Evolutionary response of marine bacteria to the co-occurring pnictogens phosphorus and arsenic

Jaclyn K. Saunders

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Gabrielle Rocap, Chair

John Baross

James E. Gawel

Program Authorized to Offer Degree:

School of Oceanography

University of Washington

Abstract

Evolutionary response of marine bacteria to the co-occurring pnictogens phosphorus and arsenic

Jaclyn K. Saunders

Chair of the Supervisory Committee:

Dr. Gabrielle Rocap

School of Oceanography

Abstract

Phosphorus is an essential component of biomolecules, and is believed to be the limiting nutrient in the marine environment on geologic timescales. In marine waters, phosphorus co-occurs with arsenic, an element with many similar chemical properties. The occurrence of these two elements has shaped the evolutionary trajectory of the microorganisms that thrive in marine waters, and conversely, biochemical transformations by marine microorganisms influence the global cycling of these elements. Chapter 1 demonstrates the response of the marine Picocyanobacterium *Prochlorococcus* strain MED4 to phosphorus stress through a proteomic analysis of cells exposed to nutrient starvation and long-term nutrient limitation. A biogeographic analysis of some of the most differentially expressed proteins under phosphorus stress highlights the selective force of phosphorus scarcity on the genomic capacity of environmental *Prochlorococcus* populations. In Chapter 2, the inextricable link between the cycling of phosphorus and arsenic is highlighted in the evolutionary response of

Prochlorococcus arsenic detoxification mechanisms, where populations which experience greater phosphorus scarcity, and therefore greater risk of cellular arsenic exposure, show greater genomic capacity for arsenic detoxification. The arsenic detoxification mechanisms outlined in Chapter 2 utilize one of the few physicochemical differences between P and As – the different reduction potentials – in the process of detoxification. Chapter 3 describes a different group of marine microorganisms which also exploit the reduction potential of arsenic, but in this case for bioenergetic gains in anoxic marine waters. The genomic capacity for a complete arsenic metabolic cycle is described, with gene presence and expression of dissimilatory arsenate reductase and a chemoautotrophic form of arsenite oxidase in anoxic pelagic waters. This thesis highlights the selective pressures of phosphorus and arsenic availability on the evolution of marine microbial communities.

Acknowledgements

It is nearly impossible to thank all the amazing people who have helped to make this achievement possible. I must first thank my advisor, Gabrielle Rocap, for her years of support, vision, and guidance. I am in awe of the sheer breadth of your knowledge and your consistently insightful feedback. I am also extremely grateful to the rest of my committee members. You have all shaped my critical thinking skills and research in innumerous ways, and you have done so with kindness. John Baross – your passion for understanding the origins of life is infectious. Thank you for our many discussions and your support over the years. You have been an inspiring presence throughout my years in graduate school. A big thanks to Jim Gawel for your sincerity and support. It has been great to have you on my committee and to work with you at UW Tacoma. I deeply admire your mentorship of undergraduates and hope to emulate your mentoring style. Steve Emerson, I could always count on you to pull me out of my narrow focus to look at the broader scientific questions. Thank you, Steve, for asking the tough holistic questions in a way which I found challenging, but never disparaging. To Bob Morris, it has been wonderful to have your insight as a member of my committee, but even more so I have thoroughly enjoyed teaching with you. Your demeanor with your students is admirable – the students visibly notice your honest sincerity and enthusiasm. Thank you to Bruce Nelson, I greatly appreciate your generosity and am sincerely thankful for you taking the time to help make my defense a reality. I am so fortunate to have had such a wonderful group of people to mentor me throughout graduate school.

I also want to thank all the other amazing people I have had the honor to work with or learn from. The entire School of Oceanography has been an incredible place to grow and learn these last few years. I would like to thank Mikelle Nuwer, who has been such an amazing

teaching and life mentor over the years. Mikelle, you are such an amazing and kind person. I am so thankful for the opportunity to work in the School of Oceanography alongside such a passionate and brilliant group of individuals. I also want to thank the UW Astrobiology Program for being so welcoming and inspirational. The interdisciplinary work done by the Astrobiology Program has captivated me. I am so thankful to the many students, faculty, and staff of the Astrobiology Program which have left an enduring influence on me. In particular, I would like to thank John Baross, Roger Buick, Rika Anderson, Eva Stüeken, Vikki Meadows, and David Catling for the mentorship and fascinating discussions. A special thanks to Roger and Eva who provided me the opportunity to try my hand at Geology. Thanks to Jim Gawel and Aron Rigg for teaching me the ways of the ICP-MS at UW Tacoma. I would also like to thank the greater administration of the College of the Environment. Dean Graumlich and her administration clearly care about the student experience within the College of the Environment, and for that I thank you. Thank you for the opportunity to represent my fellow students through the Student Advisory Council.

I must thank all the amazing people I have worked with at the Center for Environmental Genomics. CEG has been my second home. Thanks to Gabrielle Rocap, Ginger Armbrust, and Bob Morris who are the PI's running the giant mixed group of enthusiastic researchers and students. Thanks to the wonderful Rita Peterson who has always watched out for me. You are so good to us, Rita. Rhonda Morales, Megan Schatz, and Bryn Durham – you are lab wizards who have always been so helpful and kind with my numerous questions. Specifically, I thank everyone in the Rocap Lab, both past and present, for your comradery and assistance. Thanks to Michael Carlson, Clara Fuchsman, Christina Miller, and Cedar McKay. I dearly thank all the

amazing, brilliant, energetic, and spirited students that I have been so fortunate to mentor in the lab. You have all made such a tremendous impact on both my research and me as a person. I cannot imagine working in the lab without you all. To all my students: Jennefer Lopez, Myesa Legendre-Fixx, Claire Knox, Kelsey Gibbons, Michael Lee, and Natalie Kellogg; I wish you all the best in your future endeavors and I cannot wait to see what amazing things you accomplish. I must also thank my numerous office mates who have become good friends and always made the office a fun place to be: Sara Bender, Gwenn Hennon, Katie Marshall Lalish, Michael Carlson, Jeff Turner, and Sophie Clayton – you guys rock.

Thank you to all the amazing mentors who helped me grow as a young scientist before graduate school. Thanks to Susan Kalisz for her amazing Evolution and Population Genetics classes at the University of Pittsburgh and taking me into her lab as an undergraduate researcher. Thanks to the entire Kalisz lab group, specifically April Randle who was an amazing graduate student mentor. I also want to thank the wonderful researchers who I had the honor to work with at the Smithsonian Environmental Research Center, specifically my direct mentors Melissa McCormick and Mario Sengco. You were influential forces on my scientific training and you both are just totally awesome people. Finally, I thank my family. Your love and support has provided me the stability to pursue my dreams. Ginny and Taleb, I am so thankful for your support throughout college. Mom and Jennifer – you have always been my biggest champions, and for that I am forever grateful. Hans, I am so lucky to have you. You and Edie are my family, my home.

Dedication

To Mom.

For your unwavering love and support.

Table of Contents

List of Figures

List of Tables

Introduction

Phosphorus, hydrogen, carbon, nitrogen, oxygen, and sulfur are the major macronutrients comprising biomolecules essential for life on Earth. Phosphorus is primarily found in the pentavalent state in the form of phosphate and phosphate esters. Phosphate esters are incredibly stable, undergoing slow abiotic hydrolysis, but hydrolyzing readily in the presence of enzymes making them an ideal component for biomolecules (Westheimer, 1987). This capacity makes biomolecules stable enough to maintain structure, but labile enough to catabolize when needed. Phosphate esters are a critical component of many biomolecules: the addition and removal of phosphate groups in the process of oxidative phosphorylation acts as the chemical energetic currency driving cellular function, multiple intermediate metabolites are phosphate esters, phospholipids are a major component of cellular membranes, and the sugar-phosphate backbone of DNA and RNA provides the necessary structure to maintain genetic material.

In the marine environment, primary production is generally limited by the availability of the macronutrients phosphorus or nitrogen (Howarth, 1988). Surface phosphate concentrations are quite variable globally, ranging from <0.1 μmol L^{-1} in oligotrophic subtropical gyres to > 2.0 μmol L^{-1} in the Southern Ocean (Karl, 2007). Within central regions of oligotrophic gyres, phosphate concentrations can range from 0.2-10 nmol L^{-1} (Wu *et al.*, 2000; Ammerman, 1993). Not only does phosphorus vary spatially in availability, but it has also varied in abundance in the ocean over geologic time scales with early oceans likely undergoing periods of extreme phosphorus scarcity (Planavsky *et al.*, 2010; Bjerrum and Canfield, 2002). While biologically available nitrogen may be supplied through biological fixation of atmospheric N_2 gas, phosphorus is ultimately sourced through the weathering of continental rock (Paytan and McLaughlin, 2007). Since the ultimate origin of phosphorus in the marine environment is from

continental sources, it is believed that over geologic timescales phosphorus is the primary limiting nutrient for primary productivity (Tyrrell, 1999). A byproduct of phosphorus scarcity is the increased risk of arsenic toxicity for microbes. Arsenic, similar to phosphorus, is also predominately found in modern marine waters in the pentavalent form arsenate. Arsenate is present in marine oligotrophic surface waters at a range of 10-15 nmol L^{-1} (Cutter and Cutter, 2006; Cutter *et al.*, 2001). A mechanism of arsenic cellular toxicity is rooted in its similarity to phosphorus, where it actively competes phosphate in uptake and biosynthetic processes. The relative abundance of these two elements has not been stable throughout Earth's history, rather microbial communities have had to respond to varying availabilities of P:As (Planavsky *et al.*, 2010; Bjerrum and Canfield, 2002; Chi Fru *et al.*, 2015; Chi Fru *et al.*, 2016). P and As are linked through microbial pathways, and biogeochemical cycles, as organisms strive to gather the necessary resources for life.

Nutrient scarce oligotrophic subtropical gyres are the largest ecosystem on the planet comprising about 40% of the surface of the earth (Polovina *et al.*, 2008). These oligotrophic regions are intensely vertically stratified, with nutrient input occurring either on seasonal cycles or through wind-driven mixing events (Karl, 1999; Lee *et al.*, 1994). In the euphotic zone, nutrients are actively taken up and recycled in the microbial loop (Azam *et al.*, 1983; Caron, 1994). In cases where iron is available, for example through Aeolian dust deposition from the Sahara Desert (Duce and Tindale, 1991), iron-intensive biological nitrogen fixation by marine cyanobacteria may provide a biologically available nitrogen source driving production to phosphorus limitation (Wu *et al.*, 2000). Nutrient scarcity places a bottom-up control on the amount of primary production that can be supported by a system, as well as influences the community structure of microbial assemblages in a region. As global temperatures continue to

increase due to climate change, vertical stratification of surface waters will intensify driving reduced nutrient availability and therefore a reduction in primary productivity in tropical and subtropical oligotrophic waters (Behrenfeld *et al.*, 2006). Not only are these oligotrophic regions of the ocean expected to be driven to greater nutrient scarcity, but they are rapidly increasing in expanse at a rate of 1-4% per year (Polovina *et al.*, 2008). It is imperative to understand and predict the controls on nutrient availability, and therefore system productivity, in the face of rising global temperatures.

The marine picocyanobacterium *Prochlorococcus* is the main primary producer in these oligotrophic regions (Campbell and Nolla, 1994; Campbell *et al.*, 1997; DuRand *et al.*, 2001; Scanlan, 2012), accounting for 13-48% of primary production (Johnson *et al.*, 2006). The sheer abundance of *Prochlorococcus*, about 10^{27} cells globally (Flombaum *et al.*, 2013), means that it has a significant impact on global biogeochemical cycles. *Prochlorococcus* is a successful minimalist, small in cell size with a streamlined cellular architecture (Ting *et al.*, 2007), an ability to dramatically reduce its cellular phosphorus quota when nutrient stressed (Bertilsson *et al.*, 2003; Heldal *et al.*, 2003), and maintenance of a minimal streamlined genome (Rocap *et al.*, 2003). Part of the success of *Prochlorococcus* is through individual sacrifice of phenotypic plasticity via gene loss, but maintenance of a large pan-genome (Kettler *et al.*, 2007) which enables niche specific success among strains. *Prochlorococcus* ecotypes follow the phylogenetic divisions of the 16S rRNA; however, the accessory genes associated with phosphate acquisition are not congruent with the 16S phylogeny (Martiny *et al.*, 2006). Rather, the occurrence of phosphate acquisition genes appears to correspond to the environmental phosphate conditions experienced by the *Prochlorococcus* populations (Martiny *et al.*, 2009). *Prochlorococcus* strains and populations which dominate the phosphate scarce regions of the surface ocean maintain the

highest abundance and most numerous types of phosphorus acquisition genes. Chapter 1 focuses on the physiological response of *Prochlorococcus* to phosphorus stress through the analysis of proteomic profiles of the axenic *Prochlorococcus* strain MED4 cultured under different nutrient stress conditions. Under P-stress, the loci PMM1409, PMM1414, and PMM1416 are some of the most dramatically upregulated proteins, and are therefore targeted for a biogeographic analysis of their occurrence in global *Prochlorococcus* populations captured in marine surface water metagenomes in oceanic regions with varying degrees of phosphate availability. Analysis of the relative abundance of these genes was conducted using a bioinformatics pipeline which assigns function and taxonomic origin to metagenomic reads using a phylogenetic inference placement method (Berger *et al.*, 2011; Berger and Stamatakis, 2011; Stamatakis, 2014) outlined in Appendix I. The occurrence pattern of three loci in *Prochlorococcus* populations under various environmental phosphate conditions further illustrating the selective force phosphorus scarcity has on shaping global *Prochlorococcus* populations.

When phosphate stressed, *Prochlorococcus* will increase its phosphate acquisition capabilities and therefore phosphate uptake rate and affinity (Krumhardt *et al.*, 2013). While aiding in the acquisition of the essential macronutrient, this increase in acquisition capacity under phosphate scarce conditions introduces a greater risk of uptake of the chemically and structurally similar arsenate anion. *Prochlorococcus* maintains the high affinity phosphate uptake system *pst.* The *pst* system is unable to completely differentiate between arsenate and phosphate, resulting in the indiscriminate uptake of arsenate in the *Escherichia coli* model system (Rosenberg *et al.*, 1977; Tawfik and Viola, 2011). The ratio of available phosphate to arsenate can be driven extremely low, below 1 in some instances (Wurl *et al.*, 2013; Karl and Tien, 1997), resulting in a greater risk of competitive arsenate uptake. Once inside the cell, arsenate can become toxic by

competing with phosphate, for example through the decoupling of oxidative phosphorylation, the process that produces ATP (Mandal and Suzuki, 2002; Oremland and Stolz, 2003). The reduced form, arsenite, is even more toxic because it interferes with enzyme activity by bonding to –SH and –OH groups in enzymes (Mandal and Suzuki, 2002; Akter *et al.*, 2005). Microorganisms possess a range of cellular mechanisms to moderate the toxic effects of the arsenic molecule.

This risk of arsenic uptake is not restricted to modern stratified surface waters, but rather has been a dominant force shaping the evolutionary trajectory of life throughout geologic time. Microbial arsenic resistance mechanisms are numerous, taxonomically widespread, and subject to frequent horizontal transfer (Stolz *et al.*, 2006; Páez-Espino *et al.*, 2009). Phylogenetic analyses of many of the enzymes acting in arsenic resistance indicate a likely ancient origin, prior to the divergence of Bacteria and Archaea (Lebrun, 2003; Jackson and Dugas, 2003). The general efflux detoxification pathway involves the reduction of arsenate to arsenite, and then subsequent expulsion of arsenic from the cell through arsenite specific transporters (Carlin *et al.*, 1995; Ghosh *et al.*, 1999). The genes sufficient for a complete efflux pathway were previously identified in *Prochlorococcus* genomes (Scanlan *et al.*, 2009) including the arsenate reductase, *arsC*, the *acr3* arsenite efflux transporter and *arsR*, an arsenite-binding *trans*-acting repressive regulator (Xu *et al.*, 1996). This efflux detoxification was believed to be the sole arsenic detoxification strategy for *Prochlorococcus*.

In Chapter 2, a second putative arsenic detoxification mechanism in *Prochlorococcus* is identified, a strategy where arsenic is methylated into more innocuous organic compounds. This pathway involves the repeated methylation and reduction of arsenical compounds to less toxic states mediated by the enzyme ArsM, arsenite *S*-adenosylmethyltransferase (Qin *et al.*, 2006; Yin *et al.*, 2011). The role arsenic toxicity, as a function of the P:As ratio, has had on the

evolution of *Prochlorococcus* arsenic resistance mechanisms and their distribution in environmental *Prochlorococcus* populations is also demonstrated in Chapter 2. The varied arsenic detoxification strategies maintained in global *Prochlorococcus* populations can result in very different impacts on global biogeochemical cycling of arsenic. Findings in Chapter 2 demonstrate that some populations of *Prochlorococcus* maintain the genomic capacity for arsenic efflux which results in trivalent arsenite as a byproduct, whereas the putative arsenic methylation strategy appears to be maintained ubiquitously by *Prochlorococcus* populations and results in the production of organoarsenicals. These two pathways have implications for the global arsenic cycle as arsenite is oxidized back to arsenate on the order of hours-days (Cutter, 1992) whereas simple organoarsenicals have a residence time on the order of years (Cutter and Cutter, 2006).

While arsenic is often considered a toxin in marine systems, due to its competitive uptake and cellular toxicity effects, it can be of beneficial use to marine microbes. Phosphorus and arsenic have very different oxidation-reduction potentials: arsenic is electrochemically positive with a potential of +139mV, whereas phosphorus is electrochemically negative with a potential of -690mV (Rosen *et al.*, 2011). Cellular arsenate detoxification systems exploit this difference in redox potential to reduce arsenate to arsenite. The cellular mechanisms can then differentiate an arsenite molecule from a phosphate molecule and act accordingly to detoxify the molecule via expulsion or methylation. A range of microorganisms exhibit the capacity to turn the potentially toxic arsenic into a beneficial source of bioenergetic gains by exploiting the reduction potential of arsenic (Amend *et al.*, 2014). Multiple bacteria have demonstrated the capacity to use arsenate as a terminal electron acceptor in dissimilatory respiratory chains (Oremland and Stolz, 2003; Ahmann *et al.*, 1994). Some microbes have also demonstrated the ability to reap energetic gains

through the oxidation of arsenite. A few arsenite oxidizing bacteria have shown the ability to use arsenite oxidation to fuel carbon fixation in aerobic and anaerobic conditions, with the discovery of a microorganism which uses arsenite as a reductant in photosynthesis (Oremland and Stolz, 2003; Rosen *et al.*, 2011; Amend *et al.*, 2014). Some chemoautotrophic arsenite oxidizers have been shown to utilize nitrate in place of oxygen as the terminal electron acceptor (Oremland *et al.*, 2002; Rhine *et al.*, 2007; Rhine *et al.*, 2006) with a few representatives demonstrating complete reduction of nitrate to N_2 (Rhine *et al.*, 2006) , highlighting the impact these organisms may have on global nitrogen cycling. A full metabolic arsenic cycle – chemoautotrophic oxidation of arsenite cycled with dissimilatory arsenate reduction – has been identified as an active cycle in modern systems like the arsenic loaded Mono Lake (Oremland and Stolz, 2003). A complete metabolic arsenic cycle is likely ancient, existing since the Archaean (Lebrun, 2003; Duval *et al.*, 2008) with fossil evidence of arsenic-rich organic globules supporting a complete metabolic arsenic cycle 2.72 billion years ago (Sforna *et al.*, 2014) and phylogenetic analyses indicating a likely ancient origin, prior to the Last Universal Common Ancestor, of the bioenergetic arsenite oxidase enzyme which was likely coupled to the reduction of nitrogen oxyanions (van Lis *et al.*, 2013).

Oxygen Deficient Zones (ODZs) are functionally anoxic regions of the water column where oxygen concentrations are below 4 nmol L^{-1} (Ulloa *et al.*, 2012; Tiano *et al.*, 2014). These mid-layer oceanic oxygen deficient zones, sandwiched between oxygenated surface and deep layers, are caused through a combination of organic matter respiration and slow mid-layer depth horizontal circulation of water masses common in eastern ocean basins along the subtropics (Wyrtki, 1962). These regions are major sources of nitrogen loss (removal of bioavailable forms) from the marine system as numerous anaerobic metabolisms rely on nitrogen oxyanions as

electron acceptors (Lam and Kuypers, 2011). These anoxic marine waters are also modern day proxies for early anoxic oceans (Ulloa *et al.*, 2012). Arsenic is 14th most abundant element in seawater (Mandal and Suzuki, 2002), and arsenic concentrations were likely much greater in early anoxic oceans (Chi Fru *et al.*, 2015; Oremland *et al.*, 2009; Kulp, 2014). In Chapter 3, the genomic capacity for a complete anaerobic arsenic metabolic cycle in a modern global oxygen deficient zone is observed. Genes specific for dissimilatory arsenate reduction, *arrAB,* and chemoautotrophic arsenite oxidation, *aioAB,* are identified in a metagenome from the anoxic waters of the Eastern Tropical North Pacific (ETNP). These genes are shown to not only be present, but also actively transcribed in modern oxygen deficient waters in the Eastern Tropical North Pacific and Eastern Tropical South Pacific. Additionally, an abundance of metagenomic and metatranscriptomic reads of an enzyme closely related to the bioenergetic arsenite oxidase enzyme, *aioA*, are identified. A long contiguous DNA sequence (>82 kilobases) which contains this "arsenite oxidase like", or *aioA-like,* sequence was assembled from the ETNP metagenome. This stretch of genomic sequence also included a sequence of a fumarate reductase, a key enzyme in the reverse TCA cycle of carbon fixation (Hügler and Sievert, 2011), as a well as a nitrate reductase (*napA*) sequence. These arsenic, and arsenic related, metabolisms have the capacity to dramatically influence global arsenic, carbon, and nitrogen cycling in modern anoxic marine waters and may have dramatically influenced these cycles in early oceans where arsenic was likely more plentiful and anoxic conditions pervaded.

The chapters and appendices of this thesis show the selective pressure that phosphorus and arsenic have had on shaping the evolution and metabolic capacity of marine microorganisms, and conversely the potential these microbial metabolisms have on influencing the global cycling of P, As, C, and N today. The group 15 elements phosphorus and arsenic are physicochemically

similar, being taken up through the same cellular uptake pathways and causing toxicity through competitive replacement in biomolecule formation. The reduction potential of these two elements is exploited by marine microbes: in detoxification mechanisms cells exploit this feature in order to identify the toxic arsenic molecule, and some marine microbes are able to exploit this reduction potential for beneficial energetic gains. All three chapters demonstrate some of the evolutionary pressures imposed by these two elements. Chapter 1 shows the physiological response of *Prochlorococcus* strain MED4 to phosphorus stress through a proteomics analysis, and the evolutionary pressure phosphorus stress has on shaping the phosphorus acquisition capacity of *Prochlorococcus* populations. Chapter 2 also shows the streamlining of *Prochlorococcus* genomes, but here the inextricable link between phosphorus and arsenic competitive uptake is highlighted, evidenced by how it has shaped arsenic detoxification strategies in *Prochlorococcus*. And finally, Chapter 3 demonstrates an adventitious use of arsenic by marine microorganisms for the reaping of bioenergetic gains by exploiting the redox properties of arsenic in anoxic marine waters. As a whole, this thesis demonstrates microbial evolutionary ingenuity in response to phosphorus and arsenic availability in the marine environment.

References:

1 Westheimer FH (1987). Why nature chose phosphates. *Science* **235:** 1173-1178.

2 Howarth RW (1988). Nutrient limitation of net primary production in marine ecosystems. *Annual review of ecology and systematics***:** 89-110.

3 Karl DM (2007). The marine phosphorus cycle. ASM Press: Washington. pp 523-539.

4 Wu J, Sunda W, Boyle EA, Karl DM (2000). Phosphate Depletion in the Western North Atlantic Ocean. *Science* **289:** 759-762.

5 Ammerman J (1993). Microbial cycling of inorganic and organic phosphorus in the water column. *Handbook of methods in aquatic microbial ecology Lewis Publishers, Boca Raton, FL***:** 649-660.

6 Planavsky NJ, Rouxel OJ, Bekker A, Lalonde SV, Konhauser KO, Reinhard CT *et al.* (2010). The evolution of the marine phosphate reservoir. *Nature* **467:** 1088-1090.

7 Bjerrum CJ, Canfield DE (2002). Ocean productivity before about 1.9 Gyr ago limited by phosphorus adsorption onto iron oxides. *Nature* **417:** 159-162.

8 Paytan A, McLaughlin K (2007). The Oceanic Phosphorus Cycle. *Chem Rev* **107:** 563-576.

9 Tyrrell T (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature* **400:** 525-531.

10 Cutter GA, Cutter LS (2006). Biogeochemistry of arsenic and antimony in the North Pacific Ocean. *Geochemistry Geophysics Geosystems* **7**.

11 Cutter GA, Cutter LS, Featherstone AM, Lohrenz SE (2001). Antimony and arsenic biogeochemistry in the western Atlantic Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* **48:** 2895-2915.

12 Chi Fru E, Arvestål E, Callac N, El Albani A, Kilias S, Argyraki A *et al.* (2015). Arsenic stress after the Proterozoic glaciations. *Scientific Reports* **5:** 17789.

13 Chi Fru E, Hemmingsson C, Holm M, Chiu B, Iñiguez E (2016). Arsenic-induced phosphate limitation under experimental Early Proterozoic oceanic conditions. *Earth and Planetary Science Letters* **434:** 52-63.

14 Polovina JJ, Howell EA, Abecassis M (2008). Ocean's least productive waters are expanding. *Geophysical Research Letters* **35**.

15 Karl DM (1999). A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems* **2:** 181-214.

16 Lee D-K, Niiler P, Warn-Varnas A, Piacsek S (1994). Wind-driven secondary circulation in ocean mesoscale. *Journal of Marine Research* **52:** 371-396.

17 Azam F, Fenchel T, Field JG, Gray J, Meyer-Reil L, Thingstad F (1983). The ecological role of water-column microbes in the sea. *Estuaries* **50**.

18 Caron D (1994). Inorganic nutrients, bacteria, and the microbial loop. *Microbial Ecology* **28:** 295-298.

19 Duce RA, Tindale NW (1991). Atmospheric transport of iron and its deposition in the ocean. *Limnology and Oceanography* **36:** 1715-1726.

20 Behrenfeld MJ, O'Malley RT, Siegel DA, McClain CR, Sarmiento JL, Feldman GC *et al.* (2006). Climate-driven trends in contemporary ocean productivity. *Nature* **444:** 752-755.

21 Campbell L, Nolla HA (1994). The importance of *Prochlorococcus* to community structure in the central north pacific ocean. *Limnology and Oceanography* **39:** 954-961.

22 Campbell L, Liu H, Nolla HA, Vaulot D (1997). Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991–1994 ENSO event. *Deep Sea Research Part I: Oceanographic Research Papers* **44:** 167-192.

23 DuRand MD, Olson RJ, Chisholm SW (2001). Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **48:** 1983-2003.

24 Scanlan D (2012). Marine Picocyanobacteria. In: Whitton BA (ed). *Ecology of Cyanobacteria II*. Springer Netherlands. pp 503-533.

25 Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311:** 1737-1740.

26 Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N *et al.* (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proceedings of the National Academy of Sciences of the United States of America* **110:** 9824-9829.

27 Ting CS, Hsieh C, Sundararaman S, Mannella C, Marko M (2007). Cryo-Electron Tomography Reveals the Comparative Three-Dimensional Architecture of *Prochlorococcus*, a Globally Important Marine Cyanobacterium. *Journal of Bacteriology* **189:** 4485-4493.

28 Bertilsson S, Berglund O, Karl DM, Chisholm SW (2003). Elemental composition of marine *Prochlorococcus* and *Synechococcus*: Implications for the ecological stoichiometry of the sea. *Limnology and Oceanography* **48:** 1721-1731.

29 Heldal M, Scanlan DJ, Norland S, Thingstad F, Mann NH (2003). Elemental composition of single cells of various strains of marine *Prochlorococcus* and *Synechococcus* using X-ray microanalysis. *Limnology and Oceanography* **48:** 1732-1743.

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

31 Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:** e231.

32 Martiny AC, Coleman ML, Chisholm SW (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proceedings of the National Academy of Sciences* **103:** 12552-12557.

33 Martiny AC, Huang Y, Li W (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environmental Microbiology* **11:** 1340-1347.

34 Berger SA, Krompass D, Stamatakis A (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology* **60:** 291-302.

35 Berger SA, Stamatakis A (2011). Aligning short reads to reference alignments and trees. *Bioinformatics* **27:** 2068-2075.

36 Stamatakis A (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*.

37 Krumhardt KM, Callnan K, Roache-Johnson K, Swett T, Robinson D, Reistetter EN *et al.* (2013). Effects of phosphorus starvation versus limitation on the marine cyanobacterium *Prochlorococcus* MED4 I: uptake physiology. *Environmental Microbiology* **15:** 2114-2128.

38 Rosenberg H, Gerdes RG, Chegwidden K (1977). Two systems for the uptake of phosphate in *Escherichia coli*. *J Bacteriol* **131:** 505-511.

39 Tawfik DS, Viola RE (2011). Arsenate Replacing Phosphate: Alternative Life Chemistries and Ion Promiscuity. *Biochemistry* **50:** 1128-1134.

40 Wurl O, Zimmer L, Cutter GA (2013). Arsenic and phosphorus biogeochemistry in the ocean: Arsenic species as proxies for P-limitation. *Limnology and Oceanography* **58:** 729-740.

41 Karl DM, Tien G (1997). Temporal variability in dissolved phosphorus concentrations in the subtropical North Pacific Ocean. *Mar Chem* **56:** 77-96.

42 Mandal BK, Suzuki KT (2002). Arsenic round the world: a review. *Talanta* **58:** 201-235.

43 Oremland RS, Stolz JF (2003). The Ecology of Arsenic. *Science* **300:** 939-944.

44 Akter KF, Owens G, Davey DE, Naidu R (2005). Arsenic speciation and toxicity in biological systems. *Rev Environ Contam Toxicol* **184:** 97-149.

45 Stolz JF, Basu P, Santini JM, Oremland RS (2006). Arsenic and Selenium in Microbial Metabolism*. *Annual Review of Microbiology* **60:** 107-130.

46 Páez-Espino D, Tamames J, Lorenzo V, Cánovas D (2009). Microbial responses to environmental arsenic. *BioMetals* **22:** 117-130.

47 Lebrun E (2003). Arsenite Oxidase, an Ancient Bioenergetic Enzyme. *Molecular Biology and Evolution* **20:** 686-693.

48 Jackson CR, Dugas SL (2003). Phylogenetic analysis of bacterial and archaeal *arsC* gene sequences suggests an ancient, common origin for arsenate reductase. *BMC Evol Biol* **3**.

49 Carlin A, Shi W, Dey S, Rosen BP (1995). The *ars* operon of Escherichia coli confers arsenical and antimonial resistance. *J Bacteriol* **177:** 981-986.

50 Ghosh M, Shen J, Rosen BP (1999). Pathways of As(III) detoxification in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **96:** 5001-5006.

51 Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological Genomics of Marine Picocyanobacteria. *Microbiology and Molecular Biology Reviews* **73:** 249-299.

52 Xu C, Shi W, Rosen BP (1996). The chromosomal *arsR* gene of *Escherichia coli* encodes a trans-acting metalloregulatory protein. *The Journal of biological chemistry* **271:** 2427-2432.

53 Qin J, Rosen BP, Zhang Y, Wang G, Franke S, Rensing C (2006). Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. *Proc Natl Acad Sci U S A* **103:** 2075-2080.

54 Yin XX, Chen J, Qin J, Sun GX, Rosen BP, Zhu YG (2011). Biotransformation and volatilization of arsenic by three photosynthetic cyanobacteria. *Plant Physiol* **156:** 1631-1638.

55 Cutter GA (1992). Kinetic controls on metalloid speciation in seawater. *Mar Chem* **40:** 65-80.

56 Rosen BP, Ajees AA, McDermott TR (2011). Life and death with arsenic. *Bioessays* **33:** 350-357.

57 Amend JP, Saltikov C, Lu G-S, Hernandez J (2014). Microbial Arsenic Metabolism and Reaction Energetics. *Reviews in Mineralogy and Geochemistry* **79:** 391-433.

58 Ahmann D, Roberts AL, Krumholz LR, Morel FMM (1994). Microbe grows by reducing arsenic. *Nature* **371:** 750-750.

59 Oremland RS, Hoeft SE, Santini JM, Bano N, Hollibaugh RA, Hollibaugh JT (2002). Anaerobic oxidation of arsenite in Mono Lake water and by a facultative, arsenite-oxidizing chemoautotroph, strain MLHE-1. *Appl Environ Microbiol* **68:** 4795-4802.

60 Rhine ED, Ni Chadhain SM, Zylstra GJ, Young LY (2007). The arsenite oxidase genes (*aroAB*) in novel chemoautotrophic arsenite oxidizers. *Biochemical and Biophysical Research Communications* **354:** 662-667.

61 Rhine ED, Phelps CD, Young LY (2006). Anaerobic arsenite oxidation by novel denitrifying isolates. *Environmental Microbiology* **8:** 899-908.

62 Duval S, Ducluzeau AL, Nitschke W, Schoepp-Cothenet B (2008). Enzyme phylogenies as markers for the oxidation state of the environment: The case of respiratory arsenate reductase and related enzymes. *BMC Evol Biol* **8**.

63 Sforna MC, Philippot P, Somogyi A, van Zuilen MA, Medjoubi K, Schoepp-Cothenet B *et al.* (2014). Evidence for arsenic metabolism and cycling by microorganisms 2.7 billion years ago. *Nature Geosci* **7:** 811-815.

64 van Lis R, Nitschke W, Duval S, Schoepp-Cothenet B (2013). Arsenics as bioenergetic substrates. *Biochim Biophys Acta-Bioenerg* **1827:** 176-188.

65 Ulloa O, Canfield DE, DeLong EF, Letelier RM, Stewart FJ (2012). Microbial oceanography of anoxic oxygen minimum zones. *Proceedings of the National Academy of Sciences of the United States of America* **109:** 15996-16003.

66 Tiano L, Garcia-Robledo E, Dalsgaard T, Devol AH, Ward BB, Ulloa O *et al.* (2014). Oxygen distribution and aerobic respiration in the north and south eastern tropical Pacific oxygen minimum zones. *Deep Sea Research Part I: Oceanographic Research Papers* **94:** 173-183.

67 Wyrtki K (1962). The oxygen minima in relation to ocean circulation. *Deep Sea Research and Oceanographic Abstracts* **9:** 11-23.

68 Lam P, Kuypers MMM (2011). Microbial Nitrogen Cycling Processes in Oxygen Minimum Zones. In: Carlson CA, Giovannoni SJ (eds). Annual Reviews: Palo Alto. pp 317-345.

69 Oremland RS, Saltikov CW, Wolfe-Simon F, Stolz JF (2009). Arsenic in the Evolution of Earth and Extraterrestrial Ecosystems. *Geomicrobiol J* **26:** 522-536.

70 Kulp TR (2014). Early earth: Arsenic and primordial life. *Nature Geosci* **7:** 785-786.

71 Hügler M, Sievert SM (2011). Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean. *Annual review of marine science* **3:** 261-289.

Chapter 1

Effects of phosphorus starvation versus limitation on the marine cyanobacterium *Prochlorococcus* MED4: Protein expression

1.1 Abstract:

The marine picocyanobacterium *Prochlorococcus* is the dominant primary producer in open ocean oligotrophic waters. Nutrient scarcity places a limit on primary production, with phosphate scarcity being a major control on production in these nutrient depleted regions. Phosphate scarcity has shaped the evolutionary trajectory of *Prochlorococcus*, evidenced by the variable occurrence of P-acquisition genes among the strains and global populations of *Prochlorococcus*. Here we present a proteomic analysis of axenic *Prochlorococcus* MED4 cultured under nutrient replete, long-term nutrient limitation, and nutrient starvation conditions, evaluating both phosphorus and nitrogen stress responses by the proteome. The proteomic profiles revealed 167 proteins that are statistically differentially expressed under nutrient stress conditions. A general nutrient stress response observed is the reduction in structural proteins associated with RNA polymerase and ribosomes, with nutrient starvation showing a greater reduction in these proteins than nutrient limitation. Reduction in ATP synthase appears to be a hallmark of the onset of nutrient starvation, with stabilization of ATP synthase protein levels over long-term nutrient stress exposure. Long-term nutrient limitation is also associated with a decrease in photosystem II proteins, and nutrient starvation showing a greater decline in proteins associated with chlorophyll biosynthesis. Notably, the PstS high affinity phosphate binding protein, which is commonly used as a biomarker, was found to be significantly upregulated under P-stress and significantly downregulated under N-stress conditions, highlighting the need to use caution when utilizing this locus as a biomarker. The P-acquisition genes PhoA and the organic-P poring PMM0709 were also upregulated under P-stress. We identified a significant

upregulation of uncharacterized loci associated with P-stress inducible genomic islands, with PMM1409 and PMM0719 under P-starvation specifically and PMM1414 and PMM1416 under P-stress in general. The loci PMM1409, PMM1414, and PMM1416 are among the most highly upregulated proteins identified under conditions of phosphorus stress. In order to better understand these very active, but functionally uncharacterized loci, a biogeographic analysis of the presence of these genes in environmental *Prochlorococcus* populations was conducted. PMM1409 and PMM1416 occurrence is restricted to regions associated with the most extreme phosphorus scarcity and PMM1414 shows a greater relative abundance in regions with lower phosphate concentrations when compared with regions of greater phosphate availability. The variable occurrence of the P-stress inducible loci in global metagenomes highlights the powerful selective force of phosphorus on *Prochlorococcus* populations.

1.2 Introduction:

Subtropical oligotrophic gyres account for about 40% of the surface of the Earth (Polovina *et al.*, 2008) making them one of the largest ecosystems on the planet. The marine picocyanobacterium *Prochlorococcus* is the main primary producer in these regions (Campbell and Nolla, 1994; Campbell *et al.*, 1997; DuRand *et al.*, 2001; Scanlan, 2012), accounting for 13-48% of primary production (Johnson *et al.*, 2006). The sheer abundance of *Prochlorococcus*, about 10^{27} cells globally (Flombaum *et al.*, 2013), makes it a key player in biogeochemical cycles. *Prochlorococcus* is extremely successful in these nutrient limited waters due to several adaptations to survival in oligotrophic conditions. This organism is a successful minimalist – only maintaining the cellular machinery and genetic material required for survival (Partensky and Garczarek, 2010). *Prochlorococcus* is small in cell size with a streamlined cellular architecture (Ting *et al.*, 2007) and can reduce its cellular phosphorus requirements through

substitution of sulfolipids for phospholipids in the cell membrane (Van Mooy *et al.*, 2006). As a result *Prochlorococcus* under P-replete conditions has a cellular N:P content over 20, well above the Redfield ratio of 16:1, and can exceed 100:1 under P starvation (Bertilsson *et al.*, 2003; Heldal *et al.*, 2003).

As a genus, *Prochlorococcus* maintains small genomes which are thought to provide enhanced fitness through reduced nutrient requirements in oligotrophic systems (Giovannoni, 2005). Part of the success of *Prochlorococcus* is through individual sacrifice of phenotypic plasticity via gene loss, but maintenance of a large pan-genome (Kettler *et al.*, 2007) with numerous accessory genes variable in presence among *Prochlorococcus* which enables niche specific success among strains. *Prochlorococcus* can be divided up into physiologically and genetically distinct ecotypes (Moore *et al.*, 1998) which reflect adaptations to environmental parameters like light intensity (Moore *et al.*, 1998; Moore and Chisholm, 1999), temperature optima (Johnson *et al.*, 2006), and nitrogen utilization capabilities (Moore *et al.*, 2002) amongst others with further genetic divisions beyond ecotypes at the subpopulation level likely in response to biotic environmental pressures, especially the arms race between predator/prey and virus/host (Kashtan *et al.*, 2014).

Many of these physiological differences follow the phylogenetic divisions of the 16S rRNA; however, the accessory genes associated with phosphate acquisition genes are not congruent with the 16S phylogeny. As a genus, *Prochlorococcus* maintains the high-affinity phosphate uptake pathway *pstABCS*, but does not possess a homolog of the low-affinity phosphate uptake system (*pitAB*) common in other prokaryotes (Kettler *et al.*, 2007; Rocap *et al.*, 2003; Scanlan *et al.*, 2009). Some strains of *Prochlorococcus* are also capable of utilizing organic P sources (Moore *et al.*, 2005) through the potential uptake of organic-P by the outer

membrane porin PMM0709 (Reistetter *et al.*, 2013) and by cleaving inorganic P via alkaline phosphatases like *phoA* or possibly by 5'-nucelotidase-like enzymes (Ammerman and Azam, 1985; Krumhardt *et al.*, 2013). Many of these P-acquisition genes, such as *pstABCS, phoA* and, porin PMM0709, are found on the genomic island spanning from PMM0705-PMM0725 in MED4 which is also referred to as the "*phoB* region" because of the location of the 2-component histidine kinase regulator *phoBR* near many of the P-acquisition genes in *Prochlorococcus* (Coleman *et al.*, 2006; Martiny *et al.*, 2006). An additional genomic island in MED4, ranging from PMM1403-PMM1416, contains a suite of uncharacterized genes which show environmental genomic linkage with many of the characterized P-acquisition genes (Coleman *et al.*, 2006; Martiny *et al.*, 2006). Both of these genomic islands show enhanced gene expression under P-starvation culturing conditions (Reistetter *et al.*, 2013; Martiny *et al.*, 2006; Fuszard *et al.*, 2010; Fuszard *et al.*, 2012) as well as under a previous proteomic analysis which showed upregulated of PstS, PhoA, porin PMM0709, and PMM1416 after long-term phosphorus starvation (Fuszard *et al.*, 2010).

Variation in environmental phosphate availability appears to influence the genomic capacity for phosphorus acquisition among *Prochlorococcus.* Surface phosphate concentrations are quite variable globally, ranging from < 0.1 μmol L^{-1} in oligotrophic subtropical gyres to > 2.0 μmol L^{-1} in the Southern Ocean (Karl, 2007). *Prochlorococcus* populations and strains from P-scarce regions possess a greater complement of P-acquisition genes, indicating environmental selection for these P-acquisition genes when phosphate is in scarce supply (Martiny *et al.*, 2006; Martiny *et al.*, 2009; Coleman and Chisholm, 2011). Nutrient scarcity places a bottom-up control on the amount of primary production that can be supported, as well as influences the community structure of the microbial assemblage in a region. The oligotrophic regions of the ocean where

Prochlorococcus dominates are expected to become even more nutrient stressed through enhanced vertical stratification of the water column (Behrenfeld *et al.*, 2006) and spread in geographic space (Polovina *et al.*, 2008) due to rising global temperatures. The intensification of nutrient stress in tropical and subtropical gyres may favor the presence of many of these phosphorus stress related accessory genes in *Prochlorococcus* populations which thrive there, and greater expression of these loci may indicate physiological response of the cells to these stressors.

The high affinity phosphate binding transporter *pstS* has been targeted previously to detect signatures of P-stress in the environment (Fuller *et al.*, 2005). Environmental expression of *pstS* has been suggested as a biomarker of P-stress because it is significantly upregulated under P-stress in culture (Reistetter *et al.*, 2013; Martiny *et al.*, 2006; Fuszard *et al.*, 2012) and it is ubiquitously present in both picocyanobacterial genomes (Scanlan *et al.*, 2009) and environmental metagenomes with *Prochlorococcus* populations (Martiny *et al.*, 2009). However, complications arise with *pstS* as a biomarker as some strains have multiple copies, in addition some cyanophages carry *pstS* (Sullivan *et al.*, 2005) and express it when the host is P-stressed (Zeng and Chisholm, 2012).

Here, we utilize proteomics to detect proteins produced by axenic (without contaminating bacteria) cultures of the *Prochlorococcus* strain MED4, a High Light strain isolated from P-scarce waters in the Mediterranean with a host of accessory P-acquisition genes (Rocap *et al.*, 2003; Moore *et al.*, 1995), in response to nutrient stress. Previous studies have focused on cellular response to P-stress using only batch culturing conditions to elicit a P-starvation inducible response (Martiny *et al.*, 2006; Fuszard *et al.*, 2010; Fuszard *et al.*, 2012). Nutrient status in the environment does not generally function like batch culturing conditions where there

is an abrupt shift of nutrients over such short timescales, rather environmental nutrient stress is likely a more long-term exposure of moderate injection of nutrients which maintains limited growth rates. Here we present proteomics evidence to deconvolute signatures of nutrient limitation and nutrient starvation, as well as further deconvolute a P-specific or N-specific nutrient stress response. A biogeographic analysis was conducted of the occurrence of the most differentially expressed loci under P-stress in global surface metagenomes (Rusch *et al.*, 2007; Sunagawa *et al.*, 2015), showing the selective force of phosphorus availability on *Prochlorococcus* populations.

1.3 Results:

Identification of Prochlorococcus *MED4 peptides*

Batch cultures of axenic *Prochlorococcus* MED4 were grown in duplicate under phosphorus limiting or nitrogen limiting conditions. The P batch cultures grew exponentially until day 10 with a growth rate of 0.68 day^{-1} when they entered stationary phase with a growth rate of 0 day^{-1}; N batch cultures grew exponentially until day 14 when they entered stationary phase (Fig. 1). Cells were harvested for proteomics analysis during exponential growth on day 7 for P cultures and during stationary phase on days 11 and 16 for P and N cultures, respectively. Chemostat cultures were grown in duplicate at a dilution rate controlling the MED4 growth rate at a stable 0.2 day^{-1}. Previous analyses of these cultures have demonstrated that the MED4 cells have distinct physiological responses to different forms of P-stress, whether it is the onset of P-starvation or long-term P-limitation under chemostat conditions. The cells show physiological response to P-stress by decreasing their cellular P quotas under limitation, and even more dramatically under P-starvation. MED4 also shows increased uptake kinetics for P through an

enhanced specific affinity and V_{max} for PO_4 & ATP(Krumhardt *et al.*, 2013), and a higher Michaelis-Menten constant (K_m) for ATP when cells were P-stressed. While the alkaline phosphatase PhoA enzyme was upregulated, P-uptake experiments with these cultures under P-stress indicate that MED4 likely cleaves orthophosphate groups from organic molecules, like ATP, with an additional, albeit unknown, enzyme (Krumhardt *et al.*, 2013). A targeted gene expression of these cultures demonstrated that the P-stress related genes *phoA, pstS,* and a probable organic-P porin are all upregulated under both P-stress conditions, but that the *phoR* P-sensing histidine kinase is specifically upregulated under long-term P-limitation.

Proteomic profiles for each condition were identified through tandem mass spectral abundance identification, and comparison of identified peptides to the fully sequenced metagenome of *Prochlorococcus* MED4 (Rocap *et al.*, 2003). Three analytical runs of each biological sample were evaluated per nutrient condition. A cluster analysis of the proteomic profiles from these analytical replicates was conducted in order to evaluate binning of runs (Figure 2). Analytical replicates of biological replicates were closely related, and so were binned for further analysis as proteomic profiles representing each nutrient status. Relative abundance proteomic profiles of cultures under P- and N- limitation and starvation were compared to the proteomic profile of cells under nutrient replete exponential growth from day 7 phosphorus batch cultures in order to identify significantly differentially expressed proteins present under the different nutrient stress conditions.

Differential expression of proteins by nutrient status

This comparison of differentially expressed proteins under different nutrient stressors – N or P stress – and under different physiologically stressed states inducing either limited or

stationary growth, has enabled us to identify loci associated with general nutrient stress response as well as nutrient and physiological state-specific responses. The proteomic profile of each condition was assessed, with the nutrient stress states compared against the proteomic profile of the nutrient replete cells. *Prochlorococcus* MED4 contains a possible 1,942 protein coding genes as determined by analysis of the genome. Of these possible proteins, 936 different proteins were identified in the mass spectral analysis. Of the proteins identified, 167 proteins were identified as statistically differentially expressed under nutrient stress (88, 73, 61, and 98 proteins differentially expressed under P-limitation, P-starvation, N-limitation, and N-starvation, respectively). Some general trends were observed, differentiating characteristics of the onset of nutrient-starvation compared to long-term nutrient limitation. Structural proteins associated with RNA polymerase and the Ribosome were significantly downregulated under nutrient limitation, and even more so under nutrient starvation, corresponding with a decline in growth rates from 0.2 day^{-1} under limitation to 0 day^{-1}at the onset of starvation. In addition, long-term nutrient limitation showed a greater decline in proteins associated with photosystem II, which the onset of starvation showed a greater reduction in chlorophyll biosynthetic proteins. At the onset of starvation, structural ATP synthase proteins were significantly downregulated, but appear to stabilize in abundance under long-term nutrient limitation conditions. Phosphorus and nitrogen stress specific response were most notable in the accessory proteins associated with acquisition of these elements.

Phosphate related

Loci PstS, PhoA, and the organic phosphorus porin PMM0709 previously observed to be upregulated under P-starvation in MED4, both in gene expression (Reistetter *et al.*, 2013; Martiny *et al.*, 2006) and in a proteomics set (Fuszard *et al.*, 2010), were also significantly

upregulated under both P-starvation and long term P-limitation in our study (Figure 3). Notably, the PstS high affinity uptake phosphate binding transport protein is significantly downregulated under both N-starvation and long-term N-limitation when compared with nutrient replete conditions. Protein PMM0719, significantly upregulated under P-starvation only, is an uncharacterized protein encoded by a gene located in the P-stress inducible genomic island that ranges from PMM0705-PMM0725. Uncharacterized proteins PMM1409, PMM1414, and PMM1416 are located in the P-stress inducible genomic island from PMM1403-PMM1416 (Coleman *et al.*, 2006; Martiny *et al.*, 2006). PMM1416 is one of the most highly differentially expressed proteins under both P-stress conditions at a log2 fold change > 4.9 under P-limitation and 5.2 under P-starvation. PMM1414 was upregulated under both P-stress conditions (log2 fold change > 3 for both). PMM1409 was only upregulated under P-starvation conditions (log2 fold change > 4).

