. S4). The *Deltaproteobacteria,* which constituted the most abundant microorganisms in the sulfate-rich zone (Fig. S3), have reducing-type *dsr* genes and complete sulfate reduction pathways (with the exception of the two *Myxococcales* bins). We also found the bin SG8-35-4, similar to *Thermodesulfovibrio* spp. within the *Nitrospira* phylum, is capable of sulfate reduction based on the presence of *dsr* genes.

Surprisingly, a complete sulfate reduction pathway was identified in the *Gemmatiomonadetes*-like bin SG8-17. The *dsrAB* genes from this bin fall within a phylogenetically deeply branched clade (Fig. S4). This clade (designated DSR-J, 22) includes *dsrAB* sequences from an intertidal sand flat, Hydrate Ridge, deep sea, and estuary sediments, suggesting these sulfate reducers are widespread in coastal and marine sediments. Genes of this clade were hypothesized to have been horizontally transferred (Mussmann et al. 2005). We did not identify *dsr* genes in any of the other *Gemmatiomonadetes*-like bins; therefore it is likely that these genes were recently acquired in that organism. Since all of the sequences for this clade had previously been recovered from large-insert (fosmid) clones that lacked 16S rRNA genes, their taxonomic affiliation had been uncertain. These *dsrAB* genes are located on a 9.6kb contig that is confidently binned with SG8-17. This bin contains additional genes for sulfate reduction on other contigs, including *aprAB*, SAT, and *dsrC*. Thus, this member of the *Gemmatimonadetes* is likely capable of sulfate reduction.

Up to 95% of the sulfide and thiosulfate generated by sulfate reduction is re-oxidized to sulfate in marine sediments (Jørgensen, 1990). Interestingly, the genomes of the most abundant *Gammaproteobacteria* (SG8-11, -15, -45, -47 -50, SMTZ1-46, and SMTZ-46) include genes for sulfide oxidation (*rdsr, apr,* and SAT) and pathways for thiosulfate oxidation (*soxABDZY* genes). We did not identify any sulfide quinone or flavocytochrome c reductases which are common to sulfide oxidizers. All of these genomes contain genes for nitrate reduction. Many bins (SG8-30, -31, -41, -45, and -46) also have *nirS* and *norBC* genes for the reduction of nitrite to nitrous oxide, N_2O (Fig. 3). *nosZ* genes were only found in *Gemmatimonadetes* (SG8-23) and

Myxococcales (SG-38 and -38-1) bins, suggesting they are reducing the N_2O produced by *Gammaproteobacteria* to N_2 (Fig. 2). Since all of these genomes are nearly complete (Table S2), the lack of genes for denitrification is unlikely to reflect incomplete sequence coverage. Two gammaproteobacterial genomes (SG8-30 and -31) lack the sox pathway and *rdsr* genes but have genes for Ni-Fe-hydrogenases, suggesting they utilize H_2 rather than reduced sulfur as an electron donor for nitrate reduction.

In several *Gammaproteobacteria*, the gene clusters for nitrate reduction and sulfur oxidation are mutually intertwined. The *rdsr* gene cluster of bins SG8-11, -15, -41, -45, -47, and -50, includes a gene with homology to a nitrate sensor (*narX*) and a *luxR*-like transcription regulator (Fig. 4). This type of sensor protein has been implicated in gene expression in response to changes in nitrate/nitrite concentrations (Nohno et al. 1989). Further, the gene cluster for nitrate reduction (*napABCDGH*) in the gammabacterial bins SG8-45 and SG8-50 contains two *dsrC* genes, which have been suggested to regulate *rdsr* gene expression (Grimm et al. 2010). The presence of two *dsrC* genes in the *nap* operon suggests that regulation of sulfur oxidation and nitrate reduction is coordinated, consistent with of the coupling of these processes in their close relative *Thioalkalivibrio* spp. (Sorokin et al. 2004 and 2008). The preference of these *Gammaproteobacteria* in shallow sediment samples, where the sulfide porewater concentrations decrease to the detection limit (Lazar et al. in review), is consistent with nitrate-dependent oxidation of sulfur compounds. We did identify cytochrome c oxidases in SG8-11, -31, -45, and -50, suggesting that some of the *Gammaproteobacteria* are also capable of O_2 reduction.

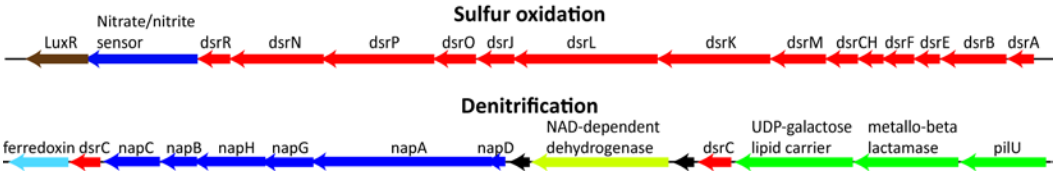

Figure 5.4. Operons for sulfur oxidation and nitrate reduction present in the dominant Gammaproteobacteria genotypes. Those shown here are present in the SG8-45 bin. However, syntenous operons are also present in several other Gammaproteobacteria bins (SG8-11, 15, 45, 47, 50, STMZ1-46, and SMTZ-46).

In addition to denitrification, several members of the community are putatively capable of dissimilatory nitrate reduction to ammonia (DNRA). *Deltaproteobacteria* (SG8-35-2, SM-61), *Gemmatimonadetes* (SG8-23), *Chloroflexi* (SM-28-1), *Nitrospira* (SG8-35-4), and WS3 (DG-63)

have formate-dependent nitrite reductase (*nrfA)*, which is the signature gene for DNRA (Einsle, 2011). We also found nitrogen fixation genes (*nifH*) in WS3 (DG-63) and *Nitrospira* (SG8-35-1).

Iron cycling. While the microbial cycling of iron in marine sediments has been commonly documented by geochemical approaches (Moeslund et al. 1994), well-documented biochemical pathways for iron oxidation and reduction now enable environmental genetic studies of these processes as well. For example, periplasmic and outer-membrane-anchored c-type cytochromes and a beta-barrel protein within the metal reduction (Mtr) respiratory pathway are essential for respiratory electron transport across the outer cell membrane to iron minerals in *Shewanella oneidensis* (Coursolle and Gralnick, 2012, White et al. 2013). Homologues of these enzymes catalyze lithotrophic iron oxidation in *Sideroxydans lithotrophicus* (Liu et al. 2012) and phototrophic iron oxidation in *Rhodopseudomonas palustri*s (Jiao and Newman, 2007). Therefore, we searched our datasets for genes homologous to *mtr* genes and found them in 7 different genomic bins. Betaproteobacterial genomes reconstructed from the sulfate-rich zone (SG8-39, 40, and 41) have *mtrABC* genes and are closely related (>98% at the 16S rRNA gene level) to *Sideroxydans lithotrophicus* (Fig. 1, S2), suggesting that they are capable of iron oxidation. *Gemmatimonadetes* bin SG8-38-2 has a putative *mtrABC* gene cluster and may be involved in iron cycling as well. Consistent with Zixibacteria genomes obtained from groundwater, bin SM-73-2 contains *mtrAB* but lacks the genes for extracellular cytochromes implicated in iron reduction (Castelle et al. 2013). Only the *Gammaproteobacteria* bin SM-47 contains both the genes homologous to *mtrABC* and a cytoplasmic membrane-associated c-type cytochrome (*cymA*) required for iron reduction (Coursolle and Gralnick, 2012, Myers and Myers, 2007). The closely related bacteria SG8-11 and SG8-30 have all these genes except *mtrC* and *cymA*, respectively. Bacteria capable of iron reduction have multiple multi-heme cytochromes (Carlson et al. 2012); SG8-11 and SG8-30 also each have genes encoding six unique types of these cytochromes.

5.4 Conclusions

The highly-resolved genomic reconstruction of estuary sediment microbial populations conducted here enabled us to piece together physiological pathways of individual community members, including several recently defined (eg. KD3-62, TA06, Zixibacteria, and BRC1) and two newly described (WOR-1 and BRC2) uncultured candidate phyla. Hence, this study provides the first genomic information on microorganisms that are widespread and mediate key biogeochemical processes in marine and estuary sediments. The 17 genomes belonging to WOR-1, KSB1, KD3-62, TA06 (groups 1 and 2), BRC1, and BRC2 are the first to be constructed from these phyla. Based on their genome sequences, several of these groups are capable of hydrolysis and fermentation of a variety of organic compounds, greatly expanding the range of bacterial phyla that hydrolyze and ferment biopolymers (sugars and proteins) to low molecular weight substrates. These genomes have an average of 14 carbohydrate hydrolases per genome, with WOR-1 and BRC2 having the most with 20 and 39, respectively. *Planctomycetes* and *Bacteroides* are also among the most versatile carbohydrate degrading bacteria in the White Oak River sediments with an average of 33 and 100 hydrolytic genes per genome, respectively.

This study identified new bacterial capabilities in sulfur cycling. One uncultured *Gemmatimonadetes* bacterium was linked to a previously taxonomically unassigned *dsr* gene clade, suggesting that this group is capable of sulfate reduction, a process that is commonly catalyzed by *Deltaproteobacteria*. Sequences for this group have been recovered from sediments throughout the world, suggesting that it is widespread. The oxidation of sulfide and thiosulfate is essential to sulfur cycling in marine and coastal sediments (Jørgensen, 1990), but little is known about the organisms involved. Recent evidence suggested that *Gammaproteobacteria* mediate sulfur oxidation in coastal sediments (Lenk et al. 2010). Several of the most abundant *Gammaproteobacteria* in the sulfate-rich zone described here have the genetic potential for coupling anaerobic sulfur or hydrogen oxidation to denitrification. Many sediment bacteria are commonly considered to be specialized to certain niches. However, our findings suggest that some can toggle between different oxidative and reductive functionalities depending on the dynamic redox state and the availability of different nitrogen and sulfur species in their sedimentary habitat.

