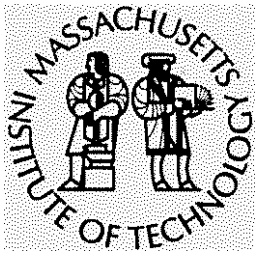
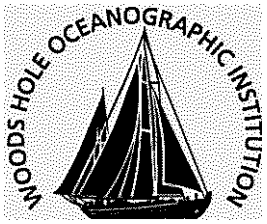
MIT/WHOI

Massachusetts Institute of Technology
Woods Hole Oceanographic Institution



Joint Program in Oceanography/ Applied Ocean Science and Engineering



DOCTORAL DISSERTATION

Protein regulation in *Trichodesmium* and other marine bacteria: Observational and interpretive biomarkers of biogeochemical processes

by

Noelle Adriana Held

February 2020

Protein regulation in *Trichodesmium* and other marine bacteria: Observational and interpretive biomarkers of biogeochemical processes

by

Noelle Adriana Held

B.S. Stetson University, 2013

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
and the
WOODS HOLE OCEANOGRAPHIC INSTITUTION

February 2020

Signature of Author.....
Joint Program in Oceanography/Applied Ocean Science and Engineering
Massachusetts Institute of Technology
and Woods Hole Oceanographic Institution
December 20, 2019

Certified by .....
Dr. Mak. A. Saito
Thesis Supervisor
Woods Hole Oceanographic Institution

Accepted by .....
Dr. Colleen M. Hansel
Chair, Joint Committee for Chemical Oceanography
Woods Hole Oceanographic Institution

Protein regulation in *Trichodesmium* and other marine bacteria: Observational and interpretive biomarkers of biogeochemical processes

by

Noelle Adriana Held

Submitted to the Massachusetts Institute of Technology and Woods Hole Oceanographic Institution on December 20, 2019 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Oceanography and Microbial Biogeochemistry

Abstract

Marine microbes play key roles in global biogeochemistry by mediating chemical transformations and linking nutrient cycles to one another. A major goal in oceanography is to predict the activity of marine microbes across disparate ocean ecosystems. Towards this end, molecular biomarkers are important tools in chemical oceanography because they allow for both the observation and interpretation of microbial behavior. In this thesis, I use molecular biomarkers to develop a holistic, systems biology approach to the study of marine microbes. I begin by identifying unique patterns in the biochemical sensory systems of marine bacteria and suggest that these represent a specific adaptation to the marine environment. Building from this, I focus on the prevalent marine nitrogen fixer *Trichodesmium*, whose activity affects global nitrogen, carbon, phosphorus, and trace metal cycles. A metaproteomic survey of *Trichodesmium* populations identified simultaneous iron and phosphate co-stress throughout the tropical and subtropical oceans, demonstrating that this is caused by the biophysical limits of membrane space and nutrient diffusion. Tackling the problem at a smaller scale, I investigated the metaproteomes of individual *Trichodesmium* colonies captured from a single field site, and identified significant variability related to iron acquisition from mineral particles. Next, I investigated diel proteomes of cultured *Trichodesmium erythraeum* sp. IMS101 to highlight its physiological complexity and understand how and why nitrogen fixation occurs in the day, despite the incompatibly of the nitrogenase enzyme with oxygen produced in photosynthesis. This thesis develops a fundamental understanding of how *Trichodesmium* and other organisms affect, and are affected by, their surroundings. It indicates that a reductionist approach in which environmental drivers are considered independently may not capture the full complexity of microbe-chemistry interactions. Future work can focus on benchmarking and calibration of the protein biomarkers identified here, as well as continued connection of systems biology frameworks to the study of ocean chemistry.

Thesis supervisor: Dr. Mak A. Saito
Title: Senior Scientist, Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution

Acknowledgements

This work was supported by an MIT Walter A. Rosenblith Presidential Fellowship and a National Science Foundation Graduate Research Program Fellowship under grant number 1122274 [N.Held]. This work was also supported by the WHOI Ocean Ventures fund [N.Held], Gordon and Betty Moore Foundation grant number 3782 [M.Saito], National Science Foundation grant numbers OCE-1657766 [M.Saito], EarthCube-1639714 [M.Saito], OCE-1658030 [M.Saito], and OCE-1260233 [M.Saito], and funding from the UK Natural Environment Research Council (NERC) under grants awarded to C.M. (NE/N001079/1) and M.L. (NE/N001125/1). This thesis was completed during a writing residency at the Turkeyland Cove Foundation.

I am so fortunate to have been surrounded by wonderful people during my time in the joint program and throughout my life.

My advisor, Mak, took a chance on 21-year-old me and started my path to oceanography. Mak's optimism, creativity, and gentle way of encouragement is inspiring, and is something I will strive to emulate in my own interactions as a mentor. Because of Mak, I emerge from graduate school not just a better scientist, but also a better person. I didn't always know that I wanted to be an oceanographer, but I think Mak knew. Perhaps the thing I admire most about him is that he let me figure this out for myself.

Among Mak's talents is his ability to attract fantastic scientists to his group. I am extremely fortunate to have been surrounded by individuals who are not only amazing scientists, but also plain good people. Matt and Dawn have taught me almost every laboratory and field skill I know. In addition, Dawn has shown me what it really means to have a giving spirit, and Matt has shown me how to remain positive despite any and all challenges. I am lucky to have been under their wing in the formative years of my career.

My committee has gone above and beyond the call of duty. I especially want to thank Mike Laub and Forest White for providing their time, perspective, and encouragement on a project outside their normal line of work. John Waterbury taught me how to culture *Trichodesmium*, and has been a wealth of information throughout the writing process. Ben Van Mooy and Tracy Mincer have provided crucial input, brainstorming sessions, and contacts with others in the field. Additionally, I extend my thanks to Dennis McGillicuddy, whose work motivated many aspects of this research, for chairing my defense.

Other collaborators, both near and far, have provided key input on this work. These include Eric Webb, who answered all my texts about *Trichodesmium*, Korrina Kunde, my partner-in-quantitation who became a very good friend, Clare Davis, a fantastic sea-going buddy, and Dave Hutchins, Claire Mahaffey, and Maeve Lohan.

I made many life-long friends during graduate school. My lab-mates Natalie Cohen, Jaclyn Saunders, Michael Mazzotta, Becca Chmiel, and Marissa Kellogg have provided

support, advice, and camaraderie throughout this process. Former lab-mate Nick Hawco has and continues to me a mentor to me in many aspects. I am also lucky to have the example and friendship of two amazing women, Gabriela Farfan and Erin Black, to look up to in the last few years. Classmate Kevin Sutherland has been with me through it all and has become a very good friend. Brittany Widner supplied me with unlimited whit, yoga tips, and fermented foods. Finally, Marianne Acker has somehow managed not to get sick of me despite being with me 24/7. I have seen many times her willingness to go the extra mile for others, myself included. She has taught me a lesson in confidence for which I am extremely grateful.

I had some excellent teachers growing up who helped direct me to this point. Ms. Currie, my middle school language arts teacher, encouraged me to write, a skill that is so crucial to scientific career. Mrs. Brennan, my chemistry teacher at Seminole High School, is the reason I was a chemistry major and therefore probably the instigator of it all. Michael Denner and Harry Price at Stetson University, as well as Eric Brown at Boise State University, provided me with early opportunities and guidance that have served me well.

I grew up surrounded by many very strong, motivated women. Without their example I would never have had the courage to step foot on a research vessel. These include my mother, Jennifer Held, grandmother Barbara Held, Oma Trudy Mader, and Aunt Lynn Krugman. My best friend Rachel Homza, who has the most integrity of any person I have ever met, has kept me grounded throughout graduate school.

I met my partner, A.J. at the start of graduate school, and I am so happy that I did. A.J. has been with me through every up and down, and has gone above and beyond to facilitate this research and my career. He taught me how to perform Western blots, made buffers in bulk on the weekends, and even rescued a lost electrode from Logan Airport. I am so grateful to have such a supportive partner. Into my life A.J. also brought a wonderful family – Elaine Colokathis and Stephen and Deborah Devaux, who made me feel like one of their own from the beginning. Thank you for making the last five years the best of my life. I can't wait to see what we will do with the rest.