Nitrogen related

Only two nitrogen-specific loci were identified as significantly differentially expressed under nutrient stress conditions. The cyanate transporter CynA was significantly upregulated at a log2 fold change of 2.7 greater than the replete conditions. The urea transporter UrtA was not significantly differentially expressed in either N-stress condition or under P-starvation, but was significantly downregulated under P-starvation at a log2 fold change of -4.3.

Transport & binding proteins

Additional transport and binding proteins, aside from those previously associated with P- or N- uptake, were significantly differentially expressed. The iron acquisition protein IdiA was significantly upregulated under both N-starvation and N-limitation when compared to nutrient

replete conditions (log2 fold change > 2). The probable ferritin, PMM0804, was significantly upregulated under long term P-limitation, but at a lower level where the log2 fold change is only ~ 0.5 when compared to replete cells. The transport protein PMM0214, annotated as a likely sulfate transmembrane transporter, was significantly upregulated under both N-stress conditions, with the probable magnesium transport protein MgtE upregulated under N-starvation only.

Energy metabolism

An indicator specific to nutrient starvation appears to be the reduction in ATP-synthase cellular machinery. Loci associated with ATP synthase, AtpACDH, are significantly downregulated under both N- and P-starvation, but they are not significantly downregulated under long term nutrient limitation. This reduction in ATP synthase proteins is in line with a proteomic analysis of *Synechocystis* sp. PCC6803 which found a similar reduction under P-starvation (Gan, 2006); however, a previous proteomic analysis of MED4 after multiple days of phosphorus starvation by Fuszard *et al.* (2010) did not find any differential change in these proteins under P-stress. It should be noted that Fuszard *et al.,* (2010) sampled for the P-starvation proteomic profile after the cells had been in stationary phase for a longer period of time; it is possible that this differential expression of ATP synthase proteins was not observed as it may be a hallmark of the onset of starvation with stabilization of levels under long-term nutrient stress.

The Large subunit of RuBisCo (RbcL) was identified as significantly downregulated in response to long term P-limitation and under both N-stress conditions which is in line with Tolonen *et al.* (2006), but not under P-starvation condition which somewhat conflicts with the results of Fuszard *et al.,* (2010) as they identified significantly increased expression of RbcL under P-starvation. It is possible that we did not see a significant increase in RbcL as we sampled

earlier into the onset of stationary phase (Figure 1). Curiously, the small subunit of RuBisCo, RbcS, is significantly upregulated at a high level (log2 fold change > 3) under both N-stress conditions.

While Fuszard *et al.,* (2010) did not identify any proteins associated with the pentose phosphate pathway (Calvin) cycle as upregulated, we observed Gnd, RpiA, and Tal as upregulated under both P-stress conditions and N-starvation, P-limitation and both N-stress conditions, and P-starvation conditions, respectively. GlpX was identified as downregulated under P-starvation conditions, whereas Fuszard *et al.,* (2010) observed no significant change. We also identified a significant reduction in the protein CsoS1 and CsoS2 associated with the carbon concentrating mechanism in response to P-limitation similar to Fuszard *et al.* (2010), as well as under N-stress conditions. In general, we identify a reduction in carbon fixation in response to P-stress but with a higher resolution and more complex variability in the expression pattern than previously reported for MED4 (Fuszard *et al.*, 2010).

Photosynthesis

Cellular response to nutrient stress is generally to downregulate proteins associated with photosynthesis. Proteins associated with pigment production, like the chlorophyll biosynthetic protein HemCL and ChlLP, are generally downregulated, especially under nutrient starvation conditions. Many of the proteins associated with the oxygen evolving Photosystem II are significantly downregulated under long term nutrient limitation (P: Psb28, PsbABCDO; N:PsbCDO). This reduction in the oxygen evolving complex PSII may further be reflected by the lack of expression change in the peroxidase Tpx (PMM0856) which is in line with a previous proteomic profile of MED4 under P-starvation (Fuszard *et al.*, 2010), however, we found that the

antioxidant TrxA is significantly upregulated under both P-stress conditions whereas Fuszard *et al.* (2010) found no significant change.

Transcription

All proteins under nutrient stress conditions that are associated with transcription show either no change or a significant reduction in expression when compared with replete cells. Nutrient starvation has the strongest response with the greatest reduction in RNA polymerase associated structural proteins like RpoABC1C2. This shows a reduction in the cellular production which is further reflected by the reduction of ribosomal structural proteins like RpsA1A2CGM and RplAKN associated with the large and small subunits of the ribosome, respectively. The cellular response to nutrient stress is to reduce the production of new proteins, which is reflected by the reduction in growth rate of the cultures.

Cellular Processes

There is no clear related pattern in the differential expression of DNA replication loci (DnaAN and GyrAB) among either Nutrient-specific response or physiological state response. The same is generally true for cell division related proteins HimA and MinD, aside from an upregulation of the cell division protein FtsH2 under N-stress, and FtsH3 under N-limitation and downregulation of FtsH3 under P-limitation. A methyltransferase of unknown function, PMM1232, appears to be associated with a P-specific response as it is significantly downregulated under both P-limitation & P-starvation. The putative nickel-containing superoxide dismutase SodA is significantly downregulated under N-starvation conditions only.

Amino acid synthesis

Many loci associated with amino acid synthesis were differentially expressed under nutrient stressed conditions when compared to the profile of nutrient replete exponentially growing cells. Multiple amino acid synthesis proteins were upregulated in response to stress, which is contrary to an expectation of a reduction in synthesis of N-rich biomolecules when cells are N-stressed. The glycine synthesis protein GlyA is significantly upregulated under all nutrient stressed conditions.

Protein synthesis

In general, most of the differentially expressed proteins associated with the ribosome are downregulated in response to nutrient stress. Exceptions to this are the ribosomal structural proteins RplL and RpsE which are upregulated under long term N-limitation and show no significant change under the other nutrient stress conditions. In addition, all the translation factors that were found to be significantly differentially expressed (FusA, Tsf, and TufA) are all downregulated under nutrient stress. Conversely, all the proteins significantly differentially expressed relating to tRNA aminoacylation (AspS, ProS, MetG) are all upregulated under nutrient stress conditions.

Protein fate

Loci associated with protein fate, including protein folding chaperones, degradating proteases, and protein localization proteins are differentially expressed among the various nutrient stress conditions. One general observation is that the Clp degradation pathway appears to be downregulated under nitrogen starvation conditions (ClpC, P3, P4) while this pathway does not appear to be impacted under long term N-limitation. There is also an upregulation of some of these loci under P-stress, specifically P-starvation (ClpP2,X) and P-limitation (ClpP4).

Generally, the significantly differentially expressed proteins associated with protein folding and stabilization (DnaK2, GroEL1, GroEL2, GrpE, HtpG, PMM1293, and Tig) are downregulated under nutrient-stress especially N-starvation, aside from GroES and PMM0894 which are upregulated under P-starvation and P-limitation, respectively.

Metabolic intermediates biosynthesis

Many of the proteins associated with metabolic intermediates biosynthesis which are identified as significantly differentially expressed under nutrient stress conditions when compared to a nutrient replete state are upregulated under nutrient stress. However, the cobalamin synthesis protein PMM1033 is downregulated under P-stress conditions and the protein AhcY which is involved in the regeneration of the *S*-adenosylmethionine methyl donor compound is significantly downregulated under N-starvation. The loci PMM0491 is upregulated with a high log2 fold change (3.68) in only N-starvation conditions and may be specific to nitrogen stress response.

Fatty acid and lipid metabolism

All the loci identified and significantly differentially expressed for fatty acid and lipid metabolism are significantly down-regulated under nutrient stress conditions. Notably, this includes the sulfolipid synthesis related protein SqdB which is significantly downregulated under P-starvation and P-limitation as well as N-starvation.

Signal transduction

We did not identify many two-component signal transduction pathways as significantly differentially expressed under nutrient stress when compared to nutrient replete cells. Absent

from the list of significantly differentially expressed proteins are the loci associated with the phosphate two component signaling PhoBR system. Also absent, is the nitrogen response regulator NtcA. PhoB was identified as present in all nutrient stress conditions and PhoR was identified as present in P-starvation and both N-stress conditions while PhoBR was absent from nutrient replete conditions. However, it should be noted that while these loci were present under stress conditions they were not expressed at levels deemed as significantly differentially present (Supplementary Table 1). NtcA was not identified at all in nutrient replete conditions, and was only detected in one technical proteomics replicate under N-starvation. The signaling pathway protein PMM0169 appears to be a general nutrient-stress inducible protein as it is significantly upregulated under all nutrient stress conditions at a log2 fold change greater than 2 for all stress states.

Biogeographic analysis

The loci PMM1409, PMM1414, and PMM1416 are some of the most significantly upregulated proteins under P-stress conditions, but their function is unknown. They are located on a P-stress inducible genomic island, with PMM1416 showing genomic linkage to a *phoA Prochlorococcus* genomic fragment from a metagenome in the Sargasso Sea (Martiny *et al.*, 2006) and has been shown to respond to both P-starvation and light stress culture conditions (Coleman *et al.*, 2006). PMM1416 is a conserved hypothetical protein 694 amino acids in length with no identifiable domains (no cluster of orthologous groups (COG), protein family (pFam), enzyme commission (EC), or InterPro (IPR) identification. PMM1409 and PMM1414 are both lacking in identifiable domains, but do commonly share the InterPro domain 21734 for the domain of unknown function DUF3303. These two proteins are smaller at 106 and 112 amino acids in length for PMM1409 and PMM1414, respectively. These uncharacterized loci appear to

play a role in phosphorus stress response in MED4, as they are so highly upregulated under P-stress conditions.

In order to better characterize these highly expressed proteins of unknown function, these loci were targeted for a biogeographic analysis in surface water (~ 5m depth) microbial metagenomes which capture *Prochlorococcus* populations from the Global Ocean Sampling (GOS) (Rusch *et al.*, 2007) survey and the TARA Oceans metagenomes (Sunagawa *et al.*, 2015). These loci were assessed to see if they show the same presence/absence pattern in environmental populations of *Prochlorococcus* as characterized P-acquisition genes, where P-stress inducible genes are present at higher frequencies in populations which experience extreme P-scarcity (Martiny *et al.*, 2009; Saunders and Rocap, 2016). *Prochlorococcus*-linked metagenomic reads were identified using a phylogenetically informed placement approach (Saunders and Rocap, 2016; Berger *et al.*, 2011; Berger and Stamatakis, 2011). The metagenomes span various levels of P availability (Figure 5), with the Atlantic (GOS) and Mediterranean (TARA) locations have the lowest annual average phosphate concentrations. The Red Sea (TARA) has a slightly higher average (>0.1 µmol/L) with the Indian sites (GOS) slightly higher than the Red Sea. The Indian Ocean TARA locations and the Pacific (GOS) locations have the highest and most variable annual average phosphate concentrations.

Upon construction of the phylogenetic tree for PMM1414, loci PMM1414, PMM1409, and PMM0364 were identified as incredibly closely related and present on the same tree (Supplementary Figure 1). We therefore analyzed sequences relating to all three loci, even though they display different patterns in response to nutrient stress with PMM1409 upregulated under P-starvation only, PMM1414 unregulated under both P-stress states, and PMM0364 displaying no differential expression in response to P- or N- stress. These three loci appear to be

Prochlorococcus specific paralogs, being more closely related to each other than to sequences in other taxa, including *Synechococcus*. These tree loci appear to be unique to the genus *Prochlorococcus* with an ancestral origin in a Picocyanobacterial precursor gene.

Prochlorococcus reads associated with PMM1414 appear in every location analyzed (Figures 6 & 7). However, the relative abundance of these reads differs greatly from location to location. The relative abundance of PMM1414 is greatest in the Atlantic (GOS) and Mediterranean (TARA) locations where they are closest to the value of 1, where 1 represents one copy of PMM1414 per *Prochlorococcus* genome sampled in the metagenomes. Reads associated with the P-starvation loci PMM1409 appear in the Atlantic (GOS) and Mediterranean (TARA) locations (Figures 6 & 7), and are virtually absent everywhere else with the exception of a few outliers in the Indian Ocean locations in the TARA dataset. The locus PMM0364 - which is closely related to PMM1409 & PMM1414, but displays no response to nutrient stress - was identified in a portion of the *Prochlorococcus* populations in every location, but at varying relative abundance (Figures 6 & 7) indicating that it is likely not under the same selective influence of environmental phosphorus conditions.

The phylogenetic tree for PMM1416 shows that representative strains of both *Prochlorococcus* and *Synechococcus* maintain a copy of this uncharacterized gene, unlike PMM1409 and PMM1414 which appear to be specific to *Prochlorococcus* (Supplementary Figure 2). Intriguingly, this gene is also found in Alphaproteobacteria and eukaryotic phytoplankton like *Emiliana huxleyi* and *Ostreococcus*, suggesting this gene has likely been horizontally transferred. *Prochlorococcus*-linked PMM1416 reads were of greatest relative abundance in the Mediterranean (TARA) and Atlantic (GOS) locations (Figures 6 & 7), but

maintained by only a small portion of the populations there. PMM1416 reads were virtually absent from the *Prochlorococcus* populations in all other locations.

1.4 Discussion:

Culturing *Prochlorococcus* MED4 under batch and chemostat conditions has enabled a thorough study of the cellular response to not only nutrient replete conditions and a rapid onset of nutrient starvation conditions, but also under long term nutrient limitation which better simulates the environmental state than rapidly shifting boom and bust periods in oligotrophic waters. The addition of chemostats has enabled the study of cellular response at three physiological growth conditions: exponential growth, limited growth, and stationary phase. To further distinguish a phosphorus-stress specific response from a general nutrient stress response, culturing experiments were conducted under nitrogen limiting conditions as well. Previous work on these same culturing experiments focused on a detailed analysis of phosphorus uptake kinetics and cellular P quota change under nutrient stress (Krumhardt *et al.*, 2013) as well as a gene expression analysis of specific lipid and P-stress related genes (Reistetter *et al.*, 2013). The proteomics analysis presented here enabled a holistic approach where all possible proteins produced by MED4 had the opportunity to be observed, as opposed to the targeted selection of a few specific genes.

An overall response to nutrient stress was observed in the production ribosomal and RNA polymerase structural proteins. A reduction of these proteins was observed under nutrient limitation conditions, with an even greater reduction in these structural proteins observed under nutrient starvation conditions. The reduction in these transcription and translation structural proteins followed the reduction in growth rates from 0.2 day^{-1} to 0 day^{-1} further supporting the

prior studies conducted on these experiments reflecting the distinct physiological differences between cells which are experiencing exponential growth, limited growth, and those which are in stationary phase. These physiologically distinct states were further reflected in the reduced production of photosystem II proteins under long-term nutrient limitation, and the greater reduction in both chlorophyll biosynthetic proteins and ATP synthase structural proteins at the early stages of nutrient starvation.

This holistic proteomics analysis further reinforces what had already been observed about some loci in response to P-stress, like the increased expression under P-stress of the alkaline phosphatase PhoA, the P-org porin PMM0709, and the high affinity P-transport protein PstS (Reistetter *et al.*, 2013; Martiny *et al.*, 2006; Fuszard *et al.*, 2010). However, it also revealed some surprises and complications – like the significantly lower expression of PstS under N-stress. PstS has been previously used as an indicator of P-stress in the field (Fuller *et al.*, 2005). We have also previously suggested PstS as a potential biomarker of P-stress as it is present in all picocyanobacterial genomes (Scanlan *et al.*, 2009) and is found in global *Prochlorococcus* populations (Martiny *et al.*, 2009), but we also noted multiple drawbacks for its use as a marker, including that multiple copies are found in some picocyanobacterial genomes and it is found in cyanophage (Sullivan *et al.*, 2005; Zeng and Chisholm, 2012). The expression pattern we identify here, where PstS is significantly upregulated under P-stress and significantly downregulated under N-stress, adds a further note of caution for targeted use of this locus as a biomarker in the field. For example, when transecting from N-stressed waters to P-stressed waters, one might see an inflated signal when targeting PstS alone as there is the potential to move from populations which are down-regulating PstS to populations which are upregulating PstS. A notable discrepancy between our previous targeted gene expression studies on the same

cultures is that of the locus PhoR. In the previous gene expression study, *phoR* was shown to be one of the best targets for indication of P-limitation (Reistetter *et al.*, 2013), however in the proteomic profiles we did not identify a peptide for PhoR in any spectral run under any condition. This discrepancy in the identification of PhoR highlights some of the differences between gene expression patterns and protein expression.

The proteomics profile provides further support for the increased expression of the loci of unknown function PMM0719, PMM1409, PMM1414, and PMM1416 which are found in two different genomic islands (Coleman *et al.*, 2006) containing a suite of genes which have previously been shown to increase expression in response to P-starvation, with PMM1416 showing increased expression under both P-stress and light-stress (Coleman *et al.*, 2006). PMM1416 is the most significantly upregulated protein under P-stress conditions. P-uptake experiments on these cultures indicated that there is a likely unknown enzyme, in addition to PhoA, which is involved in the cleavage of orthophosphate form organic-P sources in MED4 (Krumhardt *et al.*, 2013). A metagenomic analysis of *Prochlorococcus* in the Sargasso Sea found a genomic linkage between the locus PMM1416 and PhoA (Martiny *et al.*, 2006). Given the expression pattern of PMM1416 and its presence in populations in relation to environmental phosphate concentrations, it is possible that this is a likely P-cleavage enzyme.

The inclusion of chemostat cultures in this analysis enabled us to identify a fine-scale response by PMM0719 and PMM1409 where increased expression is specifically during the onset of P-starvation, and not under long-term P-limitation conditions. These loci of unknown function were not targeted in our previous gene expression analysis, but by using this holistic proteomics approach these loci were identified as having some of the most dramatic differential expression in response specifically to P-stress. Due to the large differential expression of the loci

PMM1409, PMM1414, and PMM1416 and the P-starvation specific response by PMM1409, we further identified the presence of these genes in environmental metagenomes capturing global *Prochlorococcus* populations in order to gain further insight into the role of these P-stress loci of unknown function.

In addition, previous field analyses have targeted the nitrogen transcriptional regulator *ntcA*, which is negatively controlled by the presence of ammonium (Lindell *et al.*, 1998; Tolonen *et al.*, 2006; Saito *et al.*, 2014), showing differential expression in *Synechococcus* under N-stress specifically (Lindell and Post, 2001) and increased abundance of *Prochlorococcus*-linked NtcA proteins in environmental samples from nitrogen-scarce regions (Saito *et al.*, 2014). However, we did not identify significant differential expression of NtcA (PMM0246) and only identified two NtcA peptides in one spectral run of an N-starved replicate. In general, this holistic proteomics approach appears to be biased against transcriptional regulators which may not occur at high abundances relative to other proteins in the cell and are therefore missed in this analysis and may be better captured by gene expression. The cyanate transporter, CynA, which is believed to be under control of the global nitrogen regulator NtcA (Tolonen *et al.*, 2006) responded in an expected manner, where N-starved cells showed a significant upregulation of the organic-N transporter. The urea uptake transporter UrtA, which was also predicted to be under control of the NtcA regulator (Tolonen *et al.*, 2006) displayed an interesting pattern of expression, contrary to what would have been expected. If UrtA was under control of the NtcA regulator, an expression pattern similar to CynA would be expected with expected upregulation of UrtA under N-stress conditions. However, UrtA protein expression was more stable across conditions, showing no significant difference between nutrient replete conditions and N-stress conditions, with a significant reduction observed under long-term P-limitation. In environmental

Prochlorococcus populations the UrtA proteins were identified as being more abundant in regions associated with nitrogen scarcity (Saito *et al.*, 2014). However, our culturing of axenic MED4 indicates that expression of the urea transporter *urtA* is probably not under control of the ammonium-induced repression of the nitrogen regulator NtcA. Once again, this may be an over simplified analysis in the response of this nitrogen transporter due to growth on a simple N-source. Intriguingly, the iron transport protein IdiA was significantly upregulated under both N-stress conditions although the cultures were not grown under iron scarce conditions. This expression of an iron uptake protein under N-stress may hint at a deep hardwired cellular response to the often co-limiting nitrogen and iron resources in the environment.

Biogeographic Analysis:

Prochlorococcus environmental populations have been previously shown to maintain a higher relative abundance of P-acquisition genes(Martiny *et al.*, 2009) and P-stress related genes(Saunders and Rocap, 2016) in regions associated with phosphorus scarcity, like the Sargasso Sea. PMM1416 is only maintained by a small portion of the *Prochlorococcus* populations captured in the Atlantic (GOS) and Mediterranean (TARA) metagenomes, and is virtually absent from the metagenomes in all other locations analyzed. PMM1414 appears in all of the global *Prochlorococcus* populations with only the Atlantic (GOS) and the Mediterranean (TARA) show relative abundances around the value of 1 copy per *Prochlorococcus* genome. Intriguingly, while the Red Sea is one of the locations with some of the lowest phosphate concentrations, these populations have a generally lower relative abundance of PMM1414 as well as all the other loci assessed. It is possible that these *Prochlorococcus* populations do not show a similar relative abundance of these loci as the Atlantic (GOS) and Mediterranean (TARA) locations partly due to the relatively higher phosphate concentrations, but this may also

be due to population biogeographic separation as the Red Sea is rather isolated, experiencing generally a general unidirectional flow of water, and the biota it contains, from the Red Sea into the Mediterranean through the Suez Canal(Por, 2012) and seasonal mixing events constricted through the strait of Bab el Mandeb which connects the Gulf of Aden with the Indian Ocean(Sofianos and Johns, 2002).

Not only do these loci display this P-specificity in physiology through protein expression, but the environmental selection forces exerted by P-stress appear to have shaped the occurrence of PMM1409, PMM1414, and PMM1416 in global *Prochlorococcus* populations. As global temperatures continue to rise, vertical stratification of the water column will continue to intensify resulting in greater nutrient scarcity (Behrenfeld *et al.*, 2006) with continued geographic expansion of the most low nutrient waters (Polovina *et al.*, 2008). The accessory genes associated with nutrient acquisition, like *phoA*, *pstS*, porin PMM0709, PMM1409, PMM1414, and PMM1416 will likely confer advantages to the *Prochlorococcus* which maintain them. It will be interesting to see if *Prochlorococcus* population level gene frequencies for the P-stress accessory genes increase in the most P-scarce regions of the world in response to rising temperatures which would show an evolutionary response by a major global primary producer to climate change.

1.5 Summary:

The proteomic profiles of axenic *Prochlorococcus* MED4 under nutrient replete conditions, long term nutrient limitation, and starvation conditions show a varied and complex response in cellular protein production under both P-stress and N-stress. In general, the cellular response to nutrient stress is to reduce transcription and translation of proteins, which is further

reflected by the reduction in growth rate. Long term nutrient limitation shows a greater reduction in the photosynthetic apparatus, specifically PSII, with a general reduction in proteins associated with chlorophyll production under nutrient starvation conditions. The phosphorus acquisition proteins were upregulated, further supporting previous studies showing upregulation of these loci under P-stress conditions (Reistetter *et al.*, 2013; Coleman *et al.*, 2006; Martiny *et al.*, 2006; Fuszard *et al.*, 2010). The nitrogen stress response exhibited by the cells was not directly in line with previously reported studies, especially with regard to NtcA and UrtA(Tolonen *et al.*, 2006; Saito *et al.*, 2014), however this may be a result of culturing on a simple ammonium nitrogen source. Specifically, a new detail was identified with regard to PstS expression being significantly upregulated under P-stress and significantly downregulated under N-stress, as this was the first study of MED4 to look at both P- and N- stress in tandem. We identified loci of unknown function which are dramatically upregulated under P-stress, with PMM1414 and PMM1416 upregulated under both P-stress conditions and PMM1409 upregulated under P-starvation. A biogeographic analysis of these loci indicates that their presence is selected for in global *Prochlorococcus* populations experiencing the most extreme phosphorus scarcity. The physiological upregulation of P-acquisition accessory proteins under P-stress and the variable presence of the phosphorus accessory genes in global metagenomes highlights the powerful influence phosphorus availability has on shaping the picocyanobacterium *Prochlorococcus*.

1.6 Methods:

Culturing conditions

Batch and chemostat cultures were conducted as described by Krumhardt *et al.* (2013). Cultures were grown on modified AMP1C artificial seawater media with P-stress nutrient

concentrations of NaH_2PO_4 at 1.15 µM and $(NH_4)SO_4$ at 400 µM, and N-stress nutrient concentrations of NaH_2PO_4 at 10 µM $(NH_4)SO_4$ at 25 µM to produce N:P ratios of 348:1 and 5:1 for P-stress and N-stress, respectively. Batch cultures of axenic *Prochlorococcus* MED4 were grown in duplicate under phosphorus limiting or nitrogen limiting conditions. The P batch cultures grew exponentially until day 10 when they entered stationary phase; N batch cultures grew exponentially until day 14 when they entered stationary phase (Fig. 1). Chemostat cultures were acclimated for a minimum of 30 days, with cultures performed in duplicate at a dilution rate controlling the MED4 growth rate at a stable 0.2 day^{-1}. Growth of cultures was monitored by direct cell counts through flow cytometry. Between 35-50 ml of culture was filtered onto 0.2 µm polycarbonate filters and frozen at -80°C, collecting between ~$2x10^8$-$4.5x10^9$ cells, for proteomics analysis during exponential growth on day 7 for P cultures, during stationary phase on days 11 and 16 for P and N cultures, respectively, with chemostat cultures sampled after 30 days of acclimation. Replete samples were extracted in the exponential phase of the batch cultures, thus phosphorus replete is an adequate proxy for the nitrogen replete condition and represents cells cultured under general replete conditions.

Mass spectrometry

Cells were removed from filters by adding 1.2 ml of Tris and bead beating without beads. Cells were physically disrupted using a French press and centrifuged on a tabled top centrifuge at 39,000 g for one hour. Whole cells crude fraction was desalted, prepped, and trypsin digested in a manner similar to method previously reported (Morris *et al.*, 2010). About 200 ng of each sample was introduced in triplicate for the LC-MS/MS analysis. Mass spectrometry was performed on an LTQ-Orbitrap hybrid mass spectrometer (Thermo Fisher, San Jose, CA, USA). Data-dependent scans were completed by precursor ion selection in the fourier transform (FT)-

based analyzer (Orbitrap) followed by collision induced dissociation (CID) in the linear ion trap (LTQ) as described previously (Morris *et al.*, 2010).

Protein identification

The spectral results were compared across all potential tryptic peptides possible from the MED4 genome and interpreted using a local copy of SEQUEST (Eng *et al.*, 1994). Minimum threshold for identity were set to peptide matched to CID spectra of at least 95%. Full runs of spectral counts and unique peptides for all nitrogen and phosphorus conditions were acquired resulting in two biological replicates and three analytical replicates for each nutrient condition. The spectral counts were normalized first by dividing by the protein length to provide a spectral abundance factor (SAF). They were then divided by the sum of all SAF values in the corresponding technical run to produce normalized spectral abundance factors (NSAF) (Pavelka *et al.*, 2008). This normalization process accounts for the fact that larger proteins contribute more peptides and allows protein abundances to be compared across different runs.

Before conducting significance analyses, a stringent cut off was applied to each nutrient condition individually to insure false peptide detections were removed and only the most differently produced proteins were considered. Spectral counts were only evaluated for differential expression if there was at least one unique peptide in each technical replicate with one technical replicate having a minimum of two unique peptides.

Cluster analysis of replicates

A dendrogram was plotted to examine the hierarchical relationship between the proteomic profiles of all runs. To do this NSAF values for all runs, with the stringent cut off of unique peptide presence in all technical replicates with one technical replicate having at least two

unique peptides, were used. Using the statistical package R, a Euclidean distance matrix was calculated to compare the dissimilarity between the runs (Team, 2014). The data was hierarchically clustered, using the complete linkage method, a type of agglomerative hierarchical clustering (Maimon and Rokach, 2005).

Proteomic profile comparison

To determine proteins that were statistically significantly differentially expressed between biological replicates and nutrient conditions, the Power Law Global Error Model (PLGEM) and QSpec were used. PLGEM (version 1.36.0) is a package for R (version 3.1.1) used to determine differentially expressed genes or proteins in microarray or proteomics data (Team, 2014; Pavelka *et al.*, 2004). The model was run in a step-by-step mode for the phosphorus (P-replete, P-starved, P-limited) and nitrogen (P-replete, N-starved, N-limited) data. For each data set, the model was fit to the best fitting condition: replete for the phosphorus data and starved for the nitrogen data. All comparisons were run at a p-value of 0.01 to determine significantly differently produced proteins. The same parameters were used to compare the biological replicates for each nutrient stress condition individually. In addition to PLGEM, QSpec (version 1.2.2), a software to analyze label-free quantification data, was utilized to calculate false discover rates (FDRs) of peptides (Choi *et al.*, 2008). Proteins were only considered significantly differently produced if they were considered significant by PLGEM, had a FDR less than 0.90, and were not significantly differentially expressed between the biological replicates. For all proteins identified as significantly differentially expressed, the average log2 fold change of the NSAF values for nutrient stress conditions were calculated as compared to the NSAF values of the nutrient replete condition.

Heat maps

Heat maps were made to compare the expression frequency of the significant genes across all four conditions compared with P-replete, which acts as a baseline of exponentially growing replete cells. With this common baseline, comparisons across all four nutrient stress conditions can be made. Heat maps were produced in R using heatmap2 in the gplots package (version 2.17.0) (Team, 2014; Warnes *et al.*, 2015). Along the vertical axis, the 167 significant proteins are partitioned by cell functional category, and are then clustered using a Euclidean distance matrix and agglomerative hierarchical clustering using the complete linkage method (Maimon and Rokach, 2005). Clustering is based on the protein expression, specifically the log2 fold change where fold change is the ratio of average NSAF values for each condition compared to P-replete. The color gradient visually representing the log2 fold change was produced using the RColorBrewer package (version 1.1-2) (Neuwirth, 2014). General categorical listings pulled from the annotations of the MED4 proteome were pulled from the UniProt database (www.uniprot.org) with categorical groupings loosely based off of TIGR Fam annotations.

Phylogenetic trees

Reference sequences of single copy core housekeeping genes *glnA, rpsD,* and *tyrS* were gathered by searching a list of reference organisms (including cyanobacteria and common marine bacteria identified in the GOS metagenomes (Rusch *et al.*, 2007)) for genes associated with the relevant Clusters of Orthologous Groups (COGs) (Tatusov *et al.*, 2000) identification using the online tool www.MicrobesOnline.org (Dehal *et al.*, 2009). The ortholog gene tree tool from MicrobesOnline was also used to gain additional phylogenetic resolution on the tree through the collection of a broader range of taxonomic sequences. The amino acid sequences of identified

genes were downloaded locally for further analysis. The proteins PMM1409, PMM1414, and PMM1416 do not have identified COGs. For these sequences, we used the ortholog gene tree on MicrobesOnline to gather sequences for phylogenetic analysis. For PMM1409/PMM1414, and PMM1416 we also conducted a blastp (Altschul *et al.*, 1990; Altschul *et al.*, 1997) search of the NCBI nr database using the MED4 sequence as a query and added additional amino acid sequences from the results to enhance the phylogenetic resolution on the tree.

The amino acid sequences were aligned using the alignment program Muscle v. 3.8.31 (Edgar, 2004). A maximum likelihood tree was inferred from the best of 20 starting trees using the phylogenetic tree program RAxML v. 8.2.8 (Stamatakis, 2006; Stamatakis, 2014) using the amino acid substitution model WAG for all trees except for PMM1416 where model VT was used. Empirical character frequencies and a gamma model of rate heterogeneity with an estimated alpha value were used for construction of all trees. Bootstrap analyses were conducted on the PMM1409/PMM1419 and PMM1416 trees with an n=100 maximum support.

Placement of metagenomics reads

Metagenomes from the GOS Survey, specifically those collected from the 0.1- 0.8 µm filter size fraction collected at ~5m depth, were downloaded locally on 3 of February 2011 from the CAMERA web portal (Sun *et al.*, 2011). Metagenomes from the TARA Oceans dataset, using the filter size fraction 0.22-1.6 µm at about 5m depth, were downloaded locally in May 2016 from the EMBL-EBI Metagenomics repository (https://www.ebi.ac.uk/metagenomics/). For TARA Oceans metagenomes, the processed nucleotide reads were downloaded and used for analysis (Mitchell *et al.*, 2015). In order to be considered for complete biogeographic analysis each metagenome sample was required to contain at least one *Prochlorococcus*-linked read for

each of the single copy core housekeeping genes so as to designate metagenomes capturing *Prochlorococcus* populations, with 34 GOS metagenomes and 12 TARA metagenomes analyzed.

For each gene (*glnA, rpsD, tyrS, PMM0364, PMM1409, PMM1414, PMM1416*), metagenomics reads likely to be *Prochlorococcus*-linked were recruited by conducting a tblastn search of the metagenomes using representative *Prochlorococcus* and *Synechococcus* sequences as the query (Altschul *et al.*, 1990; Altschul *et al.*, 1997). Sequences recruited with e-values of $\leq$ 1 and $\leq$ 20 from GOS and TARA metagenomes, respectively, were used for further phylogenetic placement analysis. Sequence reads were trimmed according to their best hit with a query sequence in order to ensure translation in the proper reading frame and to avoid overlap at the end of the target gene sequence in gene-specific alignments. Trimmed sequences $\geq$ 120 nucleotides were translated into amino acid space and aligned to the reference alignment used to construct the protein trees using the program PaPaRa (Berger and Stamatakis, 2011). Metagenomic reads from the PaPaRa alignment were then assigned to nodes on the reference phylogenetic tress using EPA: Evolutionary Placement Algorithm (Berger *et al.*, 2011; Stamatakis, 2014). All reads which were placed with appropriate *Prochlorococcus* nodes were deemed as *Prochlorococcus*-linked metagenomics reads.

For each gene, the number of metagenomic reads deemed as *Prochlorococcus*-linked were counted per metagenome sample/run. This number was then length normalized according to the length of the MED4 query gene to help with sampling bias for longer gene sequences. In order to calculate the relative frequency of target biomarkers in environmental field populations, we calculated the average of the length normalized abundance of the single copy core housekeeping genes as a baseline for the relative number of *Prochlorococcus* genomes sampled in each metagenome sample/run.

Nutrient concentration

Annual average surface phosphate concentrations were collected for GOS sites from the World Ocean Atlas, 2009 (Garcia *et al.*, 2010) as described previously (Saunders and Rocap, 2016). Annual average surface phosphate concentrations were collected from the World Ocean Atlas, 2013 (Garcia *et al.*, 2014) in a 1 degree x 1 degree grid around the sampling locations for the TARA metagenomes.

1.7 Acknowledgements:

I would like to thank Gabrielle Rocap for her guidance on these analyses and editorial feedback and Claire Knox who had a major role in the statistical analysis of this data. Original design of the experiment is credited to Gabrielle Rocap and Lisa Moore. All culturing work was conducted by Lisa Moore and her lab group at the University of Southern Maine, including Kristen Krumhardt and others. Brook Nunn from the University of Washington conducted the spectral analyses for proteomics data. This work was made possible through support from the National Science Foundation OCE-0453029, OCE-0723866, and OCE-1138368 grants to Gabrielle Rocap, a Mary Gates Scholarship to Claire Knox, and a NASA Earth and Space Sciences Fellowship and National Science Foundation Graduate Research Fellowship to Jaclyn K. Saunders.

1.8 References:

1 Polovina JJ, Howell EA, Abecassis M (2008). Ocean's least productive waters are expanding. *Geophysical Research Letters* **35**.

2 Campbell L, Nolla HA (1994). The importance of *Prochlorococcus* to community structure in the central north pacific ocean. *Limnology and Oceanography* **39:** 954-961.

3 Campbell L, Liu H, Nolla HA, Vaulot D (1997). Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991–1994 ENSO event. *Deep Sea Research Part I: Oceanographic Research Papers* **44:** 167-192.

4 DuRand MD, Olson RJ, Chisholm SW (2001). Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **48:** 1983-2003.

5 Scanlan D (2012). Marine Picocyanobacteria. In: Whitton BA (ed). *Ecology of Cyanobacteria II*. Springer Netherlands. pp 503-533.

6 Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311:** 1737-1740.

7 Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N *et al.* (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proceedings of the National Academy of Sciences of the United States of America* **110:** 9824-9829.

8 Partensky F, Garczarek L (2010). *Prochlorococcus*: advantages and limits of minimalism. *Annual review of marine science* **2:** 305-331.

9 Ting CS, Hsieh C, Sundararaman S, Mannella C, Marko M (2007). Cryo-Electron Tomography Reveals the Comparative Three-Dimensional Architecture of *Prochlorococcus*, a Globally Important Marine Cyanobacterium. *Journal of Bacteriology* **189:** 4485-4493.

10 Van Mooy BAS, Rocap G, Fredricks HF, Evans CT, Devol AH (2006). Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *Proceedings of the National Academy of Sciences* **103:** 8607-8612.

11 Bertilsson S, Berglund O, Karl DM, Chisholm SW (2003). Elemental composition of marine *Prochlorococcus* and *Synechococcus*: Implications for the ecological stoichiometry of the sea. *Limnology and Oceanography* **48:** 1721-1731.

12 Heldal M, Scanlan DJ, Norland S, Thingstad F, Mann NH (2003). Elemental composition of single cells of various strains of marine *Prochlorococcus* and *Synechococcus* using X-ray microanalysis. *Limnology and Oceanography* **48:** 1732-1743.

13 Giovannoni SJ (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science* **309:** 1242-1245.

14 Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:** e231.

15 Moore LR, Rocap G, Chisholm SW (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393:** 464-467.

16 Moore LR, Chisholm SW (1999). Photophysiology of the Marine Cyanobacterium *Prochlorococcus*: Ecotypic Differences among Cultured Isolates. *Limnology and Oceanography* **44:** 628-638.

17 Moore LR, Anton FP, Rocap G, Chisholm SW (2002). Utilization of Different Nitrogen Sources by the Marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnology and Oceanography* **47:** 989-996.

18 Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A *et al.* (2014). Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*. *Science* **344:** 416-420.

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

20 Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological Genomics of Marine Picocyanobacteria. *Microbiology and Molecular Biology Reviews* **73:** 249-299.

21 Moore LR, Ostrowski M, Scanlan DJ, Feren K, Sweetsir T (2005). Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. *Aquat Microb Ecol* **39:** 257-269.

22 Reistetter EN, Krumhardt K, Callnan K, Roache-Johnson K, Saunders JK, Moore LR *et al.* (2013). Effects of phosphorus starvation versus limitation on the marine cyanobacterium *Prochlorococcus* MED4 II: gene expression. *Environmental Microbiology* **15:** 2129-2143.

23 Ammerman JW, Azam F (1985). Bacterial 5-nucleotidase in aquatic ecosystems: a novel mechanism of phosphorus regeneration. *Science* **227:** 1338-1340.

24 Krumhardt KM, Callnan K, Roache-Johnson K, Swett T, Robinson D, Reistetter EN *et al.* (2013). Effects of phosphorus starvation versus limitation on the marine cyanobacterium *Prochlorococcus* MED4 I: uptake physiology. *Environmental Microbiology* **15:** 2114-2128.

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

26 Martiny AC, Coleman ML, Chisholm SW (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proceedings of the National Academy of Sciences* **103:** 12552-12557.

27 Fuszard MA, Wright PC, Biggs CA (2010). Cellular acclimation strategies of a minimal picocyanobacterium to phosphate stress. *FEMS Microbiol Lett* **306:** 127-134.

28 Fuszard MA, Wright PC, Biggs CA (2012). Comparative quantitative proteomics of *Prochlorococcus* ecotypes to a decrease in environmental phosphate concentrations. *Aquatic Biosystems* **8:** 7-7.

29 Karl DM (2007). The marine phosphorus cycle. ASM Press: Washington. pp 523-539.

30 Martiny AC, Huang Y, Li W (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environmental Microbiology* **11:** 1340-1347.

31 Coleman ML, Chisholm SW (2011). Phosphorus-related gene content is similar in *Prochlorococcus* populations from the North Pacific and North Atlantic Oceans Reply. *Proceedings of the National Academy of Sciences of the United States of America* **108:** E64-E66.

32 Behrenfeld MJ, O'Malley RT, Siegel DA, McClain CR, Sarmiento JL, Feldman GC *et al.* (2006). Climate-driven trends in contemporary ocean productivity. *Nature* **444:** 752-755.

33 Fuller NJ, West NJ, Marie D, Yallop M, Rivlin T, Post AF *et al.* (2005). Dynamics of community structure and phosphate status of picocyanobacterial populations in the Gulf of Aqaba, Red Sea. *Limnology and Oceanography* **50:** 363-375.

34 Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005). Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biology* **3:** 790-806.

35 Zeng Q, Chisholm SW (2012). Marine viruses exploit their host's two-component regulatory system in response to resource limitation. *Current biology : CB* **22:** 124-128.

36 Moore LR, Goericke R, Chisholm SW (1995). Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Marine Ecology Progress Series* **116**.

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

38 Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G *et al.* (2015). Structure and function of the global ocean microbiome. *Science* **348**.

39 Gan CS (2006). Response of Synechocystis sp. PCC 6803 to photoperiod and phosphate alterations using functional proteomics approaches, University of Sheffield.

40 Saunders JK, Rocap G (2016). Genomic potential for arsenic efflux and methylation varies among global *Prochlorococcus* populations. *ISME J* **10:** 197-209.

41 Berger SA, Krompass D, Stamatakis A (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology* **60:** 291-302.

42 Berger SA, Stamatakis A (2011). Aligning short reads to reference alignments and trees. *Bioinformatics* **27:** 2068-2075.

43 Lindell D, Padan E, Post AF (1998). Regulation of *ntcA* expression and nitrite uptake in the marine *Synechococcus* sp. strain WH 7803. *Journal of Bacteriology* **180:** 1878-1886.

44 Tolonen AC, Aach J, Lindell D, Johnson ZI, Rector T, Steen R *et al.* (2006). Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Molecular Systems Biology* **2:** 53.

45 Saito MA, McIlvin MR, Moran DM, Goepfert TJ, DiTullio GR, Post AF *et al.* (2014). Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **345:** 1173-1177.

46 Lindell D, Post AF (2001). Ecological Aspects of *ntcA* Gene Expression and Its Use as an Indicator of the Nitrogen Status of Marine *Synechococcus* spp. *Applied and Environmental Microbiology* **67:** 3340-3349.

47 Por FD (2012). *Lessepsian migration: the influx of Red Sea biota into the Mediterranean by way of the Suez Canal*, vol. 23. Springer Science & Business Media.

48 Sofianos SS, Johns WE (2002). An Oceanic General Circulation Model (OGCM) investigation of the Red Sea circulation, 1. Exchange between the Red Sea and the Indian Ocean. *Journal of Geophysical Research: Oceans* **107:** 17-11-17-11.

49 Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *The ISME Journal* **4:** 673-685.

50 Eng JK, McCormack AL, Yates JR (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5:** 976-989.

51 Pavelka N, Fournier ML, Swanson SK, Pelizzola M, Ricciardi-Castagnoli P, Florens L *et al.* (2008). Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Molecular & cellular proteomics : MCP* **7:** 631-644.

52 Team RC (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.

53 Maimon O, Rokach L (2005). *Data mining and knowledge discovery handbook*, vol. 2. Springer.

54 Pavelka N, Pelizzola M, Vizzardelli C, Capozzoli M, Splendiani A, Granucci F *et al.* (2004). A power law global error model for the identification of differentially expressed genes in microarray data. *BMC Bioinformatics* **5:** 203.

55 Choi H, Fermin D, Nesvizhskii AI (2008). Significance analysis of spectral count data in label-free shotgun proteomics. *Molecular & cellular proteomics : MCP* **7:** 2373-2385.

56 Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A *et al.* (2015). gplots: Various R programming tools for plotting data. R package version 2.17.0. Available online at: http://CRAN.R-project.org/package=gplots.

57 Neuwirth E (2014). RColorBrewer: ColorBrewer palettes. R package version 1.1-2: Available online at: http://CRAN.R-project.org/package=RColorBrewer.

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

59 Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D *et al.* (2009). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Research* **38:** D396-D400.

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

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

62 Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32:** 1792-1797.

63 Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

64 Stamatakis A (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*.

65 Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S *et al.* (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39:** D546-551.

66 Mitchell A, Bucchini F, Cochrane G, Denise H, Hoopen Pt, Fraser M *et al.* (2015). EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*.

67 Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Zweng MM, Baranova OK *et al.* (2010). World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate). *NOAA Atlas NESDIS 71*. U.S. Government Printing Office: Washington, D.C. p 398.

68 Garcia H, Locarnini R, Boyer T, Antonov J, Baranova O, Zweng M *et al.* (2014). World Ocean Atlas 2013, volume 4: Dissolved inorganic nutrients (phosphate, nitrate, silicate). *NOAA Atlas NESDIS* **76:** 25.

1.9 Figure Legends:

Figure 1.1 Growth chart of batch cultures showing cell concentration determined through flow cytometry by day for P-stressed batch cultures (a) and N-stressed batch cultures (b). P-stressed batch cultures were sampled on day 7 for nutrient-replete conditions, and on day 11 for P-starvation conditions. N-stressed batch cultures were sampled on day 16 for N-starvation conditions.

Figure 1.2 A clustering analysis of the proteomic profiles of NSAF values for each spectral run using a Euclidian distance matrix. The clustering analysis shows the similarity in the profiles among the technical replicates (1, 2, or 3) and the biological replicates (A and B) supporting the use of grouping of all spectral runs per nutrient stress condition for statistical analysis.

Figure 1.3 Log2 fold change in expression of proteins in *Prochlorococcus* MED4 under P-starvation, P-limitation, N-starvation, and N-limitation culturing conditions when compared to

the proteomic profile of nutrient replete cells. Only proteins which were significantly differentially expressed under a nutrient stress condition are displayed. Proteins are displayed according to their general cellular function – depicted here are proteins associated with Phosphorus acquisition, Nitrogen Acquisition, Transport & Binding Protein, Energy Metabolism, Photosynthesis, Transcription, and Cellular Processes.

Figure 1.4 Log2 fold change in expression of proteins in *Prochlorococcus* MED4 under P-starvation, P-limitation, N-starvation, and N-limitation culturing conditions when compared to the proteomic profile of nutrient replete cells. Only proteins which were significantly differentially expressed under a nutrient stress condition are displayed. Proteins are displayed according to their general cellular function – depicted here are proteins associated with Amino Acid Synthesis, Protein Synthesis, Protein Fate, Metabolic Intermediates Synthesis, Fatty Acid & Lipid Metabolism, Signal Transduction, and Uncategorized proteins.

Figure 1.5 Box plot of annual average phosphate surface concentration in a 1 degree grid around the locations of the metagenomics samples analyzed taken from the World Ocean Atlas 2009 for GOS and 2013 for TARA.

Figure 1.6 Boxplots of the relative occurrence of *Prochlorococcus* reads associated with potential biomarker loci when compared to a composite of single copy core housekeeping genes in GOS surface metagenomes. The shaded region represents one copy of a target gene per *Prochlorococcus* genome; the gray box around the value of one represents the error in the estimation of the number of *Prochlorococcus* genomes per grouped location as determined by the spread in the single copy core housekeeping genes. Outliers in relative gene abundance are marked by black circles.

Figure 1.7 Boxplots of the relative occurrence of *Prochlorococcus* reads associated with potential biomarker loci when compared to a composite of single copy core housekeeping genes in TARA surface metagenomes. The shaded region represents one copy of a target gene per *Prochlorococcus* genome; the gray box around the value of one represents the error in the estimation of the number of *Prochlorococcus* genomes per grouped location as determined by the spread in the single copy core housekeeping genes. Outliers in relative gene abundance are marked by black circles.