Placing the metabolic capabilities of individual populations within the framework of other community members provided a wiring diagram of the potential geochemical interactions at a system level. This information provides a foundation for more fully understanding the

pathways of carbon degradation and cycling of sulfur, nitrogen, and iron in marine sediments. Several metabolic processes are present in multiple bacteria in the community. Further investigations assessing gene expression, and additional locations, will provide insights about the differential activity of these redundancies. Prior to this study our understanding of ecological roles of sediment bacteria was limited by low-resolution approaches. The comprehensive genomic reconstruction of these sediments provides a valuable link between function and diversity and reveals potential geochemical interactions within the communities. Full knowledge of the microbial players will provide a realistic road map that follows the flow on carbon and nutrients through the multiple layers of microbial processing, assimilation and remineralization in the estuarine environment.

5.5 Materials and Methods

Genomic analyses. Illumina (HiSeq) shotgun genomic reads were screened against Illumina artifacts (adapters, DNA spike-ins) with a sliding window with a kmer size of 28 and a step size of 1. Reads with 3 or more N's or with average quality score of less than Q20 and a length <50 bps were removed. Screened reads were trimmed from both ends using a minimum quality cutoff of 5 using Sickle (https://github.com/najoshi/sickle). Trimmed, screened, paired-end Illumina reads were assembled using IDBA-UD (39) with the following parameters (--pre_correction --mink 55 --maxk 95 --step 10 --seed_kmer 55). To maximize assembly reads from different sites were co-assembled (see supporting information).

Genes were called and putative function was assigned using the JGI IMG/MER system (Markowitz et al. 2012). Initial binning of the assembled fragments was done using tetra-nucleotide frequencies signatures and ESOM mapping as detailed in Dick et al. (2009) and binning was enhanced by incorporating coverage signatures for all of the assembled contigs (Sharon et al. 2013, Wrighton et al. 2012). The completeness of the genomes within bins were then estimated by counting universal single copy genes as defined by Raes et al. (2007). Several of these bins were then shown to contain multiple (2-5), closely-related genomes based on the number of single copy genes. Those bins were then further separate by plotting differential coverage between two libraries. Distinct clusters of scaffolds on the coverage plots were

manually delineated into new bins. Coverage was determine by recruiting reads to scaffolds by BLASTN (bitscore >75). Binning was also manual curated based on GC content, top blast hits, and mate-pairings.

Acknowledgements

This article is funded by grants from the Gordon and Betty Moore Foundation (GBMF2609) to GJD, European Research Council 'DARCLIFE' grant 247153 (granted to Kai-Uwe Hinrichs, University of Bremen, who is gratefully acknowledged for funding for Cassandre Lazar), and by a University of Michigan Rackham Pre-doctoral fellowship to Brett Baker. The sequencing was funded by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 to BJB. We thank Drs. Karen G. Lloyd and Cody Sheik for helpful suggestions and comments on an earlier draft of this manuscript.

Author Contributions

B.J.B. performed the genomic assembly and binning, taxonomic, and metabolic analyses. B.J.B., C.L., and A.T. designed the research. C.L. collected the sampled and extracted DNA. B.J.B., A.T., and G.J.D. wrote the paper.

5.6 Appendix C

CHAPTER IV Supplementary Information

Contents

6. **Supplementary Figure 5.7**

7. **Supplementary Figure 5.8**

Supplementary Text

Sample collection and processing. Six 1 meter plunger cores were collected from ~1.5 m water depth in three mid-estuary locations (two cores per site) of the White Oak River, North Carolina in October 2010 (site 1 at 34˚44.592N, 77˚07.435W; site 2 at 34˚44.482N, 77˚07.404W, and site 3 at 34˚44.141N, 77˚07298W). Cores were stored at 4˚C overnight and processed 24 hours after sampling. Each core was sectioned into 2 cm intervals. From each site one core was subsampled for geochemical analyses and the other were subsampled for DNA extractions. DNA was extracted using the UltraClean Mega Soil DNA Isolation Kit (MoBio, CA), using 6 g of sediment, and stored at -80˚C until use.

Genomic assembly. The shallow assembly was a combination of high-quality reads (474,179,948) from sites 2 (8-12 cm), and 3 (8-10 cm). The SMTZ assembly was generated from a combination of reads (698,574,240) from site 2 (30-32 cm) and 3 (24-28 cm). The deep assembly was generated from high-quality reads (378,027,948) of site 1 (52-54 cm). Since we were not able to co-assemble all samples from the SMTZ, one of the samples (site 1 26-30 cm) was assembled separately. The contigs from this sample were co-binned with the assembly of the other two samples (site 2 and 3, described above). This resulted in some closely related, but unique bins, for example the *Gammaproteobacteria* bins SMTZ1-46 and SMTZ-46.

Table 5.1. Summary of genome reconstruction completeness. For more detailed information about individual genomic bins see Table S1.

Phylum	# of genomes	Bins with >1 genome	Genomes >90%	Genomes >70%
Betaproteobacteria	3	0	2	3
Gammaproteobacteria	10	1	6	7
Deltaproteobacteria	5	1	4	5
BRC2	3	1	2	3
Chloroflexi	8	0	4	5
Nitrospira	4	0	2	4
Gemmatiomonadetes	5	0	3	5
Planctomycetes	11	0	9	11
Spirochetes	1	0	0	1
Myxococcales	2	0	2	2
Chlamydiae	1	0	1	1
KSB1	2	0	0	2
Bacteroidetes	3	2	3	3

OD1	5	1	5	5
WS3	2	0	2	2
Zixibacteria	5	0	3	5
TA06	9	1	6	6
WOR-1	1	1	1	1
KD3-62	1	0	0	0
BRC1	1	0	0	1
Total	82	8	55	72

Table 5.2. General characteristics of all bacterial genomic bins.

bin	# of scaffolds	# of genes	Total length	Taxonomy	longest scaffold	Completeness	%GC
SG8-41	147	3889	3747630	Betaproteobacteria	202275	106.25	62.884
SG8-39	233	3307	3137175	Betaproteobacteria	98955	100	68.078
SG8-40	260	2744	2433805	Betaproteobacteria	33740	75	60.292
SG8-19	437	4886	4668865	Chloroflexi	119936	84.375	51.572
SG8-51-3	204	2080	1759043	Chloroflexi;Dehalococcoidia	51164	118.75	50.385
SG8-35-2	292	4049	3649796	Deltaproteobacteria; Desulfobacterales	185839	106.25	46.079
SG8-35	206	2740	2591650	Deltaproteobacteria; Desulfobacterales	58461	78	48.839
SG8-13	125	3814	3800553	Deltaproteobacteria; Desulfobacterales	155941	100	56.914
SG8-31	266	6175	5968366	Gammaproteobacteria	339997	190.625	62.583
SG8-11	464	4390	3882126	Gammaproteobacteria	12335	56.25	46.909
SG8-30	113	2780	2820123	Gammaproteobacteria	111313	96.875	68.743
SG8-15	302	3171	2741304	Gammaproteobacteria	36497	68.75	47.905
SG8-47	197	2931	2696743	Gammaproteobacteria	55750	100	60.614
SG8-45	98	2517	2283082	Gammaproteobacteria;Acidithiobacillales	97626	100	55.949
SG8-50	366	4182	3645285	Gammaproteobacteria;Thiothrichales	50482	90.625	57.244
SG8-28	430	3820	3856507	Gemmatiomonas	46899	81.25	67.178
SG8-17	456	3949	3691854	Gemmatiomonas	61136	71.875	59.953
SG8-23	310	3180	3209539	Gemmatiomonas	55555	103.125	68.084
SG8-38-2	307	3039	2814545	Gemmatiomonas	143644	90.625	61.536
SG8-38-1	234	3900	3848836	Myxococcales	133715	106.25	63.032
SG8-38	105	3453	3623660	Myxococcales	147758	96.875	63.406
SG8-3	733	6668	5957751	Deltaproteobacteria	34170	106.25	49.975
SG8-35-1	357	3808	3333117	Nitrospira/ like Thermodesulfovibrio	79085	87.5	45.817
SG8-35-4	292	3028	2692108	Nitrospira/ like Thermodesulfovibrio	33354	75	46.426
SG8-24	67	1231	1116563	OD1	62749	96.875	61.437
SG8-4	479	4333	4458874	Plantomycetes/Phycisphaerae	39007	96.875	54.804
SMTZ							
61	286	4504	4241461	Deltaproteobacteria	62586	103.125	56.081