I especially want to thank my parents, John and Jennifer Held, who raised me to be the person I am today. I don't know if they always understood what I was doing or why I wanted to do it, but they always supported me and made me know that they were proud. Mom – thank you for encouraging me to do my best, for making sure I am well-dressed, for introducing me to yoga, for picking up all my calls and for giving me your work ethic. Dad – thank you for teaching me how to use power tools, for encouraging my creativity even when it is messy, and for giving me your sense of humor and your dance skills. Most of all, I thank my parents for letting me fly on my own, despite how hard it is for them at times.

This thesis is dedicated to Theodore Mader, who dreamed of going to MIT, and to Lynn Held, lover of Earth's creatures.

"Vast quantities of the little substances mentioned yesterday floating upon the water in large lines a mile or more long and 50 or 100 yards wide, all swimming either immediately upon the surface of the water or not many inches under it. The seamen…began to call it Sea Sawdust, a name certainly not ill adapted to its appearance."

Sir Joseph Banks
HMS Endeavor, Captain James Cook
August 28th, 1770
South Coast of New Guinea

"Sea Sawdust," now known as *Trichodesmium*, collected by the author
R.R.S. James Cook
June 29th, 2017
North Atlantic subtropical gyre

Protein regulation in *Trichodesmium* and other marine bacteria: Observational and interpretive biomarkers of biogeochemical processes

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix 1

Appendix 2

Appendix 3

List of Tables

Chapter 2

Chapter 3

Appendix 3

Appendix 4

List of Supplementary Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Supplementary Tables

Supplementary Tables for Chapter 2

Supplementary Tables for Chapter 3

Supplementary Tables for Chapter 4

Supplementary Tables for Chapter 5

Supplementary Tables for Appendix 1

Supplementary Tables for Appendix 2

CHAPTER 1. Introduction

1.1 MOTIVATION FOR THIS WORK

Microbes play crucial roles in the Earth system, working as mediators of chemical transformations that underlie element cycles on the global scale. These biogeochemical cycles fuel all life on Earth – human life included. The ocean is the largest home for microbes on Earth. It is not just one habitat, but many, organized geographically, temporally, and with depth. A major goal in oceanography is to predict the activity of marine microbes across these disparate ecosystems. This requires an understanding of the ocean's chemical, physical, and geological topography, plus an understanding of how microbes will respond to these features.

Towards this goal, molecular biomarkers are increasingly important tools in oceanography. They are valuable because they can be used both to observe and interpret microbial behavior. For example, distributions of biomarkers can be used to map areas of nutrient stress in the ocean, and can also be used to understand the biological and biophysical mechanisms by which organisms become nutrient stressed in the first place. The latter understanding requires a holistic approach to microbial physiology sometimes referred to as "systems biology." This is fundamentally different from a reductionist approach that seeks to predict microbial behavior in response to individual variables.

This thesis develops a framework for the integration of molecular biomarker and systems biology approaches to answer lingering questions in chemical oceanography. It builds on a rich foundation of observational research, which has identified major drivers of marine microbe activity, such as whether enough nutrients are available for the organisms to grow. In some ways it follows in this example but in other ways it differs, particularly by attempting to highlight complex microbial behaviors that result from the intersection of multiple environmental drivers. This thesis benefits from very recent advances in tandem mass spectrometry, gene sequencing, and bioinformatics, and couples these with traditional techniques such as microscopy. It features at least four major field expeditions, and is therefore facilitated by widespread access to the sea that has only been feasible, especially for women, in the last 50 years. Finally, this thesis is possible only due to a culture of precise, inter-calibrated measurement developed during the GEOTRACES program. Thus, this thesis is made possible by the coalescence of historical observations, technological advancement, a culture of precise oceanographic measurements, and access to the sea.

Much of this thesis is focused on *Trichodesmium*, a marine cyanobacterium of great importance to biogeochemical cycles. Being more complex than most other cyanobacteria, understanding *Trichodesmium* presents an intellectual challenge worth of a career. This has made it a formidable yet exciting topic for a PhD thesis. This thesis makes significant progress towards understanding *Trichodesmium* and how it interacts with its environment, but also identifies a suite of new questions, which have relevance to biogeochemical understandings of the past, current, and future oceans. Most importantly, the work on *Trichodesmium* presented in this thesis provides proof-of-concept for how systems biology approaches can be used to probe microbial biogeochemistry more generally.

A non-scientific goal of this work is to generate appreciation for the complexity of the Earth system. I hope that by gaining knowledge of the intricate biogeochemical

cycles that make our existence possible, we may be inspired to treat the world, and each other, with more respect.

1.2 A BRIEF HISTORY OF MOLECULAR BIOMARKERS OF BIOGEOCHEMICAL PROCESSES

The field of chemical oceanography uses measured distributions of chemicals to infer biological, physical, geological, and chemical processes in the ocean. The intersection of these processes, most specifically the cycling of elements in space or in time, is called biogeochemistry. Large scale, intercalibrated measurement programs such as GEOTRACES have enabled the profiling of chemical elements in the ocean. Because life depends on these elements, these data can be used to predict the behavior of marine organisms. For instance, the growth of marine phytoplankton is often limited by the availability of macro and micronutrients.[1–5] In turn, microbial activity impacts element cycles, not least because microbes catalyze chemical transformations such as carbon and nitrogen fixation.

Chemical oceanography has historically focused on inorganic compounds or small molecules such as metal-binding ligands, which may be biologically derived but are not currently part of a living cell. This provides remarkable information and predictive power about microbial activity, both in the present and past ocean.[6,7] This thesis builds on this theme but flips the question – asking the distribution of living materials, themselves also chemicals, to help us to understand chemistry of the sea. This is possible because molecular biomarkers provide information not only about the organism in question, but also about the chemical and physical context in which it lives.

Any biologically derived compound can be considered a biomarker; here, I focus on molecular biomarkers such as DNA, RNA, and proteins, which are the basis of cellular reproduction and function. There is a specific emphasis on proteins, the molecular machinery responsible for responding to stimuli, transporting elements, and performing chemical transformations in the cell. Proteins are of particular interest because they are the enzymes that ultimately perform biogeochemical reactions such as nitrogen and carbon fixation. In this work, I use protein biomarkers to infer current processes in the ocean, which are in turn the integrated result of past processes occurring from the geologic to the microscale. Development of these biomarkers as indicators of future processes, while a long-term goal of this work, is discussed briefly.

1.3 WHAT MAKES A GOOD BIOMARKER?

The term "biomarker" is used most often in the clinical sciences, where the ideal biomarker is one that can be used as a surrogate for a clinical endpoint. Two classes of biomarkers exist - the predictive and the prognostic.[8] Predictive biomarkers provide information about a current or future outcome based on a treatment. By contrast, a prognostic biomarker provides information on an outcome regardless of treatment.[9] From an oceanographic standpoint, predictive biomarkers are observational tools and include those for nutrient stress, which indicate that a community *is* experiencing starvation. By contrast, a prognostic biomarker is an interpretive tool, such as a sensory system for the

nutrient, the presence of which determines *whether* a cell can directly respond to an input. Both predictive/observational and prognostic/interpretive biomarkers are used in this thesis.

I suggest that a good biomarker has three main characteristics – 1) it should be sensitive considering the dynamic range and time scale of the process being observed, 2) it should be simple, i.e. the biomarker's biochemical role must be understood, and 3) it should be measurable. The last requirement may seem trivial but is challenging because we generally seek to compare biomarker distributions across disparate habitats. Throughout this work I return to these themes in evaluating the utility of a given biomarker.

1.4 INTRODUCTION TO PROTEOMICS

Proteomics is the study of proteins, analogous to genomics, the study of genes. High-throughput, modern proteomics is rooted in the field of analytical chemistry because it identifies proteins from mass spectra. Two basic types of mass spectrometry experiments were conducted in this work – global proteomics/data dependent acquisition (DDA)/global mode, and targeted proteomics/parallel reaction monitoring (PRM)/quantitative mode. Global proteomics experiments provide a general profile of the proteome, specifically the relative abundance of a protein relative to total protein content. Protein abundances can be compared across samples in a dataset, but due to differences in peptide ionization efficiency cannot be compared to another protein or to another dataset.[10] Global proteomics experiments do not require a prior hypothesis and can therefore provide unexpected, hypothesis generating information.[10,11] Global proteomics mass spectra can be viewed as records of an environment because they capture rich information that can be mined in future bioinformatics efforts. By contrast, quantitative proteomics uses isotopically labeled standards to correct for variability in peptide ion efficiency, so the absolute concentration of the protein can be determined. Once in absolute units (e.g. fmol protein/L seawater), comparison with different proteins and across datasets is possible. The workflows for each type of experiment are summarized in Figures 1.1 and 1.2.