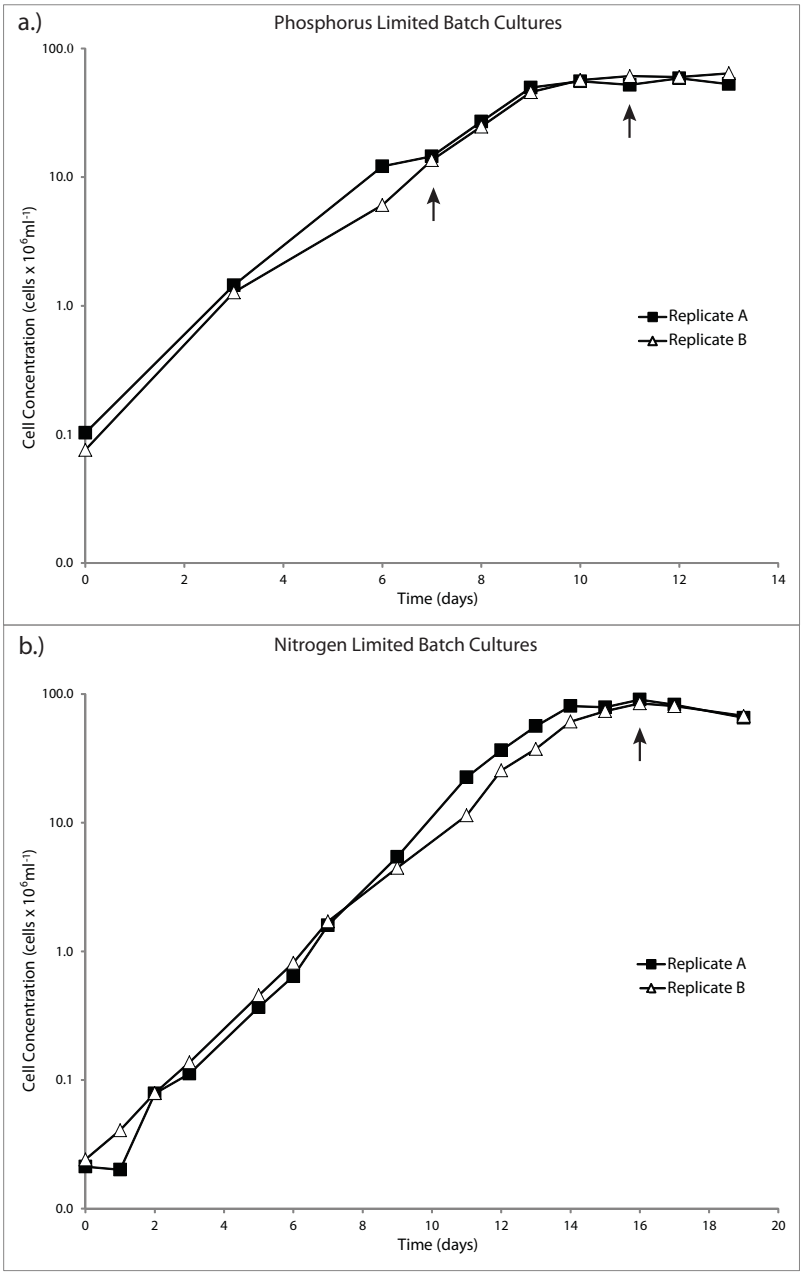

a.)
Phosphorus Limited Batch Cultures
100.0
10.0
1.0
0.1
0.0
Cell Concentration (cells x 10^6 ml^-1)
0
2
4
6
8
10
12
14
Time (days)
Replicate A
Replicate B
b.)
Nitrogen Limited Batch Cultures
100.0
10.0
1.0
0.1
0.0
Cell Concentration (cells x 10^6 ml^-1)
0
2
4
6
8
10
12
14
16
18
20
Time (days)
Replicate A
Replicate B

Figure 1.1

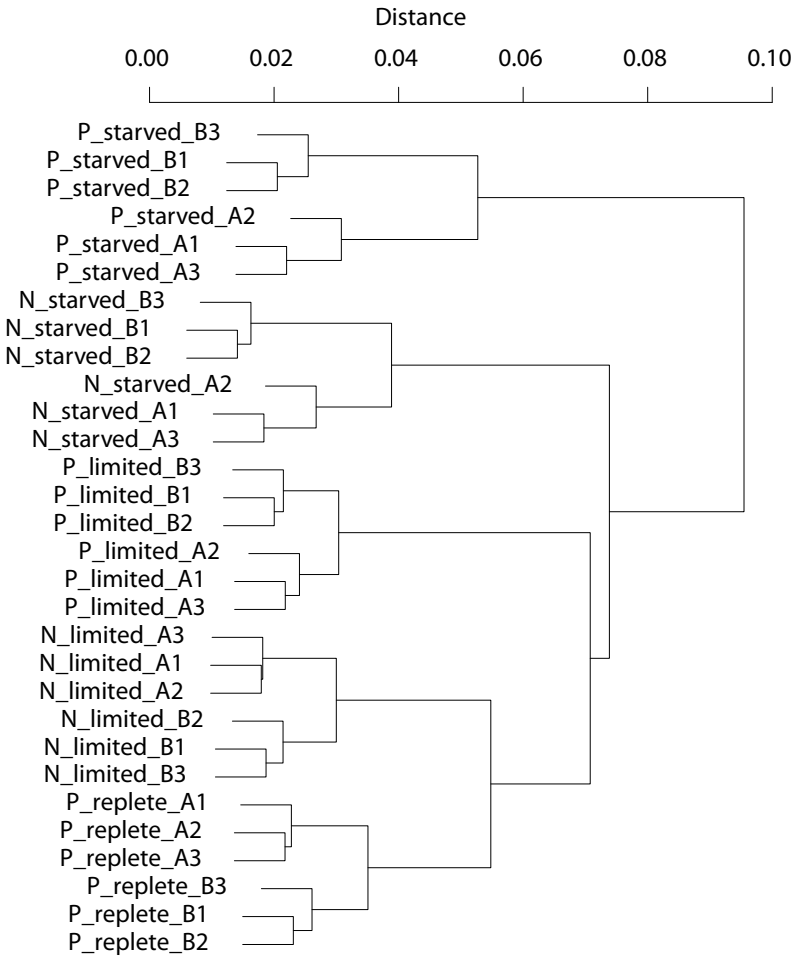

Distance
0.00
0.02
0.04
0.06
0.08
0.10
P_starved_B3
P_starved_B1
P_starved_B2
P_starved_A2
P_starved_A1
P_starved_A3
N_starved_B3
N_starved_B1
N_starved_B2
N_starved_A2
N_starved_A1
N_starved_A3
P_limited_B3
P_limited_B1
P_limited_B2
P_limited_A2
P_limited_A1
P_limited_A3
N_limited_A3
N_limited_A1
N_limited_A2
N_limited_B2
N_limited_B1
N_limited_B3
P_replete_A1
P_replete_A2
P_replete_A3
P_replete_B3
P_replete_B1
P_replete_B2

Figure 1.2

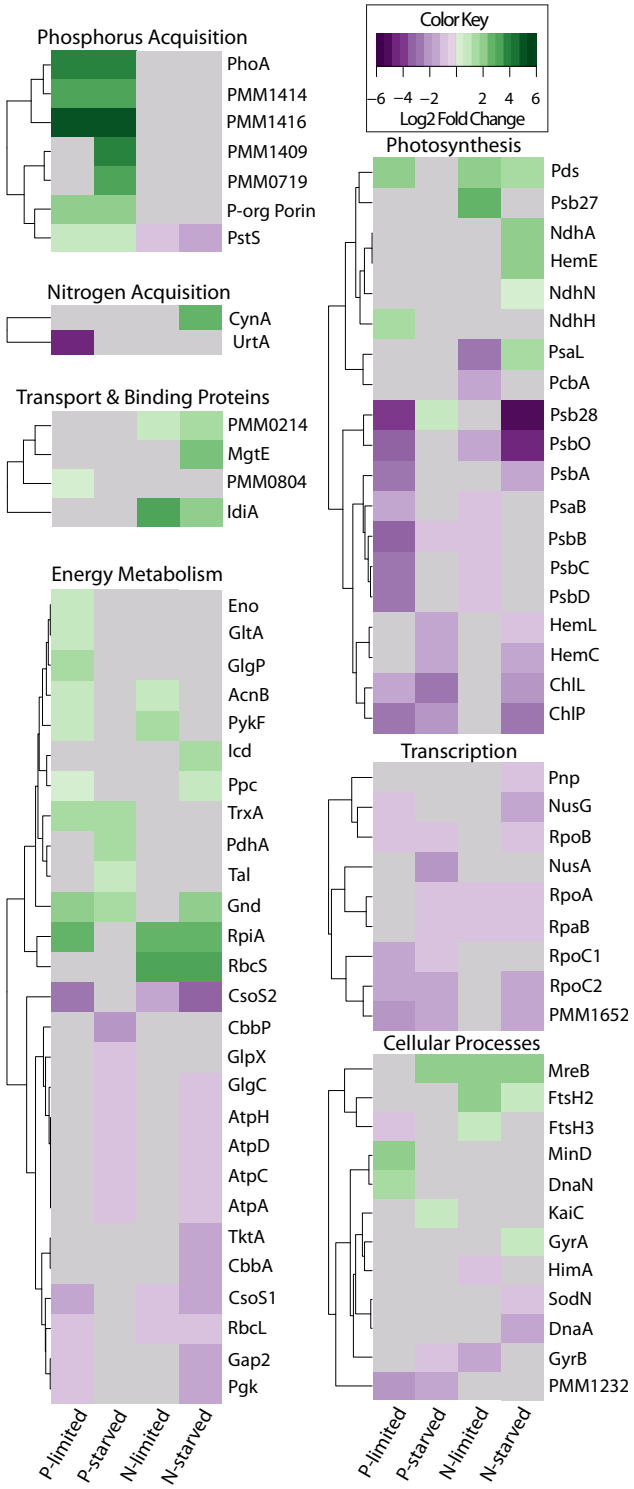

Color Key
−6 −4 −2 2 4 6
Log2 Fold Change
Phosphorus Acquisition
PhoA
PMM1414
PMM1416
PMM1409
PMM0719
P-org Porin
PstS
Nitrogen Acquisition
CynA
UrtA
Transport & Binding Proteins
PMM0214
MgtE
PMM0804
IdiA
Energy Metabolism
Eno
GltA
GlgP
AcnB
PykF
Icd
Ppc
TrxA
PdhA
Tal
Gnd
RpiA
RbcS
CsoS2
CbbP
GlpX
GlgC
AtpH
AtpD
AtpC
AtpA
TktA
CbbA
CsoS1
RbcL
Gap2
Pgk
P-limited
P-starved
N-limited
N-starved
Photosynthesis
Pds
Psb27
NdhA
HemE
NdhN
NdhH
PsaL
PcbA
Psb28
PsbO
PsbA
PsaB
PsbB
PsbC
PsbD
HemL
HemC
ChlL
ChlP
Transcription
Pnp
NusG
RpoB
NusA
RpoA
RpaB
RpoC1
RpoC2
PMM1652
Cellular Processes
MreB
FtsH2
FtsH3
MinD
DnaN
KaiC
GyrA
HimA
SodN
DnaA
GyrB
PMM1232
P-limited
P-starved
N-limited
N-starved

Figure 1.3

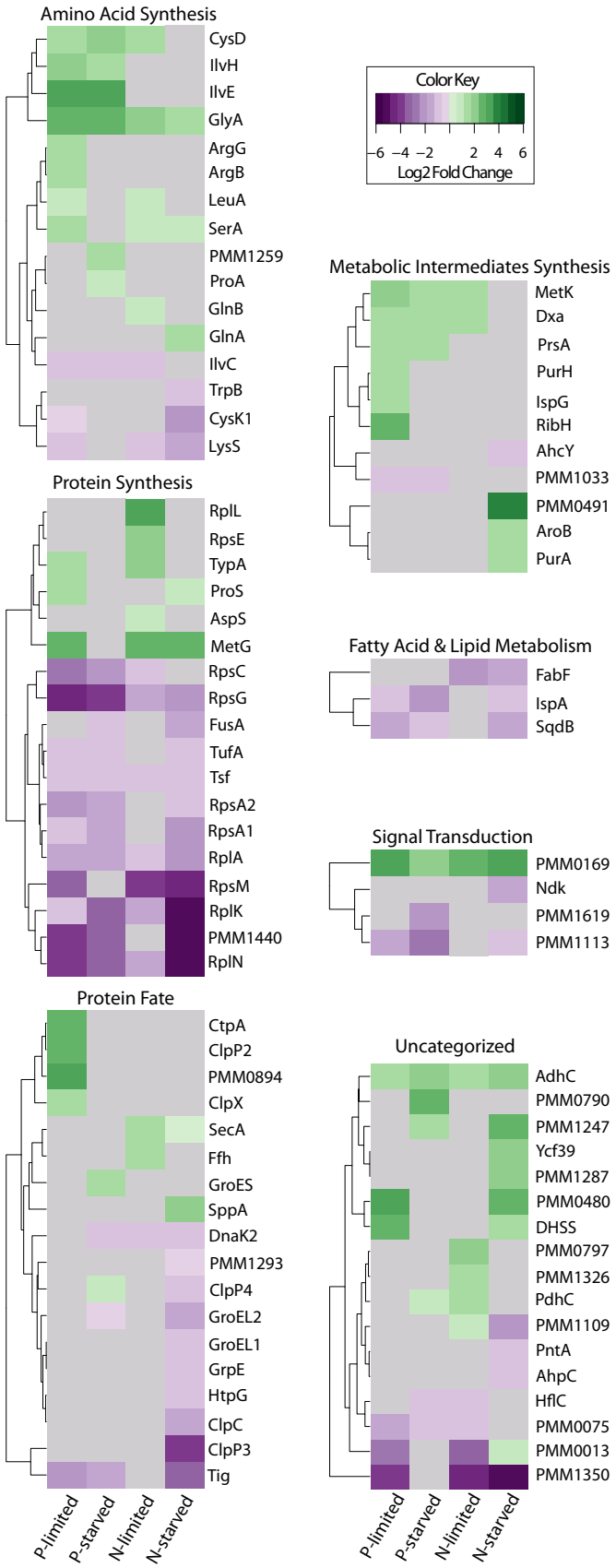

Amino Acid Synthesis
CysD
IlvH
IlvE
GlyA
ArgG
ArgB
LeuA
SerA
PMM1259
ProA
GlnB
GlnA
IlvC
TrpB
CysK1
LysS
Color Key
−6 −4 −2 2 4 6
Log2 Fold Change
Metabolic Intermediates Synthesis
MetK
Dxa
PrsA
PurH
IspG
RibH
AhcY
PMM1033
PMM0491
AroB
PurA
Protein Synthesis
RplL
RpsE
TypA
ProS
AspS
MetG
RpsC
RpsG
FusA
TufA
Tsf
RpsA2
RpsA1
RplA
RpsM
RplK
PMM1440
RplN
Fatty Acid & Lipid Metabolism
FabF
IspA
SqdB
Signal Transduction
PMM0169
Ndk
PMM1619
PMM1113
Protein Fate
CtpA
ClpP2
PMM0894
ClpX
SecA
Ffh
GroES
SppA
DnaK2
PMM1293
ClpP4
GroEL2
GroEL1
GrpE
HtpG
ClpC
ClpP3
Tig
Uncategorized
AdhC
PMM0790
PMM1247
Ycf39
PMM1287
PMM0480
DHSS
PMM0797
PMM1326
PdhC
PMM1109
PntA
AhpC
HflC
PMM0075
PMM0013
PMM1350
P-limited
P-starved
N-limited
N-starved
P-limited
P-starved
N-limited
N-starved

Figure 1.4

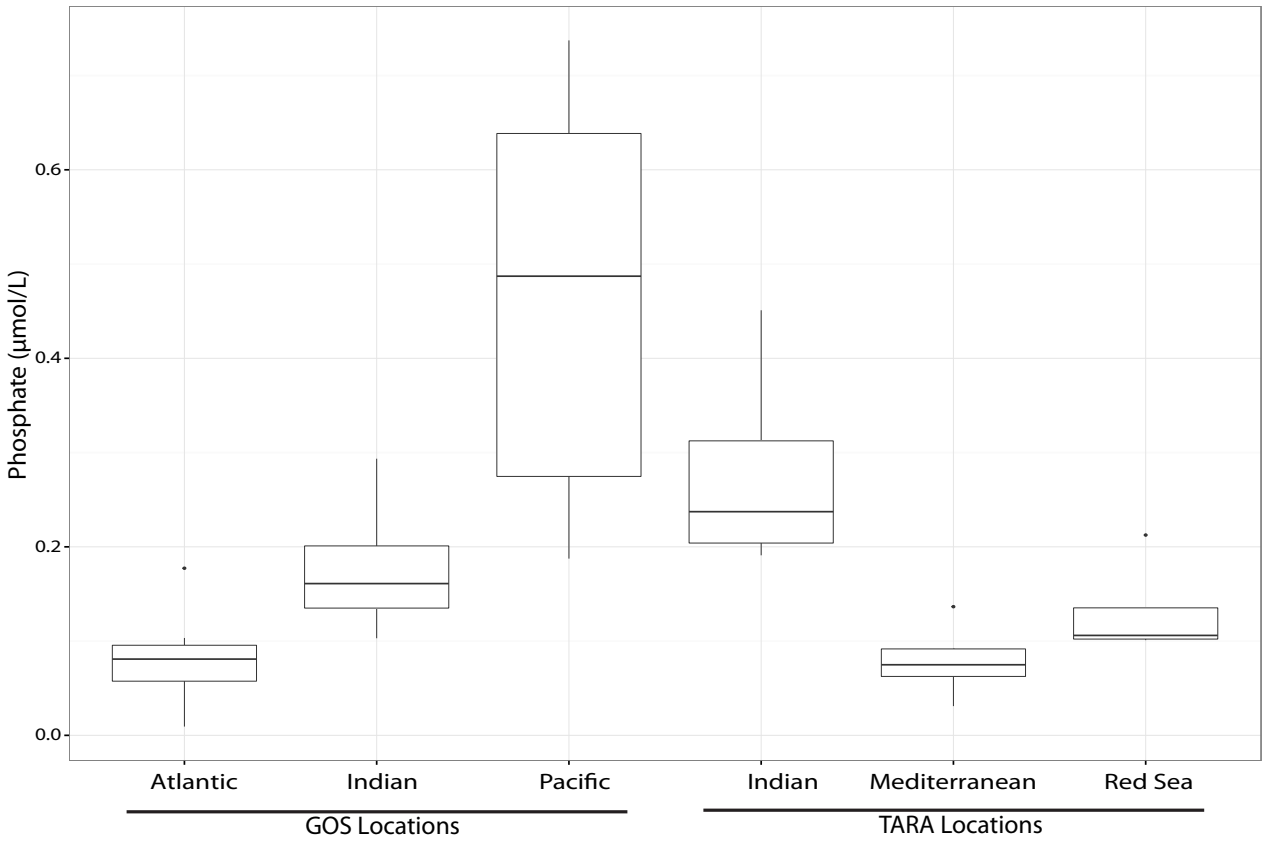

Phosphate (µmol/L)
0.0
0.2
0.4
0.6
Atlantic
Indian
Pacific
Indian
Mediterranean
Red Sea
GOS Locations
TARA Locations

Figure 1.5

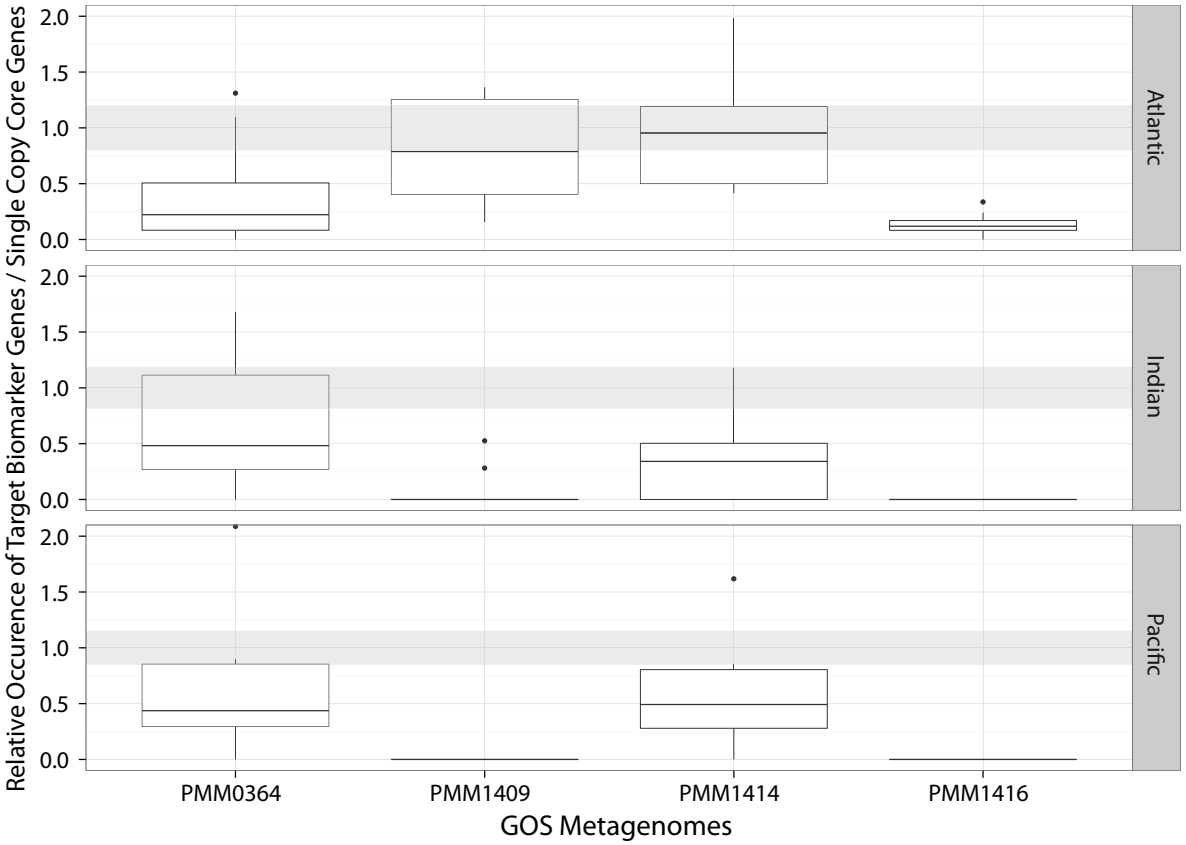

Relative Occurence of Target Biomarker Genes / Single Copy Core Genes
2.0
1.5
1.0
0.5
0.0
Atlantic
Indian
Pacific
PMM0364
PMM1409
PMM1414
PMM1416
GOS Metagenomes

Figure 1.6

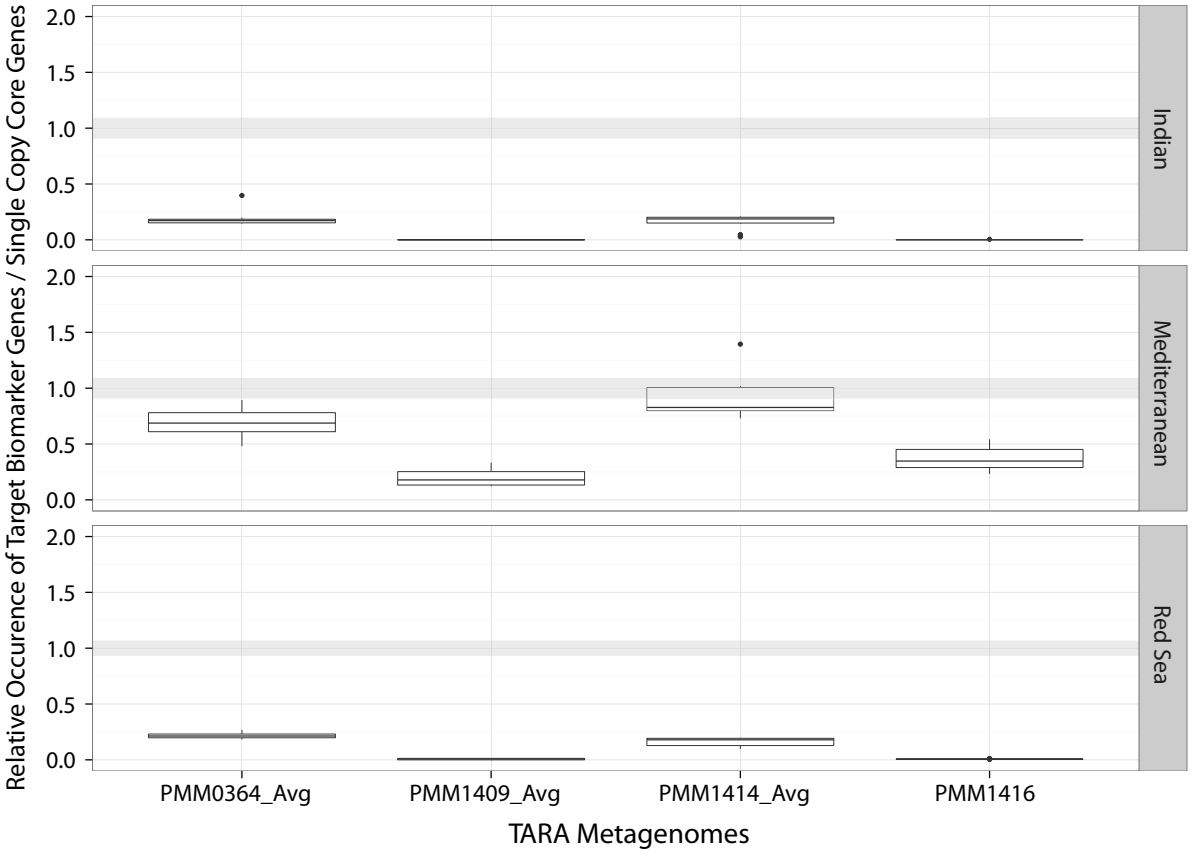

Relative Occurence of Target Biomarker Genes / Single Copy Core Genes
2.0
1.5
1.0
0.5
0.0
Indian
Mediterranean
Red Sea
PMM0364_Avg
PMM1409_Avg
PMM1414_Avg
PMM1416
TARA Metagenomes

Figure 1.7

1.10 Supplementary Figure Legends & Tables:

Supplementary Figure 1.1 Maximum likelihood tree of loci PMM0364, PMM1409, and PMM1414 made with RAxML from the best of 20 starting trees. Bootstraps (n=100) displayed.

Supplementary Figure 1.2 Maximum likelihood tree of locus PMM1416 made with RAxML from the best of 20 starting trees. Bootstraps (n=100) displayed.

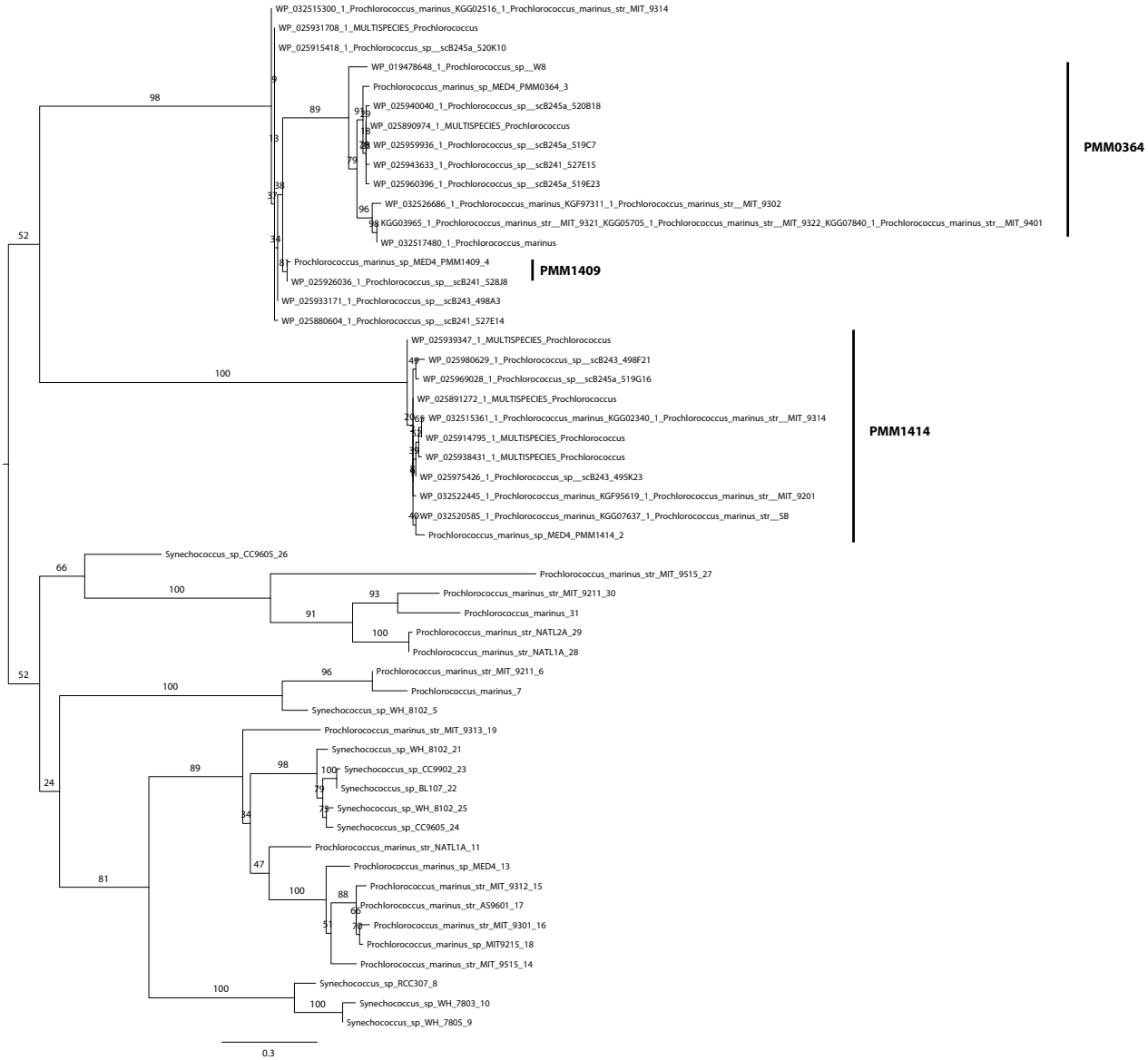
Supplementary Table 1.1 The log2 fold change in the proteomic expression of the NSAF (Normalized Spectral Abundance Factor) for each locus ID deemed significantly differentially expressed in the nutrient stressed condition when compared to the proteomic profile of the nutrient replete condition.

Gene	PMM Number	Major Group	Subcategory	Start	EC Number	Log2FC_P_limited	Log2FC_P_starved	Log2FC_N_limited	Log2FC_N_starved
trpB	PMM0164	Amino acid synthesis	Aromatic amino acid fan	160666	4.2.1.20	0	0	0	-1.157199345
cysD	PMM0642	Amino acid synthesis	aspartate family	610958	2.5.1.49	1.647716298	2.015260612	1.484653785	0
proA	PMM0590	Amino acid synthesis	glutamate family	558202	1.2.1.41	0	1.106799827	0	0
leuA	PMM1066	Amino acid synthesis	pyruvate	1007864	2.3.3.13	1.109777003	0	0.905244001	0
ilvC	PMM1315	Amino acid synthesis	pyruvate family	1268128	1.1.1.86	-0.617261113	-0.856902416	-0.666921071	0
ilvE	PMM0878	Amino acid synthesis	pyruvate family	837108	2.6.1.42,2.6.1.57	3.140854461	3.062699423	0	0
ilvH	PMM0526	Amino acid synthesis	pyruvate family	497565	2.2.1.6	2.07298737	1.535202629	0	0
serA	PMM1354	Amino acid synthesis	serine family	1304746	1.1.1.95	1.568251819	0	1.161603619	0.96628301
lysS	PMM1618	Amino acid synthesis	apartate family	1548943	6.1.1.6	-1.162348888	0	-1.010026884	-1.577812142
PMM1259	PMM1259	Amino acid synthesis		1212680	None	0	1.292400693	0	0
glnA	PMM0920	Amino-acid synthesis	Glutamate family	881367	6.3.1.2	0	0	0	1.332076974
glnB	PMM1463	Amino-acid synthesis	Glutamate family	1396237	None	0	0	1.16498513	0
cysK1	PMM0407	Amino-acid synthesis	serine family	387157	2.5.1.47	-0.508730594	0	0	-1.797631413
glyA	PMM0258	Amino-acid synthesis	serine family	249266	2.1.2.1	2.546183263	2.637342123	2.084111819	1.793004092
argB	PMM0499	Amino-acid synthesis		474822	2.7.2.8	1.256610384	0	0	0
argG	PMM1707	Amino-acid synthesis		1646147	6.3.4.5	1.37010918	0	0	0
ftsH2	PMM0226	Cellular processes	Cell division	221616	3.4.24.-	0	0	2.039412654	1.198173308
ftsH3	PMM1264	Cellular processes	Cell division	1219777	3.4.24.-	-1.091434128	0	1.156225397	0
himA	PMM1321	Cellular processes	Cell division	1272002	None	0	0	-0.612642932	0
minD	PMM0321	Cellular processes	Cell division	306371	None	1.933547534	0	0	0
dnaA	PMM0565	Cellular processes	DNA replication	531716	None	0	0	0	-1.289816006
dnaN	PMM0001	Cellular processes	DNA replication	174	2.7.7.7	1.557000027	0	0	0
gyrA	PMM1063	Cellular processes	DNA replication	1002199	5.99.1.3	0	0	0	0.640884129
gyrB	PMM1634	Cellular processes	DNA replication	1562820	5.99.1.3	0	-0.605728418	-1.552639613	0
kaiC	PMM1342	Cellular processes		1295509	None	0	0.937817915	0	0
mreB	PMM1622	Cellular processes		1552259	None	0	1.974185047	2.383768853	1.947813582
PMM1232	PMM1232	Cellular Processes		1183956	2.1.1.-	-2.037857883	-1.521144216	0	0
SodN	PMM1294	Cellular processes		1246173	1.15.1.1	0	0	0	-1.132903894
atpA	PMM1451	Energy metabolism	ATP synthase	1388581	3.6.3.14	0	-0.88470554	0	-1.032311976
atpC	PMM1450	Energy metabolism	ATP synthase	1388581	3.6.3.14	0	-0.88470554	0	-1.032311976
atpD	PMM1438	Energy metabolism	ATP synthase	1388581	3.6.3.14	0	-0.88470554	0	-1.032311976
atpH	PMM1452	Energy metabolism	ATP synthase	1388581	3.6.3.14	0	-0.88470554	0	-1.032311976
glgC	PMM0769	Energy metabolism	Biosynthesis and degrad	732893	2.7.7.27	0	-0.857113613	0	-0.74139467
glgP	PMM1601	Energy metabolism	Biosynthesis and degrad	1530381	2.4.1.1	1.368099952	0	0	0
gltA	PMM0161	Energy metabolism	Biosynthesis and degrad	157452	2.3.3.1	0.90904935	0	0	0
ppc	PMM1575	Energy metabolism	Carbon fixation	1505707	4.1.1.31	0.465583778	0	0	0.820956691
prkB, cbbP	PMM0785	Energy metabolism	Carbon fixation	747649	2.7.1.19	0	-1.924050263	0	0
rbcL	PMM0550	Energy metabolism	Carbon fixation	519087	4.1.1.39	-0.793030009	0	-0.631507616	-0.954641001
rbcS	PMM0551	Energy metabolism	Carbon fixation	520592	4.1.1.39	0	0	3.580822611	3.424500334
ccmK aka csoS1	PMM0549	Energy metabolism	Carbon fixation	518725	None	-1.537488889	0	-0.946664344	-1.660864843
csoS2	PMM0552	Energy metabolism	Carbon fixation	521024	None	-2.915815407	0	-1.268994286	-3.291026972
cbbA	PMM0781	Energy metabolism	Glycolysis	744048	4.1.2.13,4.1.2.40	0	0	0	-1.514003976
eno	PMM0208	Energy metabolism	Glycolysis	Energy metab	4.2.1.11	0.951910166	0	0	0
gap2	PMM0023	Energy metabolism	Glycolysis	24150	1.2.1.12	-1.087053367	0	0	-1.610957177
pgk	PMM0195	Energy metabolism	Glycolysis	188883	2.7.2.3	-1.149746513	0	0	-1.626193013
pykF	PMM0912	Energy metabolism	Glycolysis	872358	2.7.1.40	0.916013735	0	1.601173174	0
glpX	PMM0767	Energy metabolism	Pentose phosphate path	730434	3.1.3.11,3.1.3.37	0	-0.826600595	0	0
gnd	PMM0770	Energy metabolism	Pentose phosphate path	734287	1.1.1.44	2.307234446	1.252350994	0	1.892635935
tktA	PMM1610	Energy metabolism	Pentose phosphate path	1539744	2.2.1.1	0	0	0	-1.443388758
rpiA	PMM1489	Energy metabolism	Pentose phosphate path	1426747	5.3.1.6	2.594415957	0	2.716695638	2.698449476
tal	PMM0519	Energy metabolism	Pentose phosphate path	490726	2.2.1.2	0	0.748706301	0	0

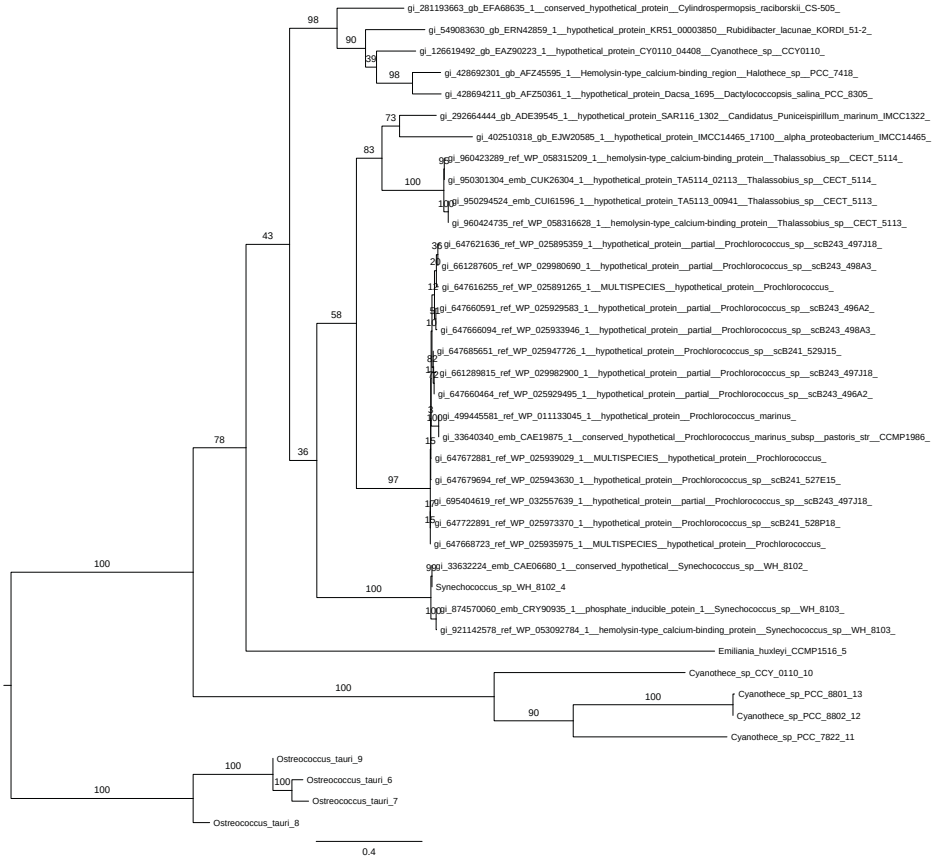
Gene	PMM Number	Major Group	Subcategory	Start	EC Number	Log2FC_P_limited	Log2FC_P_starved	Log2FC_N_limited	Log2FC_N_starved
acnB	PMM1700	Energy metabolism	TCA Cycle	1635543	4.2.1.3	0.842220079	0	0.683185834	0
icd	PMM1596	Energy metabolism	TCA Cycle	isocitrate deh	1.1.1.42	0	0	0	1.463440352
acoA, pdhA	PMM1288	Energy metabolism		1239069	1.2.4.1	0	1.524684123	0	0
trxA	PMM1061	Energy metabolism		1000647	1.-.-.-	1.280337142	1.740831463	0	0
fabF	PMM1609	Fatty acid and phospholipid metabolism		1538457	2.3.1.179,2.3.1.41	0	0	-2.303151917	-1.332020166
ispA, crtE	PMM1070	Fatty acid and phospholipid metabolism		1012480	2.5.1.1,2.5.1.10	-0.933968843	-2.093959649	0	-1.049883347
sqdB	PMM1665	Fatty acid and phospholipid metabolism		1596609	3.13.1.1	-1.660866431	-1.018789624	0	-1.399484746
ispG	PMM0676	Metabolic intermediates biosynthesis	Biosynthesis of cofactors,	643694	1.17.7.1,1.17.4.3	1.57043229	0	0	0
PMM0491	PMM0491	Metabolic intermediates biosynthesis	Biosynthesis of cofactors,	465254	4.2.1.96	0	0	0	3.676563983
PMM1033	PMM1033	Metabolic intermediates biosynthesis	Biosynthesis of cofactors,	977075	None	-0.883294979	-0.864309582	0	0
ribH	PMM1643	Metabolic intermediates biosynthesis	Biosynthesis of cofactors,	1571363	2.5.1.-,2.5.1.9	2.838092757	0	0	0
prsA	PMM1080	Metabolic intermediates biosynthesis	Purine ribonucleotide bic	1023474	2.7.6.1	1.724758397	1.433640704	0	0
purA	PMM0506	Metabolic intermediates biosynthesis	Purine ribonucleotide bic	479516	6.3.4.4	0	0	0	1.267900836
purH	PMM0266	Metabolic intermediates biosynthesis	Purine ribonucleotide bic	257588	2.1.2.3,3.5.4.10	1.775976712	0	0	0
ahcY	PMM1625	Metabolic intermediates biosynthesis		1554840	3.3.1.1	0	0	0	-1.158328266
aroB	PMM0682	Metabolic intermediates biosynthesis		651281	4.2.3.4	0	0	0	1.504288266
Dxs	PMM0907	Metabolic intermediates biosynthesis		869431	2.2.1.7	1.390228335	1.493339091	1.666258169	0
metK	PMM0311	Metabolic intermediates biosynthesis		299024	2.5.1.6	2.132261344	1.778597738	1.602343786	0
cynA	PMM0370	Nitrogen acquisition	Nitrogen acquisition	352975	None	0	0	0	2.706931697
urtA	PMM0970	Nitrogen acquisition	Nitrogen acquisition	926568	None	-4.428977275	0	0	0
phoA	PMM0708	Phosphate acquisition	Phosphate acquisition	673975	None	3.874910326	4.024731464	0	0
PMM1409	PMM1409	Phosphate acquisition	Phosphate acquisition	1352411	None	0	4.055035702	0	0
PMM1414	PMM1414	Phosphate acquisition	Phosphate acquisition	1355592	None	3.197165092	3.316315286	0	0
PMM1416	PMM1416	Phosphate acquisition	Phosphate acquisition	1357735	None	4.935229978	5.296869554	0	0
P-org Porin	PMM0709	Phosphate acquisition	Phosphate acquisition	675495	None	2.071918473	2.083773621	0	0
pstS	PMM0710	Phosphate acquisition	Phosphate acquisition	675854	None	1.117126768	1.070780707	-0.741786741	-1.470780693
PMM0719	PMM0719	Phosphate acquisition	Phosphate acquisition	683454	None	0	3.195549071	0	0
chlI	PMM1055	Photosynthesis	Pigment production	997537	6.6.1.1	-1.301261462	-2.784688961	0	-2.295215204
chlP	PMM0760	Photosynthesis	Pigment production	724558	1.3.1.83,1.3.1.-	-2.588427232	-2.071713564	0	-2.828020704
hemC	PMM0495	Photosynthesis	Pigment production	468968	2.5.1.61	0	-1.538148791	0	-1.409781576
hemE	PMM0583	Photosynthesis	Pigment production	548918	4.1.1.37	0	0	0	1.87654736
hemL	PMM0483	Photosynthesis	Pigment production	457939	5.4.3.8	0	-1.573931704	0	-0.785954436
pds	PMM0144	Photosynthesis	Pigment production	141867	1.14.99.-,1.14.99.	2.35880682	0	2.282418833	1.309193511
ndhA	PMM0160	Photosynthesis		156231	1.6.5.3	0	0	0	1.891284534
ndhH	PMM0172	Photosynthesis		165925	1.6.5.3	1.697375226	0	0	0
pcbA	PMM0627	Photosynthesis		596691	None	0	0	-1.644606115	0
ndhN	PMM1559	Photosynthesis		1487908	1.6.5.3	0	0	0	0.276573961
psaB	PMM1523	Photosynthesis		1464133	None	-1.783696193	0	-1.102445914	0
psaL	PMM1519	Photosynthesis		1460205	None	0	0	-2.639779751	1.625880621
psb27	PMM0507	Photosynthesis		480035	None	0	0	2.734315923	0
psb28	PMM0926	Photosynthesis		886571	None	-3.776726896	0.749362786	0	-5.040703697
psbA	PMM0223	Photosynthesis		216807	None	-2.502112987	0	0	-1.287600382
psbB	PMM0315	Photosynthesis		302587	None	-3.080972303	-0.744001623	-0.969176827	0
psbC	PMM1158	Photosynthesis		1108673	None	-2.835001737	0	-0.793274929	0
psbD	PMM1157	Photosynthesis		1107613	None	-2.76623121	0	-1.125013598	0
psbO	PMM0228	Photosynthesis		223707	None	-3.014221066	0	-1.227000926	-4.278197867
sppA	PMM1180	Protein fate	Degradation of proteins,	1128287	3.4.21.-	0	0	0	2.11241622
clpC	PMM1088	Protein fate	Degradation of proteins,	1034552	3.4.21.92	0	0	0	-1.23653542
clpP2	PMM1656	Protein fate	Degradation of proteins,	1587595	3.4.21.92	2.467426956	0	0	0
clpP3	PMM1314	Protein fate	Degradation of proteins,	1267406	3.4.21.92	0	0	0	-3.603151585
clpP4	PMM1313	Protein fate	Degradation of proteins,	1266710	3.4.21.92	0	0.77894102	0	-1.00665032
clpX	PMM1657	Protein fate	Degradation of proteins,	1588346	3.4.21.92	1.376943332	0	0	0
ctpA	PMM0324	Protein fate	Degradation of proteins,	309735	3.4.21.102	2.567879367	0	0	0
ffh	PMM1286	Protein fate	Protein and peptide secr	1236761	3.6.5.4	0	0	1.514780922	0

Gene	PMM Number	Major Group	Subcategory	Start	EC Number	Log2FC_P_limited	Log2FC_P_starved	Log2FC_N_limited	Log2FC_N_starved
secA	PMM1639	Protein fate	Protein and peptide secr	1569146	None	0	0	1.229160833	0.541104469
tig	PMM1655	Protein fate	Protein folding & stabiliz	1586108	5.2.1.8	-1.863129691	-1.346416025	0	-3.127106492
dnaK2	PMM1704	Protein fate	Protein folding & stabiliz	1643380	3.6.1.-	0	-1.09647064	-0.985090204	-0.776031349
groEL2 / groL2	PMM1436	Protein fate	Protein folding & stabiliz	1375274	3.6.4.9	0	-0.586497993	0	-1.24394052
groES	PMM1437	Protein fate	Protein folding & stabiliz	1375637	1.3.1.-	0	1.515928982	0	0
groEL1 / groL1	PMM0452	Protein fate	Protein folding & stabiliz	432301	3.6.4.9	0	0	0	-0.720356346
grpE	PMM0016	Protein fate	Protein folding & stabiliz	18283	None	0	0	0	-0.7398024
PMM0894	PMM0894	Protein fate	Protein folding & stabiliz	857531	5.2.1.8	3.117160195	0	0	0
PMM1293	PMM1293	Protein fate	Protein folding & stabiliz	1245526	5.2.1.8	0	0	0	-0.033568164
htpG	PMM0901	Protein fate	Protein folding & stabiliz	863832	None	0	0	0	-1.138658281
PMM1440	PMM1440	Protein synthesis	Ribosomal proteins	1378251	None	-3.776311274	-3.259597608	0	-5.040288075
rplA	PMM0203	Protein synthesis	Ribosomal proteins	196868	None	-1.271503531	-1.58856479	-1.121833213	-1.951544362
rplK	PMM0204	Protein synthesis	Ribosomal proteins	197360	None	-0.991676519	-3.10322803	-1.389152683	-4.883918496
rplL	PMM0201	Protein synthesis	Ribosomal proteins	195412	None	0	0	3.06928333	0
rplN	PMM1548	Protein synthesis	Ribosomal proteins	1482739	None	-3.6514574	-3.134743738	-1.483155363	-4.915434201
rps1a, rpsA1	PMM0312	Protein synthesis	Ribosomal proteins	300250	None	-0.911605134	-1.447304062	0	-2.214919307
rps1b, rpsA2, nbp1	PMM0530	Protein synthesis	Ribosomal proteins	502658	None	-1.905211895	-1.388498226	0	-0.869074376
rpsC	PMM1552	Protein synthesis	Ribosomal proteins	1484451	None	-2.841768233	-2.325054569	-1.099174283	0
rpsE	PMM1542	Protein synthesis	Ribosomal proteins	1480042	None	0	0	2.230507382	0
rpsG	PMM1510	Protein synthesis	Ribosomal proteins	1450323	None	-4.242938471	-3.726224801	-1.376049089	-1.932461254
rpsM	PMM1537	Protein synthesis	Ribosomal proteins	1476854	None	-3.211549459	0	-3.641746987	-4.475526261
fusA	PMM1509	Protein synthesis	Translation factors	1449748	3.6.5.3	0	-1.139000743	0	-1.260070453
tsf	PMM0754	Protein synthesis	Translation factors	714227	None	-0.82348524	-0.939304046	-0.736214553	-1.169761608
tufA	PMM1508	Protein synthesis	Translation factors	1447629	3.6.5.3	-0.753843742	-0.829375749	0	-1.008296825
metG	PMM0867	Protein synthesis	tRNA aminoacylation	824575	6.1.1.10	2.494334917	0	2.795244601	2.641677371
proS	PMM0508	Protein synthesis	tRNA aminoacylation	481864	6.1.1.15	1.625593478	0	0	0.937772613
aspS	PMM1688	Protein synthesis	tRNA aminoacylation	1621326	6.1.1.12	0	0	0.756295875	0
typA	PMM0762	Protein synthesis		725513	3.6.5.3	1.746391814	0	2.040162725	0
ndk	PMM0046	Signal transduction		49017	2.7.4.6	0	0	0	-1.342242954
PMM0169	PMM0169	Signal transduction		163770	None	3.428128039	2.181952723	2.516921553	3.047264324
PMM1113	PMM1113	Signal transduction		1059346	None	-1.297995972	-2.731619909	0	-0.787611368
PMM1619	PMM1619	Signal transduction		1549740	3.1.1.61	0	-2.288583106	0	0
rnj	PMM1652	Transcription	Degredation of RNA	1583967	3.1.-.-,3.-.-.-	-1.968229115	-1.451515446	0	-1.263416722
pnp	PMM1191	Transcription	Degredation of RNA	1136852	2.7.7.8	0	0	0	-0.683378102
nusA	PMM1492	Transcription	Transcription factors	1428605	None	0	-1.967489521	0	0
nusG	PMM0205	Transcription	Transcription factors	198035	None	-1.047649843	0	0	-1.266180242
rpaB	PMM0134	Transcription		132148	3.1.1.61	0	-1.043869832	-0.6316442	-1.155231083
rpoA	PMM1535	Transcription		1476001	2.7.7.6	0	-1.129526553	-0.753109097	-0.841371037
rpoB	PMM1485	Transcription		1423313	2.7.7.6	-0.716691855	-0.776244886	0	-1.087301878
rpoC1	PMM1484	Transcription		1419981	2.7.7.6	-1.367791397	-0.829508968	0	0
rpoC2	PMM1483	Transcription		1418043	2.7.7.6	-1.487715844	-1.346959826	0	-1.442843787
idiA	PMM1164	Transport & binding proteins	Iron acquisition	1117144	None	0	0	3.059266361	2.368603219
PMM0804	PMM0804	Transport & binding proteins	Iron acquisition	763008	None	0.528618806	0	0	0
mgtE	PMM1630	Transport & binding proteins		1559783	3.6.1.-	0	0	0	2.405269671
PMM0214	PMM0214	Transport & binding proteins		208916	None	0	0	1.088683193	1.495651253
adhC	PMM1234	Uncategorized		1186069	1.1.1.1,1.1.1.284,	1.730170403	2.23115067	1.785019587	2.021495786
ahpC aka tpx	PMM0856	Uncategorized		813527	1.11.1.15,1.6.4.-	0	0	0	-1.084693698
DHSS	PMM0035	Uncategorized		33697	1.12.1.2	2.512733558	0	0	1.512047966
hflC	PMM0482	Uncategorized		455827	None	0	-0.677433858	-0.716347568	0
pdhC	PMM0405	Uncategorized		383670	2.3.1.12,2.3.1.61	0	1.197227854	1.66912264	0
PMM0013	PMM0013	Uncategorized		16018	None	-2.779440825	0	-3.209638352	0.882040337
PMM0075	PMM0075	Uncategorized		79845	None	-1.334252409	-0.917714297	-1.184284632	0
PMM0480	PMM0480	Uncategorized		454404	None	3.527479231	0	0	2.828265493
PMM0790	PMM0790	Uncategorized		752004	None	0	2.622736049	0	0

Gene	PMM Number	Major Group	Subcategory	Start	EC Number	Log2FC_P_limited	Log2FC_P_starved	Log2FC_N_limited	Log2FC_N_starved
PMM0797	PMM0797	Uncategorized		757849	None	0	0	2.023593313	0
PMM1109	PMM1109	Uncategorized		1057810	None	0	0	0.772192621	-1.960155063
PMM1247	PMM1247	Uncategorized		1199985	None	0	1.332200532	0	2.962539434
PMM1287	PMM1287	Uncategorized		1238919	None	0	0	0	1.807858275
PMM1326	PMM1326	Uncategorized		1277408	None	0	0	1.60803048	0
PMM1350	PMM1350	Uncategorized		1300763	None	-3.969700995	0	-4.399898522	-5.233677796
ycf39	PMM1152	Uncategorized		1104474	None	0	0	0	1.833793722
pntA	PMM1147	Uncategorized		1100323	1.6.1.2	0	0	0	-0.869620803



Supplementary Figure 1.1



Supplementary Figure 1.2

Chapter 2

Arsenic detoxification strategies vary with environmental phosphate concentrations in global *Prochlorococcus* populations

2.1 Abstract:

The globally significant picocyanobacterium *Prochlorococcus* is the main primary producer in oligotrophic subtropical gyres. When phosphate concentrations are very low in the marine environment, the mol:mol availability of phosphate relative to the chemically similar arsenate molecule is reduced, potentially resulting in increased cellular arsenic exposure. To mediate accidental arsenate uptake some *Prochlorococcus* isolates contain genes encoding a full or partial efflux detoxification pathway, consisting of an arsenate reductase (*arsC*), an arsenite-specific efflux pump (*acr3)*, and an arsenic related repressive regulator (*arsR*). This efflux pathway was the only previously known arsenic detox pathway. We have identified an additional putative arsenic mediation strategy in *Prochlorococcus* driven by the enzyme arsenite *S*-adenosylmethionine methyltransferase (ArsM) which can convert inorganic arsenic into more innocuous organic forms and appears to be a more widespread mode of detoxification. We used a phylogenetically informed approach to identify *Prochlorococcus* linked arsenic genes from both pathways in the Global Ocean Sampling survey. The putative arsenic methylation pathway is nearly ubiquitously present in global *Prochlorococcus* populations. In contrast, the complete efflux pathway is only maintained in populations which experience extremely low PO_4:AsO_4, such as regions in the tropical and subtropical Atlantic. Thus, environmental exposure to arsenic appears to select for maintenance of the efflux detoxification pathway in *Prochlorococcus*. The differential distribution of these two pathways has implications for global arsenic cycling, as their associated end products, arsenite or organoarsenicals, have differing biochemical activities and residence times.