62	403	5339	5784682	Bacteroides	142533	106.25	45.753
62-1	315	4729	5049351	Bacteroides	73049	131.25	40.025
62	307	2552	2576725	Bacteroides	142533	125	45.236
51	306	2831	2764529	BRC1	55748	71.875	59.783
31	300	3403	2821327	KSB1	34851	84.375	40.109
63	171	4302	4281581	Chloroflexi;Anaerolineae	197251	112.5	50.906
84	362	3995	3734778	Chloroflexi;Anaerolineae	50007	103.125	59.288
28-2	107	1566	1469117	Chloroflexi;unclassified	53467	96.875	63.75
28-1	89	845	732023	Chloroflexi;unclassified	22368	59.375	63.645
39	67	1253	1126604	Chlymidae	49818	100	26.251
46	263	3271	2826034	Gammaproteobacteria;Acidithiobacillales	54660	78.125	61.379
46	182	2960	2616358	Gammaproteobacteria;Acidithiobacillales	54660	81.25	61.645
52	263	4251	4296877	Gemmatiomonas	93174	96.875	65.496
35	140	1633	1416608	Nitrospira/ like Thermodesulfovibrio	46553	90.625	46.747
25	418	3775	3770408	Planctomyce	59260	103.125	66.54
32	237	2252	2326512	Planctomyce	27318	93.75	69.176
65	212	1773	1825793	Planctomyce	34095	71.875	62.186
30	296	3613	4021226	Plantomycetes/Phycisphaerae	50768	90.625	50.742
33	322	3284	3426709	Plantomycetes/Phycisphaerae	41894	106.25	67.777
32-1	317	2856	3094560	Plantomycetes/Phycisphaerae	44002	96.875	68.276
79	241	2878	2952060	Plantomycetes/Phycisphaerae	89332	90.625	50.321
40	189	2890	2883088	TA06.1	79299	100	60.277
40	197	2541	2559446	TA06.1	79299	43.75	60.303
60	264	3104	3369763	TA06.2	63010	109.375	47.409
42	59	2515	2532789	TA06.2	339863	100	46.22
77	150	1502	1269144	TA06.2	207686	118.75	45.774
72	137	2999	3297702	BRC2	100134	181.25	44.999
29	173	2251	2222379	BRC2	59882	118.75	41.693
81	99	2856	3074913	WS3	205519	96.875	52.817
57	285	2369	2350405	KSB1	81445	81.25	46.344
73-2	124	2356	2272909	Zixibacteria	70866	87.5	42.329
73	147	2480	2223615	Zixibacteria	46423	87.5	43.972
73-3	90	1939	1882756	Zixibacteria	85050	100	45.156
73	163	1657	1441479	Zixibacteria	46423	93.75	43.017

Deep

DG-18	129	1101	936096	Chloroflexi;Dehalococcoidia	23208	68.75	53.881
DG-22	147	1356	1135001	Chloroflexi;unclassified	21435	71.875	63.458
DG-28	199	1722	1418734	BRC2	13759	75	39.032

DG-60	128	1665	1357456	Deltaproteobacteria;Syntrophobacter	44074	121.875	39.05
DG-40	88	2555	2338617	Gammaproteobaciteria/Coxiella	273781	118.75	39.872
DG-56	144	1261	1238975	KD3-62	32772	62.5	64.886
DG-74-2	79	1584	1286363	OD1	198637	156.25	33.126
DG-72	11	874	717349	OD1	207750	96.875	34.598
DG-74-1	13	655	584570	OD1	146628	96.875	38.449
DG-74-3	21	676	543780	OD1	114447	96.875	36.836
DG-75	53	581	473645	OP11	32291	75	40.782
DG-20	331	2799	2821569	Plantomycetes	32318	118.75	67.441
DG-23	173	2069	2020164	Plantomycetes	41302	109.375	52.375
DG-58	220	1720	1746620	Plantomycetes	32019	84.375	62.069
DG-61	431	4757	4307950	Spirochete	56909	84.375	46.724
DG-26	135	1727	1682905	TA06	48904	134.375	52.046
DG-24	121	2291	2323720	TA06.1	131201	81.25	60.601
DG-78-2	95	2768	2562206	TA06.2	105856	78.25	42.199
DG-78	173	2196	2090507	TA06.2	50145	78.125	40.113
DG-54-3	279	3638	3303329	WOR-1	114009	134.375	44.77
DG-63	83	3190	3519707	WS3	259192	112.5	54.986
DG-27	133	1142	1101969	Zixibacteria	25637	93.75	53.616

Table 5.3. Number of specific organic carbon degrading genes identified in the bacterial genomic bins.

Bin	Taxonomy	Chitinase	NAG	Cellulase	Endoglucanase	Exo-beta-1,3-glucanase	Arabinofuranosidase	α-D-glucuronidase	Rhamnosidase	Endo-mannanase	α-D-xylosidase	Xylosidase	β-1,4 mannosidase	β-galactosidase	α-amylase	Gluco amylase	Pullulanase	Xylanase	Arabinogalactan endo-1,4-beta-galactosidase	Beta-hexosaminidase	Hyaluronidase 1	peptidases
SG8-41	Betaproteobacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	6	0	0	0	36
SG8-39	Betaproteobacteria	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	27
SG8-40	Betaproteobacteria	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	26
SG8-19	Chloroflexi	0	0	4	2	2	0	0	0	0	0	0	0	4	0	0	3	1	0	0	0	59
SG8-51-3	Chloroflexi;Dehalococcoidia	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	10
SG8-35-2	Deltaproteobacteria; Desulfobacterales	0	0	0	0	1	0	0	0	0	0	1	0	0	0	2	1	1	0	0	0	35
SG8-35	Deltaproteobacteria; Desulfobacterales	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3	2	1	0	0	0	40
SG8-13	Deltaproteobacteria; Desulfobacterales	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	36
SG8-31	Gammaproteobacteria	1	0	0	0	3	0	0	0	1	0	0	0	0	0	2	3	2	0	0	0	91
SG8-11	Gammaproteobacteria	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	3	2	0	2	0	41
SG8-30	Gammaproteobacteria	0	0	0	0	2	0	2	0	0	0	0	0	2	0	0	2	1	0	0	0	37
SG8-15	Gammaproteobacteria	1	0	0	0	0	0	0	0	0	0	0	1	1	2	0	1	0	0	0	0	27
SG8-47	Gammaproteobacteria	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	31
SG8-45	Gammaproteobacteria;Acidithiobacillales	0	0	2	0	0	0	0	0	0	0	0	1	0	1	0	0	4	0	0	0	33
SG8-50	Gammaproteobacteria;Thiothrichales	0	0	3	0	0	0	0	0	0	0	0	0	2	0	0	1	14	0	0	0	33
SG8-28	Gemmatiomonas	0	1	3	2	0	1	6	1	0	0	0	3	6	0	1	3	0	0	2	0	123
SG8-17	Gemmatiomonas	0	0	2	0	0	0	0	0	0	0	0	1	0	0	1	2	6	0	2	0	121
SG8-23	Gemmatiomonas	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	1	0	94
SG8-38-2	Gemmatiomonas	0	0	1	1	0	0	2	1	1	0	0	0	3	0	0	0	2	0	0	0	80
SG8-38-1	Myxococcales	0	0	2	0	2	0	2	0	0	1	0	2	4	1	0	0	4	0	0	0	30
SG8-38	Myxococcales	0	0	0	1	0	0	1	0	0	1	0	2	2	1	1	0	2	0	1	0	39
SG8-3	Deltaproteobacteria	0	0	0	0	0	0	0	0	0	0	0	1	0	1	2	1	5	0	0	0	58
SG8-35-1	Nitrospira/ like Thermodesulfovibrio	0	0	0	0	0	0	0	0	2	0	1	1	0	1	1	2	4	0	0	0	30
SG8-35-4	Nitrospira/ like Thermodesulfovibrio	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	4	0	0	0	22
SG8-24	OD1	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	19
SG8-4	Plantomycetes/Phycisphaerae	0	0	2	1	0	1	11	1	2	1	1	4	16	1	1	0	3	2	0	0	43

Bin	Taxonomy	Chitinase	NAG	Cellulase	Endoglucanase	Exo-beta-1,3-glucanase	Arabinofuranosidase	α-D-glucuronidase	Rhamnosidase	Endo-mannanase	α-D-xylosidase	Xylosidase	β-1,4 mannosidase	β-galactosidase	α-amylase	Gluco amylase	Pullulanase	Xylanase	Arabinogalactan endo-1,4-beta-galactosidase	Beta-hexosaminidase	Hyaluronidase 1	peptidases
SM23-62	Bacteroides	0	2	1	4	0	4	23	14	5	0	7	12	35	2	1	0	10	1	12	0	77
SM23-62-1	Bacteroides	0	0	3	5	0	4	9	5	3	0	6	2	13	1	0	0	10	1	5	0	69
SM23-51	BRC1	0	0	0	0	0	1	2	0	0	0	0	3	2	3	0	0	3	0	2	0	37
SM23-31	KSB1	1	0	0	0	0	1	2	0	0	0	0	0	3	0	0	0	1	0	2	0	86
SM23-63	Chloroflexi;Anaerolineae	0	0	2	0	0	0	0	0	0	0	0	2	4	1	1	0	3	1	0	0	62
SM23-84	Chloroflexi;Anaerolineae	0	0	5	1	0	0	0	0	0	0	0	1	3	0	0	0	3	0	0	0	50
SM23-28-2	Chloroflexi;unclassified	0	0	0	0	0	0	0	0	0	0	0	2	1	1	1	0	1	0	0	0	21
SM23-28-1	Chloroflexi;unclassified	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	9
SM23-39	Chlymidae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	13
SM23-61	Deltaproteobacteria	0	0	1	1	0	0	0	0	0	0	0	3	1	3	0	1	4	0	0	0	51
SM23-46	Gammaproteobacteria;Acidithiobacillales	0	0	1	0	0	0	1	1	0	0	0	1	0	1	0	1	9	0	0	0	26
SM23-52	Gemmatiomonas	0	0	4	0	0	1	6	0	0	0	0	2	7	0	0	1	1	1	1	0	127
SM23-35	Nitrospira/ like Thermodesulfovibrio	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	16
SM23-25	Planctomyce	0	0	5	0	0	3	4	6	1	0	3	0	7	0	2	0	3	1	2	0	25
SM23-32	Planctomyce	0	0	3	4	0	5	11	3	1	0	1	3	17	1	0	0	7	0	1	0	15
SM23-65	Planctomyce	0	0	3	1	0	0	3	1	0	0	0	2	11	2	0	0	3	0	1	0	11
SM23-30	Plantomycetes/Phycisphaerae	0	0	0	0	0	1	3	3	0	0	1	3	3	3	0	0	2	0	1	0	102
SM23-33	Plantomycetes/Phycisphaerae	0	0	3	1	0	1	4	2	0	0	0	3	17	1	0	0	1	0	4	0	43
SM23-32-1	Plantomycetes/Phycisphaerae	0	0	5	1	0	1	8	1	0	0	1	3	13	1	0	0	5	0	1	0	22
SM23-40	TA06.1	0	0	4	0	0	1	0	0	0	0	0	3	0	3	0	0	3	0	0	0	44
SM23-60	TA06.2	0	1	5	0	0	0	0	0	0	0	0	4	0	4	0	0	1	0	0	0	62
SM23-42	TA06.2	0	1	4	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	61
SM23-72	BRC2	0	0	1	0	0	0	3	0	0	0	0	1	5	1	0	0	3	0	0	0	23
SM23-29	BRC2	0	0	0	2	0	0	11	0	1	0	0	4	13	4	0	0	3	0	1	0	31
SM23-81	WS3	0	0	2	0	0	1	0	0	0	0	0	2	2	2	0	1	4	0	0	0	49
SM23-57	KSB1	0	0	1	0	0	0	3	2	0	0	0	2	3	1	0	0	1	0	1	0	33
SM23-73-2	Zixibacteria	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	44
SM23-73	Zixibacteria	0	0	0	0	0	0	0	0	0	0	0	3	1	3	0	0	5	0	0	0	24
SM23-73-3	Zixibacteria	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	2	0	1	0	46
SM1-62	Bacteroides	0	0	2	1	0	0	10	4	3	0	2	3	15	0	0	0	4	1	6	0	77
SM1-46	Gammaproteobacteria;Acidithiobacillales	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	10	0	0	0	26
SM1-79	Plantomycetes/Phycisphaerae	0	0	5	3	0	2	2	4	0	0	1	1	4	1	0	0	2	0	0	0	0
SM1-40	TA06.1	0	0	4	0	0	1	0	0	0	0	0	3	0	3	0	0	3	0	0	0	44
SM1-77	TA06.2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SM1-73	Zixibacteria	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	1	0	0	0	24