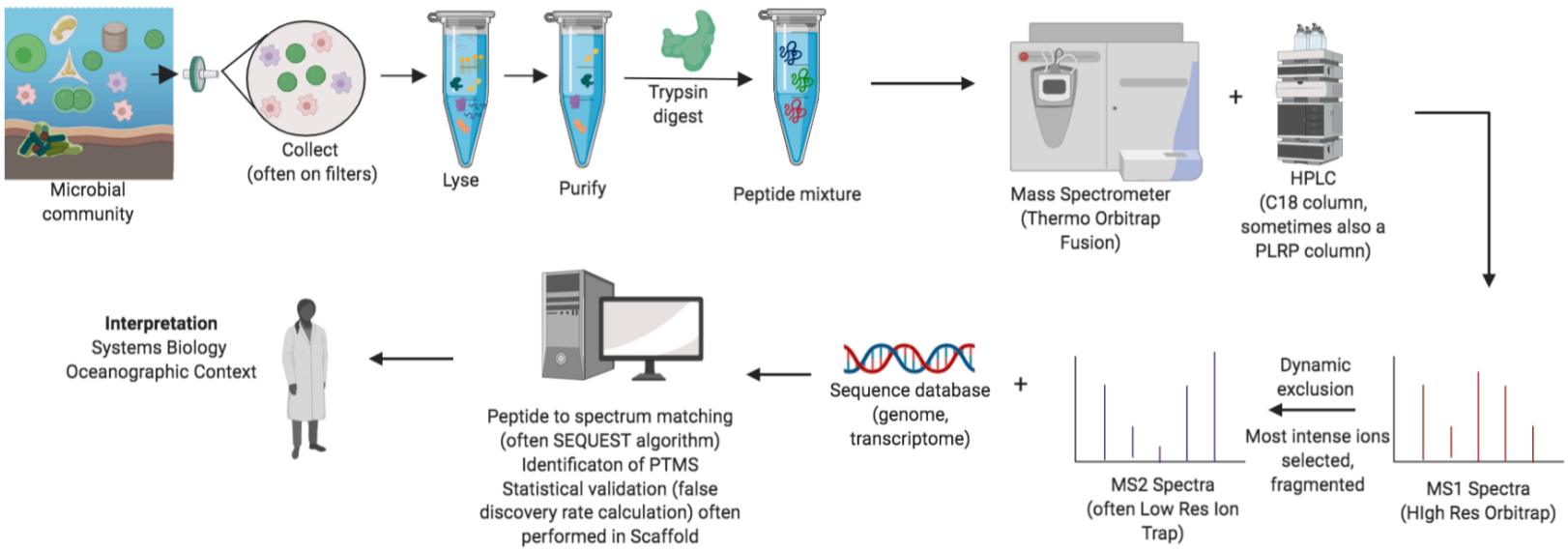


Figure 1.1. An introduction to global proteomics/metaproteomics. Microbial communities are collected, often by filtering seawater from the water column or by selecting organisms from plankton nets. The cells are lysed, and proteins purified from cellular debris. The proteins are digested with the trypsin enzyme to generate small pieces called peptides. These are analyzed by LC-MS/MS. Most of the time a single dimension of chromatography is used, but sometimes multiple dimensions are applied to separate complex protein mixtures. At defined intervals, the mass spectrometer collects an MS1 spectrum using the high resolution Orbitrap mass analyzer. From this MS1 spectrum, the most intense peaks are selected. These are fragmented, typically by collision with a high energy gas (CID or HCD), and a second, MS2 spectrum is collected. In global mode, dynamic exclusion is applied such that the mass spectrometer performs an MS2 scan at the beginning and approximately halfway through the elution peak of the MS1 ion being examined. The MS1 and MS2 spectra are combined with a sequence database, typically derived from a genome or metagenome, and matched with the SEQUEST algorithm. Post-translational modifications can be identified at this time. Because only pieces of the protein are measured, protein inference algorithms are used to match the peptides to their protein counterparts. In much of this thesis, statistical validation, or false discovery rate calculation, is performed with the Scaffold software package. The false discovery rate is calculated at the protein and peptide level. Finally, the protein and peptide identifications reach the researcher, who interprets the multivariate dataset. Often thousands of proteins and tens of thousands of peptides are identified in each metaproteome.

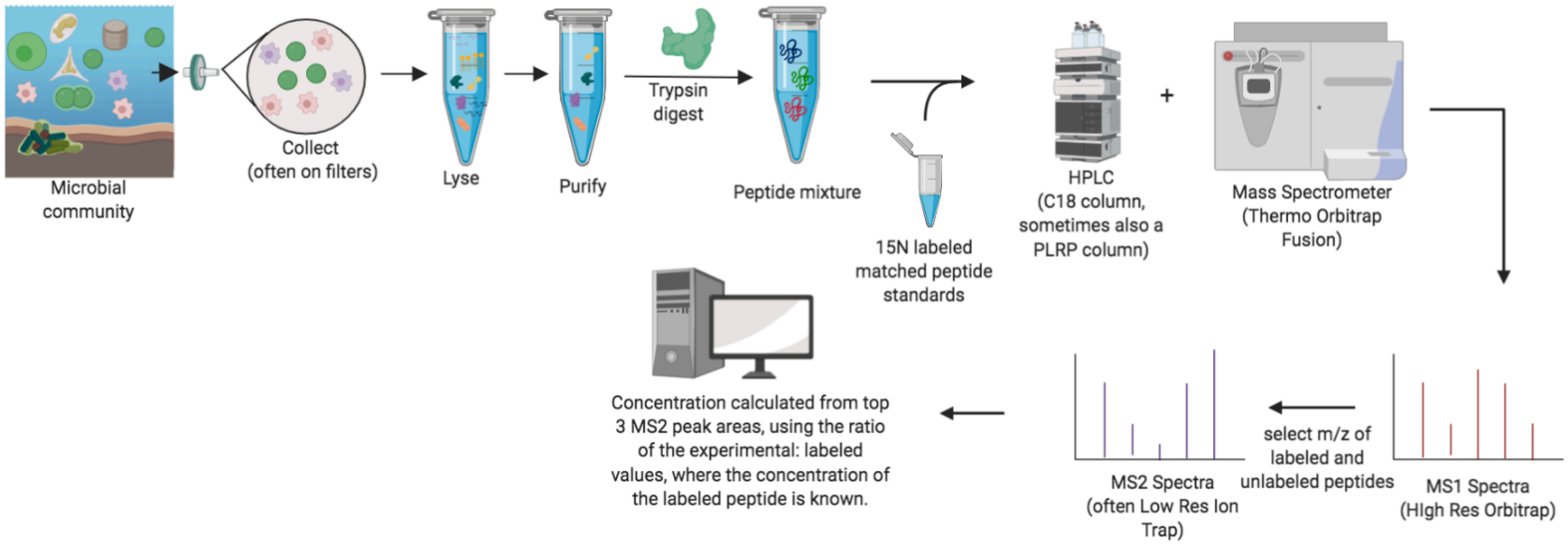


Figure 1.2. An introduction to targeted aka PRM/SRM/quantitative proteomics analyses. The analysis begins similar to global metaproteomics analyses at the start, with cells collected and lysed and proteins digested. However, before analysis a known amount of ^{15}N labeled peptide standard is added to the mixture. MS1 scans are performed as usual, however the mass spectrometer only collects MS2 spectra for masses of interest (both the labeled peptides and their unlabeled counterparts). Dynamic exclusion is not used, so MS2 spectra are collected throughout the elution of the peptide of interest. Then, the concentration of the peptide is calculated from the ratio of the experimental: labeled MS2 intensities of the top 3 most abundant peaks in the MS2 spectrum. In such experiments, the peptide(s) of interest must be pre-determined, typically from global proteomics experiments.