Journal: ISME Journal

Saunders JK, Rocap G (2016). Genomic potential for arsenic efflux and methylation varies among global Prochlorococcus populations. *ISME J* 10: 197-209. Available at: http://www.nature.com/ismej/journal/v10/n1/abs/ismej201585a.html

DOI: 10.1038/ismej.2015.85

2.2 Introduction:

The marine picocyanobacterium *Prochlorococcus* is the dominant phytoplankter in oligotrophic tropical and subtropical oceans (Campbell *et al.*, 1997; DuRand *et al.*, 2001; Scanlan, 2012). In the subtropical gyres surface phosphate concentrations are extremely low and in some cases may limit primary production (Wu *et al.*, 2000). *Prochlorococcus* has evolved several adaptations to this low P environment, including a high specific affinity for phosphate uptake and low cellular P quotas (Krumhardt *et al.*, 2013). These low cellular P quotas are achieved in part by their small genome sizes (Rocap *et al.*, 2003) and predominance of sulfolipids over phospholipids in the membrane (Van Mooy *et al.*, 2006). As a result *Prochlorococcus* under P-replete conditions have a cellular N:P content over 20, well above the Redfield ratio of 16:1, and can exceed 100:1 under P starvation (Bertilsson *et al.*, 2003; Heldal *et al.*, 2003).

Prochlorococcus can be divided up into physiologically and genetically distinct ecotypes (Moore *et al.*, 1998) which reflect adaptations to environmental parameters like light intensity (Moore *et al.*, 1998; Moore and Chisholm, 1999), temperature optima (Johnson

et al., 2006), and nitrogen utilization capabilities (Moore *et al.*, 2002) amongst others. The multiple strains of *Prochlorococcus* possess a variable assortment of accessory genes suited to the specific environments they inhabit (Rocap *et al.*, 2003; Kettler *et al.*, 2007). The complement of phosphorus acquisition genes in *Prochlorococcus* also varies among the strains (Martiny *et al.*, 2006). Many of these genes are found on genomic islands (Coleman *et al.*, 2006) and some have been acquired by horizontal transfer (Rocap *et al.*, 2003). Thus occurrence of phosphate acquisition genes is not congruent with 16S rRNA phylogeny (Martiny *et al.*, 2006), but can be correlated with phosphate concentrations in the regions these strains were isolated from. In the field, *Prochlorococcus* populations in phosphorus scarce regions contain a greater number of phosphorus acquisition genes, indicating the selective force of nutrient availability on *Prochlorococcus* genetic capacity (Martiny *et al.*, 2009).

The physiochemical similarities between inorganic phosphate and arsenate can result in the indiscriminate uptake of toxic arsenic into cells by cellular phosphate acquisition systems. The high affinity phosphate uptake system ubiquitously present among *Prochlorococcus* strains (Figure 1), consists of the periplasmic binding protein PstS and the membrane-bound ABC-type transporter PstCAB (Scanlan *et al.*, 2009). The PstCABS system is unable to completely differentiate between arsenate and phosphate in *Escherichia coli* (Rosenberg *et al.*, 1977; Tawfik and Viola, 2011). Once inside the cell, arsenate can become toxic by competing with phosphate, for example through the decoupling of oxidative phosphorylation, the process that produces ATP (Mandal and Suzuki, 2002; Oremland and Stolz, 2003). The reduced form, arsenite, is even more toxic because it interferes with enzyme activity by bonding to –SH and –OH groups in enzymes (Mandal and Suzuki, 2002; Akter *et al.*, 2005).

Microbes have evolved multiple mechanisms for cellular defense against arsenic, and the genes involved are taxonomically widespread and subject to frequent horizontal transfer (Stolz *et al.*, 2006; Páez-Espino *et al.*, 2009). The general efflux detoxification pathway involves the reduction of arsenate to arsenite, and then subsequent expulsion of arsenic from the cell through arsenite specific transporters (Carlin *et al.*, 1995; Ghosh *et al.*, 1999). The efflux system, which has evolved independently at least two different times, consists of an arsenate reductase (the analogs ArsC, ACR2, and wzb-like low molecular weight protein tyrosine phosphatase (Bennett *et al.*, 2001)) and an arsenite-specific efflux pump (ArsB or ACR3) (Páez-Espino *et al.*, 2009). Additional arsenic-related genes include an ATPase (ArsA) that couples with ArsB for expulsion of arsenite from cells, the regulatory elements ArsR and ArsD, and a gene of unknown function *arsH* (Stolz *et al.*, 2006). The genes sufficient for a complete efflux pathway were previously identified in *Prochlorococcus* genomes ((Scanlan *et al.*, 2009); Figure 1) including the arsenate reductase, *arsC,* the *acr3* arsenite efflux transporter and *arsR*, an arsenite-binding *trans*-acting repressive regulator (Xu *et al.*, 1996). This efflux detoxification was believed to be the sole arsenic detoxification strategy for *Prochlorococcus* prior to the analyses presented within this paper.

A second microbial arsenic detoxification strategy is the methylation of arsenic into more innocuous organic compounds. This pathway involves the repeated methylation and reduction of arsenical compounds to less toxic states mediated by the enzyme ArsM, arsenite *S*-adenosylmethyltransferase (Figure 1) (Qin *et al.*, 2006; Yin *et al.*, 2011). In *Rhodopseudomonas palustris* the methylation pathway includes the intermediates monomethylarsonate, MMA(V) and dimethylarsonate, DMA(V) and ultimately results in the production of the gas trimethylarsine oxide, TMAO(g), and subsequent release of arsenic from the cell (Qin *et al.*, 2006). However, the endproducts of the methylation pathway vary across taxa and it is not currently possible to determine which organoarsenicals are

synthesized through genetic analysis alone (Dembitsky and Levitsky, 2004). TMAO(g) production and presence of *arsM* genes have been observed in three freshwater cyanobacteria, *Microcystis* sp. PCC7806, *Nostoc* sp. PCC7120, and *Synechocystis* sp. PCC6803 (Yin *et al.*, 2011). A putative *arsM* was identified in the picocyanobacterium *Synechococcus* sp. WH 8102 based on sequence similarity and protein binding domain homology (Thomas *et al.*, 2010), but it has not been demonstrated whether or not this species is capable of methylating arsenic in the lab.

In the marine environment, inorganic arsenate accounts for about 85% of the standing pool of arsenic with the remaining portion of the arsenic pool comprised of the inorganic arsenite and various organoarsenicals, including two of the main products of the methylation pathway, MMA(V) and DMA(V) (Andreae, 1979; Cutter and Cutter, 2006; Wurl *et al.*, 2013; Cutter *et al.*, 2001). Arsenate has a nutrient-like depth distribution because it is a bioactive molecule which is transformed by marine biota in the surface waters via the two pathways described above (Andreae, 1979). The degree of uptake and subsequent transformation of arsenate is likely affected by the standing phosphate concentration (Cutter and Cutter, 2006; Cutter *et al.*, 2001). In oligotrophic gyres, the ratio of soluble reactive phosphorus to arsenate can shift dramatically as phosphorus is depleted, resulting in a greater abundance of arsenate relative to available phosphate (Wurl *et al.*, 2013; Karl and Tien, 1997). While inorganic arsenic concentrations are relatively stable throughout the surface waters of the open ocean ranging from 12.9-15.7 nmol/l (Andreae, 1979; Cutter and Cutter, 2006; Cutter *et al.*, 2001; Ellwood and Maher, 2002; Cutter and Cutter, 1998), surface phosphate concentrations are far more variable, ranging from < 0.1 μmol/l in oligotrophic subtropical gyres to > 2.0 μmol/l in the Southern Ocean (Karl, 2007). Thus, the ratio of available PO_4:AsO_4 also varies widely across different ocean regions, and may be a major factor driving the risk of competitive arsenic uptake.

To better understand arsenic transformations in marine cyanobacteria, we identified arsenic related genes in *Prochlorococcus* genomes. The evolutionary history of the marine picocyanobacterial arsenic genes was inferred through phylogenetic analysis. To further explore the biogeochemical connection between phosphate scarcity and arsenic transformations in marine ecosystems, we used a phylogenetically informed approach to identify *Prochlorococcus* linked arsenic related genes from the Global Ocean Sampling (GOS) metagenomic dataset spanning various ocean nutrient regimes (Rusch *et al.*, 2007; Venter *et al.*, 2004). The goal of this analysis was to elucidate the putative arsenic related mechanisms used by *Prochlorococcus*, and to understand the environmental selective pressures driving the existence of these mechanisms.

2.3 Methods:

Identification of Prochlorococcus *Arsenic Genes:*

Arsenic efflux pathway detoxification genes (*arsR,* ars*D, arsA, arsB, acr3, arsC,* and *arsH*) were queried with blastp searches of the *Prochlorococcus* (taxaid: 1218) portion of NCBI's non-redundant protein sequence database (Altschul *et al.*, 1990) using arsenic detoxification genes from other organisms as queries (Páez-Espino *et al.*, 2009). For genes identified in *Prochlorococcus* genomes (*arsR, arsC, acr3*) as well as the single copy core genes *glnA, rpsD,* and *tyrS* the Clusters of Orthologous Groups (COG) associated with blast results were identified ((Tatusov *et al.*, 2000); Table 1). The genome browser tool on MicrobesOnline.org (Dehal *et al.*, 2009) was used for identifying genomic context of the various genes within the picocyanobacteria.

To identify a *Prochlorococcus arsM*, the COGs associated with known functional *arsM* sequences from other organisms (Qin *et al.*, 2006; Yin *et al.*, 2011) were obtained (COGs 500 and 2226). Sequences of the closely related SAM-dependent mycolic acid cyclopropane synthetase (*cmaS*) were also included as an outgroup (Yuan and Barry, 1996).

Secondary structure prediction was used to confirm homology of the putatively identified *arsM* sequence in *Prochlorococcus.* Sequences identified as *arsM* in the freshwater cyanobacteria *Microcystis* sp. PCC7806, *Nostoc* sp. PCC7120, and *Synechocystis* sp. PCC 6803 (GenBank accession numbers HQ891147, HQ891148, and HM776638, respectively) were aligned with the putatively identified *arsM* sequence in *Prochlorococcus* MED4. All four sequences were submitted individually to the Phyre2 protein prediction server (Kelley and Sternberg, 2009).

Reference alignments and phylogenetic trees:

Sequence collection was performed by searching a list of reference organisms (including cyanobacteria, abundant marine bacteria in the GOS metagenomic set (Rusch *et al.*, 2007), and bacterial strains with previously identified arsenic detoxification genes (Páez-Espino *et al.*, 2009)) for genes associated with relevant COGs (Table 1) using the online portal www.MicrobesOnline.org (Dehal *et al.*, 2009). The ortholog gene tree tool on MicrobesOnline.org was used to generate a list of additional sequences of broader taxonomic range to add greater phylogenetic resolution.

Nucleic acid sequences of the genes were downloaded from www.MicrobesOnline.org and aligned using TranslatorX v 1.1 (Abascal *et al.*, 2010). The amino acid translated alignments were then used to build phylogenetic reference trees using the program RAxML v 7.2.8 (Stamatakis, 2006). A maximum likelihood tree was inferred taking the best of 20 starting trees using the amino acid substitution matrix model which

resulted in the best likelihood score during a trial run (*arsR:* WAGF; *acr3:* WAGF; *arsC:* WAGF; *arsM*: VTF; *glnA:* RTREVF; *rpsD:* RTREVF; *tyrS:* WAG), empirical character frequencies, and a gamma model of rate heterogeneity with RAxML estimated alpha value. Bootstrap analyses were conducted on all trees at n=100. These trees, which contained hundreds of taxa (*arsR:* 295; *acr3:* 112; *arsC:* 105; *arsM*: 289; *glnA:* 207; *rpsD:* 99; *tyrS:* 110), were used for phylogenetic placement of metagenomic reads.

For ease of viewing the broader phylogenetic relationships within the gene trees, smaller trees were constructed. For the genes *arsR, arsC,* and *acr3*, a subset of sequences representing the major phylogenetic groups from the original tree were used to construct a smaller tree. For *arsM*, the phylogenetic tree was focused on the clades containing taxa with known arsenite-methyltransferase activity. In addition, arsenite-methyltransferases from eukaryotic algae and mammals were obtained from the OrthoMCL database, which groups orthologous genes from eukaryotic genomes (Chen *et al.*, 2006). For all genes, alignment of the smaller subset of sequences and tree building was similar to the methods described above for the reference trees.

Recruitment and placement of metagenomic reads:

We analyzed 38 metagenomes from the Global Ocean Sampling expedition (Supplemental Table 1). Only sequences from the 0.1-0.8 μm filter size fraction were used. In order for a site to be included in the analysis, we required each single copy core housekeeping gene to be detected at least once to ensure sufficient *Prochlorococcus* sequence reads. In addition, site GS000a was excluded from analysis because of possible contamination in this sample with non-marine *Shewanella* and *Burkholderia* (DeLong, 2005). The unassembled metagenomic sequence reads from the GOS dataset were downloaded from the CAMERA web portal (http://camera.calit2.net; (Sun *et al.*, 2011)) on Februrary 3rd, 2011. All work with the dataset was conducted locally.

For each gene, (*arsR, arsC, acr3, arsM, glnA, rpsD* and *tyrS*) relevant metagenomic reads were recruited with tblastn searches using blastall v 2.2.25 (Altschul *et al.*, 1997) starting with a *Prochlorococcus* MED4, MIT9313, NATL2A, SS120, and MIT9312 as well as *Synechococcus* WH8102 protein sequence as the query (Table 1). Recruited sequences with e-values $\leq$ 1 were maintained and analyzed. Sequence reads were trimmed according to their blast alignments to ensure translation in the appropriate reading frame.

For each gene, the recruited GOS reads were aligned to the reference taxa alignment in amino acid space using PaPaRa: Parsimony-based Phylogeny-Aware Read Alignment program (Berger and Stamatakis, 2011). The metagenomic reads in the PaPaRa alignment were then assigned to nodes of the reference phylogenetic tree using EPA: Evolutionary Placement Algorithm (Berger *et al.*, 2011). All reads that were placed with appropriate *Prochlorococcus* clades in the reference trees were identified as "*Prochlorococcus*-linked metagenomic reads". The relative abundance of genes in each metagenomic sample was calculated by normalizing the number of recruited reads to the length of the MED4 query gene and then comparing the length normalized abundance of each arsenic related gene to the average abundance of length normalized single copy core housekeeping genes. One way ANOVA tests were conducted to ensure that sampling sites within basins could be grouped. Mate pairs from reads identified as *Prochlorococcus*-linked were tblastn searched against a local nr database (downloaded 4/2013). In addition, the mate pairs identified as *Prochlorococcus*-linked *arsC* sequences were recruited to the AS9601 genome using the alignment program Bowtie2 (Langmead and Salzberg, 2012) and visualized with the genomic graphical viewer Tablet (Milne *et al.*, 2013).

Nutrient Concentration Estimation:

Annual surface statistical average phosphate concentrations were collected from the World Ocean Atlas 2009 (Garcia *et al.*, 2010), values were averaged over a one degree

latitude by one degree longitude square, centered around the coordinates of the metagenomic sample sites except for station GS108a which was averaged over a two by two degree grid due to lack of available data (Supplemental Table 1). Minimum surface phosphate concentrations across the same spatial extents were also collected from the World Ocean Atlas 2009.

Phosphate binding protein specificity:

Indiscriminate arsenate uptake is likely mediated through the cellular phosphate uptake system which includes the high affinity uptake transporter *pstS*. In order to evaluate whether the specificity of the binding sites of the various copies of *pstS* among *Prochlorococcus* have changed over time, we submitted *pstS* sequences to the Phyre2 protein prediction server (Kelley and Sternberg, 2009) as well as the 3DLigandSite server (Wass *et al.*, 2010).

2.4 Results:

Efflux Pathway Arsenic Detoxification Genes in* Prochlorococcus *Genomes:

Prochlorococcus strains vary in their complement of arsenic detoxification genes (Table 1). Arsenate reductase, *arsC,* which reduces inorganic pentavalent arsenate to the trivalent arsenite state, is present at one copy per genome in all 12 strains for which a genome sequence is available. There is no indication of other known arsenate reductase homologs, such as *acr2* or the arsenate-reductase form of *wzb* (Bennett *et al.*, 2001) in any strain. In contrast, the efflux transporter *acr3* is differentially present among the *Prochlorococcus* strains, with copies only present in the strains MIT9301, MED4, NATL1A, NATL2A, and MIT9303. There are no copies of the alternative *arsB* form of arsenite transporter among the

Prochlorococcus strains. Thus, it appears not all *Prochlorococcus* have the genetic capacity to expel inorganic arsenite from the cell. The repressive regulator, *arsR,* is also differentially present among the strains. All strains that possess a copy of *acr3* also have the regulator *arsR,* in addition *arsR* is also present in MIT9312, SS120, and MIT9313.

To understand the evolutionary mechanisms responsible for the variable complements of arsenic related genes in *Prochlorococcus*, we constructed phylogenetic trees of *arsC, arsR,* and *acr3* (Supplementary Figures 1-3) and examined their genomic context. Phylogenetic analysis of all 3 genes indicates a shared ancestry among the picocyanobacteria. For *acr3* the marine picocyanobacteria form a strongly supported monophyletic group. This group is also monophyletic in *arsC* and *arsR* trees, although with less support. For all three genes the phylogeny within the picocyanobacteria is generally congruent with ribosomal RNA phylogeny. Within the genomes of strains MED4, NATL1A, and NATL2A the genes *acr3* and *arsR* are in close genomic context to one another and to P acquisition genes associated with a genomic island (Martiny *et al.*, 2006), and show synteny with each other and with marine *Synechococcus*. In MIT9301 and MIT9303 only *arsR* is part of the P island and *acr3* is elsewhere. Genes associated with genomic islands in *Prochlorococcus* often contribute to specific environmental stress responses (Coleman *et al.*, 2006), it is not surprising to find colocation of arsenic-related genes with phosphorus stress inducible genes on a genomic island, as the processes of phosphorus stress and arsenic uptake are intertwined. In all genomes *arsC* is located in a different region of the genome than *acr3* and *arsR* but this location is syntenous across both *Prochlorococcus* and *Synechococcus.* The dispersed location of *arsC* is unusual as in most bacteria, the arsenic detoxification genes are co-located on an *ars* operon (Páez-Espino *et al.*, 2009). Taken together, phylogenetic inference and genomic synteny indicate that the arsenic detoxification genes *arsC, acr3,* and *arsR* are likely ancestral to the picocyanobacteria. Thus, the differential presence of these genes among

Prochlorococcus genomes (Table 1) is most likely due to multiple independent gene loss events.

Identification of putative Prochlorococcus *ArsM*

To determine whether *Prochlorococcus* also contains the previously unknown arsenic methylation pathway, we attempted to identify an ArsM by constructing an initial phylogenetic tree of a broad array of methyltransferases belonging to COGs 500 and 2226. Within this tree, we identified a clade which includes sequences of known arsenite *S*-adenosylmethyltransferase (*arsM*) function (Páez-Espino *et al.*, 2009; Qin *et al.*, 2006; Yin *et al.*, 2011) as well as marine picocyanobacterial sequences. A more focused tree was constructed combining sequences from this clade with eukaryotic arsenite *S*-adenosylmethyltransferases (Figure 2). With the exception of *Synechococcus* WH5701, the putative *arsM* sequences of the marine picocyanobacteria form a well supported monophyletic clade among the other *arsM* sequences. Putative *arsM* sequences were identified in the majority of *Prochlorococcus* and *Synechococcus* genomes, but were not found in *Prochlorococcus* strains NATL1A, NATL2A or SS120 (Table 1). Interestingly, a few other bacteria, including *Synechococcus* WH5701, branch within a well supported clade consisting largely of eukaryotic sequences (from both eukaryotic algae and animals). This suggests that while *arsM* is ancestral for most of the picocyanobacteria, this gene has been subject to horizontal transfer events.

In order to further identify whether these *Prochlorococcus* sequences are likely functioning *arsM* orthologs, protein structure models were created for the sequence of the putative *Prochlorococcus* MED4 *arsM* as well as sequences of known arsenite *S*-adenosylmethyltransferase function from the freshwater cyanobacteria *Nostoc sp.* PCC 7120, *Synechocystis sp.* PCC 6803, and *Microcystis aeruginosa* NIES843(Yin *et al.*, 2011). The

protein model template *c3qnha* from the *arsM* sequence of the eukaryotic red algae *Cyanidioschyzon merolae sp.* 5508 (*Cm-arsM*) was identified as the most likely model for all the cyanobacterial sequences at greater than 90% confidence in homology (Figure 3). Cysteine residues integral to the binding of arsenite to ArsM protein at positions 72, 174, and 224 in *CmArsM* (Ajees *et al.*, 2012) are conserved among all the cyanobacterial sequences, including the putative *Prochlorococcus* MED4 *arsM*. Overall secondary structure and *SAM*-binding motifs are also conserved. The similarity in these key catalytic residues among the cyanobacteria and the red algae indicate likely conservation of function.

Frequency of arsenic-related genes in wild Prochlorococcus *populations*

In order to understand the selective pressures on *Prochlorococcus* arsenic related genes in the environment we examined the frequency of these genes in global *Prochlorococcus* populations. The Global Ocean Survey sampling sites used for this analysis were predominately from the Sargasso Sea, Caribbean Sea, Eastern Pacific Tropical Upwelling Region, and the Indian Ocean (Supplemental Table 1). The three oceanic basins sampled captured three significantly different phosphate regimes (Figure 4) (ANOVA p-value <0.000). The Atlantic sites have the lowest annual average phosphate conditions (0.08 μmol/l), the Indian basin sites were higher at 0.18 μmol/l. The Pacific sites had the highest average phosphate concentrations (0.49 μmol/l) and also the greatest variability among sites. The same pattern of phosphate regimes was identified when comparing phosphate minima, with the Atlantic experiencing the most extreme phosphate minima, the Indian at moderate levels, and the Pacific sites experiencing the highest minimum phosphate conditions (ANOVA p-value < 0.000). Thus, this dataset effectively captured *Prochlorococcus* populations exposed to low, moderate, and relatively high phosphate conditions.

The presence of arsenic efflux detoxification genes compared to single copy core genes varies among global *Prochlorococcus* populations (Figure 5). The average occurrence

of the single copy core genes *glnA, rpsD,* and *tyrS*, were used as a proxy for the total number of *Prochlorococcus* genomes sampled. These genes occur at a rate of one copy per genome among cultured *Prochlorococcus* (Kettler *et al.*, 2007) and their occurrence in metagenomes did not significantly vary between basins (ANOVA p-value for *glnA*, *rpsD*, and *tyrS:* 0.112, 0.056, and 0.105, respectively). The majority of environmental reads for all genes were placed within the high light *Prochlorococcus* clades. This is not surprising as most of these sites were sampled from 5m depth or shallower. Among the global *Prochlorococcus* populations, the *arsR* regulator occurs at less than one copy per genome on average. The *acr3* arsenite efflux transporter also occurs in less than one copy per genome, and was not detected at all in some populations. In contrast the *arsC* arsenate reductase appears to be present at greater than one copy per genome on average in global *Prochlorococcus* populations. The putative *arsM* arsenite *S*-adenosylmethyltransferase is present at an average of one copy per genome among global *Prochlorococcus* populations (Figure 5.a)

We observed basin scale differences in the relative occurrence of some arsenic detoxification genes (Figure 5, Table 2). In the phosphate poor Atlantic sites the *arsR* regulator appears at about one copy per *Prochlorococcus* genome (Figure 5.b), but it is significantly less frequent in Indian and Pacific populations (ANOVA p-value <0.000). *acr3*, the component of the efflux pathway responsible for expulsion of arsenite from the cell, is present in Atlantic populations at less than one copy per *Prochlorococcus* genome, still significantly more (ANOVA p-value <0.000) than in Indian and Pacific basin populations, where it was not detected in many samples. In contrast, the *arsC* arsenate reductase gene occurs in two copies per *Prochlorococcus* genome in the Atlantic and Indian oceans, but this apparent duplication is not present in Pacific populations. The *arsM* methyltransferase occurs at about one copy per genome on average in all global *Prochlorococcus* populations. Thus *Prochlorococcus* populations in the Atlantic are capable of arsenic methylation and a portion

of the population maintains the genomic potential for arsenite efflux, whereas the *Prochlorococcus* populations in the Indian and Pacific basins contain the methylation pathway but have lost the capability for arsenite efflux.

2.5 Discussion:

Arsenic is the most abundant of environmental toxins with toxicity rooted in its chemical similarity to phosphorus. We suggest that organisms can be susceptible to arsenic toxicity, even at relatively low absolute arsenic concentrations, if phosphorus concentrations are also very low. The marine cyanobacteria *Prochlorococcus* thrives in one such low P environment, the tropical and subtropical oligotrophic oceans. *Prochlorococcus* strains possess differing genomic potentials for arsenic detoxification and the presence/absence pattern of these arsenic related genes is not congruent with evolutionary relationships based on rRNA genes. Our phylogenetic reconstructions found that the marine picocyanobacteria consistently form a monophyletic clade (Figure 2 & Supplementary Figs. 1-3) suggesting that multiple independent gene loss events are responsible for the observed pattern (Table 1). We suggest that the *acr3* arsenite efflux transporter and the *arsR* regulator are being lost where they are no longer needed. The potential for efflux through aquaglycerolporins, which are also leaky to arsenite(Oremland and Stolz, 2003), does not appear possible as there are no known aquaglycerolporins among the *Prochlorococcus* genomes. Notably the strains lacking both of these genes (MIT9515, MIT9211 and AS9601) were isolated from either the Equatorial Pacific or the Arabian Sea (Table 1; (Rocap *et al.*, 2002)) where P is not typically limiting. They are also the strains observed to have fewest P acquisition genes (Martiny *et al.*, 2006) indicating that the relaxation of pressures induced by low P has allowed the loss of these genes.

The relative occurrences of arsenic related genes in *Prochlorococcus* vary in wild populations as well, where populations from low phosphate regimes maintain a greater genomic potential for arsenic detoxification. The arsenic efflux transporter, and therefore the complete efflux detoxification pathway, is only significantly present in *Prochlorococcus* populations in the Atlantic Ocean, where annual average phosphate concentrations in the sites sampled here are 0.08 μmol/l. These results are consistent with previous work which found an enrichment of *Prochlorococcus*-linked *acr3* genes at the Bermuda Atlantic Time Series station compared to the Hawaii Ocean Time Series in the North Pacific (Coleman and Chisholm, 2010). *ArsR* may also be in the process of being lost from some *Prochlorococcus* populations, as it occurs at less than one copy per genome in the Indian and Pacific basins. We estimated a PO_4:AsO_4 ratio as 5:1, 11:1 and 30:1 for the Atlantic, Indian, and Pacific sites, respectively. Because data on arsenate concentrations are sparse this calculation used an arsenate concentration of 15 nmol/l (Cutter and Cutter, 2006; Cutter *et al.*, 2001; Ellwood and Maher, 2002) combined with site-specific annual average phosphate concentrations. Thus, the arsenic efflux detoxification pathway is not present in the majority of the *Prochlorococcus* populations globally, and may be a secondary detoxification pathway that is present only where risk of indiscriminate arsenate uptake is very high, which we suggest occurs when PO_4:AsO_4 ratios approach 5:1 or lower. These results suggest that the previously known efflux detoxification pathway is not a commonly used strategy for detoxification among *Prochlorococcus*, but rather the newly identified methylation strategy appears to be a more common strategy for mitigating exposure to arsenic in the environment.

Phosphate concentrations low enough to result in P-limitation may also result in enhanced arsenic uptake. The high affinity periplasmic phosphate binding protein PstS (Figure 1) plays an important role in the uptake of arsenate into cells. *PstS* expression significantly increases under both phosphate limitation and starvation (Martiny *et al.*, 2006;

Reistetter *et al.*, 2013) resulting in an increase in phosphate uptake rates compared to P-replete growth conditions (Krumhardt *et al.*, 2013). Thus, in a low PO_4:AsO_4 environment this phosphate stress response, which increases phosphate scavenging, may result in an increase in arsenic uptake. However, it is worth noting that an additional strategy for cellular arsenic defense is the prevention of indiscriminate uptake of arsenate in the first place. *Prochlorococcus* MIT9313, which lacks the *acr3* gene, possesses two copies of *pstS,* only one of which is expressed in response to phosphate stress (Martiny 2006). Recent work on the two periplasmic phosphate binding proteins in the extremely arsenic tolerant *Halomonas* strain GFAJ-1 from Mono Lake has shown that one of the paralogs can discriminate phosphate from arsenate 4,500 fold, compared to 500-fold discrimination of the other copy (Elias *et al.*, 2012). Protein secondary structure prediction combined with binding site specificity modeling indicate that the various *Prochlorococcus* PstS binding proteins likely vary in their ligand binding specificities. Thus, it is possible that some PstS proteins may have greater capability to discriminate arsenate from phosphate than others.

Arsenate reductase, *arsC*, is required by both the efflux pathway and the methylation pathway and is present in a single copy in all isolate *Prochlorococcus* strains. Strikingly it appears to be duplicated in environmental *Prochlorococcus* populations in the Atlantic and the Indian basins, but not in the Pacific, where it occurs at an average of one copy per genome. The duplicated copies of *arsC* in these populations could confer amplified or differentially regulated expression of arsenate reductase, or may be paralogous genes of another function, although currently there are no known paralogs of this evolutionary form of *arsC* (Mukhopadhyay *et al.*, 2002). Duplicate copies of an arsenate reductase have been identified in other organisms, including the freshwater cyanobacterium *Synechocystis* PCC 6803, where they are often found on plasmids (Kaneko et al. 2003, Páez-Espino et al, 2009). It is also possible the *arsC*s we detected are on cyanophage genomes. Although no *arsC*-

containing cyanophage genomes have been observed, *Prochlorococcus* phosphate-stress inducible genes, such as *pstS* and *phoH,* are commonly found on cyanophages (Sullivan *et al.*, 2005), and there is evidence that environmental phosphorus conditions are a main selective agent on cyanophage genomes (Kelly *et al.*, 2013). However, the associated mate pairs from *arsC* reads were also identified as *Prochlorococcus*-like and recruited to the same localized region of a *Prochlorococcus* genome, thus the source of the additional copies of *arsC* is unclear.

Prior to this work, the only known arsenic detoxification pathway in *Prochlorococcus* was the efflux pathway which relies on the ACR3 arsenite efflux transporter (Scanlan *et al.*, 2009). Here we identified a putative arsenite *S*-adenosylmethyltransferase, *arsM,* which is present in most sequenced *Prochlorococcus* strains and appears to be maintained at about one copy per *Prochlorococcus* genome globally. The presence of *arsM* indicates that *Prochlorococcus* likely has the capacity to methylate inorganic arsenicals into less toxic organic forms such as MMA(V), DMA(V), or the gas TMAO(g) and potentially to biosynthesize additional organoarsenical compounds such as arsenolipids and arsenosugars, which are commonly found in eukaryotic algae (Dembitsky and Levitsky, 2004). The widespread occurrence of *arsM* in global *Prochlorococcus* populations suggests that the methylation pathway may be part of a general metabolic strategy rather than serving solely in arsenic detoxification. For example, *Prochlorococcus* uses sulfolipids in place of phospholipids in the membrane, thereby reducing the cellular phosphorus demand (Van Mooy 2006; Van Mooy et al. 2009) and it is possible that synthesis of arsenolipids allows further reduction of cellular P quotas. For example, the eukaryotic algae *Chlorella* methylates inorganic arsenicals and produces polar arsenolipids (Dembitsky and Levitsky, 2004; Lunde, 1973). Arsenosugars are the predominate form of arsenic found in algae, and the freshwater cyanobacteria *Synechocystis* sp. PCC 6803 and *Nostoc* sp. PCC 7120 are also capable of

producing arsenosugars (Miyashita *et al.*, 2012). Recent work suggests that human ingestion of arsenosugars and arsenolipids might be harmful with potential for release of toxic metabolites with arsenolipids posing the potential for toxicity at levels comparable to inorganic arsenite (Harrington *et al.*, 2008; Newcombe *et al.*, 2010; Feldmann and Krupp, 2011; Raml *et al.*, 2009; Meyer *et al.*, 2014; Leffers *et al.*, 2013a; Leffers *et al.*, 2013b). Further laboratory work is needed to verify the functionality of *arsM* and to determine to what extent *Prochlorococcus* is capable of methylating arsenicals and what organic compounds are synthesized.

The ratio of phosphate to arsenate appears to exert influence on the biological conversion of inorganic arsenate to other forms, as the cycling of these two elements are inextricably linked. Along an open ocean transect that covered varying phosphate regimes, when the PO_4:AsO_4 ratio was less than 1, the total arsenic pool consisted of a greater proportion of reduced and methylated arsenicals compared to the relatively phosphate-rich region, with an inorganic PO_4:AsO_4 ratio of 100:1 (Cutter et al. 2006). In the North Atlantic both methylated arsenicals and arsenite detoxification products, have been identified and the spatial distribution of arsenite closely corresponded with phosphate availability and can be used in conjunction with measures of alkaline phosphatase to identify regions of phosphorus limitation (Wurl *et al.*, 2013). There is also evidence that the diazotrophic cyanobacterial communities in these phosphate poor regions are responding to arsenic toxicity. The expression of *arsC, arsA,* and *arsB* were detected in a metatranscriptome of a *Trichodesmium* bloom, suggesting microbial response to arsenic toxicity associated with nitrogen fixation (Hewson *et al.*, 2009). Additionally, expression of the *Crocosphaera watsonii* WH8105 arsenite efflux transporters increases in laboratory cultures when arsenate is present and the cells are phosphate stressed. In the oligotrophic Sargasso Sea, with a DIP concentration <0.01

μmol/l, a transcript for a *C. watsonii* WH8105 arsenic efflux transporter was successfully identified indicating an *in situ* response to arsenic toxicity (Dyhrman and Haley, 2011).

The sheer abundance of *Prochlorococcus*, which can grow at densities greater than 7 x 10^5 cells ml^{-1} (Campbell *et al.*, 1998) , makes it a key player in biogeochemical cycling of arsenic in the open ocean. As warming associated with climate change causes more intense and persistent stratification of the upper water column, phosphate scarcity and corresponding arsenic exposure will be exacerbated resulting in an increased cycling of arsenicals mediated by *Prochlorococcus*. The two detoxification pathways have both different end products and intermediary substrates with a broad range of associated toxicities. The previously known but less widespread efflux pathway expels arsenite, which is more toxic than arsenate, from the cell back into the environment where it can either be oxidized back into arsenate in a matter of hours-days (Cutter, 1992) or taken up by eukaryotic organisms, such as algae and fish, via diffusion, hexose permease transporters, and aquaporins (Ventura-Lima *et al.*, 2011). The methylation pathway is more effective at sequestering the toxic effects of arsenic as the detoxified products MMA(V) and DMA(V) have a residence time on the order of years rather than days (Cutter et al. 2006), and TMAO(g) may remove arsenic altogether from the marine system. However, the methylation and storage of organoarsenicals by *Prochlorococcus* may result in bioaccumulation of complexed organoarsenicals in upper trophic levels alone (Dembitsky and Levitsky, 2004). Prior to the identification of *arsM*, it appeared that active arsenic detoxification by *Prochlorococcus* in the environment would simply contribute to the transformation of inorganic arsenate to arsenite, which would then be rapidly oxidized back to arsenate – in addition, this interaction would only occur in the most phosphate stressed regions of the ocean. Identification of the widespread *arsM* gene among global *Prochlorococcus* populations indicates a more complex interaction between this organism

and environmental arsenic, with a more complicated impact on global arsenic cycling through the production of organoarsenicals.

2.6 Conclusion:

Our analysis suggests that *Prochlorococcus* utilizes arsenic detoxification mechanisms in oceanic regions of phosphorus scarcity where the PO_4:AsO_4 ratios are low. There are two detoxification pathways in *Prochlorococcus*, the efflux pathway (Scanlan *et al.*, 2009), and the putative arsenic methylation pathway identified by this work. The methylation pathway is nearly ubiquitous in global *Prochlorococcus* populations and likely confers some basal arsenic resistance; however, the efflux pathway is only maintained in populations which experience extreme phosphorus scarcity and have a greater possibility of arsenic uptake into the cell due to reduced PO_4:AsO_4 ratios in the environment. Our analysis also suggests that the ability to detoxify arsenic is an ancestral state in marine picocyanobacteria. It is vital to understand the products of the different arsenic biochemical pathways, their global occurrence, and the rates of detoxification, as the chemical transformations of arsenic by these two pathways are vastly different and have ecological implications for arsenic biogeochemical cycling.

2.7 Acknowledgements:

We thank Cedar McKay for help setting up local computational resources. We would also like to thank Simon Berger for providing amino acid support to the PaPaRa alignment program. This work was made possible through support from an National Science Foundation Graduate Research Fellowship to JKS and NSF OCE-1138368 to GR.

Conflict of Interest:

The authors declare no conflict of interest.

2.8 References:

1 Campbell L, Liu H, Nolla HA, Vaulot D (1997). Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991–1994 ENSO event. *Deep Sea Research Part I: Oceanographic Research Papers* **44:** 167-192.

2 DuRand MD, Olson RJ, Chisholm SW (2001). Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **48:** 1983-2003.

3 Scanlan D (2012). Marine Picocyanobacteria. In: Whitton BA (ed). *Ecology of Cyanobacteria II*. Springer Netherlands. pp 503-533.

4 Wu J, Sunda W, Boyle EA, Karl DM (2000). Phosphate Depletion in the Western North Atlantic Ocean. *Science* **289:** 759-762.

5 Krumhardt KM, Callnan K, Roache-Johnson K, Swett T, Robinson D, Reistetter EN *et al.* (2013). Effects of phosphorus starvation versus limitation on the marine cyanobacterium *Prochlorococcus* MED4 I: uptake physiology. *Environmental Microbiology* **15:** 2114-2128.

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

7 Van Mooy BAS, Rocap G, Fredricks HF, Evans CT, Devol AH (2006). Sulfolipids dramatically decrease phosphorus demand by picocyanobacteria in oligotrophic marine environments. *Proceedings of the National Academy of Sciences* **103:** 8607-8612.

8 Bertilsson S, Berglund O, Karl DM, Chisholm SW (2003). Elemental composition of marine *Prochlorococcus* and *Synechococcus*: Implications for the ecological stoichiometry of the sea. *Limnology and Oceanography* **48:** 1721-1731.

9 Heldal M, Scanlan DJ, Norland S, Thingstad F, Mann NH (2003). Elemental composition of single cells of various strains of marine *Prochlorococcus* and *Synechococcus* using X-ray microanalysis. *Limnology and Oceanography* **48:** 1732-1743.

10 Moore LR, Rocap G, Chisholm SW (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393:** 464-467.

11 Moore LR, Chisholm SW (1999). Photophysiology of the Marine Cyanobacterium *Prochlorococcus*: Ecotypic Differences among Cultured Isolates. *Limnology and Oceanography* **44:** 628-638.

12 Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311:** 1737-1740.

13 Moore LR, Anton FP, Rocap G, Chisholm SW (2002). Utilization of Different Nitrogen Sources by the Marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnology and Oceanography* **47:** 989-996.

14 Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:** e231.

15 Martiny AC, Coleman ML, Chisholm SW (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proceedings of the National Academy of Sciences* **103:** 12552-12557.

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

17 Martiny AC, Huang Y, Li W (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environmental Microbiology* **11:** 1340-1347.

18 Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological Genomics of Marine Picocyanobacteria. *Microbiology and Molecular Biology Reviews* **73:** 249-299.

19 Rosenberg H, Gerdes RG, Chegwidden K (1977). Two systems for the uptake of phosphate in *Escherichia coli*. *J Bacteriol* **131:** 505-511.

20 Tawfik DS, Viola RE (2011). Arsenate Replacing Phosphate: Alternative Life Chemistries and Ion Promiscuity. *Biochemistry* **50:** 1128-1134.

21 Mandal BK, Suzuki KT (2002). Arsenic round the world: a review. *Talanta* **58:** 201-235.

22 Oremland RS, Stolz JF (2003). The Ecology of Arsenic. *Science* **300:** 939-944.

23 Akter KF, Owens G, Davey DE, Naidu R (2005). Arsenic speciation and toxicity in biological systems. *Rev Environ Contam Toxicol* **184:** 97-149.

24 Stolz JF, Basu P, Santini JM, Oremland RS (2006). Arsenic and Selenium in Microbial Metabolism*. *Annual Review of Microbiology* **60:** 107-130.

25 Páez-Espino D, Tamames J, Lorenzo V, Cánovas D (2009). Microbial responses to environmental arsenic. *BioMetals* **22:** 117-130.

26 Carlin A, Shi W, Dey S, Rosen BP (1995). The *ars* operon of Escherichia coli confers arsenical and antimonial resistance. *J Bacteriol* **177:** 981-986.

27 Ghosh M, Shen J, Rosen BP (1999). Pathways of As(III) detoxification in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* **96:** 5001-5006.

28 Bennett MS, Guan Z, Laurberg M, Su XD (2001). Bacillus subtilis arsenate reductase is structurally and functionally similar to low molecular weight protein tyrosine phosphatases. *Proc Natl Acad Sci U S A* **98:** 13577-13582.

29 Xu C, Shi W, Rosen BP (1996). The chromosomal *arsR* gene of *Escherichia coli* encodes a trans-acting metalloregulatory protein. *The Journal of biological chemistry* **271:** 2427-2432.

30 Qin J, Rosen BP, Zhang Y, Wang G, Franke S, Rensing C (2006). Arsenic detoxification and evolution of trimethylarsine gas by a microbial arsenite S-adenosylmethionine methyltransferase. *Proc Natl Acad Sci U S A* **103:** 2075-2080.

31 Yin XX, Chen J, Qin J, Sun GX, Rosen BP, Zhu YG (2011). Biotransformation and volatilization of arsenic by three photosynthetic cyanobacteria. *Plant Physiol* **156:** 1631-1638.

32 Dembitsky VM, Levitsky DO (2004). Arsenolipids. *Prog Lipid Res* **43:** 403-448.

33 Thomas DJ, Nava GM, Cai SY, Boyer JL, Hernandez-Zavala A, Gaskins HR (2010). Arsenic (+ 3 oxidation state) methyltransferase and the methylation of arsenicals in the invertebrate chordate *Ciona intestinalis*. *Toxicol Sci* **113:** 70-76.

34 Andreae MO (1979). Arsenic speciation in seawater and interstitial waters: the influence of biological-chemical interactions on the chemistry of a trace element. *Journal Name: Limnol Oceanogr; (United States); Journal Volume: 24:3***:** Medium: X; Size: Pages: 440-452.

35 Cutter GA, Cutter LS (2006). Biogeochemistry of arsenic and antimony in the North Pacific Ocean. *Geochemistry Geophysics Geosystems* **7**.

36 Wurl O, Zimmer L, Cutter GA (2013). Arsenic and phosphorus biogeochemistry in the ocean: Arsenic species as proxies for P-limitation. *Limnology and Oceanography* **58:** 729-740.

37 Cutter GA, Cutter LS, Featherstone AM, Lohrenz SE (2001). Antimony and arsenic biogeochemistry in the western Atlantic Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* **48:** 2895-2915.

38 Karl DM, Tien G (1997). Temporal variability in dissolved phosphorus concentrations in the subtropical North Pacific Ocean. *Mar Chem* **56:** 77-96.

39 Ellwood MJ, Maher WA (2002). An automated hydride generation-cryogenic trapping-ICP-MS system for measuring inorganic and methylated Ge, Sb and As species in marine and fresh waters. *Journal of Analytical Atomic Spectrometry* **17:** 197-203.

40 Cutter GA, Cutter LS (1998). Metalloids in the high latitude North Atlantic Ocean: Sources and internal cycling. *Mar Chem* **61:** 25-36.

41 Karl DM (2007). The marine phosphorus cycle. ASM Press: Washington. pp 523-539.

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

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

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

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

46 Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D *et al.* (2009). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Research* **38:** D396-D400.

47 Yuan Y, Barry CE, 3rd (1996). A common mechanism for the biosynthesis of methoxy and cyclopropyl mycolic acids in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* **93:** 12828-12833.

48 Kelley LA, Sternberg MJE (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* **4:** 363-371.

49 Abascal F, Zardoya R, Telford MJ (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* **38:** W7-13.

50 Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

51 Chen F, Mackey AJ, Stoeckert CJ, Roos DS (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research* **34:** D363-D368.

52 DeLong EF (2005). Microbial community genomics in the ocean. *Nat Rev Microbiol* **3:** 459-469.

53 Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S *et al.* (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res* **39:** D546-551.