Bin	Taxonomy	Chitinase	NAG	Cellulase	Endoglucanase		Arabinofuranosidase	α-D-glucuronidase	Rhamnosidase	Endo-mannanase	α-D-xylosidase	Xylosidase	β-1,4 mannosidase	β-galactosidase	α-amylase	Gluco amylase	Pullulanase	Xylanase	Arabinogalactan endo-1,4-beta-galactosidase	Beta-hexosaminidase	Hyaluronidase 1	peptidases
DG-18	Chloroflexi;Dehalococcoidia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	10
DG-22	Chloroflexi;unclassified	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	16
DG-28	Deferribacteres or SAR324?	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	6
DG-60	Deltaproteobacteria;Syntrophobacter	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	3	0	0	0	17
DG-40	Gammaproteobaciteria/Coxiella	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	7	0	0	0	22
DG-56	KD3-62	0	0	2	0	0	1	1	1	0	0	0	3	7	1	0	0	2	0	0	0	14
DG-74-2	OD1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	3	0	0	0	18
DG-72	OD1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	17
DG-74-1	OD1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	11
DG-74-3	OD1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	8
DG-20	Plantomycetes	0	0	3	3	0	2	3	4	0	0	0	0	6	0	2	0	5	1	1	0	17
DG-23	Plantomycetes	0	0	2	0	0	1	0	1	0	0	0	1	3	1	0	0	1	0	1	0	20
DG-58	Plantomycetes	0	0	1	1	0	0	1	3	1	0	0	1	8	1	0	0	3	0	0	0	12
DG-61	Spirochete	0	0	2	0	0	0	1	0	0	0	0	2	2	2	2	1	2	0	0	0	45
DG-26	TA06	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	35
DG-24	TA06.1	0	0	4	0	0	0	0	0	0	0	0	1	0	1	0	0	4	0	0	0	36
DG-78-2	TA06.2	0	1	4	0	0	0	0	0	0	0	0	2	0	2	0	0	1	0	0	0	42
DG-78	TA06.2	0	1	3	0	0	0	0	0	0	0	0	2	0	2	0	0	3	0	0	0	47
DG-54-3	WOR-1	0	0	2	0	0	0	2	0	2	0	0	2	4	2	0	1	4	0	1	0	61
DG-63	WS3	0	0	1	0	0	0	0	0	2	0	0	3	0	3	0	0	6	0	1	0	44
DG-27	Zixibacteria	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	0	0	24

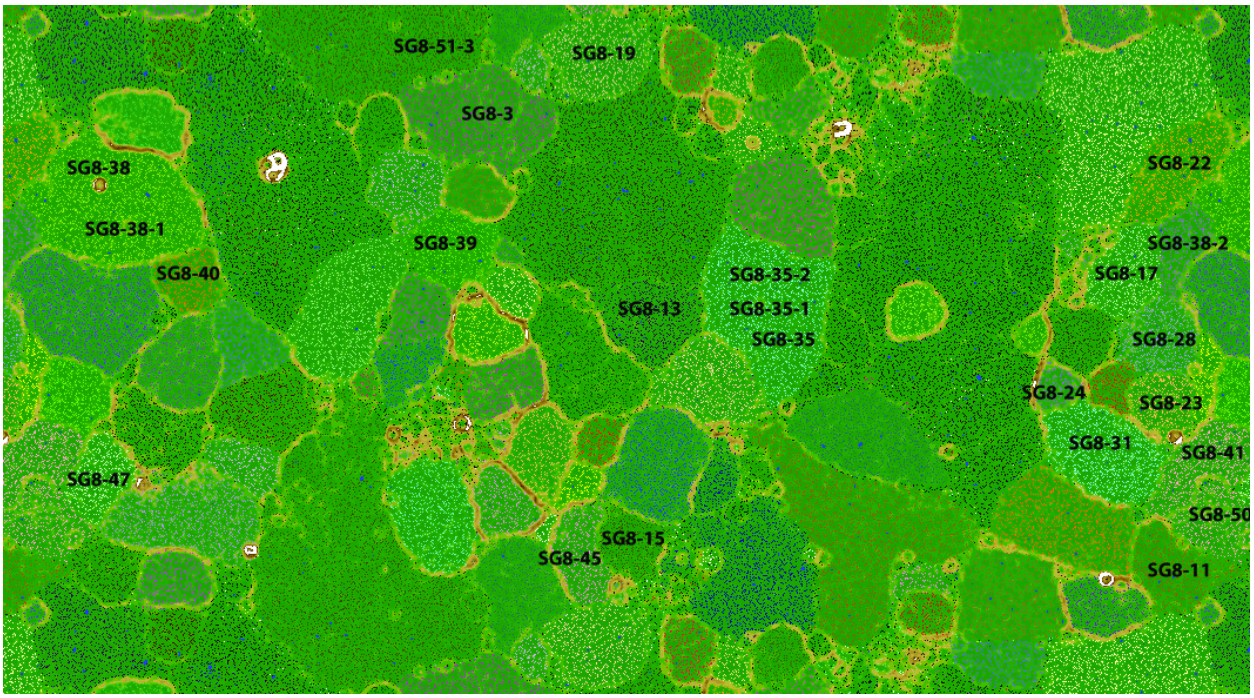

Figure 5.5. Tetra-nucleotide ESOM binning map of shallow assembly. The delineated bins are colored and labeled. Noticed that in a few cases (SG8-35, 35-1, 35-2, and SG8-38 and 38-1) multiple closely-related bins falls within one large cluster on this map. These bins were found to contain more than one genome and were therefore, further delineated based on differential coverage plots.