1.5 *TRICHODESMIUM* – A GLOBALLY IMPORTANT AND MYSTERIOUS MARINE CYANOBACTERIUM

The majority of this thesis focuses on the marine cyanobacterium *Trichodesmium*. *Trichodesmium* has global importance because it is a significant supplier of fixed nitrogen (N) to the tropical and subtropical ocean, where N availability is the limiting factor on cell growth.[12–14] Nitrogen fixation by *Trichodesmium* fuels up to half of the new production in the surface ocean.[15] This in turn fuels food webs by stimulating the transfer of organic carbon, nitrogen, and other nutrients from the surface; significant amounts of diazotrophically derived nitrogen are exported to depth.[16–18] It also has importance in certain local environments, such as in Florida where it is implicated in stimulating harmful algal blooms.[19]

For a process of such global importance, there is much uncertainty about what controls nitrogen fixation by *Trichodesmium*. It is thought that nutrient limitation plays a significant role. Because it can generate its own fixed nitrogen, *Trichodesmium* has a tendency to drive itself towards phosphate limitation.[20–26] In addition, the high iron demand of the nitrogenase enzyme means that *Trichodesmium* is often iron stressed as well.[27–29] Other potentially limiting factors include light , CO_2, and nickel availability.[30–32]

Nitrogen fixation is energetically and nutritionally costly, providing incentive for cells to closely regulate the process. The problem is exacerbated in diazotrophs like *Trichodesmium* that fix both nitrogen and carbon because the nitrogenase enzyme is susceptible by oxygen produced in photosynthesis.[33] Therefore in most organisms nitrogen and carbon fixation are separated spatially or temporally.[34,35] Both strategies have been suggested in *Trichodesmium*, but the extent to which either is necessary for nitrogen fixation is uncertain. For instance, cyclic changes in photosynthetic efficiency during the photoperiod suggested that carbon and nitrogen fixation are temporally separated.[36] Others have suggested that nitrogen fixation occurs in partially differentiated "diazocyte" cells, however the evidence for against this strategy outweighs the evidence for it (see Chapter 5).[37–39] As a result, there has not been consensus on how and why *Trichodesmium* fixes nitrogen and carbon during the day.

Trichodesmium is unique in other ways, too. First, it lives in multi-cellular filaments, which sometimes aggregate into colonies of different morphologies. This has implications for the chemistry and physics of the ocean - when *Trichodesmium* blooms, it forms dense mats that can attenuate light, heat and gas transfer.[14] *Trichodesmium* colonies have a remarkable ability to dissolve adsorbed mineral particles, likely to access the metals inside with implications for trace metal cycling.[40–42] They also modulate their biological surroundings; specifically, colonies attract a population of epibionts that is distinct from the surrounding seawater and may assist in nutrient acquisition and recycling.[43–48] A final unique behavior is that *Trichodesmium* colonies can migrate vertically in the water column. They are often observed to sink at night, sometimes below the phosphocline where they can access nutrients at depth, then rising to the surface during the day to receive optimal

exposure to light.[49,50] This migration is of chemical significance because in some cases it may represent a considerable fraction of daily phosphorus export flux.[49]

Together, the global importance and mysterious physiology of *Trichodesmium* make it an intriguing topic of study. It was selected as the focus of this work because it has complex physiology relative to other cyanobacteria, having more protein encoding genes, sensory systems, and uptake proteins than its peers (see Chapter 2). Understanding its behavior is a challenging yet achievable goal, and has implications for multiple ocean contexts.

1.6 THIS THESIS

The overarching theme of this thesis is identification and utilization of molecular biomarkers to infer microbial behavior and its impacts chemical processes. In many cases I take a systems biology approach, seeking to understand how cellular processes are connected by biochemical and biophysical restrictions on the organism. After a brief survey of marine bacterial in general, I hone in on *Trichodesmium* and attempt to understand its behavior from laboratory and field data.

Chapter 2 presents a survey of two component sensory (TCS) systems in marine bacteria. It hypothesizes that because sensory systems directly link the physiology of the cell to the surrounding environment, they can make particularly sensitive, prognostic biomarkers of biogeochemical processes. This study identified unique patterns in the sensory systems of marine bacteria that reflect direct adaptations to co-variance in the marine environment. It provides proof-of-concept that sensory proteins can predict biogeochemical parameters, specifically phosphate concentrations in the tropical Pacific ocean.

Chapter 3 provides a metaproteomic survey of field populations of *Trichodesmium*, primarily from the Atlantic Ocean. Using predictive biomarkers of nutrient limitation, it demonstrates pervasive and simultaneous iron and phosphate stress, demonstrating that co-stress is the rule rather than the exception. Using a simple cell model, I demonstrate that space availability on the cell membrane limits nutrient uptake. This may explain the prevalence of nutrient co-stress across biogeochemical contexts. Nutrient acquisition strategies are altered when *Trichodesmium* lives as a colony, and some benefits and costs of this lifestyle are explored.

Chapter 4 describes active processing of mineral particles by *Trichodesmium* colonies in the field. Using element mapping and single-colony metaproteomes, it demonstrates that some colonies associate with mineral particles, and that these associations are heterogeneous in nature. By studying the colonies individually instead of integrating the signals of many colonies as is typically done, we identified a unique response to dissolved versus particulate iron sources. This suggested that a different regulatory pathway is used depending on the metal source, and implies that direct interaction with particles is an important niche for *Trichodesmium* colonies.

Chapter 5 describes the proteome of cultured *Trichodesmium erythraeum sp.* IMS101 over the diel cycle. It seeks to answer the question of how and why

Trichodesmium fixes nitrogen during the day, despite the theoretical drawbacks of simultaneous carbon and nitrogen fixation. It indicates that daytime nitrogen fixation occurs in support of *Trichodesmium's* unique vertical migration patterns. It describes mechanisms allowing for daytime nitrogen fixation including increased respiration of oxygen, over-production of the nitrogenase enzyme, and temporal separation during the photoperiod. It also demonstrates that cellular POC and PON content can be modeled from just a handful of proteins in the dataset.

Chapter 6 characterizes a post-translational modification of the NifH enzyme, demonstrating that its potential utility as biomarker of nitrogen fixation rate over the diel cycle. This modification has been hypothesized/described before but never characterized by mass spectrometry. This chapter demonstrates that post-translational modifications can provide key information about biogeochemical processes, as well as the basic behavior of marine microbes.

This thesis answers key questions about the controls on nitrogen fixation by *Trichodesmium*, and therefore sheds light on controls on nitrogen, carbon, phosphorous, and trace metal cycling in the surface ocean. It digs deeply into the physiology of a single marine organism and in the process develops frameworks for thinking about nutrient limitation, reaction rates, and microbe-environment interactions more broadly. It demonstrates that proteins and protein post-translational modifications, captured from mass spectra, can provide detailed and layered information about ocean biogeochemistry.

1.7 REFERENCES

1. Moore, C. M. *et al*. Processes and patterns of oceanic nutrient limitation. *Nat. Geosci* **6**, 701–710 (2013).
2. Bench, S. R. *et al*. Whole genome comparison of six *Crocosphaera watsonii* strains with differing phenotypes. *J. Phycol*. **49**, 786–801 (2013).
3. Dugdale, R. & Wilkerson, F. Nutrient Limitation of New Production in the Sea. in *Primary Productivity and Biogeochemical Cycles in the Sea* (eds. Falkowski, P. G., Woodhead, A. D. & Vivirito, K.) 107–122 (Springer US, 1992). doi:10.1007/978-1-4899-0762-2_7
4. Moore, J. K., Doney, S. C., Glover, D. M. & Fung, I. Y. Iron cycling and nutrient-limitation patterns in surface waters of the world ocean. *Deep. Res. Part II Top. Stud. Oceanogr*. **49**, 463–507 (2001).
5. Smith, S. V. Phosphorus versus nitrogen limitation in the marine environment. *Limnol. Oceanogr*. **29**, 1149–1160 (1984).
6. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. W. Emergent biogeography of microbial communities in a model ocean. *Science* **315**, 1843–6 (2007).
7. Follows, M. J. & Dutkiewicz, S. Modeling diverse communities of marine microbes. *Ann. Rev. Mar. Sci*. **3**, 427–451 (2011).
8. Strimbu, K. & Tavel, J. A. What are Biomarkers? *Curr Opin HIV AIDS* **5**, 463–466 (2011).
9. Ballman, K. V. Biomarker: Predictive or prognostic? *J. Clin. Oncol*. **33**, 3968–3971 (2015).