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

55 Berger SA, Stamatakis A (2011). Aligning short reads to reference alignments and trees. *Bioinformatics* **27:** 2068-2075.

56 Berger SA, Krompass D, Stamatakis A (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology* **60:** 291-302.

57 Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9:** 357-359.

58 Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L *et al.* (2013). Using Tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14:** 193-202.

59 Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Zweng MM, Baranova OK *et al.* (2010). World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate). *NOAA Atlas NESDIS 71*. U.S. Government Printing Office: Washington, D.C. p 398.

60 Wass MN, Kelley LA, Sternberg MJ (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res* **38:** W469-473.

61 Ajees AA, Marapakala K, Packianathan C, Sankaran B, Rosen BP (2012). Structure of an As(III) S-Adenosylmethionine Methyltransferase: Insights into the Mechanism of Arsenic Biotransformation. *Biochemistry* **51:** 5476-5485.

62 Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68:** 1180-1191.

63 Coleman ML, Chisholm SW (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci U S A* **107:** 18634-18639.

64 Reistetter EN, Krumhardt K, Callnan K, Roache-Johnson K, Saunders JK, Moore LR *et al.* (2013). Effects of phosphorus starvation versus limitation on the marine cyanobacterium *Prochlorococcus* MED4 II: gene expression. *Environmental Microbiology* **15:** 2129-2143.

65 Elias M, Wellner A, Goldin-Azulay K, Chabriere E, Vorholt JA, Erb TJ *et al.* (2012). The molecular basis of phosphate discrimination in arsenate-rich environments. *Nature* **491:** 134-137.

66 Mukhopadhyay R, Rosen BP, Phung LT, Silver S (2002). Microbial arsenic: from geocycles to genes and enzymes. *FEMS Microbiol Rev* **26:** 311-325.

67 Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005). Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. *PLoS Biology* **3:** 790-806.

68 Kelly L, Ding HM, Huang KH, Osburne MS, Chisholm SW (2013). Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. *Isme J* **7:** 1827-1841.

69 Lunde G (1973). The synthesis of fat and water soluble arseno organic compounds in marine and limnetic algae. *Acta Chem Scand* **27:** 1586-1594.

70 Miyashita S-i, Fujiwara S, Tsuzuki M, Kaise T (2012). Cyanobacteria produce arsenosugars. *Environmental Chemistry* **9:** 474-484.

71 Harrington CF, Brima EI, Jenkins RO (2008). Biotransformation of arsenobetaine by microorganisms from the human gastrointestinal tract. *Chemical Speciation and Bioavailability* **20:** 173-180.

72 Newcombe C, Raab A, Williams PN, Deacon C, Haris PI, Meharg AA *et al.* (2010). Accumulation or production of arsenobetaine in humans? *Journal of Environmental Monitoring* **12:** 832-837.

73 Feldmann J, Krupp EM (2011). Critical review or scientific opinion paper: arsenosugars--a class of benign arsenic species or justification for developing partly speciated arsenic fractionation in foodstuffs? *Anal Bioanal Chem* **399:** 1735-1741.

74 Raml R, Raber G, Rumpler A, Bauernhofer T, Goessler W, Francesconi KA (2009). Individual variability in the human metabolism of an arsenic-containing carbohydrate, 2',3'-dihydroxypropyl 5-deoxy-5-dimethylarsinoyl-beta-D-riboside, a naturally occurring arsenical in seafood. *Chemical research in toxicology* **22:** 1534-1540.

75 Meyer S, Matissek M, Muller SM, Taleshi MS, Ebert F, Francesconi KA *et al.* (2014). In vitro toxicological characterisation of three arsenic-containing hydrocarbons. *Metallomics* **6:** 1023-1033.

76 Leffers L, Wehe CA, Huwel S, Bartel M, Ebert F, Taleshi MS *et al.* (2013a). *In vitro* intestinal bioavailability of arsenosugar metabolites and presystemic metabolism of thio-dimethylarsinic acid in Caco-2 cells. *Metallomics* **5:** 1031-1042.

77 Leffers L, Ebert F, Taleshi MS, Francesconi KA, Schwerdtle T (2013b). *In vitro* toxicological characterization of two arsenosugars and their metabolites. *Molecular nutrition & food research* **57:** 1270-1282.

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

79 Dyhrman ST, Haley ST (2011). Arsenate Resistance in the Unicellular Marine Diazotroph *Crocosphaera watsonii*. *Front Microbiol* **2:** 214.

80 Campbell L, Landry MR, Constantinou J, Nolla HA, Brown SL, Liu H *et al.* (1998). Response of microbial community structure to environmental forcing in the Arabian Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **45:** 2301-2325.

81 Cutter GA (1992). Kinetic controls on metalloid speciation in seawater. *Mar Chem* **40:** 65-80.

82 Ventura-Lima J, Bogo MR, Monserrat JM (2011). Arsenic toxicity in mammals and aquatic animals: A comparative biochemical approach. *Ecotoxicology and Environmental Safety* **74:** 211-218.

83 Rosen BP (2002). Biochemistry of arsenic detoxification. *FEBS Lett* **529:** 86-92.

2.9 Figure Legends & Tables:

Figure 2.1 Diagram representing an idealized cross-section of a *Prochlorococcus* cell highlighting mechanisms likely responsible for arsenic entry into the cell and detoxification. *(1)* Arsenate likely enters the cell through the inorganic high affinity phosphate transporters PstCABS. *(2)* Once arsenate is inside the cytoplasm it is reduced by the arsenate reductase ArsC to the (III) oxidation state, arsenite. *(3)*Arsenite can then bind to the *trans*-acting ArsR repressive regulator, *(4)* arsenite can be recognized by the arsenite efflux transporter ACR3 which shuttles the toxin out of the cell, or *(5)* the arsenite can be methylated by the putative arsenite *S*-adenosylmethyltransferase, ArsM. Monomethylarsonic acid (MMA) can undergo a series of oxidation and methylation steps to further methylate the species to dimethylarsinic acid (DMA); further methylation into more compound organoarsenicals may also be possible, but cannot be confirmed without laboratory analysis.

Figure 2.2 Phylogeny of arsenite *S*-adenosylmethyltransferases (*ArsM*). The phylogenetic tree was constructed using the maximum-likelihood program RAxML. The statistical significance of the branch pattern was estimated by conducting a 100 bootstrap replications

of the original amino acid alignment; bootstraps $\geq$ 50 shown. The related *SAM* dependent mycolic acid cyclopropane synthetase (*cmaS*) gene was used as an outgroup. Bold sequences represent sequences of genes with confirmed *ArsM* function (Qin *et al.*, 2006; Yin *et al.*, 2011; Rosen, 2002).

Figure 2.3 Amino acid pairwise alignment of cyanobacterial sequences to the red algae *Cyanidioschyzon merolae sp.* 5508 ArsM. Predicted secondary structures generated by the fully automated protein homology/analogy recognition engine Phyre2 (Kelley & Sternberg 2009) are depicted below the amino acids. Red arrows above amino acid alignment indicate conserved cysteine residues identified to be integral for catalysis of the arsenic methylation reaction in *C. merolae* (Ajees 2012). Red boxes indicate conserved motifs of regions involved in *SAM* binding to the protein. Purple loops in predicted secondary structure symbolize alpha helices and teal arrows symbolize predicted beta sheets.

Figure 2.4 Boxplot of the statistical average annual phosphate concentration at the surface for the GOS metagenome sites analyzed in this study according to ocean basin location. Outliers are represented by stars.

Figure 2.5 Boxplots of the relative *Prochlorococcus* linked arsenic detoxification gene abundance within the GOS metagenomes of the Atlantic, Pacific, and Indian basins. The shaded region represents one copy of a gene per genome; the spread of this shading around 1 represents the error associated with the estimation of *Prochlorococcus* genomes within the basin using single copy core gene abundance. (a) Average across all basins, (b) Atlantic sites, (c) Indian sites, and (c) Pacific sites. Outliers are represented by stars.

Table 2.1. Presence/Absence of arsenic detoxification genes identified within genomes of Prochlorococcus strains. Abbreviation: COG, Clusters of Orthologous Groups. Presence of gene marked by locus tag, absence represented by '–'. Associated COGs for relevant genes also denoted. Prochlorococcus genomes are listed in order of physiological light optima which follows the 16S rRNA phylogeny. [a]*acr3* referred to as '*arsB*' in Scanlan et al., (2009) and as '*arsA*' in Martiny et al. (2006).

			arsR	*acr3*[a]	*arsC*	*arsM*
			Regulator	*Efflux transporter*	*Reductase*	*Methyltransferase*
Strain	*Clade*	*Isolate Region*	COG640	COG798	COG1393	COG500
MED4	High Light 1	Med. Sea	PMM0714	PMM0716	PMM0512	PMM0416
MIT9515	High Light 1	Eq. Pacific	—	—	P9515_05761	P9515_04771
AS9601	High Light 2	Arabian Sea	—	—	A9601_05691	A9601_04661
MIT9301	High Light 2	Sargasso Sea	P9301_12351	P9301_12581	P9301_05391	P9301_04351
MIT9215	High Light 2	Eq. Pacific	—	—	MIT9215_706	MIT9215_808
MIT9312	High Light 2	Gulf Stream	PMT9312_0726	—	PMT9312_0513	PMT9312_0411
NATL2A	Low Light 1	N. Atlantic	PMN2A_0446	PMN2A_0447	PMN2A_1845	—
NATL1A	Low Light 1	N. Atlantic	NATL1_11581	NATL1_11591	NATL1_05701	—
SS120	Low Light 2	Sargasso Sea	Pro1578	—	Pro0511	—
MIT9211	Low Light 3	Eq. Pacific	—	—	P9211_05131	P9211_04141
MIT9313	Low Light 4	Gulf Stream	PMT1001	—	PMT1256	PMT1171
MIT9303	Low Light 4	Sargasso Sea	P9303_11041	P9303_11311	P9303_07481	P9303_08531

Table 2.2. Statistical analysis of relative abundance of arsenic detoxification genes in the various ocean basins. p-values < 0.05 in bold; p-values <0.01 with *. A one-way ANOVA was conducted on each gene set, and significantly different gene sets (p-value < 0.05) were analyzed for pairwise differences between basins using Tukey's post-hoc test.

			Tukey Test		
	ANOVA p-value	Partial Eta Squared	Atlantic vs. Indian	Atlantic vs. Pacific	Indian vs. Pacific
ArsR	**0.002***	0.306	**0.007***	**0.002***	0.737
ArsC	**0.006***	0.252	0.987	**0.044**	**0.008***
ACR3	**0.000***	0.641	**0.000***	**0.000***	0.808
ArsM	0.355	0.057	N/A	N/A	N/A
$[PO_4]$	**0.000***	0.679	0.165	**0.000***	**0.000***

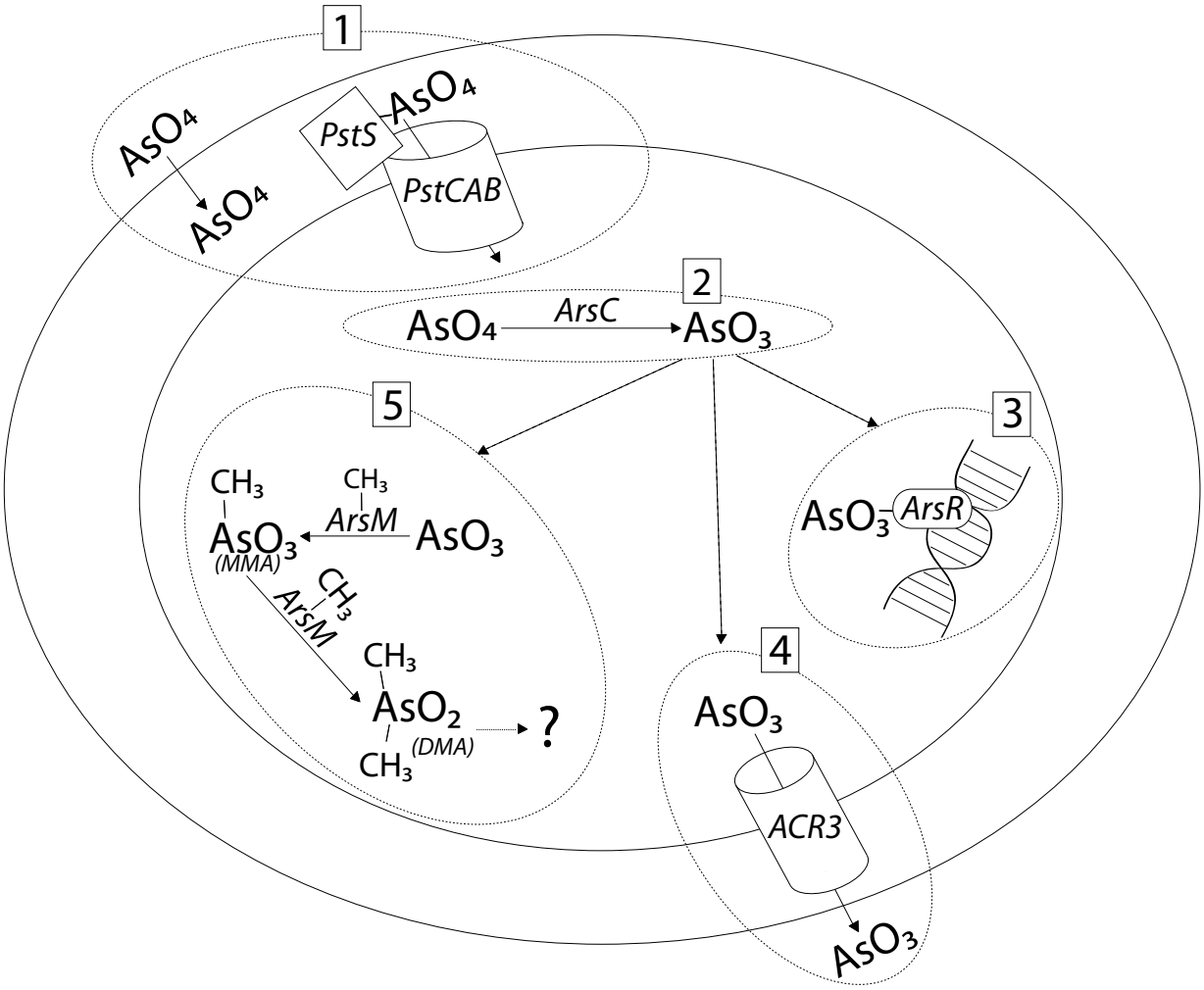

1
AsO4
AsO4
PstS
AsO4
PstCAB
2
AsO4
ArsC
AsO3
5
CH3
CH3
ArsM
AsO3
AsO3
(MMA)
CH3
ArsM
CH3
AsO2
?
CH3
(DMA)
3
AsO3
ArsR
4
AsO3
ACR3
AsO3

Figure 2.1

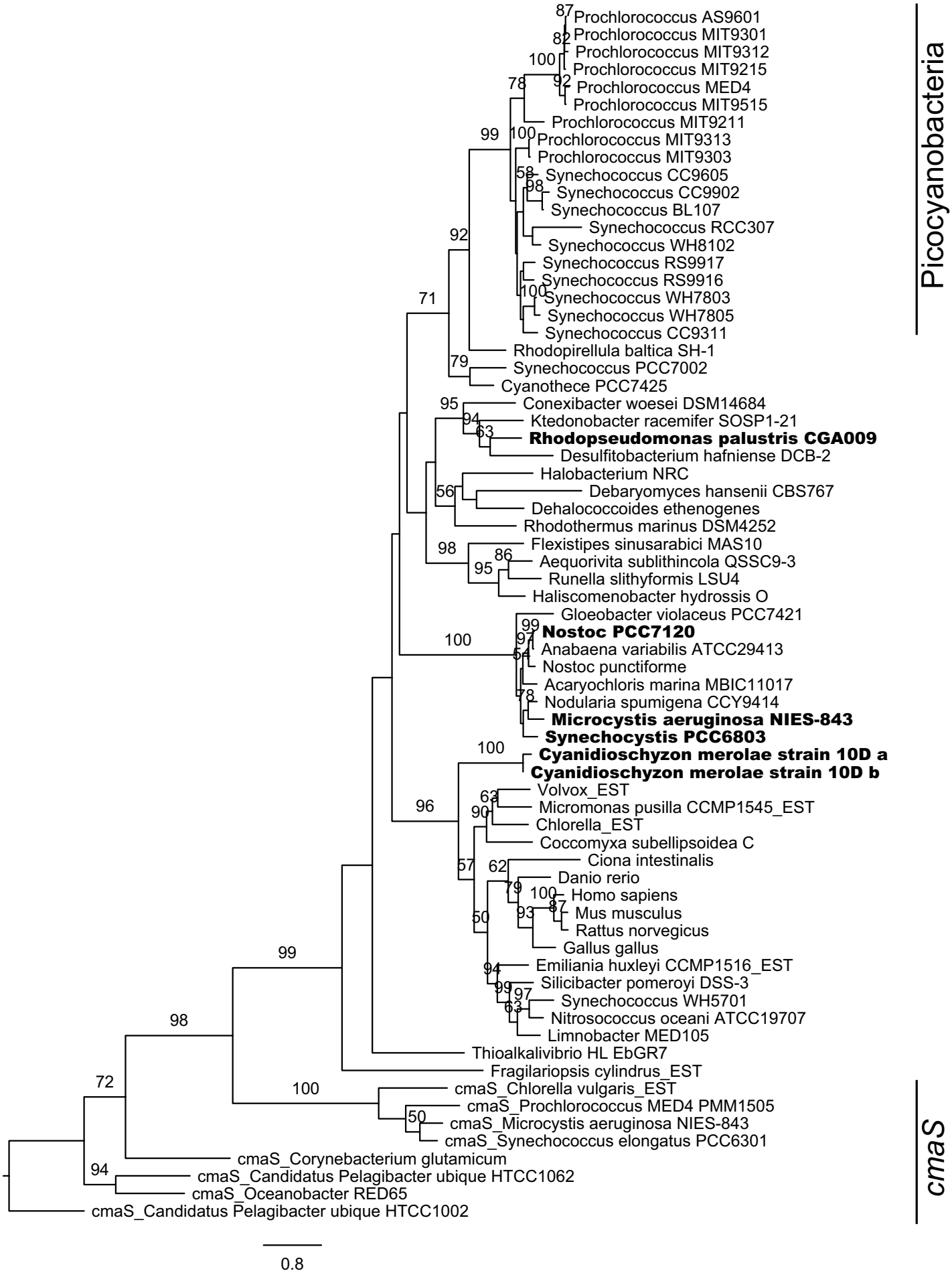

Prochlorococcus AS9601
Prochlorococcus MIT9301
Prochlorococcus MIT9312
Prochlorococcus MIT9215
Prochlorococcus MED4
Prochlorococcus MIT9515
Prochlorococcus MIT9211
Prochlorococcus MIT9313
Prochlorococcus MIT9303
Synechococcus CC9605
Synechococcus CC9902
Synechococcus BL107
Synechococcus RCC307
Synechococcus WH8102
Synechococcus RS9917
Synechococcus RS9916
Synechococcus WH7803
Synechococcus WH7805
Synechococcus CC9311
Rhodopirellula baltica SH-1
Synechococcus PCC7002
Cyanothece PCC7425
Conexibacter woesei DSM14684
Ktedonobacter racemifer SOSP1-21
Rhodopseudomonas palustris CGA009
Desulfitobacterium hafniense DCB-2
Halobacterium NRC
Debaryomyces hansenii CBS767
Dehalococcoides ethenogenes
Rhodothermus marinus DSM4252
Flexistipes sinusarabici MAS10
Aequorivita sublithincola QSSC9-3
Runella slithyformis LSU4
Haliscomenobacter hydrossis O
Gloeobacter violaceus PCC7421
Nostoc PCC7120
Anabaena variabilis ATCC29413
Nostoc punctiforme
Acaryochloris marina MBIC11017
Nodularia spumigena CCY9414
Microcystis aeruginosa NIES-843
Synechocystis PCC6803
Cyanidioschyzon merolae strain 10D a
Cyanidioschyzon merolae strain 10D b
Volvox_EST
Micromonas pusilla CCMP1545_EST
Chlorella_EST
Coccomyxa subellipsoidea C
Ciona intestinalis
Danio rerio
Homo sapiens
Mus musculus
Rattus norvegicus
Gallus gallus
Emiliania huxleyi CCMP1516_EST
Silicibacter pomeroyi DSS-3
Synechococcus WH5701
Nitrosococcus oceani ATCC19707
Limnobacter MED105
Thioalkalivibrio HL EbGR7
Fragilariopsis cylindrus_EST
cmaS_Chlorella vulgaris_EST
cmaS_Prochlorococcus MED4 PMM1505
cmaS_Microcystis aeruginosa NIES-843
cmaS_Synechococcus elongatus PCC6301
cmaS_Corynebacterium glutamicum
cmaS_Candidatus Pelagibacter ubique HTCC1062
cmaS_Oceanobacter RED65
cmaS_Candidatus Pelagibacter ubique HTCC1002
0.8
Picocyanobacteria
cmaS

Figure 2.2

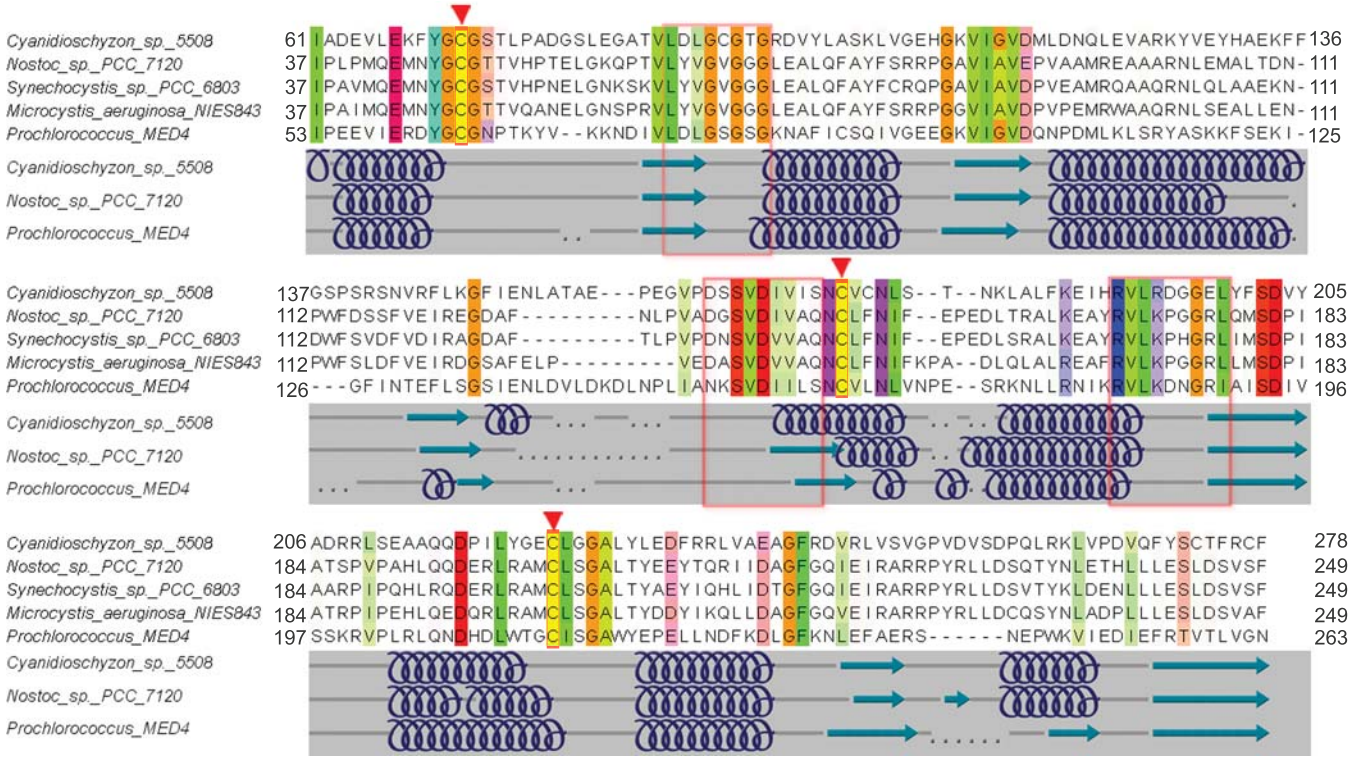

Cyanidioschyzon_sp._5508
Nostoc_sp._PCC_7120
Synechocystis_sp._PCC_6803
Microcystis_aeruginosa_NIES843
Prochlorococcus_MED4

Figure 2.3

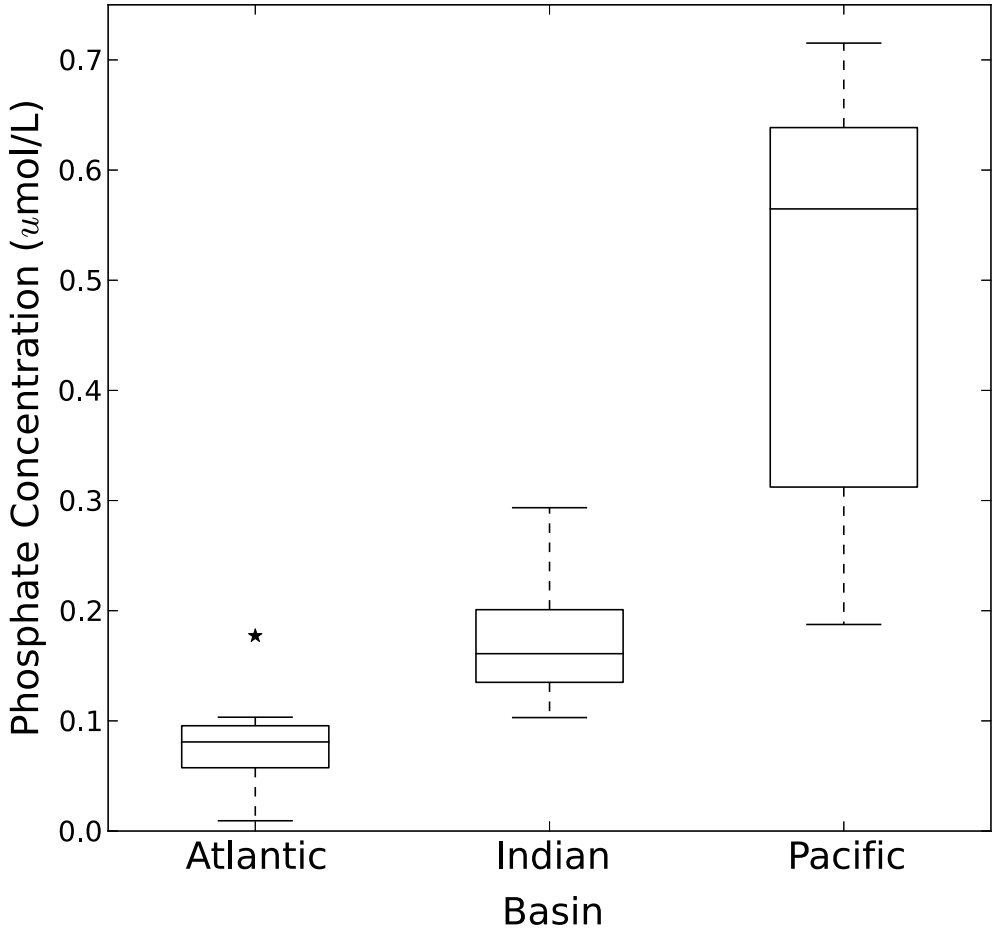

Phosphate Concentration ($u$mol/L)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Atlantic
Indian
Pacific
Basin

Figure 2.4

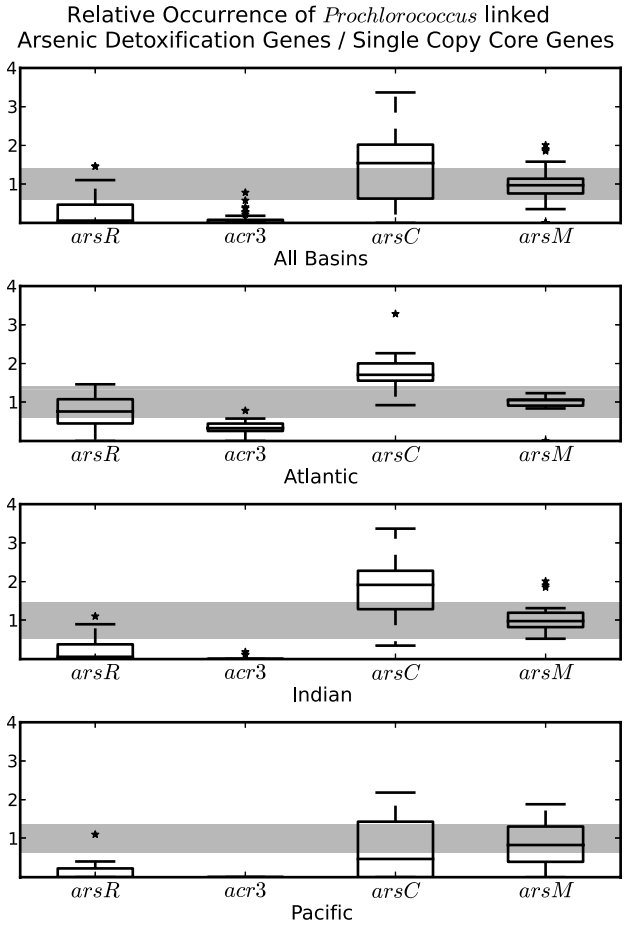

Relative Occurrence of $Prochlorococcus$ linked
Arsenic Detoxification Genes / Single Copy Core Genes
4
3
2
1
$arsR$
$acr3$
$arsC$
$arsM$
All Basins
Atlantic
Indian
Pacific

Figure 2.5

2.10 Supplementary Figure Legends & Tables:

Supplemental Figure 2.1 Phylogeny of the arsenite-binding *trans*-acting repressive regulator (*ArsR*) showing the cyanobacteria and representatives of broader bacterial taxonomic groups from the larger tree used for environmental read placements. The phylogenetic tree was constructed using the maximum-likelihood program RAxML. The statistical significance of the branch pattern was estimated by conducting a 100 bootstrap replications of the original amino acid alignment; bootstraps > 50 shown.

Supplemental Figure 2.2 Phylogeny of the arsenite efflux transporter (*ACR3*) showing the cyanobacteria and representatives of broader bacterial taxonomic groups from the larger tree used for environmental read placements. The phylogenetic tree was constructed using the maximum-likelihood program RAxML. The statistical significance of the branch pattern was estimated by conducting a 100 bootstrap replications of the original amino acid alignment; bootstraps > 50 shown.

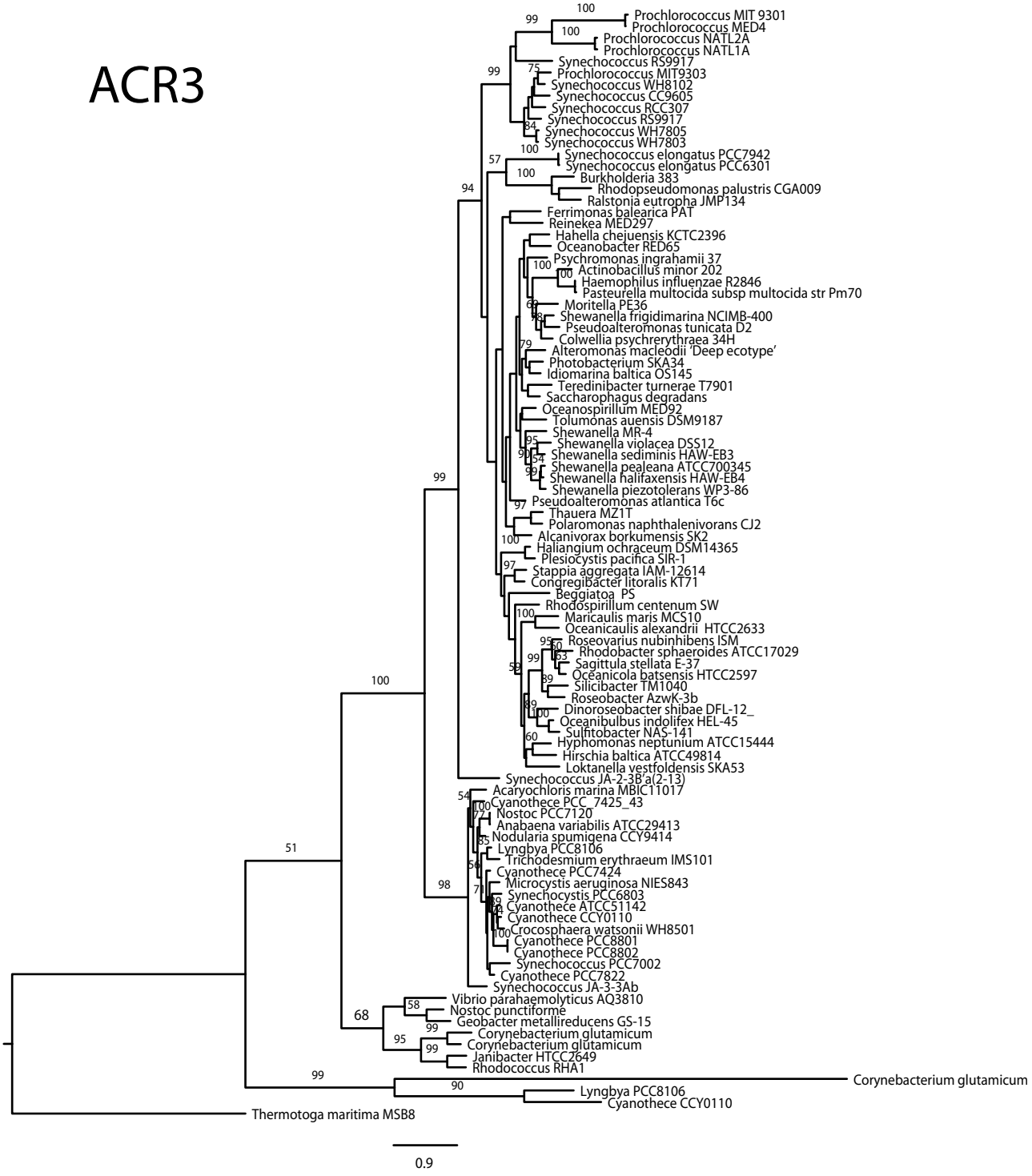
Supplemental Figure 2.3 Phylogeny of arsenate reductase (*ArsC*) showing the cyanobacteria and representatives of broader bacterial taxonomic groups from the larger tree used for environmental read placements. The phylogenetic tree was constructed using the maximum likelihood program RAxML. The statistical significance of the branch pattern was estimated by conducting a 100 bootstrap replications of the original amino acid alignment; bootstraps > 50 shown.

Supplemental Table 2.1 Global Ocean Survey (GOS) metagenomic samples used for analysis in this study. [PO4], represents the statistical average annual phosphate concentration in the surface in a 1 degree x 1 degree box around the metagenomic sample location; phosphate data gathered from the World Ocean Atlas 2009, except for GS1108a which was averaged over a 2 degree x 2 degree grid due to lack of available data.

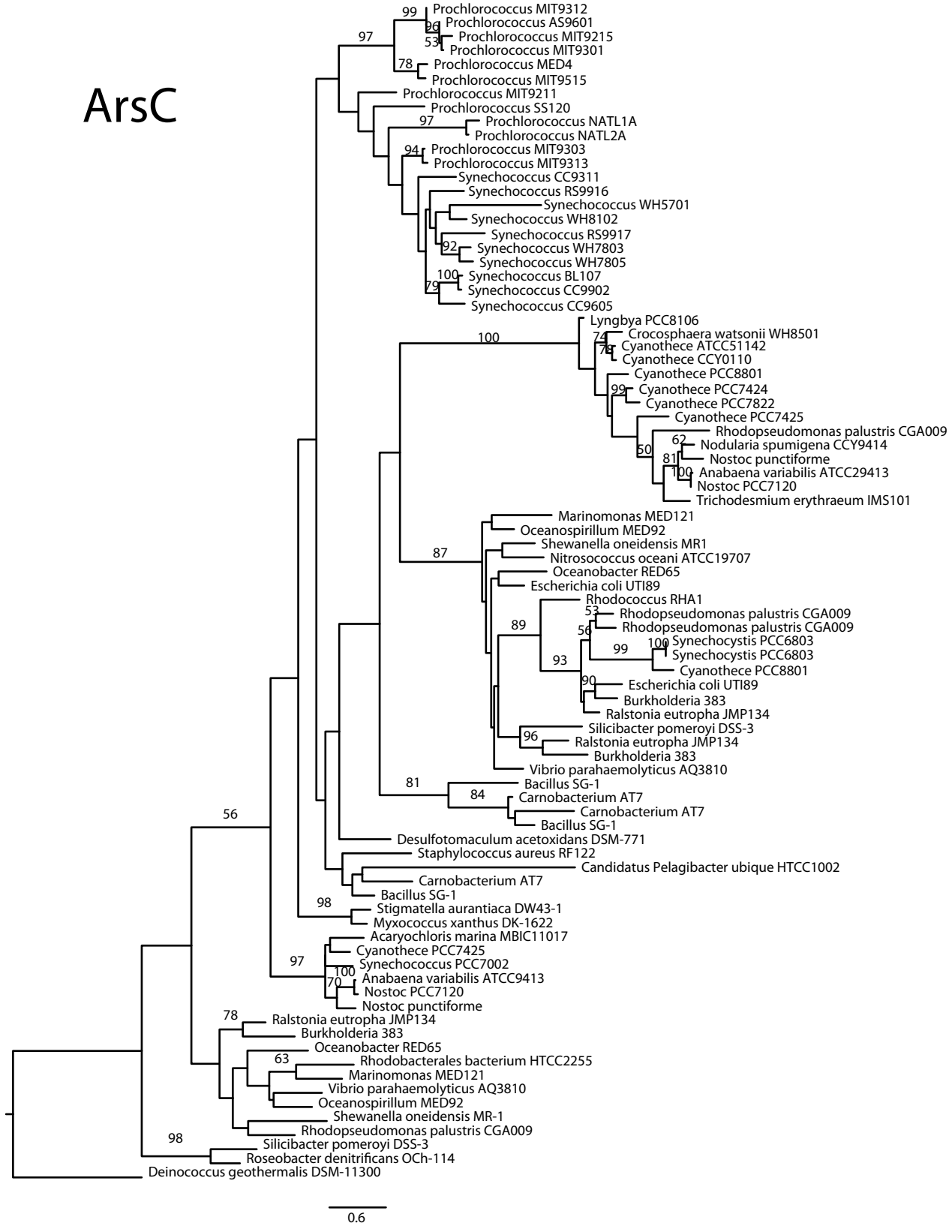
GOS_loc	Site Desc	location	[PO4]	Lat (N)	Lon (W)
GS000b	Open Ocean	Sargasso	0.0093	31.175	-64.324
GS000d	Open Ocean	Sargasso	0.0093	31.175	-64.324
GS014	Coastal	NW_Atlantic	0.0930	32.507	-79.264
GS015	Coastal	Carribbean	0.0861	24.488	-83.070
GS016	Coastal Sea	Carribbean	0.0735	24.175	-84.344
GS017	Open Ocean	Carribbean	0.1773	20.523	-85.414
GS018	Open Ocean	Carribbean	0.1033	18.037	-83.785
GS019	Coastal	Carribbean	0.0757	10.716	-80.254
GS022	Open Ocean	Equatorial_Pac	0.6393	6.493	-82.904
GS023	Open Ocean	Equatorial_Pac	0.2747	5.640	-86.565
GS025	Fringing Reef	Equatorial_Pac	0.2747	5.553	-87.088
GS026	Open Ocean	Equatorial_Pac	0.5258	1.264	-90.295
GS027	Coastal	Equatorial_Pac	0.6430	-1.216	-90.423
GS028	Coastal	Equatorial_Pac	0.6430	-1.217	-90.320
GS030	Warm Seep	Equatorial_Pac	0.6363	0.272	-91.633
GS031	Coastal upwelling	Equatorial_Pac	0.6363	-0.301	-91.652
GS032	Mangrove	Equatorial_Pac	0.7153	-0.594	-91.069
GS035	Coastal	Equatorial_Pac	0.4250	1.389	-91.817
GS036	Coastal	Equatorial_Pac	0.6038	-0.021	-91.198
GS047	Open Ocean	Equatorial_Pac	0.4485	-10.131	-135.449
GS048a	Coral Reef	Equatorial_Pac	0.1875	-17.476	-149.812
GS049	Coastal	Equatorial_Pac	0.1875	-17.453	-149.799
GS108a	Open Ocean	Indian_Ocean	0.1030	-12.093	96.882
GS109	Open Ocean	Indian_Ocean	0.1600	-10.944	92.059
GS110a	Open Ocean	Indian_Ocean	0.1725	-10.446	88.303
GS111	Open Ocean	Indian_Ocean	0.1330	-9.597	84.198
GS112a	Open Ocean	Indian_Ocean	0.1270	-8.505	80.376
GS113	Open Ocean	Indian_Ocean	0.1620	-7.008	76.331
GS114	Open Ocean	Indian_Ocean	0.1943	-4.990	64.977
GS115	Open Ocean	Indian_Ocean	0.2210	-4.663	60.523
GS116	Open Ocean	Indian_Ocean	0.2445	-4.635	56.836
GS117a	Coastal sample	Indian_Ocean	0.2935	-4.614	55.509

GS119	Open Ocean	Madagascar	0.1350	-23.216	52.306
GS120	Open Ocean	Madagascar	0.1783	-26.035	50.123
GS121	Open Ocean	Madagascar	0.1350	-29.349	43.216
GS122a	Open Ocean	Madagascar	0.1575	-30.898	40.420
GS123	Open Ocean	Madagascar	0.1553	-32.399	36.592
GS148	Fringing Reef	Madagascar	0.2550	-6.317	39.009

Supplementary Figure 2.1



Supplementary Figure 2.2



Supplementary Figure 2.3

Chapter 3

Arsenic-based metabolisms in oxygen deficient zones: potential for another cryptic metabolic cycle in marine waters

3.1 Abstract/ Introduction:

Oxygen Deficient Zones (ODZs) are functionally anoxic regions of the water column where oxygen concentrations are below 4 nmol L^{-1} (Ulloa *et al.*, 2012; Tiano *et al.*, 2014). These mid-layer oxygen deficient zones, sandwiched between oxygenated surface and deep layers, are caused through a combination of organic matter respiration and slow mid-layer horizontal circulation of water masses common in eastern ocean basins along the subtropics (Wyrtki, 1962). Redox gradients through the ODZ water column impact the distribution and respiratory processes of the associated microbial communities (Wright *et al.*, 2012). Aerobic metabolisms are replaced by alternative electron acceptors, with the next best electron acceptor being nitrate (Lam and Kuypers, 2011), followed by nitrite with evidence of a cryptic sulfur cycle, which is active but lacking in overt chemical expression (Canfield *et al.*, 2010), supported by the autotrophic oxidation of sulfide from the use of sulfate as an electron acceptor for dissimilatory sulfate reduction (Canfield *et al.*, 2010). Arsenic is another powerful redox sensitive element, with a redox potential of + 130 mV between the As(V)/As(III) couple (Oremland *et al.*, 2009), which supports a complete microbial metabolic cycle - analogous to cryptic sulfur cycling - of autotrophic oxidation of As(III) and dissimilatory heterotrophic reduction of As(V) along the chemocline of other aquatic systems, like the arsenic rich Mono Lake (Oremland and Stolz, 2003). Here we demonstrate the genetic capacity, through gene presence and expression, for a complete metabolic arsenic cycle in global marine ODZs and identify an arsenotrophic-related enzyme that may support autotrophic carbon fixation through a reductive TCA pathway in these

regions. The presence of these metabolic pathways suggests that arsenic based metabolisms support organic matter production in modern oceans, likely impact nitrogen biogeochemical cycling, and may have been a major metabolism in early Precambrian oceans when anoxia pervaded and arsenic concentrations were higher (Chi Fru *et al.*, 2015).

3.2 Results & Discussion:

Microbes utilizing arsenic for energetic gains have been studied and isolated in a range of systems with moderate to high arsenic concentrations, including polluted and unpolluted soils, sediments, hot springs, lakes, wastewater and mine drainage (Amend *et al.*, 2014). This is the first account of arsenotrophic cycling in an environment with nano-molar levels of available arsenic (Amend *et al.*, 2014). A complete metabolic arsenic cycle is likely ancient, existing since the Archaean (Duval *et al.*, 2008; Lebrun, 2003), with fossil evidence of arsenic-rich organic globules supporting a complete metabolic arsenic cycle 2.72 billion years ago (Sforna *et al.*, 2014). While the open ocean is not considered an arsenic-loaded system inorganic arsenic is still quite abundant in marine waters, ranging from 1-3 µg/L (Neff, 1997), and thermodynamically stable in the inorganic form as arsenate in oxygenated aqueous environments [As(V) as $H_2AsO_4^-$ & $HAsO_4^{2-}$] and as arsenite in anoxic environments [As(III) as $H_3AsO_3^0$ & $H_2AsO_3^-$] (Oremland and Stolz, 2003). Previous studies of arsenic in the anoxic waters of the Black Sea (Cutter, 1992) and the anoxic fjord of Saanich Inlet (Peterson and Carpenter, 1983) have demonstrated that arsenite concentrations in these marine anoxic waters are not congruent with what would be expected from thermodynamics alone. Radio-labeled arsenic experiments were carried out on the anoxic waters from Saanich Inlet with rates of arsenite oxidation greatly diminished when antibiotics were added to the seawater (Peterson and Carpenter, 1983) suggesting microbially mediated impacts on the redox state of arsenic in these anoxic waters. We hypothesize that this

thermodynamic imbalance in anoxic marine waters may be, at least partially, due to the presence of arsenotrophic microbes active in marine water columns.

An oceanographic research cruise was conducted through the Eastern Tropical North Pacific (ETNP) ODZ in April of 2012 where samples for metagenomic analysis were taken at multiple depths through the anoxic heart of the ODZ (Supplementary Figure 1) at station 136, including both free-living and particle-attached (>30 μm) fractions at station 141. These metagenomes were assembled into contigs, where complete gene sequences for arsenite oxidation (*aioAB*) and dissimilatory arsenate reduction (*arrAB*) were identified. Alpha subunits of metabolic arsenic cycling enzymes are all found within the DMSO reductase superfamily, also known as the complex iron sulfur molybdoenzymes (CISM) (Rothery *et al.*, 2008). Arsenotrophic enzymes identified in these ETNP metagenomes were all identified through phylogenetic inference within the larger CISM group of enzymes, which also contains other enzymes critical in respiratory redox transitions like the nitrate reductases Nap & Nar, formate dehydrogenases Fdn & Fdh, polysulfide reductase Psr, and tetrathionate reductase Ttr among others (Rothery *et al.*, 2008). For comparison to a well-studied analogous metabolic cycle previously known to exist in these ODZs, we also identified sequences associated with cryptic sulfur cycling, namely the bacterial forward and reverse dissimilatory sulfite reductase subunit *a* (*dsrA)*.

A complete dissimilatory arsenate reductase alpha subunit (*arrA*) sequence was identified in the ETNP ODZ metagenome assemblies (gene_44_ArrA_allDepths_contig_2) (Figure 1.a). Dissimilatory arsenate-reducing prokaryotes (DARPs) are often anaerobic heterotrophs which conserve metabolic energy from the reduction of arsenate, often using organic compounds as electron donors with occasional examples of DARPs using hydrogen or reduced sulfur as donors

(Amend *et al.*, 2014). The contig was assembled from a targeted recruitment of reads pulled from metagenomes located in anoxic waters where total free arsenic content is ~20 nmol at these depths in open ocean Pacific waters (Cutter and Cutter, 2006). Notably, no reverse arsenate reductase (*arxA*) sequences were assembled – *arxA* resembles the respiratory arsenate reductase of the DARPs *(arr*) in sequence, but functions in a reverse direction by oxidizing arsenite to arsenate (Richey *et al.*, 2009). In addition to the dissimilatory arsenate reductase alpha subunit, the beta subunit was also identified downstream of *arrA* on contig ArrA_allDepths_contig_2 (Figure 1.b) and is phylogenetically more closely related to the beta subunits of known ArrB sequences than to ArxB sequences (Supplementary Figure 2).

The arsenite oxidase enzyme AioA is evolutionarily distinct from the ArrA/ArxA group and is found elsewhere on the DMSO reductase superfamily tree (Supplementary Figure 3). The AioA enzyme can be used in both dissimilatory arsenite oxidation, believed to likely act as a detoxification strategy (Amend *et al.*, 2014), and in chemoautotrophic arsenite oxidation where metabolic gains are made from the oxidation of As(III) to As(V) (van Lis *et al.*, 2013). Most known chemoautotrophic arsenite oxidizers are strict mesophilic aerobes (Amend *et al.*, 2014; van Lis *et al.*, 2013), however there are a few known representatives which can utilize nitrate as the terminal electron acceptor (Oremland *et al.*, 2002; Rhine *et al.*, 2007; Rhine *et al.*, 2006), with some strains capable of reduction to N_2 (Rhine *et al.*, 2006). Phylogenetic analysis of assembled sequences from the ETNP ODZ identified a complete *aioA* sequence on contig AioA_allDepths_contig_0 (Figure 2.a). This environmental *aioA* sequence from the ETNP falls within a clade of sequences from taxa with known chemoautotrophic arsenite oxidation capabilities. A full beta subunit *aioB* located upstream and a complete cytochrome-c located downstream, were also captured on the *aioA*-containing contig, a similar genomic organization to

arsenite oxidizers (van Lis *et al.*, 2013). The *aioB* subunit is also phylogenetically more closely related to beta subunits of chemoautotrophic arsenite oxidizers (Supplementary Figure 4) than to dissimilatory arsenite oxidizers. Therefore, it is likely that the environmental assembled contig containing *aioA* is representative of a chemoautotrophic arsenite oxidation pathway.