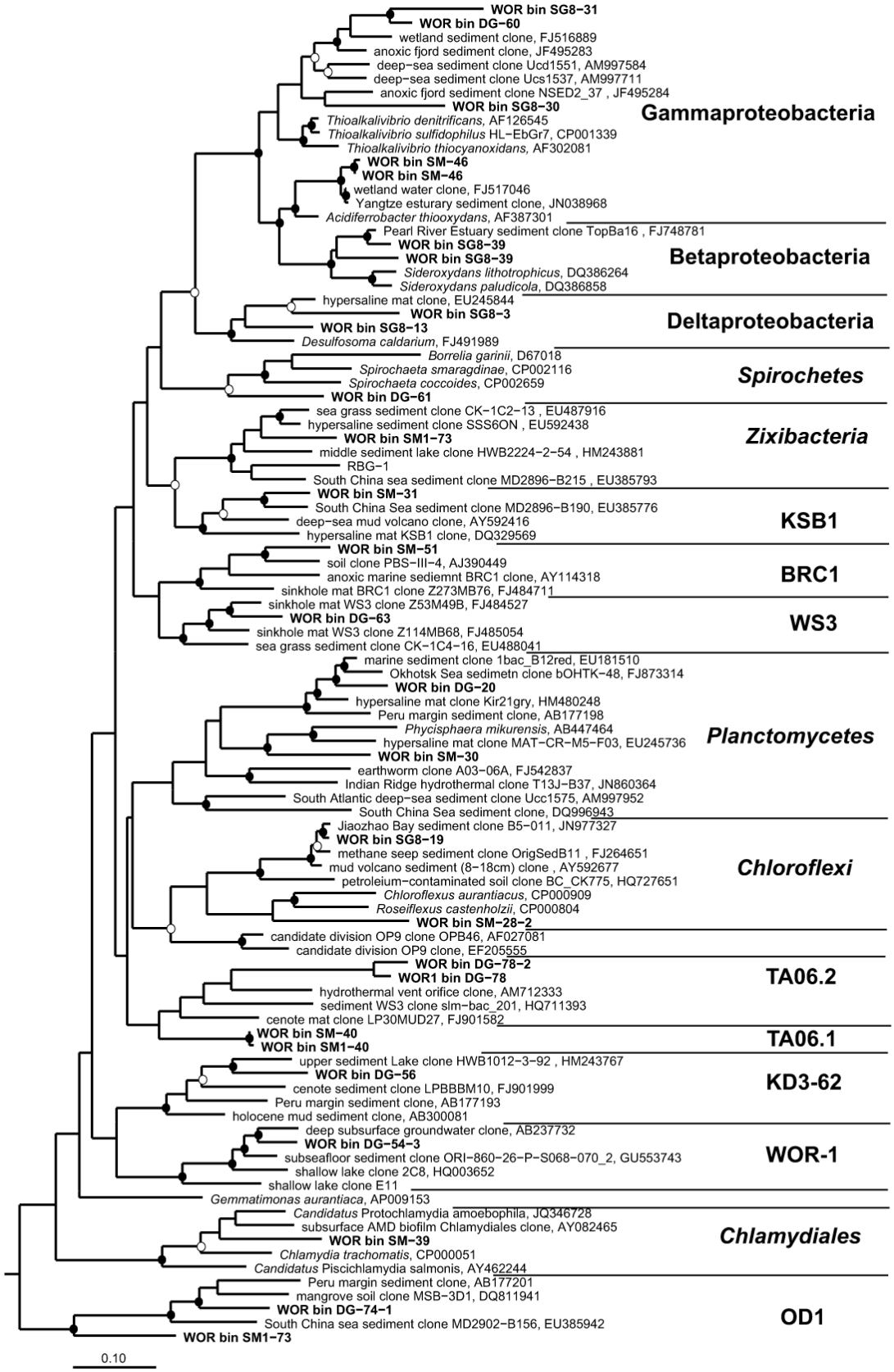
WOR bin SG8−31
WOR bin DG−60
wetland sediment clone, FJ516889
anoxic fjord sediment clone, JF495283
deep−sea sediment clone Ucd1551, AM997584
deep−sea sediment clone Ucs1537, AM997711
anoxic fjord sediment clone NSED2_37 , JF495284
WOR bin SG8−30
Gammaproteobacteria
Thioalkalivibrio denitrificans, AF126545
Thioalkalivibrio sulfidophilus HL−EbGr7, CP001339
Thioalkalivibrio thiocyanoxidans, AF302081
WOR bin SM−46
WOR bin SM−46
wetland water clone, FJ517046
Yangtze esturary sediment clone, JN038968
Acidiferrobacter thiooxydans, AF387301
Pearl River Estuary sediment clone TopBa16 , FJ748781
WOR bin SG8−39
WOR bin SG8−39
Betaproteobacteria
Sideroxydans lithotrophicus, DQ386264
Sideroxydans paludicola, DQ386858
hypersaline mat clone, EU245844
WOR bin SG8−3
Deltaproteobacteria
WOR bin SG8−13
Desulfosoma caldarium, FJ491989
Borrelia garinii, D67018
Spirochaeta smaragdinae, CP002116
Spirochetes
Spirochaeta coccoides, CP002659
WOR bin DG−61
sea grass sediment clone CK−1C2−13 , EU487916
hypersaline sediment clone SSS6ON , EU592438
Zixibacteria
WOR bin SM1−73
middle sediment lake clone HWB2224−2−54 , HM243881
RBG−1
South China sea sediment clone MD2896−B215 , EU385793
WOR bin SM−31
South China Sea sediment clone MD2896−B190, EU385776
deep−sea mud volcano clone, AY592416
KSB1
hypersaline mat KSB1 clone, DQ329569
WOR bin SM−51
soil clone PBS−III−4, AJ390449
anoxic marine sediemnt BRC1 clone, AY114318
BRC1
sinkhole mat BRC1 clone Z273MB76, FJ484711
sinkhole mat WS3 clone Z53M49B, FJ484527
WOR bin DG−63
WS3
sinkhole mat WS3 clone Z114MB68, FJ485054
sea grass sediment clone CK−1C4−16, EU488041
marine sediment clone 1bac_B12red, EU181510
Okhotsk Sea sedimetn clone bOHTK−48, FJ873314
WOR bin DG−20
hypersaline mat clone Kir21gry, HM480248
Peru margin sediment clone, AB177198
Phycisphaera mikurensis, AB447464
Planctomycetes
hypersaline mat clone MAT−CR−M5−F03, EU245736
WOR bin SM−30
earthworm clone A03−06A, FJ542837
Indian Ridge hydrothermal clone T13J−B37, JN860364
South Atlantic deep−sea sediment clone Ucc1575, AM997952
South China Sea sediment clone, DQ996943
Jiaozhao Bay sediment clone B5−011, JN977327
WOR bin SG8−19
methane seep sediment clone OrigSedB11 , FJ264651
mud volcano sediment (8−18cm) clone , AY592677
Chloroflexi
petroleum−contaminated soil clone BC_CK775, HQ727651
Chloroflexus aurantiacus, CP000909
Roseiflexus castenholzii, CP000804
WOR bin SM−28−2
candidate division OP9 clone OPB46, AF027081
candidate division OP9 clone, EF205555
WOR bin DG−78−2
WOR1 bin DG−78
TA06.2
hydrothermal vent orifice clone, AM712333
sediment WS3 clone slm−bac_201, HQ711393
cenote mat clone LP30MUD27, FJ901582
WOR bin SM−40
TA06.1
WOR bin SM1−40
upper sediment Lake clone HWB1012−3−92 , HM243767
WOR bin DG−56
KD3−62
cenote sediment clone LPBBBM10, FJ901999
Peru margin sediment clone, AB177193
holocene mud sediment clone, AB300081
deep subsurface groundwater clone, AB237732
WOR bin DG−54−3
WOR−1
subseafloor sediment clone ORI−860−26−P−S068−070_2, GU553743
shallow lake clone 2C8, HQ003652
shallow lake clone E11
Gemmatimonas aurantiaca, AP009153
Candidatus Protochlamydia amoebophila, JQ346728
subsurface AMD biofilm Chlamydiales clone, AY082465
Chlamydiales
WOR bin SM−39
Chlamydia trachomatis, CP000051
Candidatus Piscichlamydia salmonis, AY462244
Peru margin sediment clone, AB177201
mangrove soil clone MSB−3D1, DQ811941
WOR bin DG−74−1
OD1
South China sea sediment clone MD2902−B156, EU385942
WOR bin SM1−73
0.10

Figure 5.6. Phylogenetic tree (generated using maximum likelihood method in ARB software package) of 16S rRNA genes present in the bacterial genomic bins in this study. Closed circles represent maximum likelihood (ProML, ARB package) bootstraps >75% and open circles are >50% values.

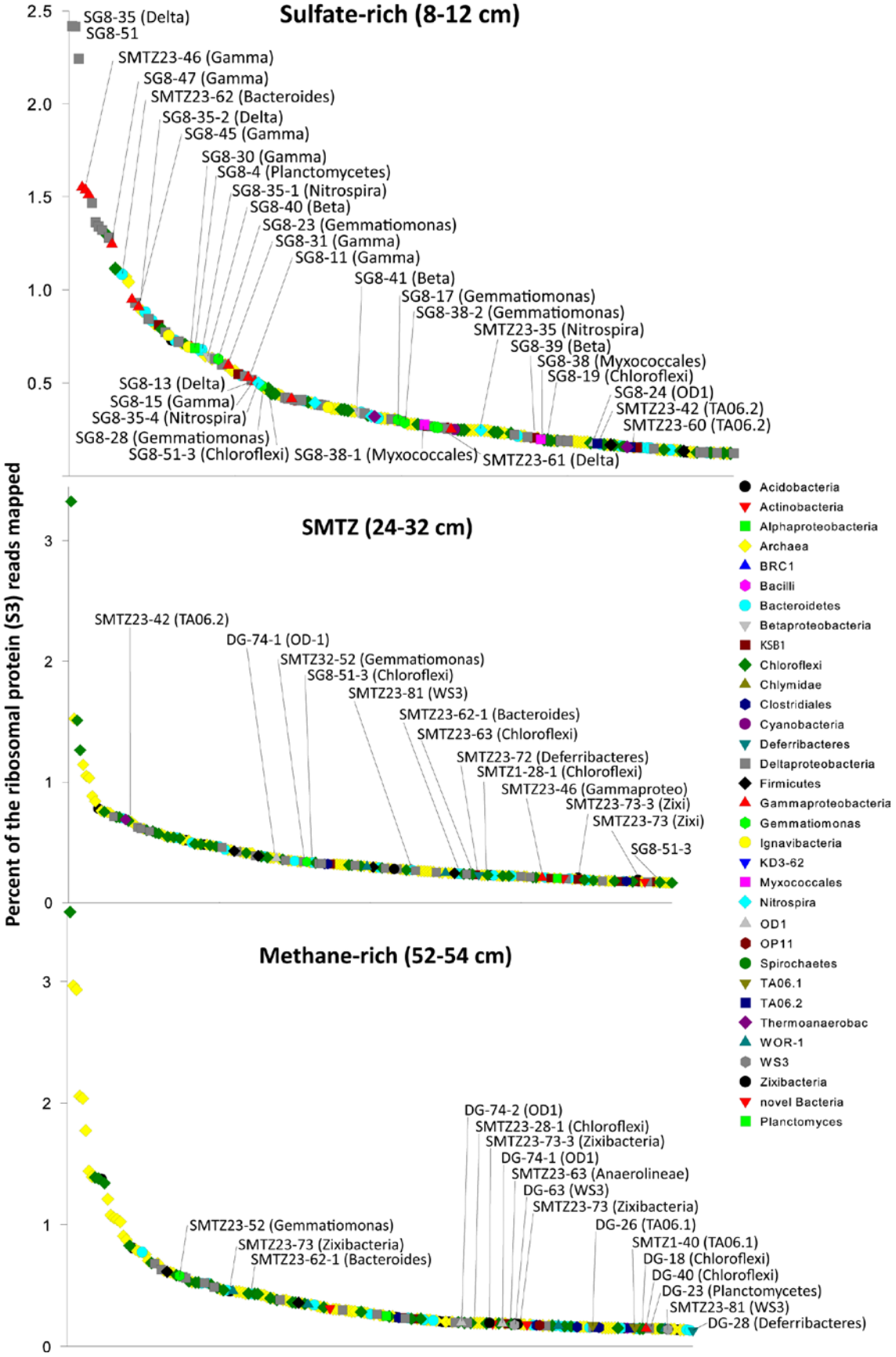
Sulfate-rich (8-12 cm)
SG8-35 (Delta)
SG8-51
SMTZ23-46 (Gamma)
SG8-47 (Gamma)
SMTZ23-62 (Bacteroides)
SG8-35-2 (Delta)
SG8-45 (Gamma)
SG8-30 (Gamma)
SG8-4 (Planctomycetes)
SG8-35-1 (Nitrospira)
SG8-40 (Beta)
SG8-23 (Gemmatiomonas)
SG8-31 (Gamma)
SG8-11 (Gamma)
SG8-41 (Beta)
SG8-17 (Gemmatiomonas)
SG8-38-2 (Gemmatiomonas)
SMTZ23-35 (Nitrospira)
SG8-39 (Beta)
SG8-38 (Myxococcales)
SG8-19 (Chloroflexi)
SG8-24 (OD1)
SMTZ23-42 (TA06.2)
SMTZ23-60 (TA06.2)
SG8-13 (Delta)
SG8-15 (Gamma)
SG8-35-4 (Nitrospira)
SG8-28 (Gemmatiomonas)
SG8-51-3 (Chloroflexi)
SG8-38-1 (Myxococcales)
SMTZ23-61 (Delta)
SMTZ (24-32 cm)
SMTZ23-42 (TA06.2)
DG-74-1 (OD-1)
SMTZ32-52 (Gemmatiomonas)
SG8-51-3 (Chloroflexi)
SMTZ23-81 (WS3)
SMTZ23-62-1 (Bacteroides)
SMTZ23-63 (Chloroflexi)
SMTZ23-72 (Deferribacteres)
SMTZ1-28-1 (Chloroflexi)
SMTZ23-46 (Gammaproteo)
SMTZ23-73-3 (Zixi)
SMTZ23-73 (Zixi)
SG8-51-3
Methane-rich (52-54 cm)
DG-74-2 (OD1)
SMTZ23-28-1 (Chloroflexi)
SMTZ23-73-3 (Zixibacteria)
DG-74-1 (OD1)
SMTZ23-63 (Anaerolineae)
DG-63 (WS3)
SMTZ23-73 (Zixibacteria)
DG-26 (TA06.1)
SMTZ1-40 (TA06.1)
DG-18 (Chloroflexi)
DG-40 (Chloroflexi)
DG-23 (Planctomycetes)
SMTZ23-81 (WS3)
DG-28 (Deferribacteres)
SMTZ23-52 (Gemmatiomonas)
SMTZ23-73 (Zixibacteria)
SMTZ23-62-1 (Bacteroides)
Percent of the ribosomal protein (S3) reads mapped
0
0.5
1.0
1.5
2.0
2.5
1
2
3
Acidobacteria
Actinobacteria
Alphaproteobacteria
Archaea
BRC1
Bacilli
Bacteroidetes
Betaproteobacteria
KSB1
Chloroflexi
Chlymidae
Clostridiales
Cyanobacteria
Deferribacteres
Deltaproteobacteria
Firmicutes
Gammaproteobacteria
Gemmatiomonas
Ignavibacteria
KD3-62
Myxococcales
Nitrospira
OD1
OP11
Spirochaetes
TA06.1
TA06.2
Thermoanaerobac
WOR-1
WS3
Zixibacteria
novel Bacteria
Planctomyces