10. Saito, M. A. *et al*. Progress and Challenges in Ocean Metaproteomics and Proposed Best Practices for Data Sharing. *J. Proteome Res*. **18**, 1461–1476 (2019).
11. Williams, T. J. & Cavicchioli, R. Marine metaproteomics: deciphering the microbial metabolic food web. *Trends Microbiol*. **22**, 248–60 (2014).
12. Karl, D. *et al*. Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57–58**, 47–98 (2002).
13. Carpenter, E. J. & Romans, K. Major Role of the Cyanobacterium *Trichodesmium* in Nutrient Cycling in the North Atlantic Ocean. **254**, 1989–1992 (1991).
14. Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B. & Carpenter, E. J. *Trichodesmium*, a globally significant marine cyanobacterium. *Science*. **276**, 1221–1229 (1997).
15. Sohm, J. A., Webb, E. A. & Capone, D. G. Emerging patterns of marine nitrogen fixation. *Nat. Rev. Microbiol*. **9**, 499–508 (2011).
16. McGillicuddy Jr., D. J. Do *Trichodesmium* spp. populations in the North Atlantic export most of the nitrogen they fix? *Global Biogeochem. Cycles* **28**, 103–114 (2014).
17. Walworth, N. G. *et al*. Nutrient-colimited *Trichodesmium* as a nitrogen source or sink in a future ocean. *Appl. Environ. Microbiol*. **84**, 1–14 (2018).
18. Dutheil, C. *et al*. Modelling N_2 fixation related to *Trichodesmium* sp.: Driving processes and impacts on primary production in the tropical Pacific Ocean. *Biogeosciences* **15**, 4333–4352 (2018).
19. Mulholland, M. R., Bernhardt, P. W., Heil, C. A., Bronk, D. A. & Neil, J. M. O. Nitrogen fixation and release of fixed nitrogen by *Trichodesmium* spp. in the Gulf of Mexico. *Limnol. Oceanogr*. **51**, 1762–1776 (2006).
20. Orchard, E. D. Phosphorus physiology of the marine Cyanobacterium *Trichodesmium*. *Massachusetts Inst. Technol*. 130 (2010). doi:10.1575/1912/3366
21. Webb, E. A., Jakuba, R. W., Moffett, J. W. & Dyhrman, S. T. Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol. Oceanogr*. **52**, 2221–2232 (2007).
22. Moutin, T. *et al*. Phosphate availability controls *Trichodesmium* spp . biomass in the SW Pacific Ocean. **297**, 15–21 (2005).
23. Hynes, A. M., Chappell, P. D., Dyhrman, S. T., Doney, S. C. & Webb, E. A. Cross-basin comparison of phosphorus stress and nitrogen fixation in *Trichodesmium*. *Limnol. Oceanogr*. **54**, 1438–1448 (2009).
24. Orchard, E. D., Webb, E. A. & Dyhrman, S. T. Molecular analysis of the phosphorus starvation response in *Trichodesmium* spp. *Environ. Microbiol*. **11**, 2400–2411 (2009).
25. Sañudo-Wilhelmy, S. A. *et al*. Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean. *Nature* **411**, 66–69 (2001).
26. Frischkorn, K. R., Krupke, A., Guieu, C., Louis, J. & Rouco, M. *Trichodesmium* physiological ecology and phosphate reduction in the western tropical South Pacific. 5761–5778 (2018).
27. Chappell, P. D., Moffett, J. W., Hynes, A. M. & Webb, E. A. Molecular evidence of iron limitation and availability in the global diazotroph *Trichodesmium*. *ISME J*. **6**, 1728–1739 (2012).
28. Nuester, J., Vogt, S., Newville, M., Kustka, A. B. & Twining, B. S. The unique

biogeochemical signature of the marine diazotroph *Trichodesmium*. *Front. Microbiol.* **3**, 1–15 (2012).
29. Rueter, J. G. Iron stimulation of photosynthesis and nitrogen fixation in *Anabaena* 7120 and *Trichodesmium* (Cyanophycease). *J.Phycol* **24,** 249–254 (1988).
30. Hutchins, D. A. *et al*. CO_2 control of *Trichodesmium* N2 fixation, photosynthesis, growth rates, and elemental ratios: Implications for past, present, and future ocean biogeochemistry. *Limnol. Oceanogr.* **52**, 1293–1304 (2007).
31. Levitan, O. *et al*. Combined effects of CO_2 and light on the N_2-fixing cyanobacterium *Trichodesmium* IMS101: a mechanistic view. *Plant Physiol.* **154**, 346–56 (2010).
32. Ho, T.-Y. Nickel limitation of nitrogen fixation in *Trichodesmium*. *Limnol. Oceanogr.* **58**, 112–120 (2013).
33. Gallon, J. R. The oxygen sensitivity of nitrogenase: a problem for biochemists and micro-organisms. *Trends Biochem. Sci.* **6**, 19–23 (1981).
34. Ernst, A., Kirschenlohr, H., Diez, J. & Böger, P. Glycogen content and nitrogenase activity in *Anabaena variabilis*. *Arch. Microbiol.* **140**, 120–125 (1984).
35. Mohr, W., Intermaggio, M. P. & LaRoche, J. Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium *Crocosphaera watsonii* WH8501. *Environ. Microbiol.* **12**, 412–421 (2010).
36. Küpper, H. *et al*. Traffic Lights in *Trichodesmium* . Regulation of Photosynthesis for Nitrogen Fixation Studied by Chlorophyll Fluorescence Kinetic Microscopy *Plant Physiology* **135**, 2120–2133 (2019).
37. Sandh, G., Xu, L. & Bergman, B. Diazocyte development in the marine diazotrophic cyanobacterium *Trichodesmium*. *Microbiology* **158**, 345–352 (2012).
38. El-Shehawy, R., Lugomela, C., Ernst, A. & Bergman, B. Diurnal expression of hetR and diazocyte development in the filamentous non-heterocystous cyanobacterium *Trichodesmium erythraeum*. *Microbiology* **149**, 1139–1146 (2003).
39. Ohki, K. Intercellular localization of nitrogenase in a non-heterocystous cyanobacterium (cyanophyte), *Trichodesmium* sp. NIBB1067. *J. Oceanogr.* **64**, 211–216 (2008).
40. Basu, S. & Shaked, Y. Mineral iron utilization by natural and cultured *Trichodesmium* and associated bacteria. *Limnol. Oceanogr.* **63**, 2307–2320 (2018).
41. Rubin, M., Berman-Frank, I. & Shaked, Y. Dust-and mineral-iron utilization by the marine dinitrogen-fixer *Trichodesmium*. *Nat. Geosci.* **4**, 529–534 (2011).
42. Rueter, J.G., Hutchins, D.A., Smith, R.W., Unsworth, N. L. Iron nutrition of *Trichodesmium*. in *Marine Pelagic Cyanobacteria: Trichodesmium and other Diazotrophs* (ed. Carpenter, E.J., Capone, D.G., Rueter, J. G.) 289–306 (Kluwer Academic Publishers, 1992).
43. Rouco, M., Haley, S. T. & Dyhrman, S. T. Microbial diversity within the *Trichodesmium* holobiont. *Environ. Microbiol.* **18**, 5151–5160 (2016).
44. Frischkorn, K. R., Rouco, M., Mooy, B. A. S. Van & Dyhrman, S. T. Epibionts dominate metabolic functional potential of *Trichodesmium* colonies from the oligotrophic ocean. *ISME-J* 2090–2101 (2017). doi:10.1038/ismej.2017.74
45. Hmelo, L. R., Van Mooy, B. A. S. & Mincer, T. J. Characterization of bacterial epibionts on the cyanobacterium *Trichodesmium*. *Aquat. Microb. Ecol.* **67**, 1–14 (2012).
46. Walworth, N. G. *et al*. Functional Genomics and Phylogenetic Evidence Suggest

Genus-Wide Cobalamin Production by the Globally Distributed Marine Nitrogen Fixer *Trichodesmium*. *Front*. *Microbiol*. **9**, 1–12 (2018).
47. Lee, M. D. *et al*. The *Trichodesmium* consortium: conserved heterotrophic co-occurrence and genomic signatures of potential interactions. *ISME J*. 1–12 (2017). doi:10.1038/ismej.2017.49
48. Lee, M. D. *et al*. Transcriptional activities of the microbial consortium living with the marine nitrogenfixing cyanobacterium *Trichodesmium* reveal potential roles in community-level nitrogen cycling. *Appl*. *Environ*. *Microbiol*. **84**, (2018).
49. White, A. E., Spitz, Y. H. & Letelier, R. M. Modeling carbohydrate ballasting by *Trichodesmium* spp. *Mar*. *Ecol*. *Prog*. *Ser*. **323**, 35–45 (2006).
50. Walsby, A. E. The properties and buoyancy providing role of gas vacuoles in *trichodesmium ehrenberg*. *Br*. *Phycol*. *J*. **13**, 103–116 (1978).