An enzyme closely related to AioA of unknown function previously identified as "AioA-Like" (Duval *et al.*, 2008) branches just outside the internal node defining the AioA bacterial and archaeal clade in the DMSO reductase superfamily tree (Fig 2). This divergent AioA-Like clade consists of a subclade made up of a handful of Alpha- and Gammaproteobacterial sequences and another subclade of sequences recently identified in the haloarchaea (Euryarchaeota phylum) (Rascovan *et al.*, 2016). Our phylogenetic analysis of assembled arsenotrophy sequences in the ETNP ODZ indicates a preponderance of *aioA-like* sequences compared to *aioA* (Figure 2). Our environmental *aioA-like* sequences grouped more closely with the bacterial subclade than with the haloarchaeal group. Rascovan *et al.* (2016) identified an abundance of these *aioA-like* sequences in an haloarchaeal red biofilm in the arsenic-loaded Diamante Lake and suggested that the presence of *aioA-like* sequences in this small group of Proteobacteria arose from a horizontal gene transfer event from a Euryarchaeotal ancestor, separate from the ancestral origin of *aioA* where phylogenetic radiation of the AioA enzyme is more congruent with the 16S rRNA phylogeny (Rascovan *et al.*, 2016). This *aioA-like* sequence may indeed be a functional homolog to the large arsenite oxidase subunit *aioA*. However, without further laboratory confirmation we cannot definitively say whether this *aioA-like* enzyme functions as an As(III) oxidase or whether its function has taken another evolutionary trajectory.

Multiple assembled contigs from within the anoxic waters of the ETNP AMZ contained *aioA-like* sequences, with a long contig over 82 kilobases assembled from a single metagenomic

sample (ETNP_120m_NODE_73101) providing additional insight into the metabolic capabilities of the host microbe (Supplemental Table 1). In addition to containing the alpha and beta subunits of *aio-like*, this contig also contains a nitrate reductase sequence (*napA*) which suggests the use of nitrate as an electron acceptor. Contig ETNP_120m_NODE_73101 also contains the enzyme fumarate reductase, a notable key enzyme for the reductive TCA carbon fixation pathway (Hügler and Sievert, 2011), suggesting the potential for a chemoautotrophic lifestyle of the microbe. The co-occurrence of these genes on a contig suggests that the AioA-Like enzyme may help support autotrophic carbon fixation in these anoxic regions.

Using a phylogenetically informed short-read placement approach, we evaluated whether there was a higher density of metagenomic short reads associated with the genes *aioA, aioA-like, arrA,* and *dsrA* in free-living (<30 μm) or particle associated (>30 μm) microbial communities (Figure 3). We also identified short-reads associated with the bacterial form of the sulfur-cycling gene *dsrA,* the forward form used for sulfur reduction and the reverse used for sulfur oxidation (Muller *et al.*, 2015). We have included this sulfur cycling gene for comparison of arsenotrophy reads to an analogous known metabolic cycle present in these regions (Canfield *et al.*, 2010). *arrA, aioA-like,* and forward bacterial *dsrA* of sulfate reducing bacteria (SRB) were more abundant in the particle associated communities when compared to the free-living fraction. It makes sense to find *arrA* and SRB *dsrA* inhabiting similar functional environments as they are likely both exploiting similar redox-driven niches in particles, coupling the heterotrophic breakdown of organic matter with the reduction of arsenate or sulfur compounds when energetically favorable (Wright *et al.*, 2012). Short reads associated with the reverse bacterial *dsrA* associated with sulfate oxidizing bacteria (SOB) do not appear to be preferentially found within either the free-living or particle-associated fractions. Reads associated with *aioA* are

preferentially found within the free-living fraction -- this is a novel finding as most known arsenotrophic microbes are found within substrate-bound and arsenic-loaded environments (Amend *et al.*, 2014). In addition, the occurrence of metagenomic reads associated with the enzyme of unknown function AioA-Like was on the same order as reads associated with the reverse *dsrA* (SOB) associated with particle-associated sulfur oxidation, indicating that the community genomic capacity for these two enzymes is roughly equivalent on particulates (Figure 3.b).

Through the metagenomic assemblies, we identified the genomic potential in the microbial community inhabiting the ETNP ODZ for dissimilatory arsenate reduction and chemoautotrophic arsenite oxidation. We analyzed publicly available metatranscriptomes in global ODZ's to identify whether these arsenotrophic & arsenotrophy-related genes are transcriptionally active in these functionally anoxic pelagic environments (Ganesh *et al.*, 2015; Schunck *et al.*, 2013; Stewart *et al.*, 2012). Using phylogenetically informed short read placement, we identified transcripts associated with *arrA, aioA,* and *aioA-like* from metatranscriptomic libraries in the ETNP (Ganesh *et al.*, 2015) and the Eastern Tropical South Pacific (ETSP) (Schunck *et al.*, 2013; Stewart *et al.*, 2012) ODZs. For comparison, we also identified transcripts associated with bacterial *dsrA,* both the forward reducing form (SRB) and the reverse oxidizing form (SOB). Of the arsenotrophic genes, transcripts associated with *aioA-like* were by far the most abundant (Figure 4) and found in both the ETNP and the ETSP ODZs. Transcript sequences associated with *aioA* and *arrA* were also found in both the ETNP and ETSP. Dissimilatory arsenate reductase *arrA* and forward bacterial *dsrA* were only found in the metatranscriptome associated with a sulfidic plume in the ETSP AMZ (Schunck *et al.*, 2013). The arsenotrophy-related transcripts indicate that these genes are indeed being expressed at a

transcriptional level comparable to the forward bacterial *dsrA* sequence by the ODZ microbial communities, suggesting the active use of these metabolic pathways. The AioA-Like enzyme of unknown function was transcriptionally active on the same order of the sulfur oxidizing form of *dsrA* (SOB) in the general ETNP & ETSP metatranscriptomes with a notable exception being the ETSP metatranscriptome associated with a sulfide plume, lending support to the hypothesis that this AioA-Like enzyme is not sulfur-related.

3.3 Conclusion:

The presence of these complete arsenotrophic gene pathways in the metagenome of the ETNP ODZ suggests microbial communities in these anoxic waters are capable of a complete bioenergetic arsenotrophic cycle. Gene expression of the aresenotrophy genes *arrA* and *aioA,* indicates that the organisms here are actively transcribing these enzymes suggesting that arsenotrophy is a viable metabolic pathway utilized in these unique anoxic water columns. Dissimilatory arsenate reduction may contribute to dissimilatory nitrate reduction to ammonia (DNRA) which may subsequently support anaerobic ammonia oxidation (anammox) processes in the ODZ (Wright *et al.*, 2012), as has been suggested for dissimilatory sulfate reduction in these regions (Canfield *et al.*, 2010). Chemoautotrophic arsenite oxidation may provide a fixed carbon source in these anoxic waters, contributing to the biochemical processes driving the reducing environment and nitrogen loss found here through the reduction of nitrogenous compounds, potentially to N_2 as has been demonstrated by some chemoautotrophic arsenite oxidizers (Rhine *et al.*, 2006). Laboratory work should be conducted to confirm the function of AioA-Like as this enzyme of unknown function is likely a major player in the cycling of carbon, nitrogen, and potentially arsenic in anoxic marine waters, potentially to a similar level as sulfur oxidation processes. Further study of marine water column arsenotrophic cycling, through incubations and

tracer experiments, is needed in order to better understand the biogeochemical implications of these pathways on not only arsenic cycling in ODZs, but also on how these metabolisms may impact carbon cycling and denitrification processes as well. Arsenic based metabolisms are ancient with molecular evidence suggesting *aioA* was present in the last universal common ancestor (van Lis *et al.*, 2013) and fossil evidence suggesting a complete microbial arsenic cycle in the Neoarchean (Sforna *et al.*, 2014) with marine arsenic concentrations fluctuating over geologic time with periods of greater arsenic load in the marine system than modern day (Chi Fru *et al.*, 2015). Arsenotrophic organisms may have flourished in anoxic marine waters during the periods of high arsenic load, with chemoautotrophic arsenite oxidation supporting carbon fixation. These arsenotrophic metabolisms may also support life on volcanically active marine systems found on exoplanetary bodies such as Europa (Oremland and Stolz, 2003). These arsenotrophic communities we identified in modern marine oxygen deficient zones can be studied as proxies for arsenic based metabolisms in early anoxic oceans as well as help us understand the potential for arsenic based metabolisms which may have implications for the biogeochemical cycling of arsenic as well as carbon and nitrogen.

3.4 Methods:

Samples were collected during a research cruise to the East Tropical North Pacific in April 2012 aboard the R/V Thompson using 10 L Niskin bottles in a 24 bottle CTD-rosette. Dissolved oxygen concentration was determined using the SBE 43 dissolved oxygen sensor attached to the CTD rosette. Nutrient samples were filtered (GF/F glass fiber) before analysis. Nutrient analyses were performed by members of the University of Washington Marine Chemistry Laboratory on board the ship using a Technicon AAII system as described by the

World Ocean Circulation Experiment (WOCE) Hydrographic Program protocol (Gordon and Olson, 1995).

DNA samples were obtained from station 136 (-106.543° W 17.043° N) in the anoxic zone at multiple depths (100m, 110m, 120m, 160m, 180m). Four liters of Niskin water was vacuum filtered onto a 0.2 μm SUPOR filter. At the offshore multi-day station BB2/141 (107.148° W 16.527° N), approximately four liters were filtered through >30 μm filters at 120m and prefiltered (<30 μm) water was then filtered onto 0.2 μm SUPOR filters.

DNA samples were extracted using freeze thaw followed by incubation with lysozyme and proteinase K and phenol/chloroform extraction. A Rubicon THRUPLEX kit was used for library prep using 50 ng of DNA per sample. 4 libraries were sequenced on a HiSeq 2500 in rapid mode (~25 million 150 bp paired-end reads per sample) at Michigan State. The other 6 libraries were sequences on a HiSeq 2500 high output mode (~40-70 million 125 bp paired-end reads per sample) at the University of Utah. Sequences were quality checked and trimmed using Trimmomatic (Bolger *et al.*, 2014).

Metagenomic reads from each depth sample were first assembled separately, processed with diginorm (Brown *et al.*, 2012) to normalize coverage and then *de novo* assembled with the VELVET assembler (Zerbino and Birney, 2008). Contigs assembled from individual depths were then functionally annotated using the prokka annotation pipeline (Seemann, 2014). Prokka alone was insufficient to identify DMSO reductase family-related sequences (*aioA, aioA-like, arrA,* and *arxA*). In order to identify these arsenotrophy-related enzymes, blast databases of all assembled contigs were created using blast version 2.2.28 (Altschul *et al.*, 1990; Altschul *et al.*, 1997) and queried with tblastn using known reference queries. Contigs which had hits to

arsenotrophy-related sequences with e-values <=1e-50 were parsed out, and submitted to MetaGeneMark for gene prediction (Besemer and Borodovsky, 1999; Zhu *et al.*, 2010). Potential genes that overlapped with the region of best blast hit were then added to the DMSO reductase tree for further identification, with full length sequences that grouped with arsenotrophy-related clades depicted on the DMSO reductase tree (Supplementary Figure 3).

In order to assemble the longest possible arsenotrophy-related contigs, a targeted assembly approach was used to pull together metagenomes from all anoxic samples. Partial and full arsenotrophy-related sequences pulled from assembled contigs from individual depths and reference regions from known arsenotrophy-related sequences (enzyme sequence with 10kb regions up- & downstream) were pulled together as queries for metagenomics read recruitment among individual depth metagenomes using the program FR-HIT (Niu *et al.*, 2011). Reads recruited with FR-HIT were then assembled with the *de novo* assembler IDBA-UD (Peng *et al.*, 2012) optimized for assembling data of uneven depth coverage. Identification of arsenotrophy-related genes from these contigs was similar to previous identification. In addition, all genes called on contigs identified as containing full-length arsenotrophy-related enzymes were annotated with InterProScan (Jones *et al.*, 2014) with general taxonomic identity on a gene-by-gene basis being inferred at the bacterial class level through best a blastn best hit to the nr/nt database.

Gene trees were constructed with amino acid translated sequences of our environmental genes and known reference sequences aligned with the program MUSCLE version 3.6 (Edgar, 2004). Phylogenetic trees were then constructed with the program RAxML version 8.0.23. A maximum likelihood tree was inferred from the best of 20 trees using the amino acid substitution matrix model which resulted in the best likelihood score during a trial run (DMSO:

BLOSUM62F; DsrA, AioB, and ArrB: WAGF) along with empirical character frequencies and a gamma model of rate heterogeneity using the RAxML estimated alpha value. Bootstrap analyses were conducted on all tree at *n*=100. The DMSO reductase family tree and DsrA tree were used as reference trees for phylogenetic placement of metagenomic and metatranscriptomic reads.

Publicly available metatranscriptomes from oxygen deficient zones in the Eastern Tropical North Pacific (Ganesh *et al.*, 2015), Eastern Tropical South Pacific (Stewart *et al.*, 2012), and within a sulfide plume in the Eastern Tropical South Pacific (Schunck *et al.*, 2013) were downloaded locally. Reads were combined into a single blast database using blast 2.2.28 (Altschul *et al.*, 1990; Altschul *et al.*, 1997) and queried with tblastn (e-value cutoff 1e-5)using full-length known reference sequences, including full length identified environmental arsenotrophy sequences assembled in this work. Reads were also recruited for identification of bacterial d*srA* for comparison. Translation of blast recruited reads of at least 33 amino acids after quality trimming were then identified using a phylogenetically-informed placement approach by comparison to known reference sequences (Appendix I, Saunders & Rocap, 2016).

A similar approach was used for the recruitment and placement of metagenomic reads from the samples at 120 m depth from station BB2/141 comparing the >30 µm (particulate) and <30 µm (free-living) fractions. In order to obtain the abundance of these reads relative to the overall prokaryotic community in the two size fractions, the length normalized number of identified read pairs for each gene was compared to the length normalized abundance of the single copy core gene RNA polymerase (*rpoB*) in the sequenced prokaryotic community.

3.5 Acknowledgements:

I would like to thank Gabrielle Rocap for her guidance and editorial feedback on this work. I would also like to thank Clara Fuchsman who collected the water samples and sequenced the metagenomes as well as provided helpful feedback. Thanks to Cedar McKay who contributed through technical support with handling the very large datasets and provided key early metagenomic *de novo* assemblies. I would also like to thank John Baross and Rika Anderson for helpful discussions and feedback on this project. This work was supported through a NASA Earth and Space Sciences Fellowship to Jaclyn K. Saunders and National Science Foundation grants OCE-1138368 and DEB-1542240 to Gabrielle Rocap.

3.6 References:

1 Ulloa O, Canfield DE, DeLong EF, Letelier RM, Stewart FJ (2012). Microbial oceanography of anoxic oxygen minimum zones. *Proceedings of the National Academy of Sciences of the United States of America* **109:** 15996-16003.

2 Tiano L, Garcia-Robledo E, Dalsgaard T, Devol AH, Ward BB, Ulloa O *et al.* (2014). Oxygen distribution and aerobic respiration in the north and south eastern tropical Pacific oxygen minimum zones. *Deep Sea Research Part I: Oceanographic Research Papers* **94:** 173-183.

3 Wyrtki K (1962). The oxygen minima in relation to ocean circulation. *Deep Sea Research and Oceanographic Abstracts* **9:** 11-23.

4 Wright JJ, Konwar KM, Hallam SJ (2012). Microbial ecology of expanding oxygen minimum zones. *Nat Rev Micro* **10:** 381-394.

5 Lam P, Kuypers MMM (2011). Microbial Nitrogen Cycling Processes in Oxygen Minimum Zones. In: Carlson CA, Giovannoni SJ (eds). Annual Reviews: Palo Alto. pp 317-345.

6 Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, Delong EF *et al.* (2010). A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. *Science* **330:** 1375-1378.

7 Oremland RS, Saltikov CW, Wolfe-Simon F, Stolz JF (2009). Arsenic in the Evolution of Earth and Extraterrestrial Ecosystems. *Geomicrobiol J* **26:** 522-536.

8 Oremland RS, Stolz JF (2003). The Ecology of Arsenic. *Science* **300:** 939-944.

9 Chi Fru E, Arvestål E, Callac N, El Albani A, Kilias S, Argyraki A *et al.* (2015). Arsenic stress after the Proterozoic glaciations. *Scientific Reports* **5:** 17789.

10 Amend JP, Saltikov C, Lu G-S, Hernandez J (2014). Microbial Arsenic Metabolism and Reaction Energetics. *Reviews in Mineralogy and Geochemistry* **79:** 391-433.

11 Duval S, Ducluzeau AL, Nitschke W, Schoepp-Cothenet B (2008). Enzyme phylogenies as markers for the oxidation state of the environment: The case of respiratory arsenate reductase and related enzymes. *BMC Evol Biol* **8**.

12 Lebrun E (2003). Arsenite Oxidase, an Ancient Bioenergetic Enzyme. *Molecular Biology and Evolution* **20:** 686-693.

13 Sforna MC, Philippot P, Somogyi A, van Zuilen MA, Medjoubi K, Schoepp-Cothenet B *et al.* (2014). Evidence for arsenic metabolism and cycling by microorganisms 2.7 billion years ago. *Nature Geosci* **7:** 811-815.

14 Neff JM (1997). Ecotoxicology of arsenic in the marine environment. *Environmental Toxicology and Chemistry* **16:** 917-927.

15 Cutter GA (1992). Kinetic controls on metalloid speciation in seawater. *Mar Chem* **40:** 65-80.

16 Peterson ML, Carpenter R (1983). Biogeochemical process affecting total arsenic and arsenic species distribution in an intermittently anoxic fjord. *Mar Chem* **12:** 295-321.

17 Rothery RA, Workun GJ, Weiner JH (2008). The prokaryotic complex iron-sulfur molybdoenzyme family. *Biochimica et biophysica acta* **1778:** 1897-1929.

18 Cutter GA, Cutter LS (2006). Biogeochemistry of arsenic and antimony in the North Pacific Ocean. *Geochemistry Geophysics Geosystems* **7**.

19 Richey C, Chovanec P, Hoeft SE, Oremland RS, Basu P, Stolz JF (2009). Respiratory arsenate reductase as a bidirectional enzyme. *Biochemical and Biophysical Research Communications* **382:** 298-302.

20 van Lis R, Nitschke W, Duval S, Schoepp-Cothenet B (2013). Arsenics as bioenergetic substrates. *Biochim Biophys Acta-Bioenerg* **1827:** 176-188.

21 Oremland RS, Hoeft SE, Santini JM, Bano N, Hollibaugh RA, Hollibaugh JT (2002). Anaerobic oxidation of arsenite in Mono Lake water and by a facultative, arsenite-oxidizing chemoautotroph, strain MLHE-1. *Appl Environ Microbiol* **68:** 4795-4802.

22 Rhine ED, Ni Chadhain SM, Zylstra GJ, Young LY (2007). The arsenite oxidase genes (aroAB) in novel chemoautotrophic arsenite oxidizers. *Biochemical and Biophysical Research Communications* **354:** 662-667.

23 Rhine ED, Phelps CD, Young LY (2006). Anaerobic arsenite oxidation by novel denitrifying isolates. *Environmental Microbiology* **8:** 899-908.

24 Rascovan N, Maldonado J, Vazquez MP, Farias ME (2016). Metagenomic study of red biofilms from Diamante Lake reveals ancient arsenic bioenergetics in haloarchaea. *Isme J* **10:** 299-309.

25 Hügler M, Sievert SM (2011). Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean. *Annual review of marine science* **3:** 261-289.

26 Muller AL, Kjeldsen KU, Rattei T, Pester M, Loy A (2015). Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. *ISME J* **9:** 1152-1165.

27 Ganesh S, Bristow LA, Larsen M, Sarode N, Thamdrup B, Stewart FJ (2015). Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. *ISME J* **9:** 2682-2696.

28 Schunck H, Lavik G, Desai DK, Großkopf T, Kalvelage T, Löscher CR *et al.* (2013). Giant Hydrogen Sulfide Plume in the Oxygen Minimum Zone off Peru Supports Chemolithoautotrophy. *PLoS One* **8:** e68661.

29 Stewart FJ, Ulloa O, DeLong EF (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* **14:** 23-40.

30 Gordon A, Olson DB (1995). WHP Cruise Summary Information of section I09N. *WOCE*.

31 Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114-2120.

32 Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv preprint arXiv:12034802*.

33 Zerbino DR, Birney E (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome research* **18:** 821-829.

34 Seemann T (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30:** 2068-2069.

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

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

37 Besemer J, Borodovsky M (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Research* **27:** 3911-3920.

38 Zhu W, Lomsadze A, Borodovsky M (2010). *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Research* **38:** e132.

39 Niu B, Zhu Z, Fu L, Wu S, Li W (2011). FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* **27:** 1704-1705.

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

41 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C *et al.* (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*.

42 Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32:** 1792-1797.

3.7 Figure Legends:

Figure 3.1 (a) A maximum likelihood tree of dissimilatory arsenate reductase (Arra) and the closely related arsenite oxidase (ArxA) including full *de novo* assembled environmental ArrA sequences from the ETNP ODZ metagenome in bold. (b) the gene complement found along the contig ArrA_allDepths_contig_2 showing the presence of a downstream *arrB* sequence.

Figure 3.2 (a) A maximum likelihood tree of the arsenite oxidase sequence (AioA) highlighted in gold and the closely related enzyme arsenite oxidase-like (AioA-Like) in red-orange, including full *de novo* assembled environmental sequences of the probable chemoautotrophic AioA and multiple AioA-Like sequences denoted by bold text. (b) The gene complement found along the contig AioA_allDepths_contig_0 showing the presence of the beta subunit sequence *aioB* and a cytochrome-C.

Figure 3.3 (a) A collapsed view of the DMSO reductase family tree (Supplemental figure 3), with short reads identified through phylogenetically informed placement as associated with arsenotrophic enzymes mapped onto their respective clades. (b) A phylogenetic tree of DsrA showing transcripts associated with forward bacterial DsrA associated with dissimilatory sulfite reduction (SRB) and the reverse DsrA associated with sulfite oxidation (SOB) mapped along the respective clades for comparison to the transcripts associated with arsenotrophy in these ODZ

regions captured by the publicly available transcriptomes from oxygen deficient zones (Ganesh *et al.*, 2015; Schunck *et al.*, 2013; Stewart *et al.*, 2012).

Figure 3.4 Phylogenetically informed short read placement of metagenomics reads among particle associated ETNP AMZ metagenomes (> 30um) and a free-living metagenome (<30 um). Reads are presented as the length normalized number of reads associated with each target gene / the number of length normalized *rpoB* reads in the metagenomes * 100% to give an estimate of the gene's % contribution to the overall microbial community metagenome. The relative proportion of a gene's contribution to the microbial community when the free-living and particulate fractions are compared broken up into panel (a) with a much smaller proportion of the community showing arsenotrophy genes arsenite oxidase, *aioA,* and dissimilatory arsenate reducatase ,*arrA,* as well as the forward version of dissimilatory sulfite reductase, *dsrA,* which is involved in sulfur reduction (b) shows a larger proportion of the community is comprised of the sequences *aioA-like* and the reverse version of *dsrA* which is involved in sulfur oxidation.

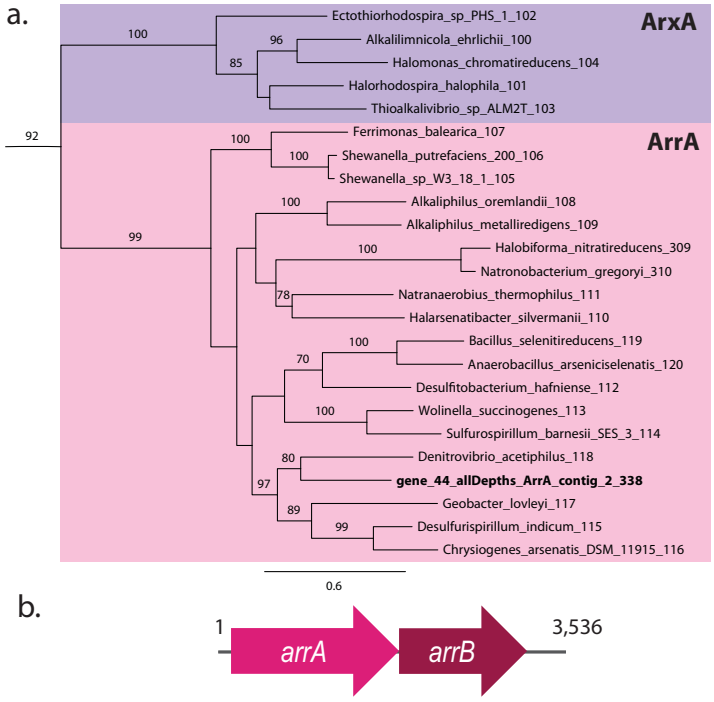

a.
ArxA
ArrA
Ectothiorhodospira_sp_PHS_1_102
Alkalilimnicola_ehrlichii_100
Halomonas_chromatireducens_104
Halorhodospira_halophila_101
Thioalkalivibrio_sp_ALM2T_103
Ferrimonas_balearica_107
Shewanella_putrefaciens_200_106
Shewanella_sp_W3_18_1_105
Alkaliphilus_oremlandii_108
Alkaliphilus_metalliredigens_109
Halobiforma_nitratireducens_309
Natronobacterium_gregoryi_310
Natranaerobius_thermophilus_111
Halarsenatibacter_silvermanii_110
Bacillus_selenitireducens_119
Anaerobacillus_arseniciselenatis_120
Desulfitobacterium_hafniense_112
Wolinella_succinogenes_113
Sulfurospirillum_barnesii_SES_3_114
Denitrovibrio_acetiphilus_118
gene_44_allDepths_ArrA_contig_2_338
Geobacter_lovleyi_117
Desulfurispirillum_indicum_115
Chrysiogenes_arsenatis_DSM_11915_116
0.6
b.
1
arrA
arrB
3,536

Figure 3.1

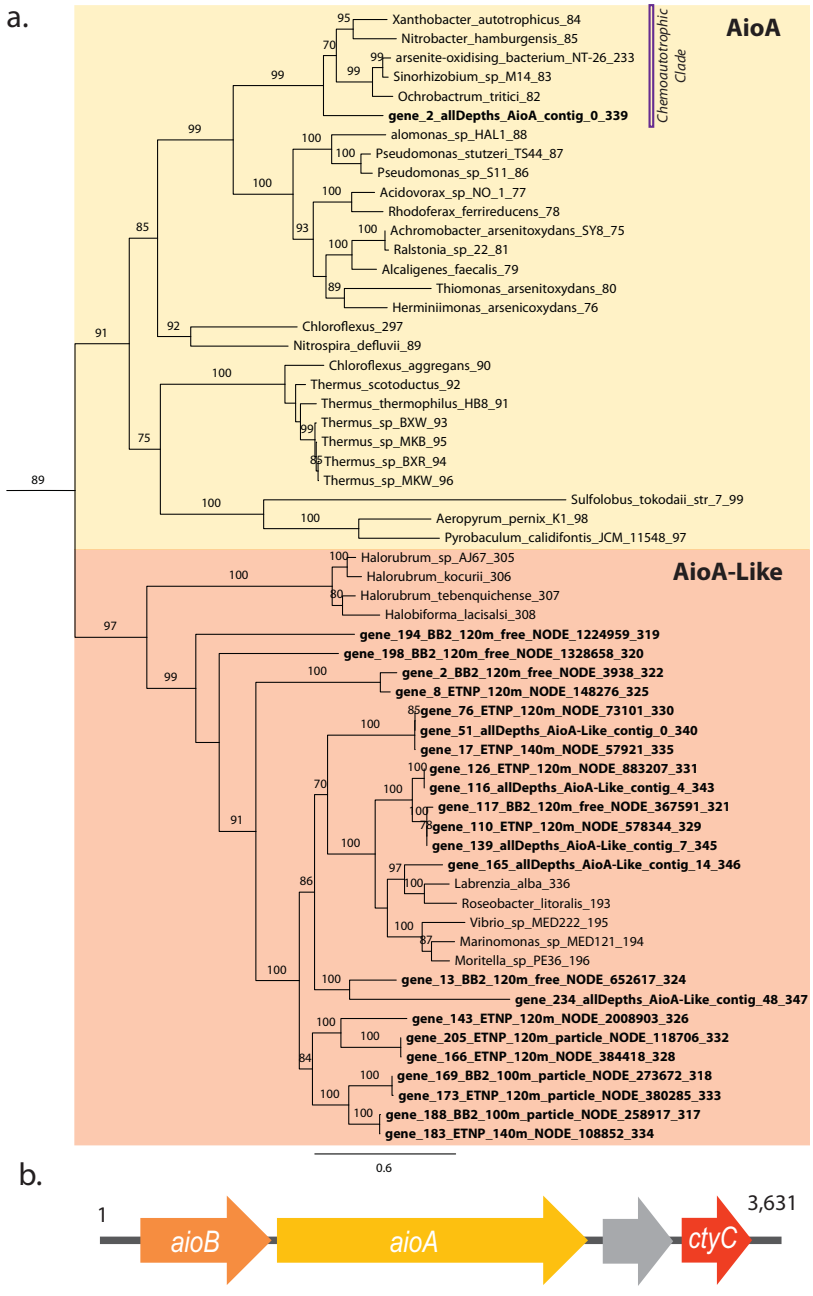

a.
AioA
Chemoautotrophic Clade
Xanthobacter_autotrophicus_84
Nitrobacter_hamburgensis_85
arsenite-oxidising_bacterium_NT-26_233
Sinorhizobium_sp_M14_83
Ochrobactrum_tritici_82
gene_2_allDepths_AioA_contig_0_339
alomonas_sp_HAL1_88
Pseudomonas_stutzeri_TS44_87
Pseudomonas_sp_S11_86
Acidovorax_sp_NO_1_77
Rhodoferax_ferrireducens_78
Achromobacter_arsenitoxydans_SY8_75
Ralstonia_sp_22_81
Alcaligenes_faecalis_79
Thiomonas_arsenitoxydans_80
Herminiimonas_arsenicoxydans_76
Chloroflexus_297
Nitrospira_defluvii_89
Chloroflexus_aggregans_90
Thermus_scotoductus_92
Thermus_thermophilus_HB8_91
Thermus_sp_BXW_93
Thermus_sp_MKB_95
Thermus_sp_BXR_94
Thermus_sp_MKW_96
Sulfolobus_tokodaii_str_7_99
Aeropyrum_pernix_K1_98
Pyrobaculum_calidifontis_JCM_11548_97
AioA-Like
Halorubrum_sp_AJ67_305
Halorubrum_kocurii_306
Halorubrum_tebenquichense_307
Halobiforma_lacisalsi_308
gene_194_BB2_120m_free_NODE_1224959_319
gene_198_BB2_120m_free_NODE_1328658_320
gene_2_BB2_120m_free_NODE_3938_322
gene_8_ETNP_120m_NODE_148276_325
gene_76_ETNP_120m_NODE_73101_330
gene_51_allDepths_AioA-Like_contig_0_340
gene_17_ETNP_140m_NODE_57921_335
gene_126_ETNP_120m_NODE_883207_331
gene_116_allDepths_AioA-Like_contig_4_343
gene_117_BB2_120m_free_NODE_367591_321
gene_110_ETNP_120m_NODE_578344_329
gene_139_allDepths_AioA-Like_contig_7_345
gene_165_allDepths_AioA-Like_contig_14_346
Labrenzia_alba_336
Roseobacter_litoralis_193
Vibrio_sp_MED222_195
Marinomonas_sp_MED121_194
Moritella_sp_PE36_196
gene_13_BB2_120m_free_NODE_652617_324
gene_234_allDepths_AioA-Like_contig_48_347
gene_143_ETNP_120m_NODE_2008903_326
gene_205_ETNP_120m_particle_NODE_118706_332
gene_166_ETNP_120m_NODE_384418_328
gene_169_BB2_100m_particle_NODE_273672_318
gene_173_ETNP_120m_particle_NODE_380285_333
gene_188_BB2_100m_particle_NODE_258917_317
gene_183_ETNP_140m_NODE_108852_334
0.6
b.
1
aioB
aioA
ctyC
3,631

Figure 3.2

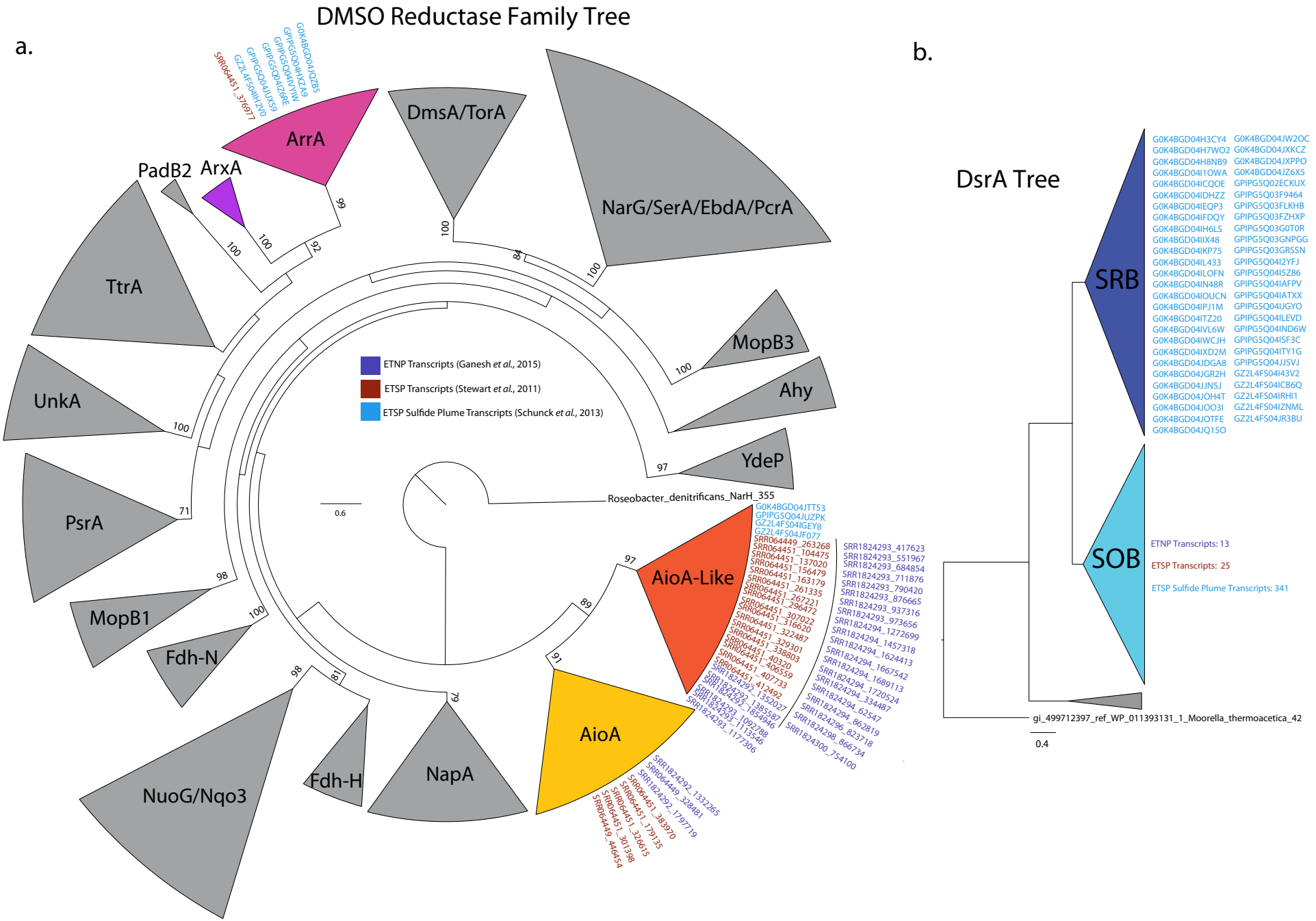

a.
DMSO Reductase Family Tree
ArrA
PadB2
ArxA
DmsA/TorA
NarG/SerA/EbdA/PcrA
TtrA
MopB3
Ahy
YdeP
UnkA
PsrA
MopB1
Fdh-N
NuoG/Nqo3
Fdh-H
NapA
AioA-Like
AioA
ETNP Transcripts (Ganesh et al., 2015)
ETSP Transcripts (Stewart et al., 2011)
ETSP Sulfide Plume Transcripts (Schunck et al., 2013)
Roseobacter_denitrificans_NarH_355
0.6
b.
DsrA Tree
SRB
SOB
ETNP Transcripts: 13
ETSP Transcripts: 25
ETSP Sulfide Plume Transcripts: 341
gi_499712397_ref_WP_011393131_1_Moorella_thermoacetica_42
0.4

Figure 3.3

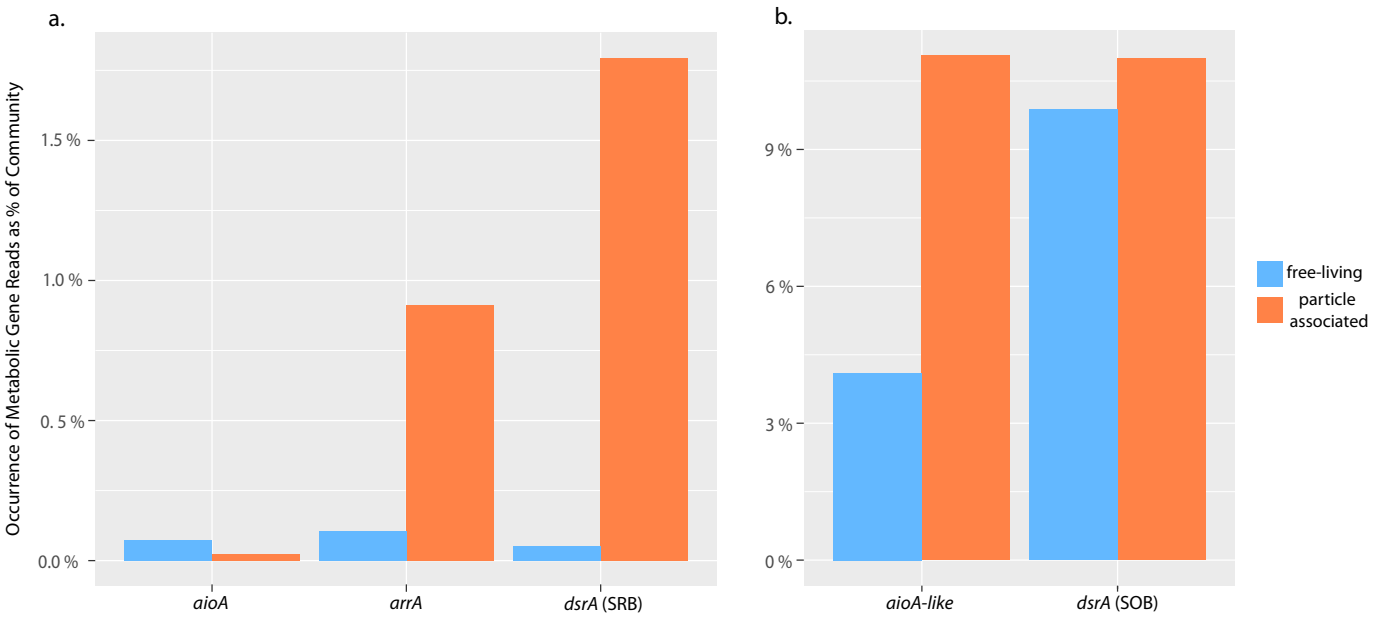

a.
b.
Occurrence of Metabolic Gene Reads as % of Community
1.5 %
1.0 %
0.5 %
0.0 %
9 %
6 %
3 %
0 %
aioA
arrA
dsrA (SRB)
aioA-like
dsrA (SOB)
free-living
particle associated

Figure 3.4

3.8 Supplementary Figure Legends & Tables:

Supplemental Figure 3.1 Ancillary data from Stations 136 & 141. Station 141 O_2 data is from STOX sensor; Station 136 O_2 data is from CTD sensor corrected with Winkler Titration.

Supplemental Figure 3.2 ArrB tree with bootstraps >= 70 showing that gene_45 is more closely related to ArrB sequences than to ArxB.

Supplementary Figure 3.3 DMSO Reductase family tree displayed with boostraps >= 70. Tree is rooted to a NarH outgroup sequence.

Supplemental Figure 3.4 AioB tree with bootstraps >= 70 displayed showing that sequence gene_1_AioA_allDepths_contig_0_14, shown in bold, is most closely related to the chemoautotrophic AioB sequences. The AioB's of the AioA-Like also group more closely with the AioB's of known AioA-Like sequences.

Supplementary Figure 3.5 The dissimilatory sulfite reductase, DsrA, tree displayed showing bootstraps >= 70.

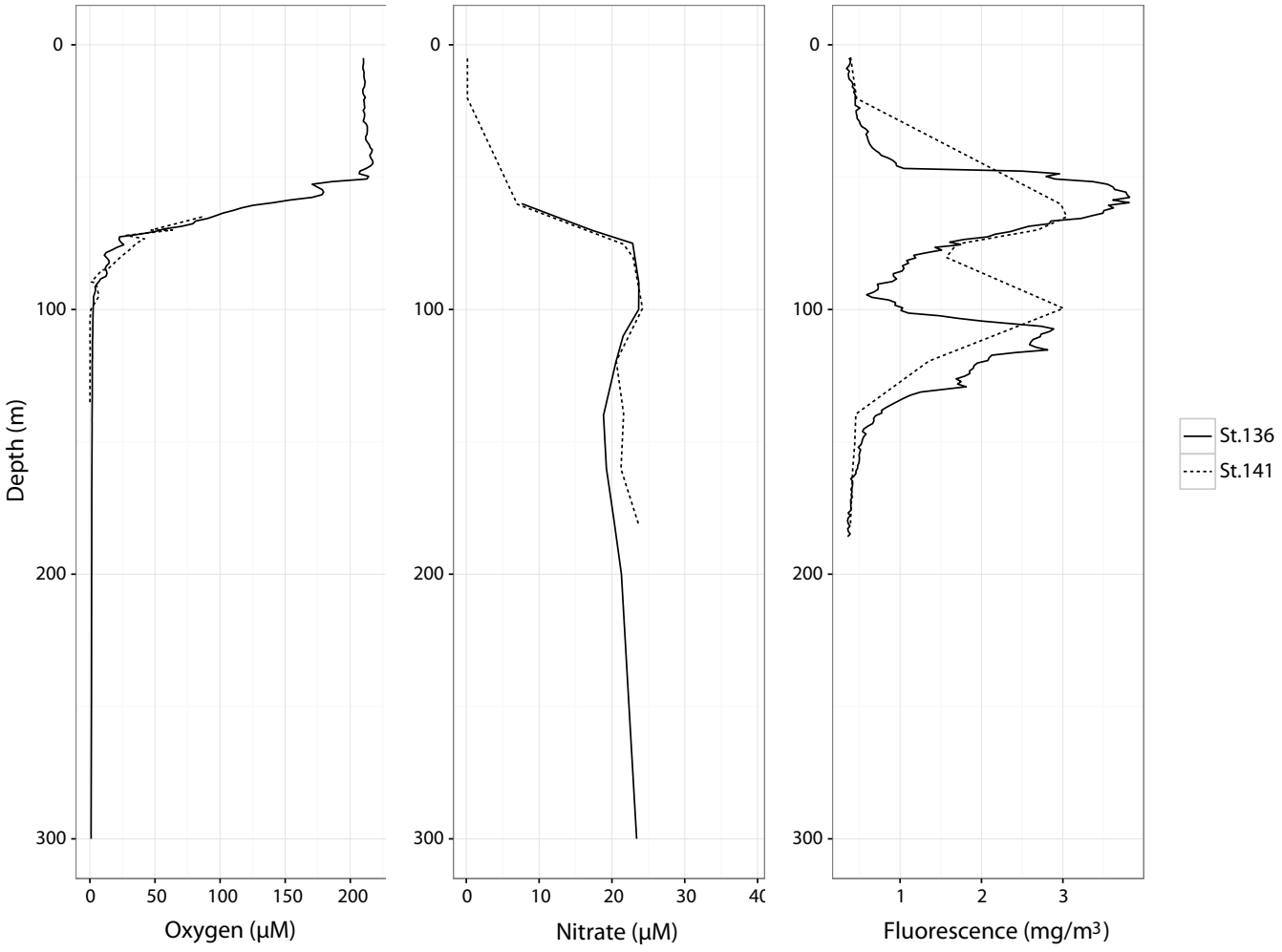
Supplemental Table 3.1 Gene annotations of contigs containing *arrA* (ArrA_allDepths_contig_2) *aioA* (AioA_allDepths_contig_0), and *aioA-like* (ETNP_120m_NODE_73101). Sequences in bold are arsenotrophy and arsenotrophy-related sequences, fumarate reductase, and the nitrate reductase sequence (NapA).