Figure 5.7. Fig. Abundances of top genotypes in the SMTZ (24-32 cm) and methane-rich (52-54 cm) layers of the sediment profiles based on the number of reads that map to all the genes for ribosomal protein S3. Those that are represented in the genomic bins are labeled.

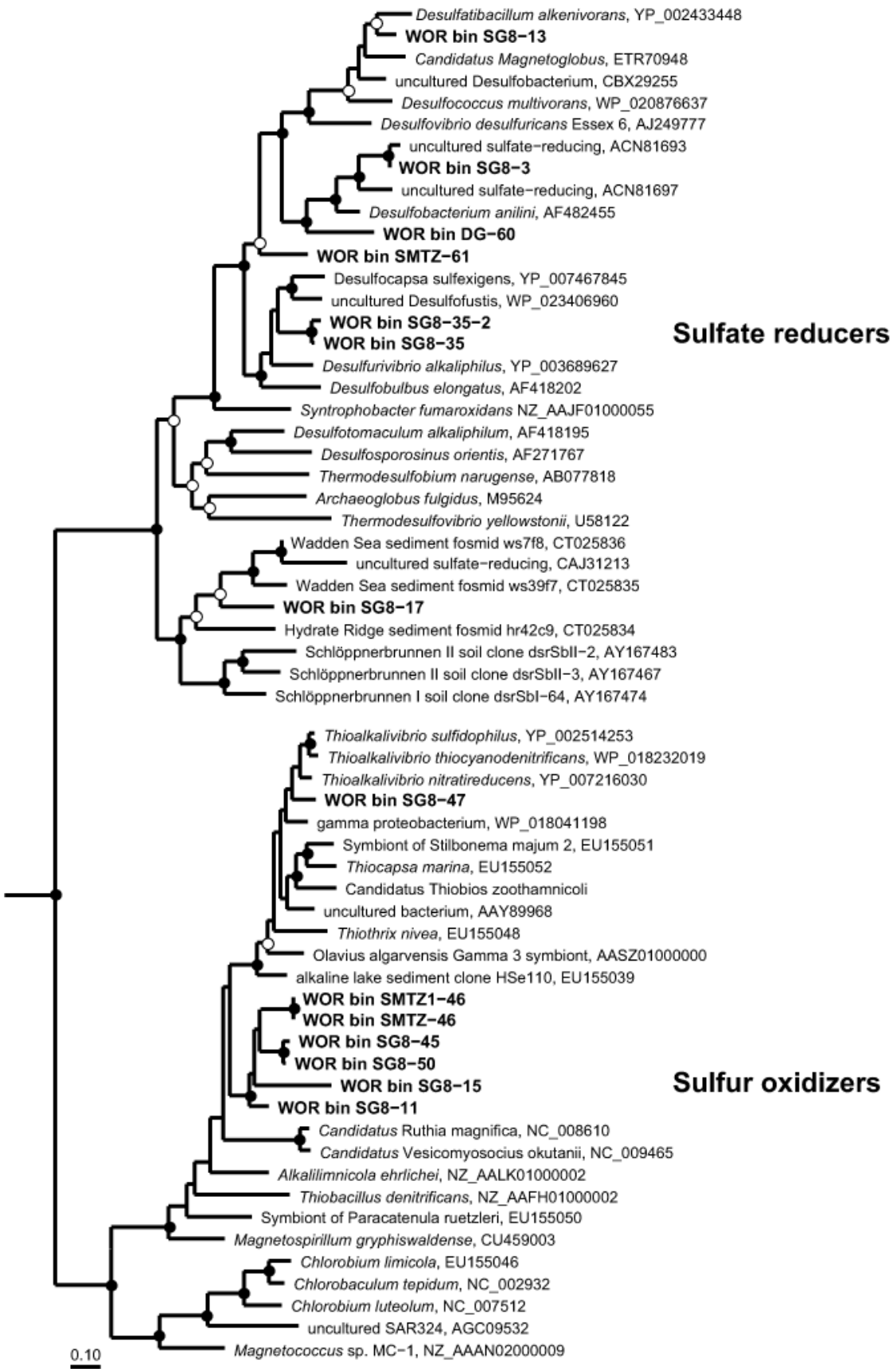
Desulfatibacillum alkenivorans, YP_002433448
WOR bin SG8−13
Candidatus Magnetoglobus, ETR70948
uncultured Desulfobacterium, CBX29255
Desulfococcus multivorans, WP_020876637
Desulfovibrio desulfuricans Essex 6, AJ249777
uncultured sulfate−reducing, ACN81693
WOR bin SG8−3
uncultured sulfate−reducing, ACN81697
Desulfobacterium anilini, AF482455
WOR bin DG−60
WOR bin SMTZ−61
Desulfocapsa sulfexigens, YP_007467845
uncultured Desulfofustis, WP_023406960
WOR bin SG8−35−2
WOR bin SG8−35
Desulfurivibrio alkaliphilus, YP_003689627
Desulfobulbus elongatus, AF418202
Syntrophobacter fumaroxidans NZ_AAJF01000055
Desulfotomaculum alkaliphilum, AF418195
Desulfosporosinus orientis, AF271767
Thermodesulfobium narugense, AB077818
Archaeoglobus fulgidus, M95624
Thermodesulfovibrio yellowstonii, U58122
Wadden Sea sediment fosmid ws7f8, CT025836
uncultured sulfate−reducing, CAJ31213
Wadden Sea sediment fosmid ws39f7, CT025835
WOR bin SG8−17
Hydrate Ridge sediment fosmid hr42c9, CT025834
Schlöppnerbrunnen II soil clone dsrSbII−2, AY167483
Schlöppnerbrunnen II soil clone dsrSbII−3, AY167467
Schlöppnerbrunnen I soil clone dsrSbI−64, AY167474
Sulfate reducers
Thioalkalivibrio sulfidophilus, YP_002514253
Thioalkalivibrio thiocyanodenitrificans, WP_018232019
Thioalkalivibrio nitratireducens, YP_007216030
WOR bin SG8−47
gamma proteobacterium, WP_018041198
Symbiont of Stilbonema majum 2, EU155051
Thiocapsa marina, EU155052
Candidatus Thiobios zoothamnicoli
uncultured bacterium, AAY89968
Thiothrix nivea, EU155048
Olavius algarvensis Gamma 3 symbiont, AASZ01000000
alkaline lake sediment clone HSe110, EU155039
WOR bin SMTZ1−46
WOR bin SMTZ−46
WOR bin SG8−45
WOR bin SG8−50
WOR bin SG8−15
WOR bin SG8−11
Candidatus Ruthia magnifica, NC_008610
Candidatus Vesicomyosocius okutanii, NC_009465
Alkalilimnicola ehrlichei, NZ_AALK01000002
Thiobacillus denitrificans, NZ_AAFH01000002
Symbiont of Paracatenula ruetzleri, EU155050
Magnetospirillum gryphiswaldense, CU459003
Chlorobium limicola, EU155046
Chlorobaculum tepidum, NC_002932
Chlorobium luteolum, NC_007512
uncultured SAR324, AGC09532
Magnetococcus sp. MC−1, NZ_AAAN02000009
Sulfur oxidizers
0.10

Figure 5.8. Phylogenetic tree of concatenated *dsrAB* genes within bacteria genomic bins from this study. Closed and open circles represent maximum likelihood (ProML, ARB package) bootstrap value >75 and >50, respectively.