CHAPTER 2. Unique patterns and biogeochemical relevance of two component sensing in marine bacteria

This material was previously published as:
Held NA, McIlvin MR, Moran DM, Laub MT, Saito MA. 2019. Unique patterns and biogeochemical relevance of two-component sensing in marine bacteria. mSystems 4:e00317-18. https://doi.org/10.1128/mSystems.00317-18.

2.1 ABSTRACT

Two component sensory (TCS) systems link microbial physiology to the environment, and thus may play key roles in biogeochemical cycles. In this study, we surveyed the TCS systems of 328 diverse marine bacteria. We identified lifestyle traits such as copiotrophy and diazotrophy that are associated with larger numbers of TCS system genes within the genome. We compared marine bacteria with 1152 reference bacteria from a variety of habitats and found evidence of "extra" response regulators in marine genomes. Examining the location of TCS genes along the circular bacterial genome, we also found that marine bacteria have a large number of "orphan" genes, as well as many hybrid histidine kinases. The prevalence of "extra" response regulators, orphan genes, and hybrid TCS systems suggests that marine bacteria break traditional understandings of how TCS systems operate. These trends suggest prevalent regulatory networking, which may allow for coordinated physiological responses to multiple environmental signals, and may represent a specific adaptation to the marine environment. We examine phylogenetic and lifestyle traits that influence the number and structure of two component systems in the genome, finding for example that lack of two component systems is a hallmark of oligotrophy. Finally, in an effort to demonstrate the importance of TCS systems to marine biogeochemistry, we examined the distribution of *Prochlorococcus/Synechococcus* response regulator PMT9312_0717 in metaproteomes of the tropical South Pacific. We found that this protein's abundance is related to phosphate concentrations, consistent with a putative role in phosphate regulation.

2.2 IMPORTANCE

Marine microbes must manage variation in their chemical, physical, and biological surroundings. Because they directly link bacterial physiology to environmental changes, TCS systems are crucial to the bacterial cell. This study surveys TCS systems in a large number of marine bacteria and identifies key phylogenetic and lifestyle patterns in environmental sensing. We found evidence that in comparison with bacteria as a whole, marine organisms have irregular TCS system constructs which might represent an adaptation specific to the marine environment. Additionally, we demonstrate the biogeochemical relevance of TCS systems by correlating the presence of the PMT9312_0717 response regulator protein to phosphate concentrations in the South Pacific. We highlight that despite their potential ecological and biogeochemical relevance, TCS systems have been understudied in the marine ecosystem. This study expands our understanding of the breadth of bacterial TCS systems and how marine bacteria have adapted to survive in their unique environment.

2.3 INTRODUCTION

A bacterium's survival is dependent on its ability to respond to changes in its environment. This is especially true for marine microbes, which experience changes in nutrient availability, light, temperature, and community structure that can occur on time scales as short as hours (1). Two component sensory systems, which modulate gene

expression on short time scales, are the most common sensory systems in prokaryotes (2). The organism's complement of two component system genes may thus reveal details about its lifestyle, ecological niche, and physiological complexity.

Canonically, two component sensory systems are composed of a single histidine kinase and response regulator protein pair (Fig. 1). The histidine kinase contains a sensory domain that is activated by a specific stimulus, which may be a small molecule, nutrient, or physical property such as light or temperature. Upon activation, the histidine kinase is autophosphorylated on a conserved histidine residue, inducing conformational changes that enhance interactions with the response regulator protein (3). Through a highly specific protein-protein interaction, the phosphate moiety is transferred from the histidine kinase to a conserved aspartate residue on the response regulator (4). This typically stimulates binding of the response regulator to a DNA promoter region, resulting in transcription of genes in the downstream operon (5). A variation on this model is the hybrid histidine kinase, in which the kinase and response regulator are located on a single protein. Intermediary protein(s) are involved in transmitting the signal, allowing for multiple levels of control and a fine tuned physiological response (3).

Two component sensory systems represent a direct link between the environment and the physiology of the prokaryotic cell, and as such may be important players in biogeochemical cycles. One well-studied example is the Pho system, composed of the histidine kinase PhoR and response regulator PhoB. The Pho system is common in marine bacteria and regulates genes that are involved in phosphate acquisition (6). In *Prochlorococcus,* activation of PhoR by low intracellular phosphate stimulates transcription of alkaline phosphatase. Alkaline phosphatase cleaves phosphate from organic matter, providing a source of phosphate that would otherwise be inaccessible (7). As the vector linking microbial physiology (i.e. the expression of alkaline phosphatase enzyme) to ocean chemistry (i.e. phosphate availability), the Pho system may be a key regulator of phosphorus cycling in the ocean. Other two component sensory systems may also play important roles in mediating microbe-environment interactions, though we do not yet understand their biogeochemical contexts.

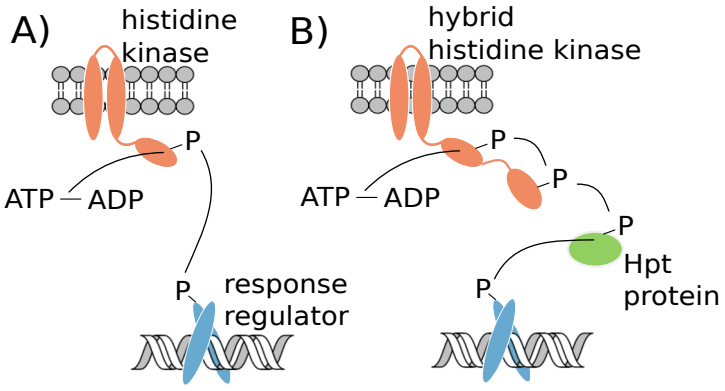


Figure 2.1 Overview of two component system signaling in A) a traditional histidine kinase – response regulator system and B) a hybrid histidine kinase system. In (A), the phosphorylation is transferred from the histidine kinase to the response regulator by a direct protein-protein interaction. In (B), the phosphorylation is transferred to an internal receiver domain on the histidine kinase, then to one or more histidine phosphotransfer (Hpt) proteins, and finally to the terminal response regulator.

In this study, we survey the two component sensory (TCS) system genes of 328 diverse marine bacteria, identifying phylogenetic and lifestyle factors that correlate with greater numbers of sensory genes. We compare these marine bacteria to a curated reference collection of 1152 bacterial genomes, most derived from the GEBA initiative (8). This allows us to identify key differences in how TCS systems are structured in marine bacteria versus bacteria in general. To demonstrate the importance of TCS systems to marine biogeochemistry, we examine the distribution of a putative phosphate sensing *Prochlorococcus* response regulator (PMT9312_0717) in metaproteomes from the tropical Pacific. We highlight gaps in our knowledge of marine TCS and emphasize the importance of TCS to our overall understanding of marine bacteria.

2.4 RESULTS

2.4.1 Lifestyle influences on TCS gene abundance

We examined the genomes of 328 diverse marine bacteria publicly available in the JGI IMG data warehouse (Table S1). The dataset emphasizes cultivable and oceanographically important organisms such as *Prochlorococcus, Synechococcus, Pelagibacterales, Alteromonas* and *Roseobacter*. In total, 15 phyla and 183 genera are represented from a variety of habitats including coastal ecosystems, the open ocean, hydrothermal vent systems, host-associated environments and marine sediments. All genomes were labeled as high quality finished or permanent drafts. TCS system genes were identified using highly conserved protein family domains for histidine kinases or response regulators as described in Materials and Methods.

We began by examining one half of the two component system – the sensory histidine kinase. The number of histidine kinases in each genome ranges from 1 (*Pelagibacter*) to 174 (*Desulfovibrio inopinatus* DSM 10711) and is related to genome size; we found on average 0.902 histidine kinases per 100 protein encoding genes (Fig. 2). A K-means clustering analysis was performed on the proportion of the genome devoted to histidine kinases (expressed for human readability as the number of histidine kinases per 100 protein encoding genes) versus genome size. This analysis revealed that the number of histidine kinases per protein encoding 100 genes is dependent on both phylogeny and lifestyle. The first cluster contains oligotrophic picoplankton such as *Prochlorococcus*, *Pelagibacterales*, and open ocean *Synechococcus*. These organisms have very small genomes and few histidine kinases per 100 protein encoding genes. Cluster 2 contains *Rhodobacter*, *Vibrio*, coastal *Synechococcus* and *Alteromonas* and *Pseudoalteromonas* genomes, which are larger and have more histidine kinases per 100 protein encoding genes. Cluster 3 is composed mainly of *Alteromonas* and *Pseudoalteromonas* genomes that have similar genome sizes as Cluster 2 organisms but more histidine kinases per 100 protein encoding genes. Cluster 4 organisms have the greatest numbers of histidine kinases per 100 protein encoding genes; many have specific lifestyle traits such as particle association, parasitism, and mat formation. A number are sulfate reducing Deltaproteobacteria.