	Gene_ID	Start	Stop	Strand	Length	Description
ArrA_allDepths_contig_2	**44**	**38**	**2569**	**+**	**2532**	**Arsenate respiratory reductase, alpha subunit**
	45	**2588**	**3442**	**+**	**855**	**Arsenate respiratory reductase, beta subunit**
AioA_allDepths_contig_0	**1**	**127**	**639**	**+**	**513**	**Arsenite Oxidase, beta subunit**
	2	**652**	**3066**	**+**	**2415**	**Arsenite Oxidase, alpha subunit**
	3	3144	3278	+	135	hypothetical protein
	4	3335	3451	+	117	Cytochrome C
ETNP_120m_NODE_73101	19	110	1231	+	1122	Mrp/NBP35 ATP-binding protein, Domain of unknown function DUF59
	20	1228	2178	-	951	Thymidylate synthase ThyX
	21	2565	3023	+	459	Stringent starvation protein B
	22	3072	4694	+	1623	Fumarate lyase, FumAB
	23	4709	4924	+	216	Ribbon-helix-helix domain
	24	4933	9021	-	4089	AsmA-like, C-terminal
	25	9245	9580	+	336	Thiamine pyrophosphate enzyme, N-terminal TPP binding domain
	26	9581	10513	-	933	Xaa-Pro dipeptidyl-peptidase-like domain
	27	10603	11100	+	498	Thioesterase-like superfamily
	28	11479	11856	+	378	Protein of unknown function DUF4389
	29	12141	13022	+	882	Voltage gated chloride channel
	30	12973	13890	+	918	Voltage gated chloride channel
	31	14459	15130	+	672	NapC/NirT cytochrome c family, N-terminal region
	32	15143	15796	+	654	Cytochrome C oxidase, cbb3-type, subunit III
	33	**15812**	**18520**	**+**	**2709**	**NapA/NarB nitrate reductase catalytic subunit**
	34	18558	19148	+	591	NapG, 4Fe-4S dicluster domain

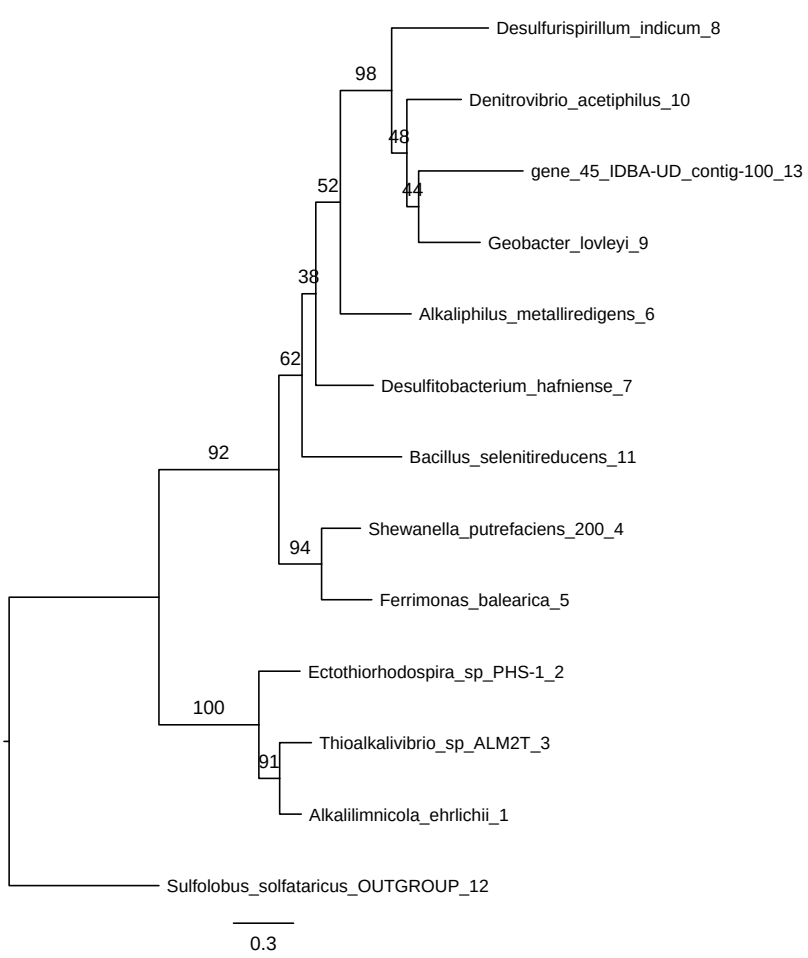
35	19149	20144	+	996	NapH, 4Fe-4S binding domain
36	20141	20734	+	594	4Fe-4S dicluster domain
37	20738	21334	+	597	hypothetical protein
38	21315	21926	+	612	4Fe-4S dicluster domain
39	21923	24013	+	2091	Protein of unknown function, DUF255
40	24105	24893	-	789	Enoyl-(Acyl carrier protein) reductase
Gene_ID	**Start**	**Stop**	**Strand**	**Length**	**Description**
41	25190	26104	-	915	Pyruvate carboxyltransferase
42	26108	28153	-	2046	Biotin carboxylase, carbamoyl-phosphate synthase
43	28219	28584	-	366	Potassium ion channel
44	28597	29388	-	792	Enoyl-CoA hydratase/isomerase
45	29393	31000	-	1608	Acetyl-coenzyme A carboxyltransferase, C-terminal
46	30997	31833	-	837	Amidohydrolase-related
47	31830	32678	-	849	Amidohydrolase-related
48	32763	33962	+	1200	CoA-transferase family III
49	34074	34979	+	906	Xylose isomerase-like TIM barrel
50	**34982**	**36475**	**+**	**1494**	**Succinate dehydrogenase/fumarate reductase flavoprotein, catalytic domain**
51	36692	37411	+	720	Short-chain dehydrogenase/reductase SDR
52	37481	38314	+	834	Phytanoyl-CoA dioxygenase (PhyH)
53	38395	39414	+	1020	TRAP transporter solute receptor DctP/TeaA
54	39428	40447	+	1020	TRAP transporter solute receptor DctP/TeaA
55	40457	40975	+	519	Tripartite ATP-independent periplasmic transporter, DctQ component
56	40999	42288	+	1290	TRAP C4-dicarboxylate transport system permease DctM subunit
57	42510	43625	-	1116	Putative bacterial porin
58	44066	45493	-	1428	Protoporphyrinogen oxidase
59	45469	46734	-	1266	Coenzyme PQQ synthesis protein E
60	46727	48376	-	1650	Universal stress protein A, UspA
61	48379	48558	-	180	hypothetical protein

62	48575	50119	-	1545	Cytochrome c-552/DMSO reductase-like, haem-binding domain
63	50112	50879	-	768	Cytochrome C oxidase, cbb3-type, subunit III
64	50876	51709	-	834	Cytochrome c
65	51706	53496	-	1791	Cytochrome bd terminal oxidase subunit I
66	53515	53676	-	162	hypothetical protein
67	53823	53927	+	105	hypothetical protein
Gene_ID	**Start**	**Stop**	**Strand**	**Length**	**Description**
68	53934	54290	-	357	Cytochrome C oxidase, cbb3-type, subunit III
69	54304	54963	-	660	Cytochrome c
70	54960	56621	-	1662	Cytochrome c-552/DMSO reductase-like, haem-binding domain
71	56621	57619	-	999	Cytochrome C oxidase, cbb3-type, subunit III
72	57603	58586	-	984	Cytochrome c
73	58586	60766	-	2181	Cytochrome bd terminal oxidase subunit I
74	60767	61183	-	417	Cytochrome C oxidase, cbb3-type, subunit III
75	61192	63342	-	2151	Hypothetical HEAT repeat protein
76	**63445**	**66138**	**-**	**2694**	**Arsenite Oxidase Like, alpha subunit**
77	**66154**	**66726**	**-**	**573**	**Arsenite Oxidase Like, beta subunt**
78	66737	67015	-	279	Cytochrome c
79	66987	68108	-	1122	Di-haem cytochrome c peroxidase
80	68124	69209	-	1086	Di-haem cytochrome c peroxidase
81	69523	70251	-	729	Polysaccharide lyase family 4, domain II
82	70897	71154	-	258	ThiamineS/Molybdopterin converting factor subunit 1
83	71164	73437	-	2274	Aldehyde oxidase/xanthine dehydrogenase, a/b hammerhead
84	73440	73916	-	477	2Fe-2S iron-sulfur cluster binding domain
85	73920	74960	-	1041	CO dehydrogenase flavoprotein-like, FAD-binding
86	75205	75324	-	120	hypothetical protein
87	75361	76152	+	792	Luciferase-like monooxygenase
88	76143	76412	+	270	Hypothetical

	89	76432	76953	-	522	AMP-dependent synthetase/ligase
	90	76995	77996	-	1002	AMP-dependent synthetase/ligase
	91	77924	79816	-	1893	Thiamine pyrophosphate enzyme
	92	79813	80808	-	996	Oxidoreductase family, NAD-binding Rossmann fold
	93	80818	81645	-	828	Xylose isomerase-like TIM barrel
	94	81945	82532	-	588	L-2-amino-thiazoline-4-carboxylic acid hydrolase



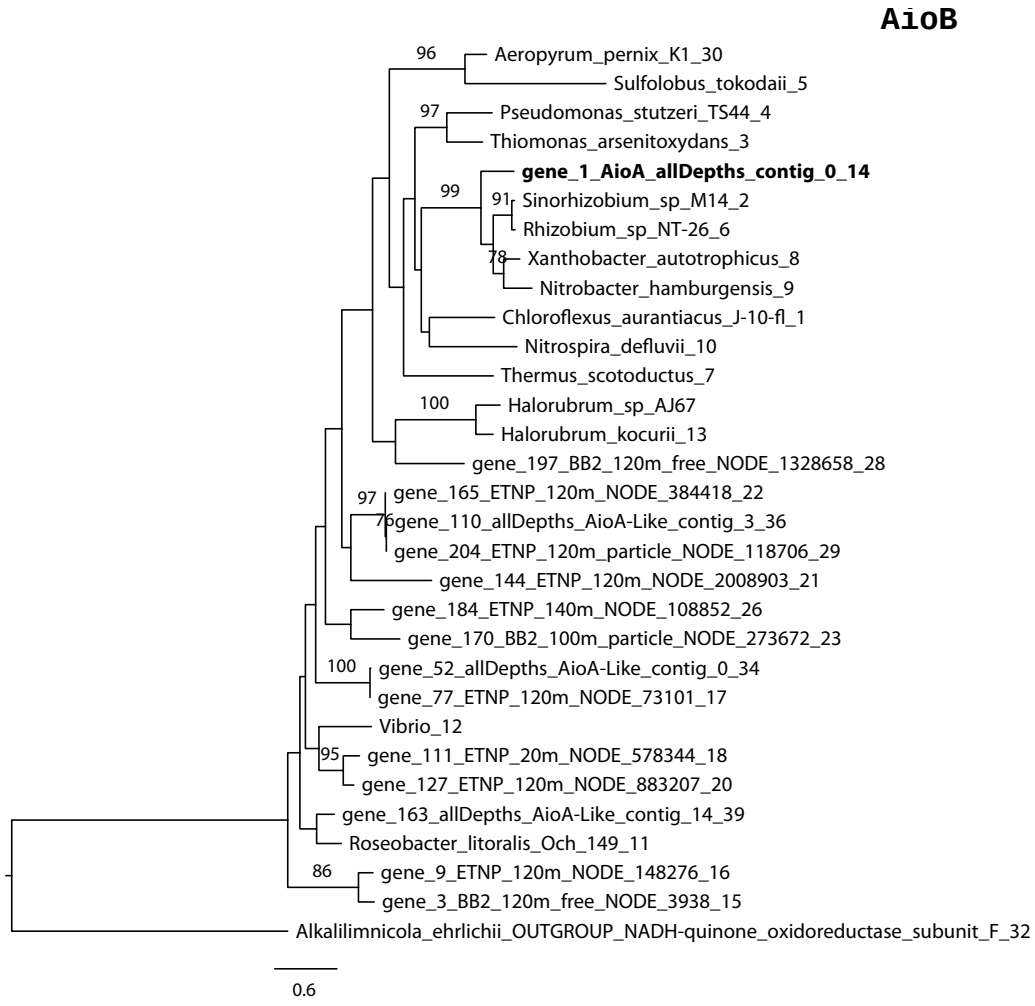
Supplementary Figure 3.1



Supplementary Figure 3.2

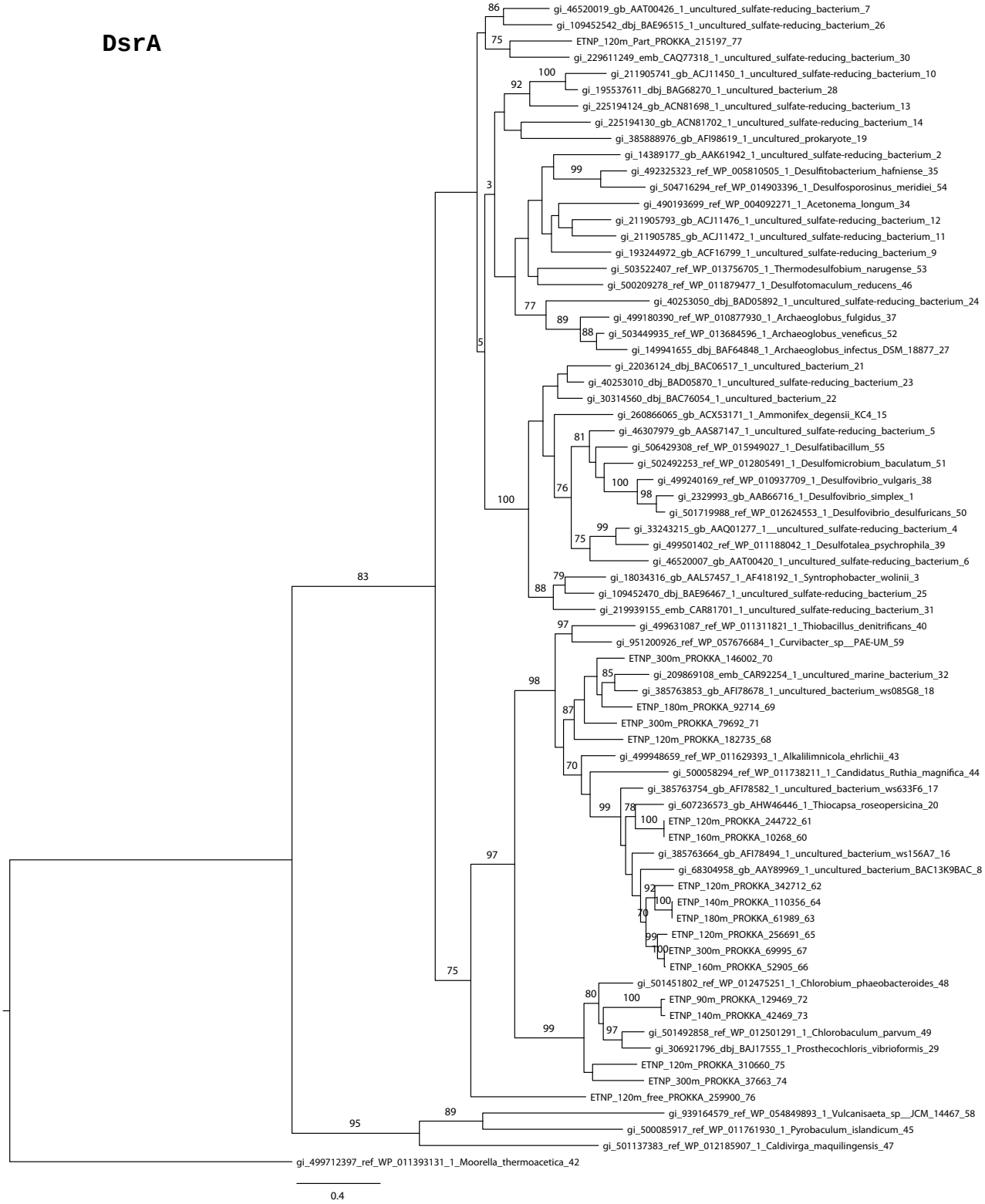
DMSO Reductase Family Tree

Supplementary Figure 3.3



Supplementary Figure 3.4

DsrA



Supplementary Figure 3.5

Conclusion

The availability of the essential macronutrient phosphorus and the closely related element arsenic has dramatically shaped marine microbial communities, with evolutionary responses to modern day shifts in availability of the nutrients as well as responses to shifts in these elements over geologic timescales. Over geologic timescales, phosphorus is believed to be the limiting nutrient for primary productivity (Tyrrell, 1999). The relative abundance of these two elements has not been stable throughout Earth's history, rather microbial communities have had to respond to varying availabilities of P:As (Chi Fru *et al.*, 2015; Chi Fru *et al.*, 2016; Bjerrum and Canfield, 2002; Planavsky *et al.*, 2010). Arsenic is toxic to cells due to its similarity to the essential biomolecule phosphorus, as the pentavalent form arsenate can become toxic by competing with the pentavalent phosphate, for example through the decoupling of oxidative phosphorylation, the process that produces ATP (Mandal and Suzuki, 2002; Oremland and Stolz, 2003). The reduced form, arsenite, is even more toxic because it interferes with enzyme activity by bonding to –SH and –OH groups in enzymes (Mandal and Suzuki, 2002; Akter *et al.*, 2005).

One response to varying availability of phosphorus has been through the evolution of various phosphorus acquisition strategies such as multiple transporters, high affinity transporters, and the ability to use multiple P-sources, like the use of alkaline phosphatases to cleave orthophosphate groups from organic molecules. These numerous phosphate uptake strategies are highlighted in the marine picocyanobacterium *Prochlorococcus*. Chapter 1 demonstrates the response of *Prochlorococcus* strain MED4 to phosphorus stress. This strain of *Prochlorococcus* was isolated from the P-stressed waters of the Mediterranean. It maintains multiple genes associated with these P-acquisition strategies, such as the high affinity transport pathway *pstCABS,* the alkaline phosphatase *phoA*, and a probable organic-P porin PMM0709(Rocap *et*

al., 2003). A proteomics analysis of MED4 under P-limitation and P-starvation culturing conditions show that these cells dramatically increase expression of many P-acquisition proteins (PhoA, PstS, PMM0709), with some of the most dramatically differentially expressed proteins (PMM1409, PMM1414, PMM1416) being uncharacterized proteins associated with a P-stress inducible genomic island (Coleman *et al.*, 2006). A biogeographic analysis of these uncharacterized, but highly expressed proteins, demonstrates the selective role of phosphorus availability on *Prochlorococcus* genomic capacity. *Prochlorococcus* populations from regions which experience the most extreme phosphorus scarcity maintain the greatest genomic occurrence of these uncharacterized P-stress inducible genes.

These extremely phosphorus scarce surface waters are also regions where the P:As ratio dramatically decreases, resulting in greater risk of cellular arsenic exposure and therefor arsenic toxicity (Wurl *et al.*, 2013; Saunders and Rocap, 2016). Shifts in the P:As ratio have occurred throughout geologic time, and so marine microbial communities have responded to the toxicity risks posed by the potential toxic competitive behavior of arsenic. Microbial arsenic resistance mechanisms are numerous, taxonomically widespread, and subject to frequent horizontal transfer (Stolz *et al.*, 2006; Páez-Espino *et al.*, 2009). Phylogenetic analyses of many of the enzymes acting in arsenic resistance indicate a likely ancient origin, prior to the divergence of Bacteria and Archaea (Lebrun, 2003; Jackson and Dugas, 2003). In Chapter 2, the arsenic detoxification strategies of *Prochlorococcus* are discussed, consisting of a previously identified efflux detoxification pathway and a putative arsenic methylation pathway identified here. The evolutionary pressure of the varying P:As ratio found in global surface *Prochlorococcus* populations is evidenced by the varying genomic capacity for arsenic detoxification, where

greater capacity for detoxification is maintained by *Prochlorococcus* populations in the most phosphorus scarce regions of the ocean.

While the *Prochlorococcus* have employed the strategies of high affinity phosphorus uptake transporters and detoxification pathways, some marine microorganisms have evolved the genomic capacity to utilize the potentially toxic arsenic for energetic gains. Chapter 3 demonstrates the genomic capacity for a complete arsenic metabolic cycle in anoxic pelagic waters of a global oxygen deficient zone. Genes specific for dissimilatory arsenate reduction, *arrAB,* and chemoautotrophic arsenite oxidation, *aioAB,* are identified in a metagenome from the anoxic waters of the Eastern Tropical North Pacific (ETNP). These genes are shown to not only be present, but also actively transcribed in modern oxygen deficient waters in the Eastern Tropical North Pacific and Eastern Tropical South Pacific. In addition, an arsenic metabolism related enzyme, *aioA-like*, was identified as numerous and may potentially support chemoautotrophic carbon fixation in these waters. The bioenergetic use of arsenic in anoxic marine waters is an excellent example of microbial ingenuity – making a beneficial use of a potential toxin. This beneficial use of arsenic in marine anoxic waters is likely an ancient strategy, with phylogenetic analyses indicating a likely ancient origin, prior to the Last Universal Common Ancestor, of the bioenergetic arsenite oxidase enzyme which was likely coupled to the reduction of nitrogen oxyanions (van Lis *et al.*, 2013).

Marine microorganisms have responded to the variable presence of phosphorus and arsenic in the marine environment on a range of timescales. *Prochlorococcus* demonstrates phenotypic plasticity in response to nutrient stress through shifted proteomic profiles. It also shows evolutionary selection through gene gains and losses associated with phosphorus and arsenic availability. Future work investigating the response of *Prochlorococcus* and other marine

microbes to shifting P:As ratios will provide further insight into the microbial response to fluctuations in these two physicochemically similar elements. As oligotrophic regions of the ocean continue to spread and intensify (Behrenfeld *et al.*, 2006; Polovina *et al.*, 2008), gene frequencies of accessory P-acquisition and As-detoxification genes among *Prochlorococcus* may act as indicators of evolutionary response to enhanced nutrient stress as a byproduct of global climate change. Further laboratory work following up on the putative arsenic methylation pathway identified in Chapter 2, through direct measurement of organoarsenical production in *Prochlorococcus* cultures, may confirm *Prochlorococcus* as a source of methylated arsenicals found in oligotrophic gyres (Wurl *et al.*, 2013). Enhanced nutrient stress as a result of rising global temperatures is likely to push P:As in oligotrophic waters lower resulting in greater risk of As uptake for the microbes there. Organoarsenical compounds like arsenosugars, DMA and arsenolipids have been identified as potential toxins (Feldmann and Krupp, 2011; Newcombe *et al.*, 2010; Raml *et al.*, 2009; Leffers *et al.*, 2013; Meyer *et al.*, 2014; Harrington *et al.*, 2008), with arsenolipids showing toxicity in human bladder and liver cell lines at the same level as inorganic arsenite (Meyer *et al.*, 2014). Bioaccumulation of complexed organoarsenicals in upper trophic levels has been observed in trophic cascades (Dembitsky and Levitsky, 2004) with the source of arsenolipids in fish likely coming from accumulation of algae-based arsenolipids (Francesconi, 2010). Greater exposure of primary producers to low P:As may result in greater organoarsenical production and intensified bioaccumulation of the potential toxin to higher order organisms. Thus, it is important to identify the types of organoarsenicals produced by primary producers, as arsenical forms vary in toxicity, and try to quantify the flux of arsenic through the food web.

Arsenic detoxification through efflux and methylation is the strategy employed by *Prochlorococcus*. However, some organisms have responded to the presence of arsenic not only through defensive detoxification strategies, and instead use the potential toxin for bioenergetic gains. The marine microbial community in the anoxic waters of the Eastern Tropical North Pacific demonstrates the capacity for the beneficial use of arsenic for bioenergetic gains. These arsenic metabolisms found here are likely ancient in origin, and potentially a remnant of early anoxic oceans. These metabolisms are likely supporting organic carbon fixation, and may contribute to global nitrogen cycling as well through the reduction of nitrogen oxyanions. Further work directly measuring chemical cycling of arsenic by these arsenic metabolizing microbes should be conducted. The arsenic metabolism related enzyme AioA-Like should be further characterized as its substrate is not currently known. These arsenotrophy and arsenotrophy-related pathways have the capacity to greatly impact biogeochemical cycles, especially AioA-Like as it is abundant and appears likely to support autotrophic carbon fixation. In addition, these metabolisms were likely of even greater importance in early anoxic oceans. The microbial responses to P and As described in this thesis have the power to dramatically shape global biogeochemical cycles in oceans of the past, present, and future as the nutrient deficient oligotrophic gyres are expanding and the anoxic waters of the Eastern Tropical North Pacific are likely a small representative of the chemical conditions which pervaded in early earth history.

References

1 Tyrrell T (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature* **400:** 525-531.

2 Chi Fru E, Arvestål E, Callac N, El Albani A, Kilias S, Argyraki A *et al.* (2015). Arsenic stress after the Proterozoic glaciations. *Scientific Reports* **5:** 17789.

3 Chi Fru E, Hemmingsson C, Holm M, Chiu B, Iñiguez E (2016). Arsenic-induced phosphate limitation under experimental Early Proterozoic oceanic conditions. *Earth and Planetary Science Letters* **434:** 52-63.

4 Bjerrum CJ, Canfield DE (2002). Ocean productivity before about 1.9 Gyr ago limited by phosphorus adsorption onto iron oxides. *Nature* **417:** 159-162.

5 Planavsky NJ, Rouxel OJ, Bekker A, Lalonde SV, Konhauser KO, Reinhard CT *et al.* (2010). The evolution of the marine phosphate reservoir. *Nature* **467:** 1088-1090.

6 Mandal BK, Suzuki KT (2002). Arsenic round the world: a review. *Talanta* **58:** 201-235.

7 Oremland RS, Stolz JF (2003). The Ecology of Arsenic. *Science* **300:** 939-944.

8 Akter KF, Owens G, Davey DE, Naidu R (2005). Arsenic speciation and toxicity in biological systems. *Rev Environ Contam Toxicol* **184:** 97-149.

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

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

11 Wurl O, Zimmer L, Cutter GA (2013). Arsenic and phosphorus biogeochemistry in the ocean: Arsenic species as proxies for P-limitation. *Limnology and Oceanography* **58:** 729-740.

12 Saunders JK, Rocap G (2016). Genomic potential for arsenic efflux and methylation varies among global *Prochlorococcus* populations. *ISME J* **10:** 197-209.

13 Stolz JF, Basu P, Santini JM, Oremland RS (2006). Arsenic and Selenium in Microbial Metabolism*. *Annual Review of Microbiology* **60:** 107-130.

14 Páez-Espino D, Tamames J, Lorenzo V, Cánovas D (2009). Microbial responses to environmental arsenic. *BioMetals* **22:** 117-130.

15 Lebrun E (2003). Arsenite Oxidase, an Ancient Bioenergetic Enzyme. *Molecular Biology and Evolution* **20:** 686-693.

16 Jackson CR, Dugas SL (2003). Phylogenetic analysis of bacterial and archaeal *arsC* gene sequences suggests an ancient, common origin for arsenate reductase. *BMC Evol Biol* **3**.

17 van Lis R, Nitschke W, Duval S, Schoepp-Cothenet B (2013). Arsenics as bioenergetic substrates. *Biochim Biophys Acta-Bioenerg* **1827:** 176-188.

18 Behrenfeld MJ, O'Malley RT, Siegel DA, McClain CR, Sarmiento JL, Feldman GC *et al.* (2006). Climate-driven trends in contemporary ocean productivity. *Nature* **444:** 752-755.

19 Polovina JJ, Howell EA, Abecassis M (2008). Ocean's least productive waters are expanding. *Geophysical Research Letters* **35**.

20 Feldmann J, Krupp EM (2011). Critical review or scientific opinion paper: arsenosugars--a class of benign arsenic species or justification for developing partly speciated arsenic fractionation in foodstuffs? *Anal Bioanal Chem* **399:** 1735-1741.

21 Newcombe C, Raab A, Williams PN, Deacon C, Haris PI, Meharg AA *et al.* (2010). Accumulation or production of arsenobetaine in humans? *Journal of Environmental Monitoring* **12:** 832-837.

22 Raml R, Raber G, Rumpler A, Bauernhofer T, Goessler W, Francesconi KA (2009). Individual variability in the human metabolism of an arsenic-containing carbohydrate, 2',3'-dihydroxypropyl 5-deoxy-5-dimethylarsinoyl-beta-D-riboside, a naturally occurring arsenical in seafood. *Chemical research in toxicology* **22:** 1534-1540.

23 Leffers L, Ebert F, Taleshi MS, Francesconi KA, Schwerdtle T (2013). *In vitro* toxicological characterization of two arsenosugars and their metabolites. *Molecular nutrition & food research* **57:** 1270-1282.

24 Meyer S, Matissek M, Muller SM, Taleshi MS, Ebert F, Francesconi KA *et al.* (2014). In vitro toxicological characterisation of three arsenic-containing hydrocarbons. *Metallomics* **6:** 1023-1033.

25 Harrington CF, Brima EI, Jenkins RO (2008). Biotransformation of arsenobetaine by microorganisms from the human gastrointestinal tract. *Chemical Speciation and Bioavailability* **20:** 173-180.

26 Dembitsky VM, Levitsky DO (2004). Arsenolipids. *Prog Lipid Res* **43:** 403-448.

27 Francesconi KA (2010). Arsenic species in seafood: Origin and human health implications. *Pure Appl Chem* **82:** 373-381.

Appendix I

From an unknown nucleotide sequence to functional and taxonomic identification: A bioinformatics pipeline for the identification and quantification of metagenomics reads using a phylogenetically informed approach

AI.1 Body:

Short-read sequence data of microbial communities have become widely available with the advent of new sequencing technologies and decline in the cost of sequencing. A key question during the analysis of these datasets is to understand what organisms are present as well as to understand the functional capacity of the microorganisms present in the microbial community. A common method for the identification of metagenomic sequences has been to report the best BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) match. However, the local alignment search strategy on which BLAST is based is limited in identification, and a global alignment is preferable especially for closely related sequences. BLAST based matches of environmental sequences are often not the closest relatives phylogenetically (Koski and Golding, 2001). BLAST-only sequence read identification has been shown to be limited in its ability to determine read assignments when the database to which unknown reads are compared does not contain reference sequences that are close enough in relation to the query (Berger *et al.*, 2011). Phylogenetic-based approaches for sequence analysis, while computationally more expensive than BLAST-only identification, is able to assign a similarity match to unknown reads with improved accuracy (Berger *et al.*, 2011).

There are many examples of Phylogenetic-based methods for sequence read identification, some examples include the maximum likelihood based RAxML Evolutionary Placment Algorithm (EPA) (Berger *et al.*, 2011; Stamatakis, 2014) and pplacer (Matsen *et al.*,

2010) as well as minimum evolution based PhyClass (Filipski *et al.*, 2015). Likelihood-based inference methods are a powerful ways to assign identity unknown microbial sequencing reads. While computationally more demanding than many other phylogentic-based approaches, recent improvements in software has made it computationally feasible to use these likelihood-based methods on large environmental datasets (Berger *et al.*, 2011). The maximum-likelihood phylogenetic tree construction software, RAxML, is capable of fast, computationally parallelized, maximum likelihood phylogenetic tree construction (Stamatakis, 2014; Stamatakis, 2006). The program PaPaRa (Berger and Stamatakis, 2011; Berger and Stamatakis, 2012) and the Evolutionary Placement Algorithm (EPA) extension of RAxML (Berger *et al.*, 2011; Stamatakis, 2014) are able to align unknown metagenomic reads to a known reference tree and then assign the unknown reads to the most similar node of a reference tree. A webserver exists for the placement of environmental sequences using RAxML EPA (Stark *et al.*, 2010), however it is not feasible to use webservers for identification of large amounts of sequence data and the webserver does not fully cover every step in sequence identification, from the parsing of target sequences from a metagenome to the identification of those sequences through placement. A scalable enclosed framework to go from meta-omic sequence data to assigned identity of unknown sequencing reads for large datasets has been lacking, especially for functional genes in amino acid space.

There are five major steps to identify unknown sequence reads through phylogenetic inference methods: 1. A reference alignment and tree of known sequences is constructed; 2. Sequencing reads of interest (queries) are parsed from a meta-omic database for phylogenetic placement analysis; 3. The recruited read queries are trimmed and formatted to the length of the target gene; 4. The unknown queries are aligned to the known reference alignment; 5. The

unknown aligned queries are phylogenetically inferred through placement to nodes on the reference gene tree. Presented here is software which integrates with previously published bioinformatics software to pass through each stage of sequencing read identification for a specified gene, with additional tools that help facilitate identification of protein coding genes and increasing phylogenetic signal by combing read pairs. We also present the results of placements from artificial metagenomic datasets where parameter optimization was performed, placement fidelity evaluated, and resolution of placements among closely related sequences under amino acid versus nucleotide space is assessed.

All software developed to link previously published bioinformatics programs is written and implemented in Python 2.7 as command-line applications. In order to evaluate the ability of this placement pipeline in the assignment of metagenomics reads, an artificial metagenome was created. The artificial metagenome consists of a total of 5,000,000 reads which were produced from a set of fully sequenced microbial genomes, with read lengths and characteristics set to reflect Illumina HiSeq sequencing technology used in the metagenomes from the anoxic waters of the East Tropical North Pacific described in Chapter 3. Read lengths were set to a maximum length of 149 bases, with paired end reads, and a small insert size producing multiple paired end reads with overlap. The bacterial taxa, and relative abundance of the taxa, in the artificial metagenome was loosely set to reflect that of the microbial community found in the marine metagenomic dataset from the Global Ocean Sampling Survey (GOS) (Rusch *et al.*, 2007).

It is possible to perform sequence read identification in either amino acid or nucleotide space. For functional protein coding genes, amino acid space may be preferred as the amino acid sequence is closer to the level at which evolution acts upon - either preserving or altering gene function. For determination between closely related sequences, nucleotide space may be favored.

Figure 1 outlines a workflow used for the determination of metagenomics reads using this pipeline with a text description of the individual programs in Table 1. Placement in amino acid space is highlighted in the figure, however nucleotide-sequence based placement is also supported.

The first stage in the bioinformatics pipeline is to identify full length reference gene sequences and construct a reference phylogenetic tree in which to evaluate metagenomics reads against (Figure 1, panel 1). Full length sequences of known origin for a gene of interest (geneX) are collected from repositories. If targeting a specific taxonomic group, it is recommended to have high resolution in the tree near the taxon of interest while also including closely related outgroups of other taxa. When targeting genes from a wide range of taxa, it is recommended to include sequences of the next closely related gene family in the reference. If a corresponding de novo assembly is available for the metagenome/metatranscriptome being analyzed, it is helpful to add identified full-length genes from these assemblies to the reference sequences for later identification. An online based database which can be helpful for gathering sequences of known reference taxa, MicrobesOnline (Dehal *et al.*, 2009), is a useful tool where genes of interest from selected genomes can be easily downloaded, both in nucleotide and amino acid space, along with ancillary data from gene carts. A tool for renaming sequence data with easily interpretable descriptors has been included with the code for this pipeline (renameMicrobesOnlineSeqs.py). Reference sequences are aligned, using the alignment program MUSCLE (Edgar, 2004). The reference alignment is then used to construct a maximum likelihood phylogenetic tree using the program RAxML (Stamatakis, 2014; Stamatakis, 2006). A manually curated fasta file of a subset of the reference sequences in amino acid space is created in order to search the metagenome – these sequences are referred to as “seeds”. If searching for a specific group, it is recommended to

include a few sequences spanning the entire clade of interest as seeds. If interested in identifying a broad taxonomic range of sequences, one should include representatives of all clades in the seeds file.

The next major step in the pipeline is to parse sequence reads from a metagenome which are related to the gene of interest, geneX (Figure 1, panel 2). An indexed blast database is created using the blast tool makeblastdb (Altschul *et al.*, 1990; Altschul *et al.*, 1997). The seeds file described previously is used as query sequences in a tblastn search which will search the amino acid sequences against a 6-frame translation of the nucleotide blast database. ParseBlastXML will then take the blast results, and return the results in a table format as well as retrieve the full metagenomic read sequence for the blast hits. TrimFormatReads then uses the blast results, including information pertaining to the reading frame in which the hit is oriented, as well as the location of the alignment of the blast hit with the query seed sequence. TrimFormatReads will create a unique list of all reads hit, choosing the blast hit with the best e-value relative to the seed sequence. The reads are oriented in the proper reading frame and trimmed so that there is no overhanging sequence from the ends of the reference alignment (this is accomplished by trimming according to the length of the sequence of the best blast seed). Optional parameters can also be input, so only reads of a certain length, e-value, or score are reported. TrimFormatReads also removes reads with ambiguous nucleotides and removes stop codons from the sequence as downstream software is unable to handle these residue markers. Statistics and information on all trimming properties is also reported. The output is a fasta file of properly formatted metagenomics reads – both amino acid and nucleotide sequence is reported – which will be identified through phylogenetic placement.

In the final stage of the pipeline, the parsed and formatted metagenomics reads are compared against the known reference sequence for high resolution identification (Figure 1, panel 3). In the case illustrated in Figure 1, the amino acid read sequences, the reference alignment used to construct the reference maximum likelihood tree, and the reference tree are combined by the program PaPaRa in order to align the unknown metagenomics reads to the reference alignment (Berger *et al.*, 2011). This combined alignment is then passed to spaceJoin, where paired end reads which are known to be linked can be joined together for placement in order to enhance the phylogenetic signal in the placement process as well as reduce double counting of sequences, as paired end reads came from the same original genetic unit. The joined alignment is then passed, along with the maximum likelihood reference tree, to RAxML's Evolutionary Placement Algorithm (RAxML-EPA) which will assign unknown metagenomics reads to the nodes on the reference tree through phylogenetic inference methods (Berger and Stamatakis, 2011). The original tree file output which is produced by RAxML-EPA can be passed to taxaDict which will create a table where each terminal numeric node is identified by the representative sequence descriptor. The "dictionary" that is produced can then be edited by hand to enable curated identification of nodes to which reads are placed. TreeLabels takes the curated taxa dictionary and the placement file in order to re-label all placements by the meaningful descriptors in the taxa dictionary. If multiple metagenomes are included in one blast database, and option to pass a file of metagenome keys for sorting of read placement by metagenome is also included. A final option to report the count of the number of reads placed to the nodes is also available in treeLabels. The final product of this pipeline is a high resolution identification of metagenomics reads for a specific gene target.

In order to test the fidelity of the placements, and compare the accuracy of placement in nucleotide versus amino acid space, an artificial metagenome was created and placements assessed for accuracy (Angly *et al.*, 2012). The core ribosomal protein, rpsD, was used as an example for the placement assessment. This is an essential protein which occurs generally at one copy per microbial genome. The taxa used to construct the reference metagenome are listed in Table 2 according to their relative abundance in the artificial metagenome. Figure 2 shows the amino acid maximum likelihood tree used for placements of the artificial metagenome reads. Representatives from each clade were used for blast querying of the metagenome. Blast results with an e-value of ≤ 1 were retained for placement with a minimum length of 40 amino acid residues or 100 nucleotide bases required for placement analysis.

Here we evaluate placements of a broad array of taxa for a single gene. Assessment of the accuracy of placements was determined by how well placements matched to broader groupings, and not strain/species specific assignments (Figure 3). All rpsD metagenomic reads placed correctly to their respective taxonomic grouping. Overall, placement in amino acid space and nucleotide space performed similarly. However, greater resolution of placement among closely related strains, like in the case of *Prochlorococcus*, was found in nucleotide placement. The relative abundance of taxonomic groups in the artificial dataset was reflected closely through the placement analysis. However, the placement analysis did show some bias in favor of groups which were more abundant in the metagenome and a bias against the more rare sequences like Planctomycetes.

Maximum likelihood phylogenetic based read identification of unknown meta-omic sequencing reads is a powerful tool for the characterization of meta-omic sequencing reads. These methods are extremely accurate (Berger *et al.*, 2011), and recent software advances has

made is computationally feasible to use these analyses on large datasets. The pipeline described here is able to be scaled for use on large datasets, such as the analysis performed in Chapter 1 on TARA Oceans metagenomes (Sunagawa *et al.*, 2015). In addition, this pipeline contains all the necessary components to go from meta-omic sequence reads to characterization and quantification of reads – an encompassing framework that can be easily scaled.

This pipeline is made available in the cloud for use, so that all components of this pipeline can be used on other datasets without need for finding and installing all components locally. An Amazon Web Services Amazon Machine Image was created which houses all the software described here, the previously existing software the pipeline depends on like RAxML (Stamatakis, 2014), as well as all software dependencies. In addition, the artificial metagenome is also included on this machine image so that an example dataset is available for trial and optimization of parameters. This pipeline enables a high resolution analysis of the abundance of target sequences in metagenomes.

A1.2 Acknowledgements:

I would like to thank Gabrielle Rocap for her guidance in developing this workflow. I would also like to thank Cedar McKay for contributing a script to this workflow and for his guidance in designing this pipeline. This work was supported through a National Science Foundation Graduate Research Fellowship and a NASA Earth and Space Sciences Fellowship to Jaclyn K. Saunders, and through National Science Foundation grants OCE-1138368 and OCE-1356779 to Gabrielle Rocap.

A1.3 References:

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

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

3 Koski LB, Golding GB (2001). The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution* **52:** 540-542.

4 Berger SA, Krompass D, Stamatakis A (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology* **60:** 291-302.

5 Stamatakis A (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*.

6 Matsen FA, Kodner RB, Armbrust EV (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11:** 538.

7 Filipski A, Tamura K, Billing-Ross P, Murillo O, Kumar S (2015). Phylogenetic placement of metagenomic reads using the minimum evolution principle. *BMC Genomics* **16:** S13-S13.

8 Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

9 Berger SA, Stamatakis A (2011). Aligning short reads to reference alignments and trees. *Bioinformatics* **27:** 2068-2075.

10 Berger SA, Stamatakis A (2012). PaPaRa 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension. *Heidelberg Institute for Theoretical Studies, http://scoh-itsorg/exelixis/publicationshtmlExelixis-RRDR-2012-2015*.

11 Stark M, Berger SA, Stamatakis A, von Mering C (2010). MLTreeMap - accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* **11:** 1-11.

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

13 Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D *et al.* (2009). MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Research* **38:** D396-D400.

14 Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32:** 1792-1797.

15 Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*.

16 Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G *et al.* (2015). Structure and function of the global ocean microbiome. *Science* **348**.

AI.4 Figure Legends & Tables:

Figure AI.1 A flowchart of the bioinformatics pipeline. Objects represent files, while connecting arrows and actions on objects – either through software or manual curation. Gray text between arrowed lines represents previously existing software while bold italicized text represents software described here. The pipeline is broken up into three major parts: part (1) is the collection and curation of the reference, part (2) is the parsing of likely reads from the metagenome, and part (3) combines the unknown reads with the reference sequences for identification through phylogenetic inference placements.

Figure AI.2 A maximum likelihood tree of the RpsD amino acid sequences of reference sequences used for identification of artificial metagenome reads.

Figure AI.3 The relative abundance of taxonomic groups determined by through placement analysis with the pipeline using either amino acid or nucleic acid sequence, as compared to the set abundance of these groups in the creation of the artificial metagenome.

Table AI.1 List of the programs used in the pipeline and a description of the function the programs serve in the pipeline.

Program	Description of Function in Pipeline	Reference
renameMicrobeOnlineSeqs.py	Renames sequence files which can be pulled from MicrobesOnline.org	This work
makeblastdb	Converts a fasta file of meta-omic sequences reads into a searchable database	[1,2]
tblastn	searches a 6-frame translation of a blast database using protein query sequences	[1,2]
parseBlastXML	parses information from blast search results and retrieves raw sequence reads from blast database	This work
trimFormatReads	Selects the best match from the blast query, ensures reads are in proper reading frame from blast match, trims reads to reduce potential overhang of up/down-stream sequence from target, removes ambiguities in sequence, and reports information on recruited reads	This work
muscle	alignment program used for reference sequences	[14]
fastaToPhyml	converts a fasta formatted file to Phyml format	This work
proteinModelSelection	RAxML-based program to find the optimal amino acid substitution model	[5,8]
RAxML	Maximum Likelihood phylogenetic tree construction	[5,8]
PaPaRa	Phylogenetically aware alignment program for aligning recruited unknown reads to the reference alignment	[9,10]
spaceJoin	Combines read pairs which are known to be linked and provides option for minimum length requirement of reads to be analyzed for placement	This work
RAxML-EPA	Maximum likelihood based assignment of unknown reads to a reference phylogeny	[4,5]
taxaDict	Creates a file where individual nodes on a tree can be manually labeled and curated for read assignment	This work
treeLabels	Labels read assignments based upon node labels determined by taxaDict; sorts and tabulates read placement by metagenome key	This work

Table AI.2 The genomes used to create an artificial metagenome of 5,000,000 reads with the relative abundance of reads of each genome listed. The broad taxonomic group that the bacterial genomes are from is also displayed.

Genome	Relative Abundance %	Group
Candidatus Pelagibacter ubique HTCC1062 NCBI taxonomyId 335992	32	Alphaproteobacteria
Bacteroides plebeius DSM 17135 NCBI taxonomyId 484018	13	Bacteroidetes
Shewanella baltica OS155 NCBI taxonomyId 325240	10	Gammaproteobacteria
Bacillus selenitireducens MLS10 NCBI taxonomyId 439292	7.5	Firmicutes
Prochlorococcus marinus str MIT 9312 NCBI taxonomyId 74546	5	Cyanobacteria
Marinomonas sp MED121 NCBI taxonomyId 314277	5	Gammaproteobacteria
Corynebacterium glutamicum ATCC 13032 NCBI taxonomyId 196627	4.6	Actinobacteria
Vibrio sp MED222 NCBI taxonomyId 314290	3.2	Gammaproteobacteria
Pseudomonas aeruginosa 2192 NCBI taxonomyId 350703	2.	Gammaproteobacteria
Deinococcus geothermalis DSM 11300 NCBI taxonomyId 319795	2.2	Deinococcus-Thermus
Prochlorococcus marinus str NATL2A NCBI taxonomyId 59920	2	Cyanobacteria
Pseudomonas aeruginosa PA7 NCBI taxonomyId 381754	2	Gammaproteobacteria
Burkholderia sp 383 NCBI taxonomyId 269483	1.7	Betaproteobacteria
Nitrosococcus halophilus Nc4 NCBI taxonomyId 472759	1.7	Gammaproteobacteria
Pseudoalteromonas atlantica T6c NCBI taxonomyId 342610	1.5	Gammaproteobacteria
Colwellia psychrerythraea 34H NCBI taxonomyId 167879	1	Gammaproteobacteria
Escherichia coli W3110 NCBI taxonomyId 316407	1	Gammaproteobacteria
Escherichia coli HS NCBI taxonomyId 331112	1	Gammaproteobacteria
Synechococcus sp RS9917 NCBI taxonomyId 221360	0.9	Cyanobacteria
Chlorobium limicola DSM 245 NCBI taxonomyId 290315	0.8	Chlorobi
Thermodesulfovibrio yellowstonii DSM 11347 NCBI taxonomyId 289376	0.8	Nitrospirae
Geobacter lovleyi SZ NCBI taxonomyId 398767	0.5	Deltaproteobacteria
Planctomyces maris DSM 8797 NCBI taxonomyId 344747	0.2	Planctomycetes
Desulfovibrio vulgaris str DP4 NCBI taxonomyId 391774	0.1	Deltaproteobacteria

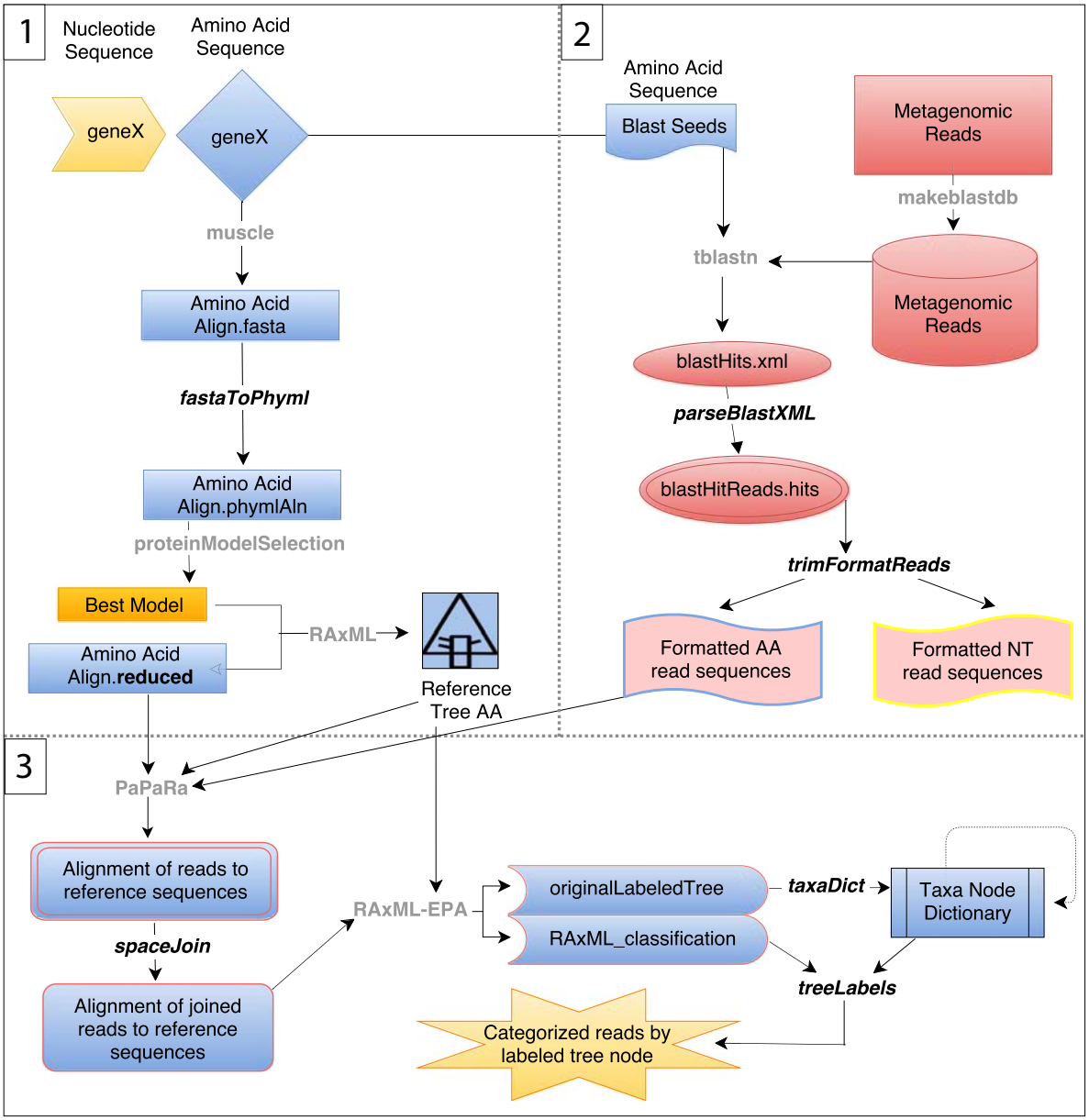

1
Nucleotide Sequence
Amino Acid Sequence
geneX
geneX
muscle
Amino Acid Align.fasta
fastaToPhyml
Amino Acid Align.phymlAln
proteinModelSelection
Best Model
Amino Acid Align.reduced
RAxML
Reference Tree AA
2
Amino Acid Sequence
Blast Seeds
Metagenomic Reads
makeblastdb
tblastn
Metagenomic Reads
blastHits.xml
parseBlastXML
blastHitReads.hits
trimFormatReads
Formatted AA read sequences
Formatted NT read sequences
3
PaPaRa
Alignment of reads to reference sequences
spaceJoin
Alignment of joined reads to reference sequences
RAxML-EPA
originalLabeledTree
RAxML_classification
taxaDict
Taxa Node Dictionary
treeLabels
Categorized reads by labeled tree node