5.7 References

Arnosti C (2011) Microbial extracellular enzymes and the marine carbon cycle. *Ann Rev Marine Sci* 3, 401-425.

Baker BJ, and Dick GJ (2013) Omic approaches in microbial ecology, charting the unknown. *Microbe*, 8, 353-359.

Bauer JE, Cai W-J, Raymond PA, Bianchi TS, Hopkinson CS, Regnier PAG (2013) The changing carbon cycle of the coastal ocean. *Nature* 504, 61-70.

Beazley MJ, Martinez RJ, Rajan S, Powell J, Piceno YM, et al. (2012) Microbial community analysis of a coastal salt marsh affected by the Deepwater Horizon oil spill. PLoS ONE 7(7), e41305. doi, 10.1371/journal.pone.0041305

Berlemont R, Martiny AC (2013) Phylogenetic distribution of potential cellulases in bacteria. *Appl Environ Microbiol* 79, 1545-54.

Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, and House CH (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc Natl Acad Sci USA* 105, 10583-10588.

Bowen J, Morrison HG, Hobbie JE, Sogin ML (2012) Salt marsh sediment diversity, a test of the variability of the rare biosphere among environmental replicates. *ISME J.* 6, 2014-2013.

Carlson HK, Iavarone AT, Gorur A, Yeo BS, Tran R, et al. (2012) Surface multiheme c-type cytochromes from *Thermincola potens* and implications for respiratory metal reduction by Gram-positive bacteria. *Proc Natl Acad Sci* 109, 1702-1707.

Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, et al. (2013) Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature Comm* 4, 2120.

Cai, W-J (2011) Estuarine and coastal ocean carbon paradox, CO_2 sinks or sites of terrestrial carbon incineration? *Annu Rev Mar Sci* 3, 123–45.

Coursolle D, Gralnick JA (2012) Reconstruction of extracellular respiratory pathways for iron(III) reduction in *Shewanella oneidensis* strain MR-1. *Front. Microbiol.* 3, 56.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, et al. (2009) Community-wide analysis of genome sequence signatures. *Gen. Biol.* 10, R85.

Einsle O (2011) Structure and function of formate-dependent cytochrome c nitrite reductase, NrfA. *Methods Enzymol* 496, 399-422.

Grimm F, Doblert N, Dahl C. (2010) Regulation of dsr genes encoding proteins responsible for the oxidation of stored sulfur in *Allochromatium vinosum*. *Microbiol* 156, 764-773.

Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, et al. (2013) Community genomic analyses constrain the distribution of metabolic traits across the *Chloroflexi* phylum and indicate roles in sediment carbon cycling. *Microbiome* 1, 22.

Jørgensen, BB (1990) A thiosulfate shunt in the sulfur cycle of marine sediments. *Science* 249, 152-154.

Jiang L, Zheng Y, Peng X, Zhou H, Zhang C, et al (2009) Vertical distribution and diversity of sulfate-reducing prokaryotes in the Pearl River estuarine sediments, South China. *Microbiol Ecol* 70, 249-262.

Jiao Y, Newman DK (2007) The pio operon is essential for phototrophic Fe(II) oxidation in *Rhodopseudomonas palustris*. *J Bacteriol*. 189, 1765-1773.

Kubo K, Lloyd KG, Biddle JF, Teske A, Amann R, Knittel K (2012) Archaea of the Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine sediments. *ISME J* 6, 1949-1965.

Lazar C, et al. Environmental controls on intragroup diversity of the uncultured benthic archaeal of the Miscellaneous Crenarchaeotal Group lineage naturally enriched in anoxic sediments of the White Oak River Estuary (North Carolina, USA). *Environ Microbiol* In review.

Lenk S, Arnds J, Zerjatke K, Musat N, Amann R, Mussmann M (2010) Novel groups of *Gammaproteobacteria* catalyse sulfur oxidation and carbon fixation in a coastal intertidal sediment. *Environ Microbiol* 13, 758-774.

Liu J, Wang Z, Belchik SM, Edwards MJ, Liu C, et al. (2012) Identification and characterization of MtoA, a decaheme c-type cytochrome of the neutrophilic Fe(II)-oxidizing bacterium *Sideroxydans lithotrophicus* ES-1. *Front Microbiol* 3, 37.

Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, et al. (2013) Predominant archaea in marine sediments degrade detrital proteins. *Nature* 496, 215-218.

Loy A, Duller S, Baranyi C, Mussmann M, Ott J, et al. (2009) Reverse dissimilatory sulfite reductase as a phylogenetic marker for a subgroup of sulfur-oxidizing prokaryotes. *Environ Microbiol* 11, 289-299.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, et al. (2012) IMG, the integrated microbial genomes database and comparative analysis system. *Nuc Acids Res* 40, D115-122.

Mussmann, M, Richter M, Lombardot T, Meyerdierks A, Kuever J, et al. (2005) Clustered genes related to sulfate respiration in uncultured prokaryotes support the theory of their concomitant horizontal transfer. *J Bacteriol* 187, 7126-7137.

Myers C, Myers J. (2000) Role of the tetraheme cytochrome CymA in anaerobic electron transport in cells of *Shewanella putrefaciens* MR-1 with normal levels of menaquinone. *J Bacteriol* 182, 67-75.

Moeslund L, Thamdrup B, Jørgensen BB (1994) Sulfur and iron cycling in a coastal sediment, Radiotracer studies and seasonal dynamics. *Biogeochemistry* 27, 129-152.

Nohno T, Noji S, Taniguchi S, Saito T (1989) The narX and narL genes encoding the nitrate-sensing regulators of *Escherichia coli* are homologous to a family of prokaryotic two-component regulatory genes. *Nucl Acids Res* 17, 2947-2957.

Oremland RS, Polcin S (1982) Methanogenesis and sulfate reduction, competitive and noncompetitive substrates in estuarine sediments. *Appl Environ Microbiol* 44, 1270-1276.

Peng, Y, Leung HCM, Yiu SM, Chin FYL (2012) IDBA-UD, a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420-1428.

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol*. 8, R10.

Sieber JR, McInerney MJ, Gunsalus RP (2012) Genomic insights into syntrophy, the paradigm for anaerobic metabolic cooperation. *Annu Rev Microbiol* 66, 429-452.

Schmidt O, Drake HL, Horn MA (2010) Hitherto unknown [Fe-Fe]-hydrogenase gene diversity in anaerobes and anoxic enrichments from a moderately acidic fen. *Appl Environ Microbiol* 76, 2027-2031.

Sekiguchi Y, Muramatsu M, Imachi H, Narihiro T, Ohashi A, et al. (2008) *Thermodesulfovibrio aggregans* sp. nov. and *Thermodesulfovibrio thiophilus* sp. nov., anaerobic, thermophilic, sulfate-reducing bacteria isolated from thermophilic and methanogenic sludge, and emended description of the genus *Thermodesulfovibrio*. *Int J Syst Evol Microbiol* 58, 2541-2548.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 23, 111-120.

Sorokin DY, Tourova TP, Kolganova TV, Sjollema KA, Kuenen JG (2002) *Thioalkalispira microaerophila* gen. nov., sp. nov., a novel lithoautotrophic, sulfur-oxidizing bacterium from a soda lake. *Int J Syst Evol Microbiol* 52, 2175-2182.

Sorokin DY, Tourova TP, Antipov AN, Muyzer G, Kuenen JG (2004) Anaerobic growth of the haloalkaliphilic denitrifying sulfur-oxidizing bacterium *Thialkavibrio thiocyandentrificans* sp. nov. with thiocyanate. *Microbiol* 150, 2435-2442.

Wasmund D, Schreiber L, Lloyd KG, Petersen DG, Schramm A, et al. (2014) Genome sequencing of a single cell of the widely distributed marine subsurface *Dehalococcoidia*, phylum *Chloroflexi*. *ISME J* 8, 383-397.

Wakeham SG, Lee C, Hedges JI, Hernes PJ, Petersen MJ (1997) Molecular indicators of diagenetic status in marine organic matter. *Geochim Cosmochim Acta* 61, 5363-5369.

White GF, Shi Z, Wang Z, Dohnalkova AC, Marshall MJ, et al. (2013) Rapid electron exchange between surface-exposed bacterial cytochromes and Fe(III) minerals. *Proc Natl Acad Sci USA* 110, 6346-6351.

Wrighton KC, Castelle CJ, Wilkins MJ, Hug LA, Sharon I, et al. (2014) Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *ISME J* in press.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, et al. (2012) Fermentation, hydrogen, and sulfur metabolisms in multiple uncultivated bacterial phyla. *Science* 337, 1661–1665.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Introduction

In this Chapter, I summarize the main findings of the chapters and I discuss what future studies might be done. My dissertation applied omic approaches to questions regarding the microbial ecology of hydrothermal plumes and estuary sediments. Some of the key questions addressed here are: Which microbes inhabit hydrothermal plumes and estuary sediments? What pathways and organisms are involved in carbon and nitrogen cycling in Guaymas Basin deep sea hydrothermal plumes and surrounding waters? What are metabolisms of novel, uncultured bacteria in sediments? How are metabolisms partitioned among community members?