Marine organisms are often classified by their nutritional preferences as copiotrophs (adapted to high nutrient conditions), oligotrophs (organisms adapted to low nutrient conditions), or in between (9-12). This provides a framework for understanding properties

such as growth rate, cell size, and genome size (Table 1). We classified marine bacteria as copiotrophs or oligotrophs based on published isolation and laboratory growth conditions and found that the copiotrophs have significantly more histidine kinases per gene than the oligotrophs ($p = 3e^{-15}$ by student's t test, Fig. 3b).

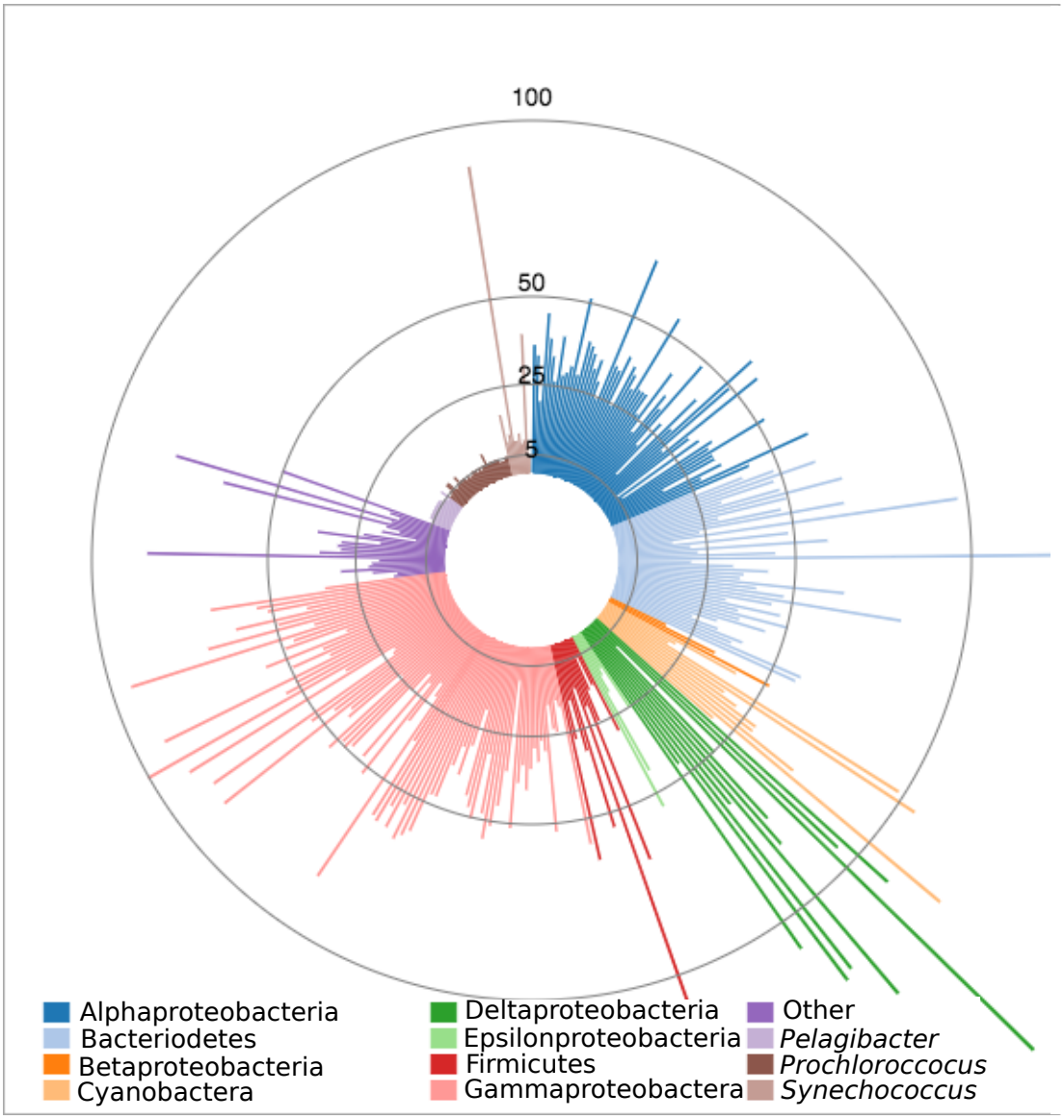


Figure 2.2 Number of histidine kinase sensory genes in the genomes of 328 diverse marine bacteria (scale provided by concentric circles). Phylogenetic groups of interest are delineated by color. The number of histidine kinases in the dataset ranges from 1 (*Pelagibacter*) to 174 (*Desulfovibrio inopinatus* DSM 10711).

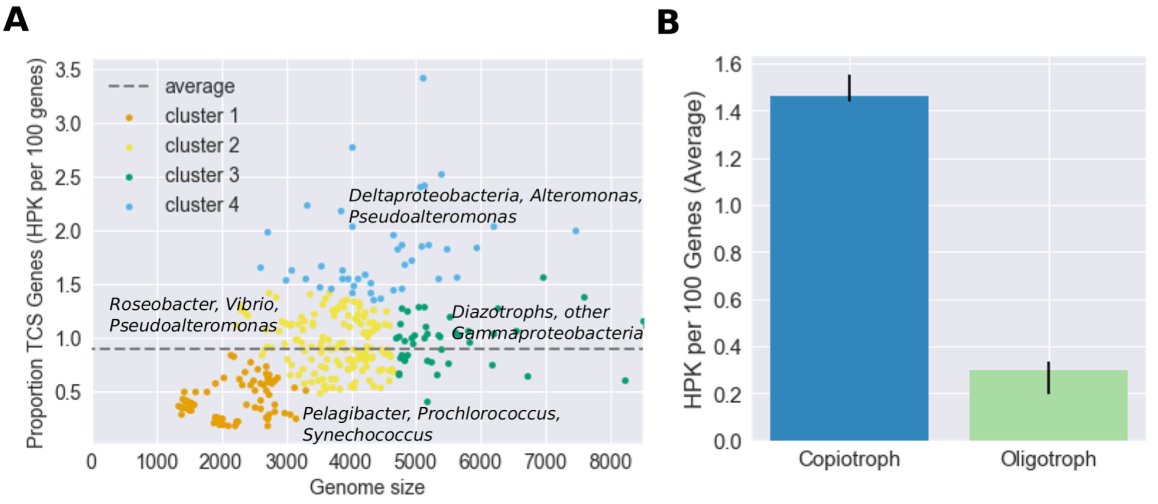


Figure 2.3 A) K-means clustering of the number of histidine kinases per 100 protein encoding genes as a function of genome size. The dashed line represents the average, 0.902 histidine kinases per 100 protein encoding genes in the genome. B) The number of histidine kinases per 100 protein encoding genes in the genome for organisms unambiguously designated as copiotrophs or oligotrophs. Error bars represent 95% confidence intervals of the average value within the copiotroph (n= 74) and oligotroph (n= 34) category. The copiotrophs have significantly more histidine kinases per gene than the oligotrophs by a student's t test ($p = 3e^{-15}$)

2.4.2 Unusual patterns in marine TCS sensing genes

We compared the TCS system genes of marine bacteria with 1152 reference bacteria (Table S2). The reference dataset includes bacteria from diverse lineages and habitats, including terrestrial, freshwater, host-associated, and marine bacteria, and is representative of trends in the two component systems of bacteria as a whole. The phylogenetic distribution of genomes in the reference dataset is broad (Figure S1). An important caveat is that both the marine and reference datasets contained mainly cultured organisms and may not be representative of natural diversity (13).

Based on the traditional understanding, the number of histidine kinase and response regulator genes is expected to be equal. Indeed, in the reference dataset we found the average response regulator: histidine kinase (RR:HPK) ratio to be 0.99 (Fig. 4). The RR:HPK ratio is slightly higher in the marine bacteria (1.03), suggesting that there are a small number of "extra" response regulators in the genomes. The difference between the marine and reference datasets is significant based on a one-way ANOVA test ($p = 0.009$, $F(372, 1151) = 6.73$).