Figure A1.1

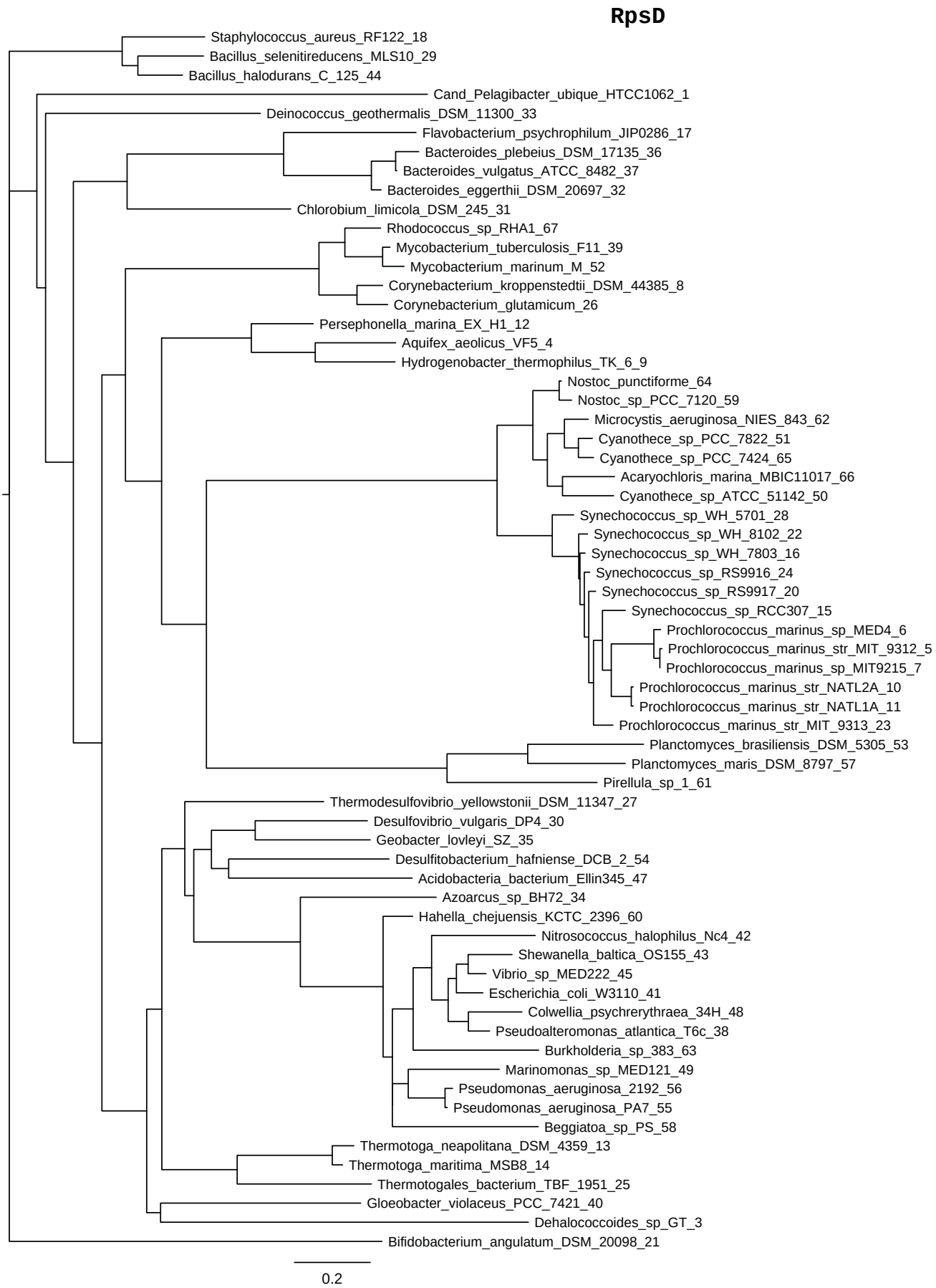

RpsD
Staphylococcus_aureus_RF122_18
Bacillus_selenitireducens_MLS10_29
Bacillus_halodurans_C_125_44
Cand_Pelagibacter_ubique_HTCC1062_1
Deinococcus_geothermalis_DSM_11300_33
Flavobacterium_psychrophilum_JIP0286_17
Bacteroides_plebeius_DSM_17135_36
Bacteroides_vulgatus_ATCC_8482_37
Bacteroides_eggerthii_DSM_20697_32
Chlorobium_limicola_DSM_245_31
Rhodococcus_sp_RHA1_67
Mycobacterium_tuberculosis_F11_39
Mycobacterium_marinum_M_52
Corynebacterium_kroppenstedtii_DSM_44385_8
Corynebacterium_glutamicum_26
Persephonella_marina_EX_H1_12
Aquifex_aeolicus_VF5_4
Hydrogenobacter_thermophilus_TK_6_9
Nostoc_punctiforme_64
Nostoc_sp_PCC_7120_59
Microcystis_aeruginosa_NIES_843_62
Cyanothece_sp_PCC_7822_51
Cyanothece_sp_PCC_7424_65
Acaryochloris_marina_MBIC11017_66
Cyanothece_sp_ATCC_51142_50
Synechococcus_sp_WH_5701_28
Synechococcus_sp_WH_8102_22
Synechococcus_sp_WH_7803_16
Synechococcus_sp_RS9916_24
Synechococcus_sp_RS9917_20
Synechococcus_sp_RCC307_15
Prochlorococcus_marinus_sp_MED4_6
Prochlorococcus_marinus_str_MIT_9312_5
Prochlorococcus_marinus_sp_MIT9215_7
Prochlorococcus_marinus_str_NATL2A_10
Prochlorococcus_marinus_str_NATL1A_11
Prochlorococcus_marinus_str_MIT_9313_23
Planctomyces_brasiliensis_DSM_5305_53
Planctomyces_maris_DSM_8797_57
Pirellula_sp_1_61
Thermodesulfovibrio_yellowstonii_DSM_11347_27
Desulfovibrio_vulgaris_DP4_30
Geobacter_lovleyi_SZ_35
Desulfitobacterium_hafniense_DCB_2_54
Acidobacteria_bacterium_Ellin345_47
Azoarcus_sp_BH72_34
Hahella_chejuensis_KCTC_2396_60
Nitrosococcus_halophilus_Nc4_42
Shewanella_baltica_OS155_43
Vibrio_sp_MED222_45
Escherichia_coli_W3110_41
Colwellia_psychrerythraea_34H_48
Pseudoalteromonas_atlantica_T6c_38
Burkholderia_sp_383_63
Marinomonas_sp_MED121_49
Pseudomonas_aeruginosa_2192_56
Pseudomonas_aeruginosa_PA7_55
Beggiatoa_sp_PS_58
Thermotoga_neapolitana_DSM_4359_13
Thermotoga_maritima_MSB8_14
Thermotogales_bacterium_TBF_1951_25
Gloeobacter_violaceus_PCC_7421_40
Dehalococcoides_sp_GT_3
Bifidobacterium_angulatum_DSM_20098_21
0.2

Figure AI.2

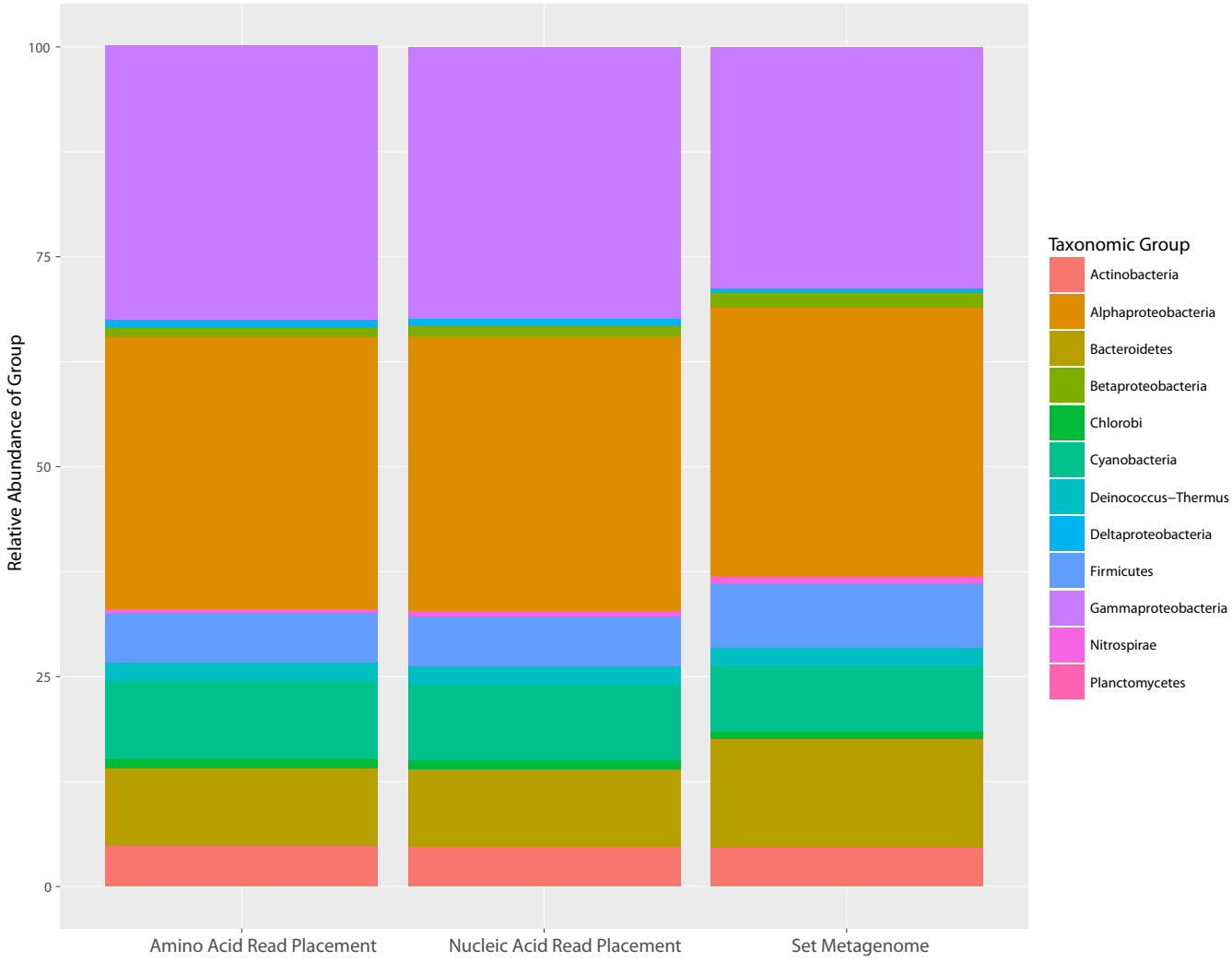

100
75
50
25
0
Relative Abundance of Group
Amino Acid Read Placement
Nucleic Acid Read Placement
Set Metagenome
Taxonomic Group
Actinobacteria
Alphaproteobacteria
Bacteroidetes
Betaproteobacteria
Chlorobi
Cyanobacteria
Deinococcus–Thermus
Deltaproteobacteria
Firmicutes
Gammaproteobacteria
Nitrospirae
Planctomycetes

Figure AI.3

Appendix II

Testing the accuracy of phylogenetic read placement in amino acid space: *Prochlorococcus* ecotype placement in a simulated metagenome

AII.1 Body:

The bioinformatics placement pipeline described in Appendix I has been used for the classification of metagenomics reads in amino acid as well as nucleotide space. Presented here is an analysis of the fidelity of phylogenetic placements in amino acid space using a simulated metagenome to test the accuracy of clade level placements of the marine picocyanobacterium *Prochlorococcus*. The original reference alignment program and likelihood-based placement methods were designed for nucleotide based sequences, a key target being 16S rDNA(Berger *et al.*, 2011; Berger and Stamatakis, 2011). In order to test the ability of these methods to place in amino acid space, the level on which evolutionary forces are more directly acting, an optimization and placement accuracy analysis is presented in order to identify optimal parameters and capabilities of assignment in amino acid space among closely related organisms.

Prochlorococcus is an excellent model system to test the fidelity of the phylogenetic placement method. The genus *Prochlorococcus* consists of strains which show >97% similarity in 16S rRNA(Moore *et al.*, 1998; Kettler *et al.*, 2007) with division among the strains according to physiological light intensity optima, breaking up into High Light and Low Light clades(Moore *et al.*, 1998). These clades can be further subdivided into genetically and physiologically distinct ecotypes(Rocap *et al.*, 2002; Ahlgren *et al.*, 2006). Sequences for twelve *Prochlorococcus* genomes, with representatives in each ecotype (High Light I, High Light II, Low Light I, Low Light II, Low Light III, and Low Light IV), have been extensively studied and were used for the creation of reference trees for metagenomics read identification(Kettler *et al.*, 2007). Additional *Prochlorococcus* genomes which were more recently sequenced(Biller *et al.*, 2014) were added

to the simulated metagenome to provide environmental variability since a perfect representative reference sequence will often not exist for environmental microbial sequences. The evaluated metric for accuracy of placements was how accurately *Prochlorococcus* metagenomics reads were placed to the correct ecotype.

The artificial metagenome consisted of *Prochlorococcus* as well as other representative bacterial groups (Figure 1). Representatives from the sister taxa *Synechococcus* were also included in simulated metagenome, as Low Light IV is the most phylogenetically basal of the *Prochlorococcus* strains, and on a gene level basis some Low Light IV genes are more closely related to *Synechococcus* than to other *Prochlorococcus.* Other representative bacterial taxa were also included in the simulated metagenome, like *Pelagibacter ubique* which is the most abundant taxonomic group in surface marine metagenomes(Rusch *et al.*, 2007). The artificial metagenome contains ~ 6 million reads which were produced from a set of sequenced microbial genomes in varying abundance (Figure 1), with read lengths and characteristics set to reflect Illumina HiSeq sequencing technology used in the metagenomes from the anoxic waters of the East Tropical North Pacific described in Chapter 3. Read lengths were set to a maximum length of 149 bases, with paired end reads, and a small insert size producing multiple paired end reads with overlap(Johnson *et al.*, 2014).

Three single copy core housekeeping genes (*glnA, rpsD, tyrS*) in *Prochlorococcus* were evaluated for accuracy of placement among ecotypes. These single copy core housekeeping genes are known to occur in one copy per *Prochlorococcus* genome and encode integral proteins for cellular function(Kettler *et al.*, 2007). The average occurrence of these three genes has been previously used as a baseline for the estimation of the abundance of *Prochlorococcus* genomes captured in various environmental genomes(Saunders and Rocap, 2016). These three genes have

been chosen to estimate *Prochlorococcus* genomes as they range in length (in MED4 - *glnA*: 1422 bp; *rpsD*: 609 bp; *tyrS*: 1239 bp), GC content (Table 1), and occur in different regions of the *Prochlorococcus* genomes. Using the bioinformatics pipeline outlined in Appendix I, the accuracy of placement of artificial metagenome reads per ecotype was assessed for the three housekeeping genes. Various length cutoffs of metagenomics reads were also assessed. Specifically, the length of reads which were passed through to the alignment step, PaPaRa(Berger and Stamatakis, 2011), and the length of reads passed to the placement step, RAxML-EPA(Berger *et al.*, 2011; Stamatakis, 2014), were evaluated designated as "first cutoff" and "second cutoff", respectively. A smaller read length for the first cutoff may be desirable if working with metagenomic read pairs as an individual read may not contain enough information for placement, but when read pairs are able to be joined together after the alignment step by spaceJoin they may be long enough and contain enough phylogenetic signal for placement, with the length cutoff for placement controlled by the second cutoff.

For each of the three housekeeping genes, *glnA, rpsD,* and *tyrS*, the accuracy with which the metagenomics reads could be placed to the correct ecotype was assessed on each phylogenetic tree (Supplemental Figures 1-3). Accuracy of placement is calculated as the number of *Prochlorococcus* metagenome reads which placed to the correct ecotype divided by the total number of *Prochlorococcus* reads that placed to the target gene in the placement step (RAxML-EPA). Various read length cutoffs were assessed for impact on accuracy of placement. A first read length cutoff of 20, 50, and 99 nucleotide length requirements for placement was assessed, as was a second length cutoff of 25, 33, 45, and 75 amino acid length cutoff requirements for the placement step. Accuracy of placements for the three housekeeping genes under varying length cutoffs is depicted in Figure 2. The variability among the three

housekeeping genes with regard to placement accuracy is visible in this graph. Overall, the general trend is the longer the read length requirement, the greater the accuracy in placement. However, there is a tradeoff for this as the total number of reads for placement analysis declines with increased read length requirement (Figure 3). There was 31% reduction in the total number of *Prochlorococcus* reads placed when comparing the least stringent length cutoffs (20 nucleotides first cutoff and 25 amino acids for second cutoff) to the most stringent length cutoff (99 nucleotides first cutoff and 75 amino acids second cutoff). Overall, more stringent length cutoffs result in a reduced number of total reads placed, but the more accurate the placements. The second cutoff (placement) has the greatest influence on the total number of reads placed. The lowest accuracy of placement, 85.3%, was for *rpsD* using a second cutoff of 25 amino acids – meaning metagenomic reads of at least 25 amino acids were placed to the reference tree. The highest accuracy of placement, 99.5%, was for *tyrS*, using a final cutoff of 75 amino acids (Figure 2). If targeting a rare organism in a metagenome, it may be favorable to use a less stringent length cutoff to boost the overall number of reads assessed for placement, but for abundant taxa one may favor the accuracy of placement provided by more stringent length cutoffs.

Abundance of *Prochlorococcus* in metagenomes has been previously estimated by using an average of the length normalized abundance, abundance of read placements normalized by the full length sequences of a target gene, of the housekeeping genes in a metagenome. The ability of placement analyses to estimate the abundance of *Prochlorococcus* genomes in a metagenome was assessed using the most stringent length cutoffs of 99 nucleotides for the first cutoff (alignment) and 75 amino acids for the second cutoff (placement). Figure 4 depicts the difference in the length normalized abundance of a gene estimated through placement when compared to

the fixed abundance in the artificial metagenome. Generally, the average of the three single copy core genes better reflects the set abundance in the artificial metagenome, this is especially evident in the High Light I ecotype where a dramatic difference in the single copy core gene *rpsD* compared with *glnA* and *tyrS* is dampened into an average of 0.0067 from the expected length normalized abundance. A noticeable trend here is the underestimation of the abundance of *Prochlorococcus* Low Light IV ecotype by placement analysis, where placement estimated abundance for all genes is < 0, with an average length normalized difference of -0.0454. This is noteworthy as this ecotype is the most basal of the *Prochlorococcus* ecotypes, and due to topology of the tree this subclade is the most difficult to identify through phylogenetically informed placement approaches. It is likely that this same behavior carries over to other basal groups which highlights the need to evaluate tree topology in the interpretation of placement analyses.

Using the model system *Prochlorococcus*, the fidelity of the phylogenetically informed placement bioinformatics pipeline described in Appendix I was evaluated and parameters optimized. A major finding of this analysis is that placement under amino acid space is capable of providing high resolution metagenomic read identification, as assessed through the accuracy of placement of reads in a simulated metagenome to respective *Prochlorococcus* ecotypes. Assessed here is placement to single copy core housekeeping genes, which are a conserved throughout the picocyanobacteria. If metagenomes are deeply sequenced, the targeted taxa is abundant, and high resolution of placements is desired (such as distinguishing reads among various strains within a genus), it may be favorable to err on the side of greater accuracy of placements by maintaining a longer read length requirement in the second cutoff (placement). However, for general purposes, read length cutoffs of 50 nucleotides for the first placement and

45 amino acids for the second placement should be sufficient, yielding accuracy in placements to specific ecotypes for all genes assessed greater than 90%. If the goal is to identify reads associated with broad groups, and if the target is low in abundance, it may be favorable to reduce length requirements to boost the total number of reads assessed albeit with the potential more inaccurate placements. Overall, the phylogenetically informed placement approach shows great value in the characterization of marine metagenomic datasets.

AII.2 References:

1 Berger SA, Krompass D, Stamatakis A (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology* **60:** 291-302.

2 Berger SA, Stamatakis A (2011). Aligning short reads to reference alignments and trees. *Bioinformatics* **27:** 2068-2075.

3 Moore LR, Rocap G, Chisholm SW (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393:** 464-467.

4 Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3:** e231.

5 Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002). Resolution of *Prochlorococcus* and *Synechococcus* Ecotypes by Using 16S-23S Ribosomal DNA Internal Transcribed Spacer Sequences. *Applied and Environmental Microbiology* **68:** 1180-1191.

6 Ahlgren NA, Rocap G, Chisholm SW (2006). Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. *Environmental Microbiology* **8:** 441-454.

7 Biller SJ, Berube PM, Berta-Thompson JW, Kelly L, Roggensack SE, Awad L *et al.* (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific data* **1:** 140034.

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

9 Johnson S, Trost B, Long JR, Pittet V, Kusalik A (2014). A better sequence-read simulator program for metagenomics. *BMC Bioinformatics* **15 Suppl 9:** S14.

10 Saunders JK, Rocap G (2016). Genomic potential for arsenic efflux and methylation varies among global *Prochlorococcus* populations. *ISME J* **10:** 197-209.

11 Stamatakis A (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*.

AII.3 Figure Legends & Tables:

Figure AII.1 The relative abundance of various microbial genomes in the artificial metagenome. *Prochlorococcus* genome colors reflect the ecotype to which they belong. The branching pattern next to the *Prochlorococcus* ecotypes in the key reflects the general branching topology of the ecotypes in 16S rRNA sequence.

Figure AII.2 Accuracy of placements as assessed by the correct placement of a *Prochlorococcus*-linked metagenomic read to its respective ecotype. Depicted are the three housekeeping genes *glnA, rpsD,* and *tyrS* under different length cutoffs: first cutoff (for alignment) set at 20, 50, or 99 nucleotides; second cutoff (for placement) set at 25, 33, 45, or 75 amino acids.

Figure AII.3 Total number of *Prochlorococcus* metagenome reads placed to various genes. Correct placements of reads to their respective ecotypes are highlighted by the gene-specific color. Reads placed to an improper clade are depicted by the gray bar. Overall, as longer reads

are required for the first cutoff (alignment) and the second cutoff (placement), the overall number of reads placed is reduced. The second cutoff length requirement has the most dramatic impact on the total number of reads placed.

Figure AII.4 Difference of length normalized abundance of *Prochlorococcus*-linked reads for the single copy core housekeeping genes for the various ecotypes estimated through placement analysis when compared to the fixed abundance in the artificial metagenome. A value of 0 indicates that the length normalized abundance of a gene estimated by placement perfectly reflects the set abundance in the artificial metagenome. Values > 1 indicate where *Prochlorococcus* genes are overestimated relative to the set proportions in the artificial metagenome, and values < 1 indicate where *Prochlorococcus* genes are underestimated relative to the set proportions. Depicted are the individual genes as well as the average length normalized abundance of the genes. Data is presented as the abundance estimated for each ecotype. Due to the phylogenetic topology, groups Low Light II and III are presented together.

Table AII.1 The GC% of *Prochlorococcus* strains representative of the various ecotypes and the GC% of the single copy core housekeeping genes found in each of those strains.

Clade	Strain	Genome GC%	glnA GC%	rpsD GC%	tyrS GC%
HL1	MED4	30.8	36.99	38.42	30.75
HL2	AS9601	31.32	36.36	39.08	31.64
LL1	NATL1A	34.98	40.3	43.84	32.69
LL2/3	SS120	36.44	39.71	44.66	37.57
LL4	MIT9313	50.74	51.83	55.83	53.95

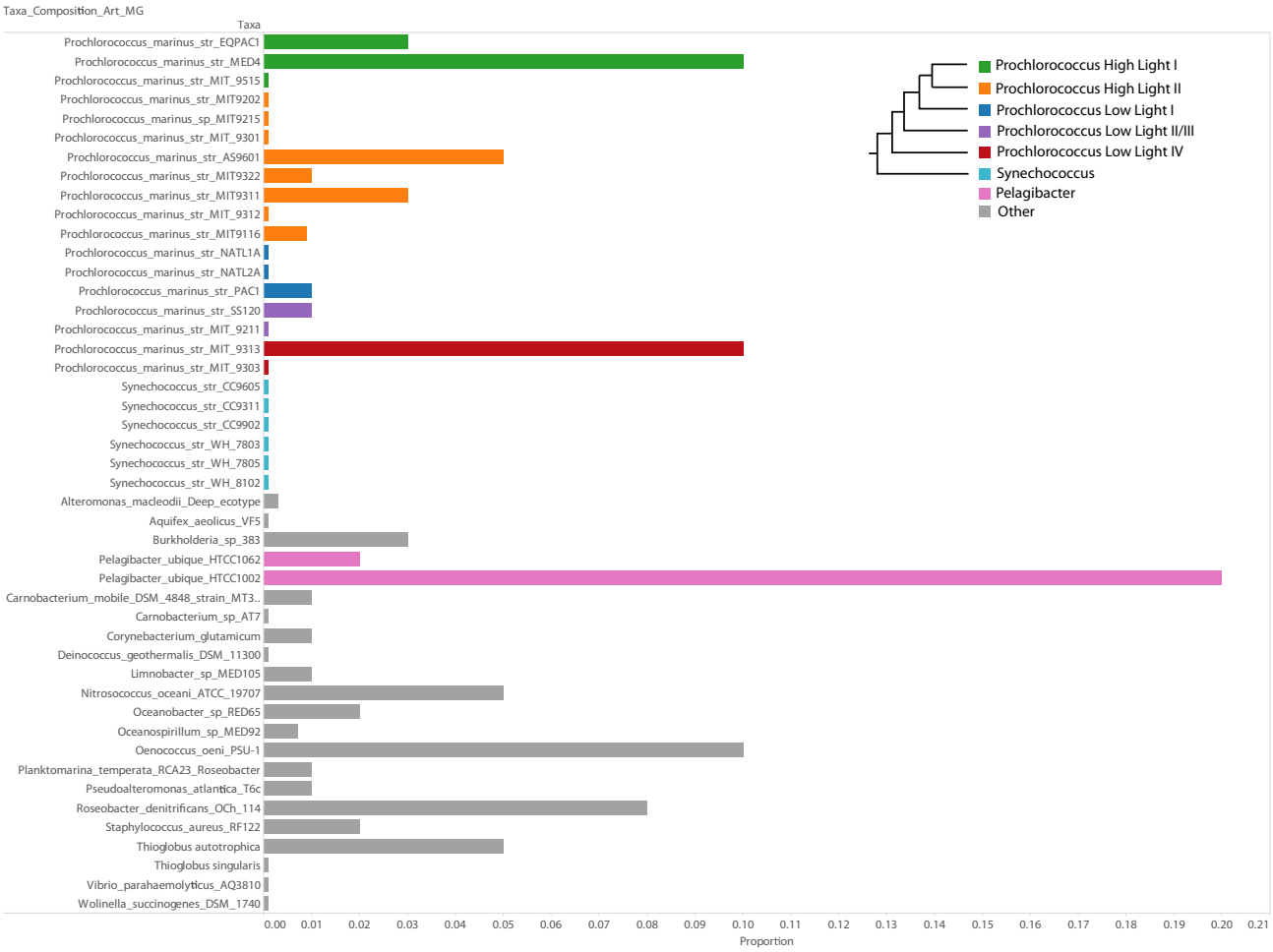

Taxa_Composition_Art_MG
Taxa
Prochlorococcus_marinus_str_EQPAC1
Prochlorococcus_marinus_str_MED4
Prochlorococcus_marinus_str_MIT_9515
Prochlorococcus_marinus_str_MIT9202
Prochlorococcus_marinus_sp_MIT9215
Prochlorococcus_marinus_str_MIT_9301
Prochlorococcus_marinus_str_AS9601
Prochlorococcus_marinus_str_MIT9322
Prochlorococcus_marinus_str_MIT9311
Prochlorococcus_marinus_str_MIT_9312
Prochlorococcus_marinus_str_MIT9116
Prochlorococcus_marinus_str_NATL1A
Prochlorococcus_marinus_str_NATL2A
Prochlorococcus_marinus_str_PAC1
Prochlorococcus_marinus_str_SS120
Prochlorococcus_marinus_str_MIT_9211
Prochlorococcus_marinus_str_MIT_9313
Prochlorococcus_marinus_str_MIT_9303
Synechococcus_str_CC9605
Synechococcus_str_CC9311
Synechococcus_str_CC9902
Synechococcus_str_WH_7803
Synechococcus_str_WH_7805
Synechococcus_str_WH_8102
Alteromonas_macleodii_Deep_ecotype
Aquifex_aeolicus_VF5
Burkholderia_sp_383
Pelagibacter_ubique_HTCC1062
Pelagibacter_ubique_HTCC1002
Carnobacterium_mobile_DSM_4848_strain_MT3..
Carnobacterium_sp_AT7
Corynebacterium_glutamicum
Deinococcus_geothermalis_DSM_11300
Limnobacter_sp_MED105
Nitrosococcus_oceani_ATCC_19707
Oceanobacter_sp_RED65
Oceanospirillum_sp_MED92
Oenococcus_oeni_PSU-1
Planktomarina_temperata_RCA23_Roseobacter
Pseudoalteromonas_atlantica_T6c
Roseobacter_denitrificans_OCh_114
Staphylococcus_aureus_RF122
Thioglobus autotrophica
Thioglobus singularis
Vibrio_parahaemolyticus_AQ3810
Wolinella_succinogenes_DSM_1740
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.20 0.21
Proportion
Prochlorococcus High Light I
Prochlorococcus High Light II
Prochlorococcus Low Light I
Prochlorococcus Low Light II/III
Prochlorococcus Low Light IV
Synechococcus
Pelagibacter
Other

Figure AII.1

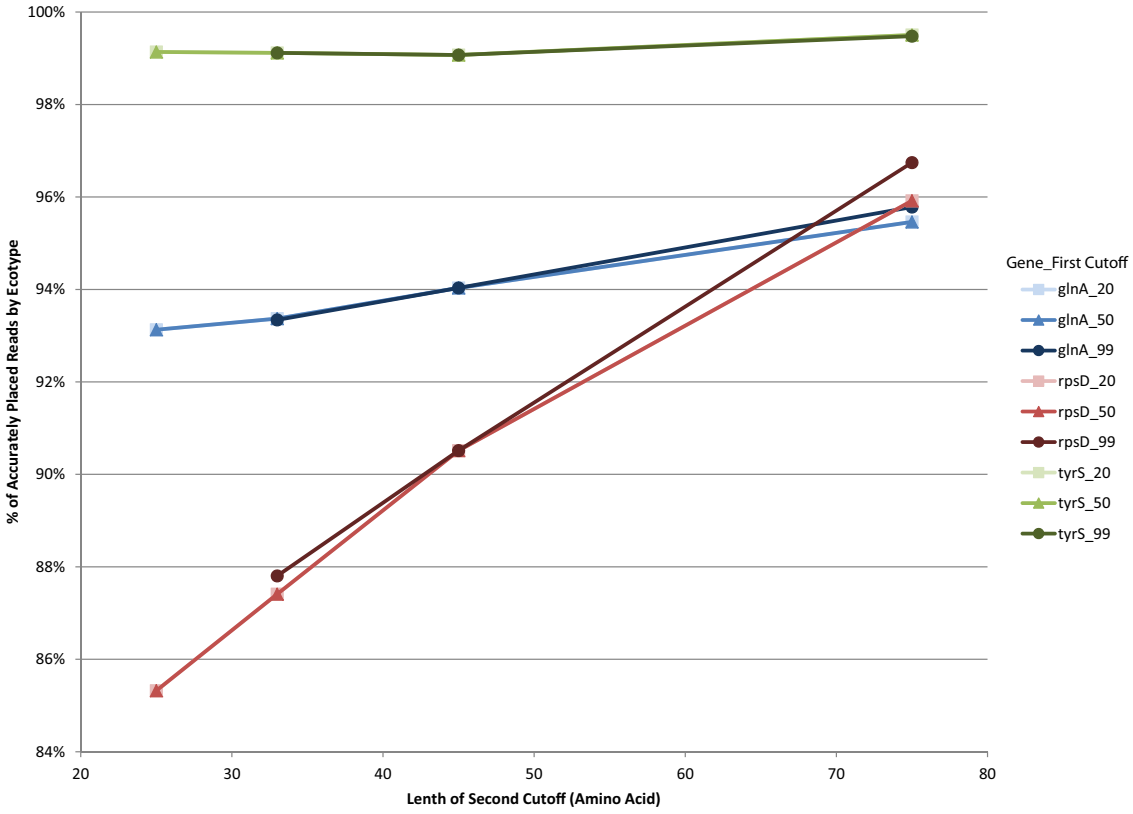

100%
98%
96%
94%
92%
90%
88%
86%
84%
% of Accurately Placed Reads by Ecotype
20
30
40
50
60
70
80
Lenth of Second Cutoff (Amino Acid)
Gene_First Cutoff
glnA_20
glnA_50
glnA_99
rpsD_20
rpsD_50
rpsD_99
tyrS_20
tyrS_50
tyrS_99

Figure AII.2

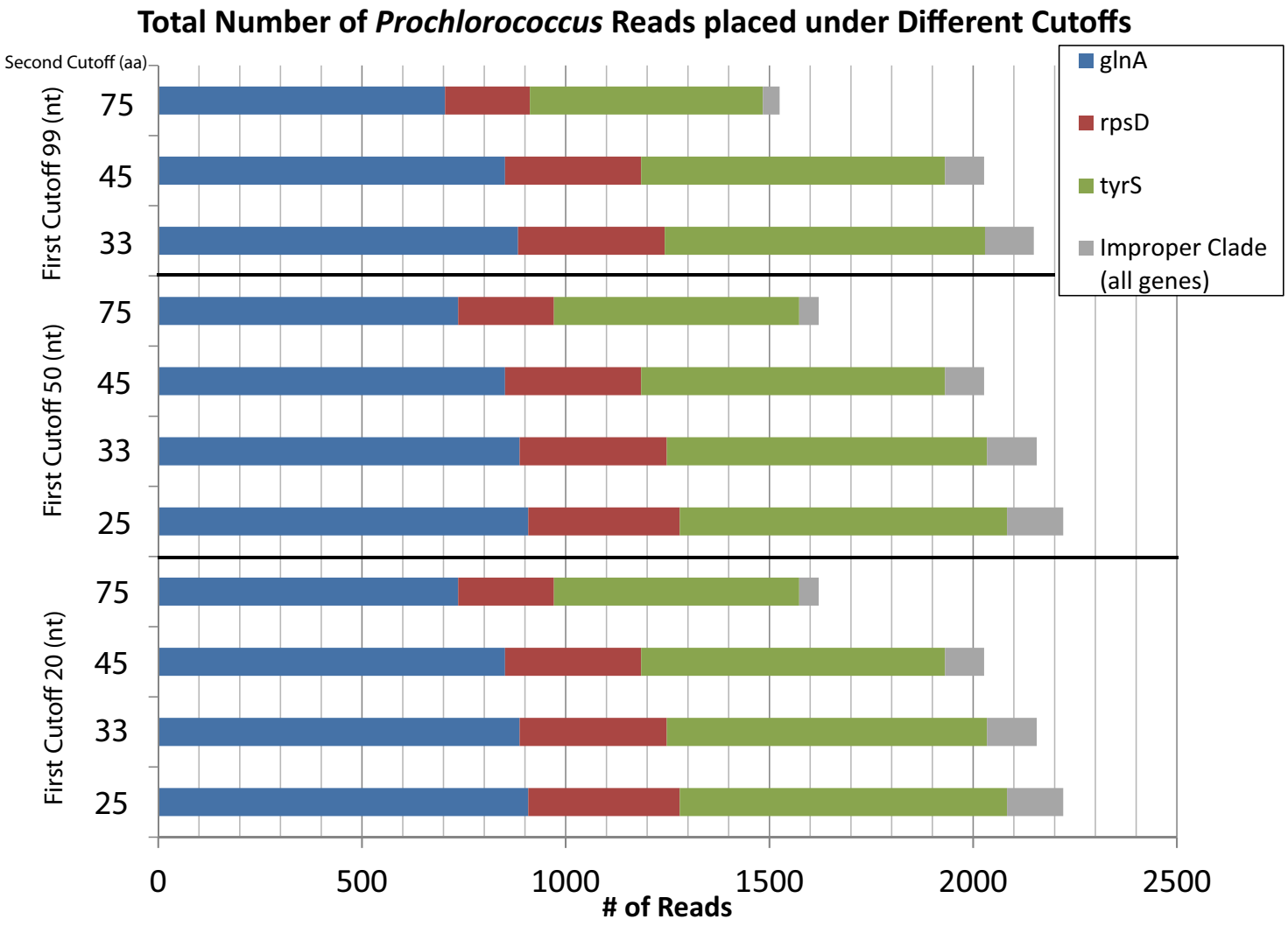

Total Number of *Prochlorococcus* Reads placed under Different Cutoffs
Second Cutoff (aa)
First Cutoff 99 (nt)
First Cutoff 50 (nt)
First Cutoff 20 (nt)
75
45
33
75
45
33
25
75
45
33
25
glnA
rpsD
tyrS
Improper Clade (all genes)
0
500
1000
1500
2000
2500
# of Reads

Figure AII.3

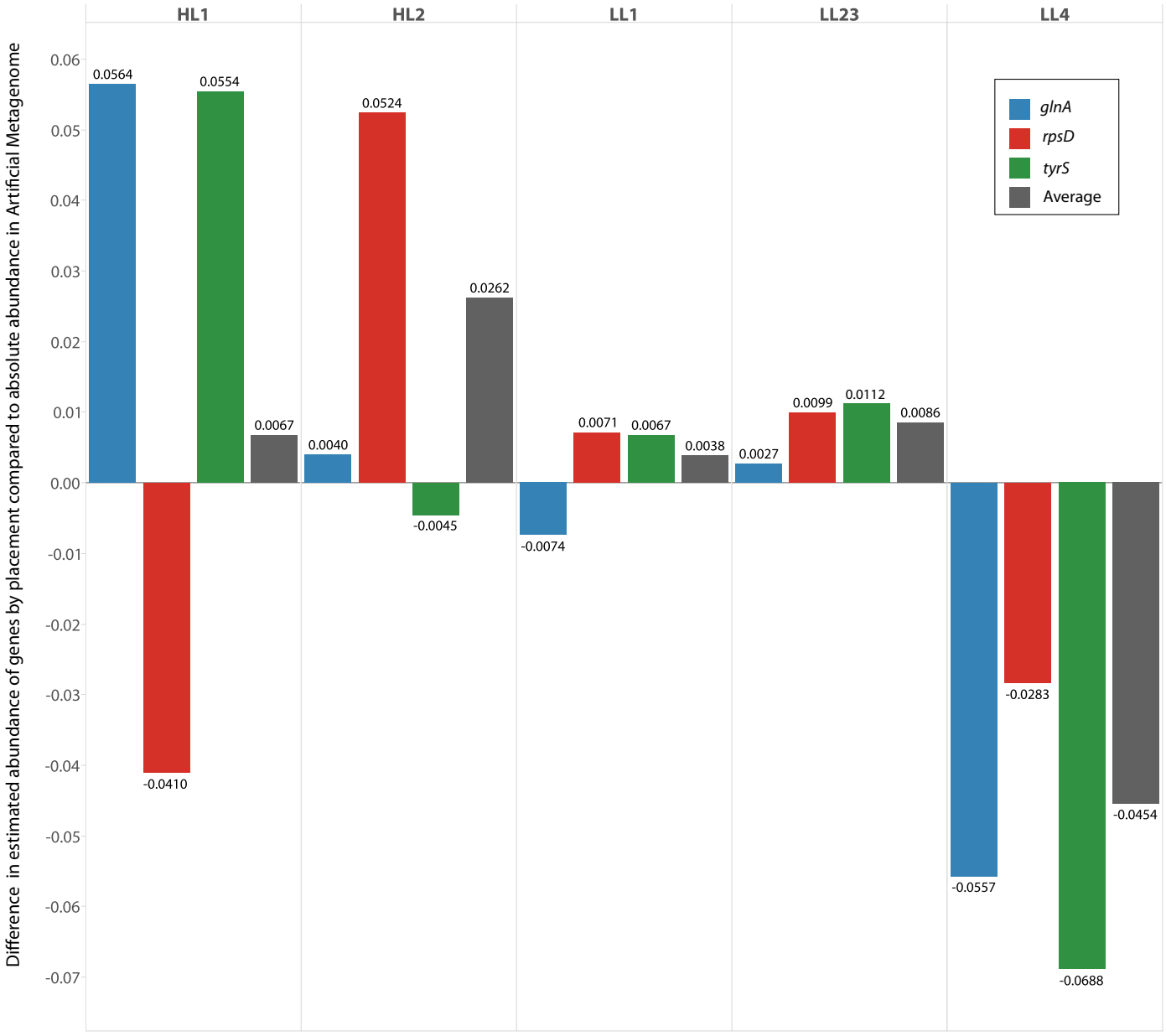

HL1
HL2
LL1
LL23
LL4
Difference in estimated abundance of genes by placement compared to absolute abundance in Artificial Metagenome
0.06
0.05
0.04
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
glnA
rpsD
tyrS
Average
0.0564
-0.0410
0.0554
0.0067
0.0040
0.0524
-0.0045
0.0262
-0.0074
0.0071
0.0067
0.0038
0.0027
0.0099
0.0112
0.0086
-0.0557
-0.0283
-0.0688
-0.0454

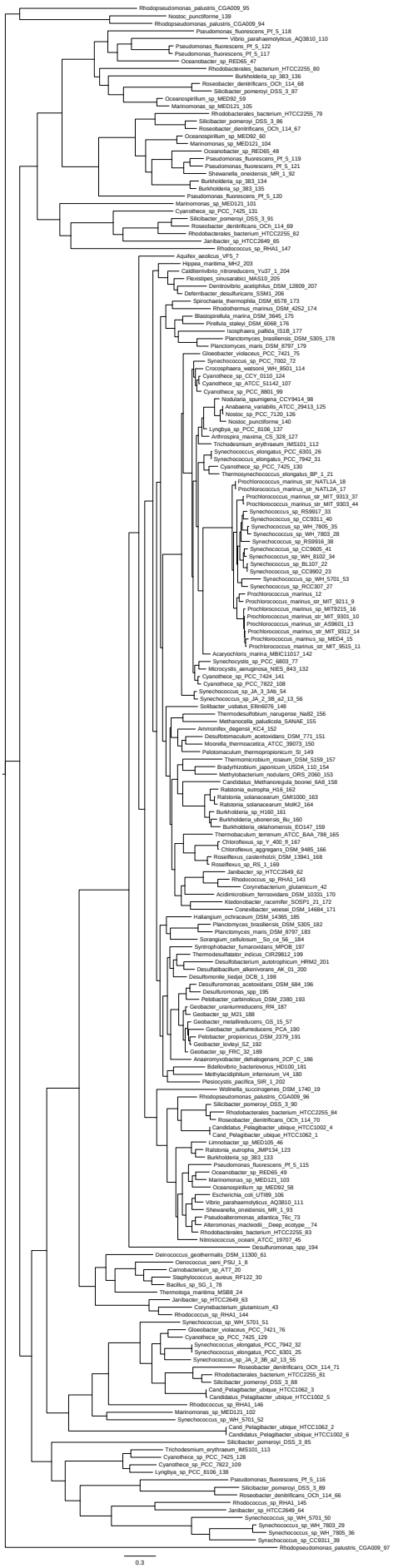
Figure AII.4

AII.4 Supplementary Figure Legends:

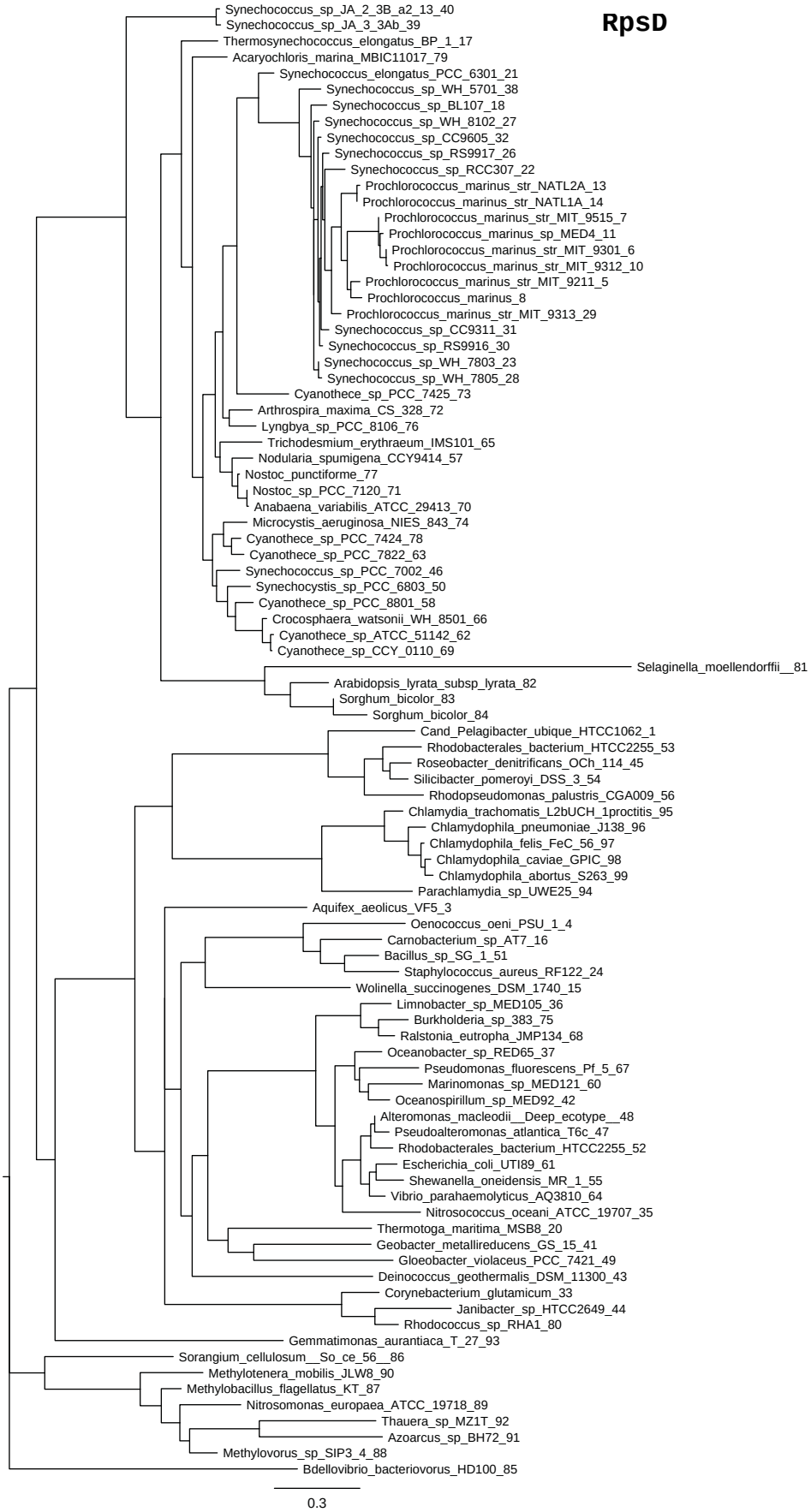
Supplementary Figure AII.1. A maximum likelihood tree of the GlnA amino acid sequences of reference sequences used for identification of artificial metagenome reads. Tree constructed from the best of 20 starting trees using the RTREV amino acid substitution model.

Supplementary Figure AII.2. A maximum likelihood tree of the RpsD amino acid sequences of reference sequences used for identification of artificial metagenome reads. Tree constructed from the best of 20 starting trees using the RTREV amino acid substitution model.

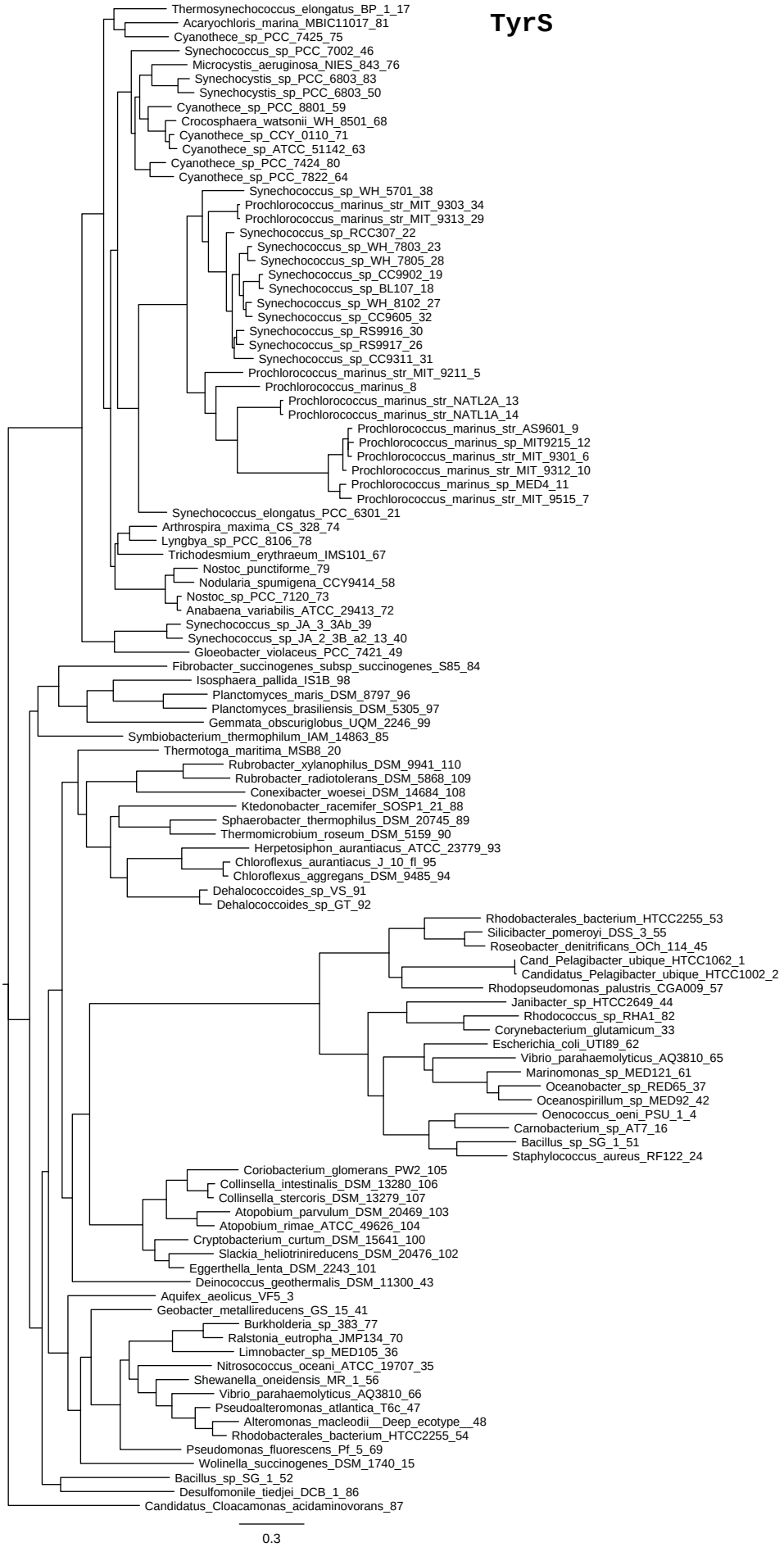
Supplementary Figure AII.3. A maximum likelihood tree of the TyrS amino acid sequences of reference sequences used for identification of artificial metagenome reads. Tree constructed from the best of 20 starting trees using the WAG amino acid substitution model.



Supplementary Figure AII.1



Supplementary Figure AII.2



Supplementary Figure AII.3