In **Chapter II** I reviewed common approaches used to analyze community genomic and transcriptomic data and discussed the limitations of relying on public databases rather than doing *de novo* assembly. **Chapter III** described the diversity, genomic content, and gene expression of ammonia-oxidizing archaea in the Guaymas Basin deep-sea and hydrothermal plumes. This was the first genomic reconstruction and assessment of transcriptional activity of entire populations of these organisms from the deep sea. Then, in **Chapter IV** I developed a novel approach of analyzing metatranscriptomic data (RNA sequence) – *de novo* assembly – and show that the next step in nitrification, nitrite oxidation, in the same Guaymas Basin is mediated by a novel member of *Nitrospirae*. In the process of doing transcriptomic assemblies of the entire Guaymas Basin communities I also looked at gene activities of uncultured microbes involved in the recycling of organic carbon, and show that these are widespread in the world oceans. **Chapter V** is the first study to reconstruct dozens of bacterial genomes, including those from novel lineages, from estuary sediments. These genomes provide insights about the metabolic capabilities of several uncultured bacteria. It also enabled me to produce the first highly resolved map of the flow of

carbon and energy through estuary sediment microbial communities, and provide links between microbes and the geochemistry of pore waters.

6.2 Microbes and metabolic pathways involved in carbon and nitrogen cycling in deep sea hydrothermal plumes

The reconstruction of multiple (~6-8) genomes belonging to ammonia oxidizing Thaumarchaea from the Guaymas Basin deep sea enabled us to investigate the full catalog of genes in the plume and background communities (Baker et al. 2012). This was the first time that a community transcriptomic dataset was analyzed using matching assembled genomes. This meant that we were able to match transcripts to genes from the same environment, which enabled us to assign function and taxonomy more accurately than previous studies that used public databases (Frias-Lopez et al. 2008, Hollibaugh et al. 2011, Stewart et al. 2011). As a result, I was able to look at the expression of different types of closely related genes involved in ammonia oxidation, ammonia monooxygenase (*amo*) genes. Surprisingly, I found that these deep-sea Thaumarchaea are closely related to species that are abundant in low-ammonia surface ocean waters. However, this cluster of organisms, referred to as "GB plume cluster", is adapted to the higher ammonia concentrations of Guaymas Basin waters, in contrast to *Nitrosopumilus maritmus*. Comparison of the Guaymas Basin genomes to *Nitrosopumilus maritmus* revealed large portions of the *Nitrosopumilus maritmus* genome were absent in the deep sea genomes. These same genomic regions were also absent in organisms from surface water in the Gulf of Maine (Tully et al. 2012), reflecting common patterns of genomic variation in Thaumarchaea from disparate environments. Also, the deep sea Thaumarchaea contain genomic regions distinct from those previous sequenced, including genes involved in the conversion of urea to ammonia. These urease genes were also highly represented in the transcript data from the Guaymas Basin, suggesting urea utilization as a source of nitrogen. Interestingly, we also found that nitrite reductase (*nirK*) genes were among the most active genes in the plume population. These nitrite reductases reduce nitrite to N_2O, a powerful greenhouse gas, and isotopic approaches recently implicated Thaumarchaea as a major source of N_2O to the atmosphere (Santoro et al. 2012).

Using publically available gene databases limits functional and taxonomic identification of transcriptomic sequencing reads from natural communities (Baker and Dick, 2013). On

average only around 15% of the mRNA reads are confidently assigned a function (Frias-Lopez et al. 2008, Lesniewski et al. 2012). Therefore, we were interested in determining if there were any geochemically important genes present in the unassigned transcripts. To do this we developed an approach of assembling these RNA reads (Baker et al. 2013). Taking this approach enabled us to reconstruct complete transcripts containing full and multiple genes (on operons) from nature for the first time. Having complete mRNAs rather than smaller portions (as are represented on individual reads) enabled better identification of function. Furthermore, it improved our ability to map transcripts back to the metagenomic data to determine which community member it originated from (Baker and Dick, 2013).

While analyzing this data we found that a significant portion of the transcripts (~1.2 Mb) belonged to bacteria of the genus *Alteromonas*. *Alteromonas* spp. are widespread in the oceans and representatives have been cultured that are able to utilize organic carbon for growth (López-Pérez et al. 2012). In the Guaymas Basin these *Altermonas* highly express genes for amino acid uptake and degradation. Interestingly, genes serving these functions are also well represented in the transcript data from uncultured Archaea referred to as Marine Group II Euryarchaeota. Also among the most abundantly represented genes in the plume and background metatranscriptomes were those encoding catalytic subunits of nitrite oxidoreductase (NxrA, NxrB, and membrane subunit). These transcripts are most similar to those found in the genome of *Nitrospira defluvii*, a nitrite oxidizer (Lücker et al. 2010). Therefore, we searched the Guaymas transcript data and found several other high expressed genes involved energy generation via the oxidation of nitrite coupled to the reduction of oxygen. Although these genes were abundant in the transcriptomic data they are present at very low coverage, or absent, in the accompanying genomic dataset. Hence, we found that this new nitrite-oxidizing bacterium was widespread in the Gulf of California, yet at low abundance in all the samples. This might be a reflection of the low energy yield that the nitrite oxidation confers relative to ammonia oxidation. This study was the first to utilize assembled transcript data to investigate the activity of a microbial community and showed that the second step of nitrification is mediated by a novel, low-abundance, uncultured group of bacteria.

6.3 Highly resolved mapping of carbon and energy flow through estuary sediment bacteria communities

The reconstruction of dozens of uncultured genomes of bacteria from the White Oak River estuary resulted in a detailed map of the flow of carbon, nitrogen, sulfur, and iron through various microbial populations in estuary sediments (Baker et al., in review). This dataset is the first of its kind from estuary sediments, and provides a foundation on which future investigations into geochemical cycling can be built. Many of the organisms identified are commonly found in marine sediments as well. Among the genomes are several from phyla that had not been genomically sampled before, including TA06.1, TA06.2, KD3-62, and KSB1. We also described the genomes belonging to newly described bacterial phyla, WOR-1 and BRC2. Many of the newly sequenced organisms were found to be capable of organic carbon degradation and fermentation. For example, one of the new groups, WOR-1, is capable of both acetate and H_2 production. This group and BRC1, BRC2, and KD3-62 have several glycoside hydrolase genes for the decomposition of organic carbon.

In addition to these new phyla, several genomes were obtained that belong to uncultured members of bacteria taxa whose ecological roles in sediments were unknown, including *Planctomycetes*, *Bacteroides*, *Spirochetes*, *Chloroflexi*, *Gammatiomonadetes*, and Proteobacteria. We were able to assign putative functions to several of these groups. For example, *Betaproteobacteria* appear to be capable of iron oxidation and, surprisingly, some of the *Gemmatiomonadetes* are capable of sulfate reduction. Genomes of the *Planctomycetes*, which are widespread in marine sediments (Schauer et al. 2009), contained the most glycoside hydrolases, suggesting these bacteria are specialized at organic carbon degradation.

Although sulfur oxidation in sediments has been measured by geochemical approaches (Schippers and Jórgensen, 2001, Jórgensen, 1990), very little was known about the microbes mediating this process. Recently, novel *Gammaproteobacteria* in coastal sediments were shown to be involved in sulfur oxidation (Lenk et al. 2011). One of the most abundant organisms in the White Oak River shallow samples (2-8 cm) were novel *Gammaproteobacteria* that are metabolically flexible. They are capable of sulfur oxidation coupled to oxygen reduction or anaerobic nitrate and nitrite reduction.

Future research at this site, and others in the oceans, will involve this level of genomic reconstruction of the archaeal community. We have already assembled several genomes belonging to archaea groups that are ubiquitous in sediments throughout the world, including Marine Benthic Groups (MBG), Miscellaneous Crenarchaeota Groups (MCG), SAGMEG, Ancient Archaeal Group (AAG), etc. Once we have a broad representation of genomes belonging to all the different groups common to sediments, we will have a better understanding of the physiologies of these uncultured microbes. This will enhance our ability to test a variety of hypothesis using additional techniques such as cultivation, enrichments on various carbon substrates, and spatial/temporal transcriptomic sampling. This will provide rich insights into linking geochemical functions to particular community members (Newman et al. 2012).

6.5 References

Baker BJ, et al. (2012) Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J* 6, 2269-2279.

Baker BJ, and Dick GJ (2013) Omic approaches in microbial ecology: charting the unknown. *Microbe* 8, 353-359.

Baker BJ, et al. (2013) Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *ISME J* 7, 1962-1973.

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

Jórgensen BB (1990) A thiosulfate shunt in the sulfur cycle of marine sediments. *Science* 249, 152-154.

Hollibaugh JT, et al. (2011) Metatranscriptomic analyses of ammonia-oxidizing organisms in an estuarine bacterioplankton assemblage. *ISME J* 5, 866–878.

López-Pérez M, et al. (2012) Genomes of surface isolates of *Alteromonas macleodii*: the life of a widespread marine opportunistic copiotroph. *Scientific Reports* 2, 696.

Lücker S, et al. (2010) A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci* 107, 13479-13484.

Newman DK, et al. (2012) Molecular biology's contributions to geobiology. *Fundamentals of Geobiology* Chapter 13.

Santoro AE, et al. (2011) Isotopic signature of N_2O produced by marine ammonia oxidizing Archaea. *Science* 333, 1282-1285.

Schauer R, et al. (2010) Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. *ISME J* 4, 159-170.

Schippers A and Jórgensen BB (2001) Biogeochemistry of pyrite and iron sulfide oxidation in marine sediments. *Geochim Cosmochim Acta* 66, 85-92.

Stewart FJ, Ulloa O, DeLong EF (2011) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* 14, 23–40.

Tully BJ, et al. (2012) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* 14, 254-267.