To better understand the origin of the high RR:HPK ratio, we examined the locations of the TCS genes in a sub-selection of marine genomes (Fig. 5). Genomes were selected for their oceanographic relevance, quality, and representation in the literature. Typically, the TCS system histidine kinase and response regulator genes are located in the same operon (14). "Orphan" genes are defined here as genes that are more than four average gene lengths away from another TCS gene. These genes may participate in regulatory networks with other TCS systems (15).

Orphan TCS genes were identified in all seven of the genomes we examined, consistent with the high RR:HPK ratios described above. For example, the *Pelagibacter ubique* HTCC1016 genome has three modular TCS systems, plus one orphan HPK (a KipI family gene) and one orphan RR (RegB). Notably, KipI is known to participate in regulatory networks (15). Larger genomes have more orphan genes. For instance, *Alteromonas sp*. Alt199 has 26 orphan histidine kinases (30% of the histidine kinase genes) and 8 orphan response regulators (14% of the response regulator genes). *Crocosphaera sp*. WH8501 has 8 orphan histidine kinases (15%) and 9 orphan response regulators (17%). It is difficult to ascertain the function of orphan genes due to lack of experimental evidence associated with them. These systems are ideal for future study.

Hybrid systems occur when the histidine kinase and response regulator are located on a single protein; in this study they were identified as genes containing both a histidine kinase HAMP and response regulator receiver domain. In marine bacteria, the percent of hybrid histidine kinases relative to total histidine kinase content ranged from zero to 61% (Fig. 6). Marine bacteria have significantly more hybrid histidine kinases than reference bacteria ($p = 3e^{-10}$ by a student's t-test).

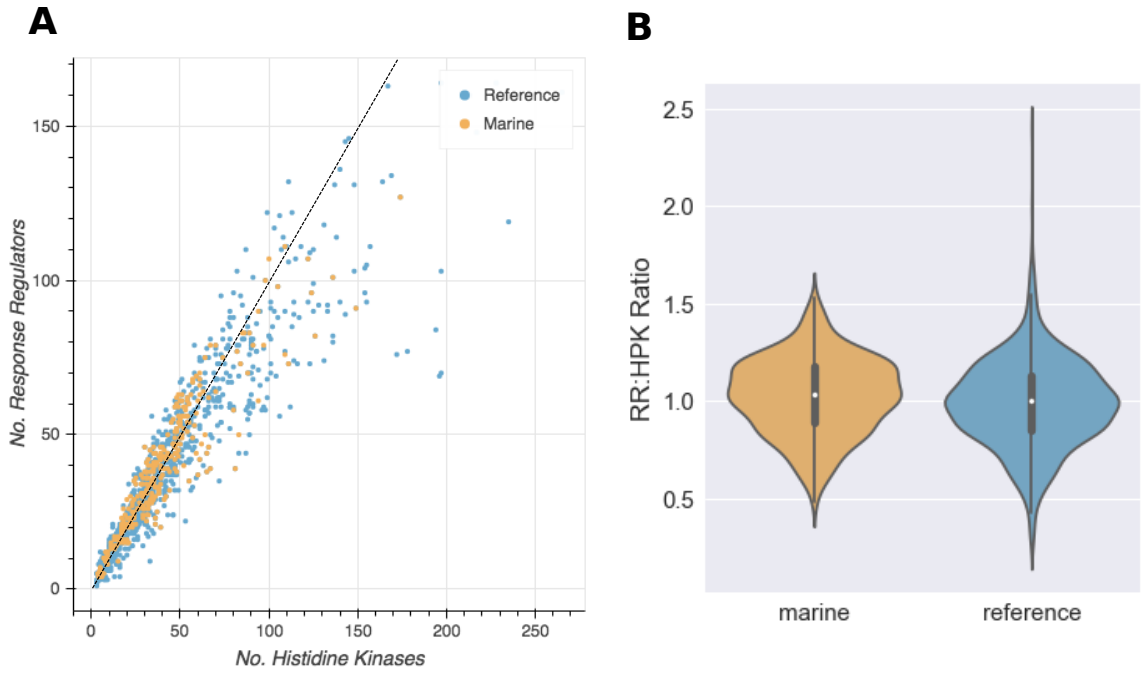


Figure 2.4 A) Number of response regulator vs. of histidine kinase genes in marine (orange) and reference (blue) bacteria. The dotted black line represents a 1:1 relationship. When the number of TCS systems is low, the ratio of RRs to HPKs is approximately 1. When the number of two component systems is large (50 or more), the RR to HPK ratio tends to be much less than 1. B) RR:HPK ratio of marine and reference bacteria. While the differences are subtle, the marine bacteria have a significantly larger RR:HPK ratio on average (1.03) than the reference bacteria (0.99) by a one-tailed ANOVA test ($p = 0.0095$, $F(327,1151) = 6.73$).

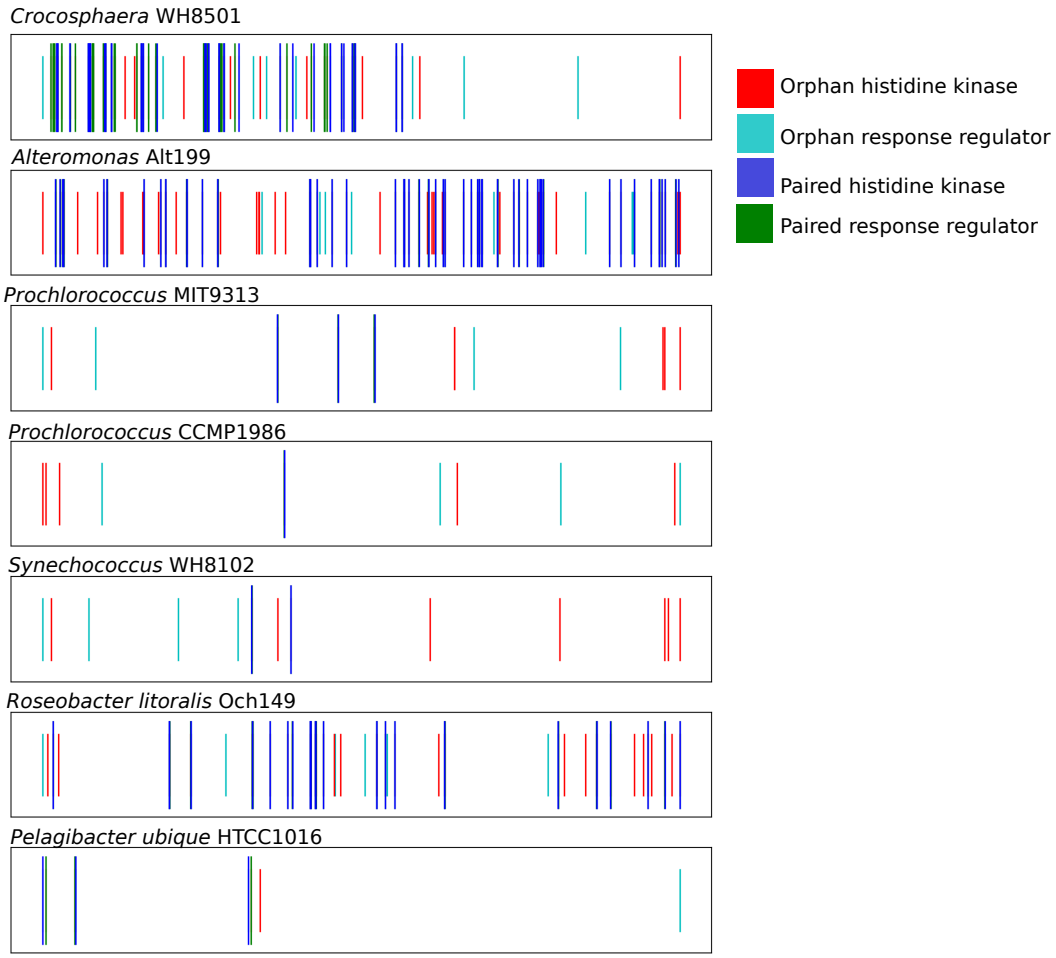


Figure 2.5 Distribution of TCS genes in various marine bacteria. The genome is linearized and depicted as a number line; genes are represented as vertical lines based on their starting location. Histidine kinases and response regulators that are within four genes of another TCS gene are represented as long blue and green lines, respectively. Orphan histidine kinases and response regulators are represented as short cyan and red lines, respectively. There are many of orphan TCS genes in marine bacteria, including in oligotrophs such as *Pelagibacter* and *Prochlorococcus*s.

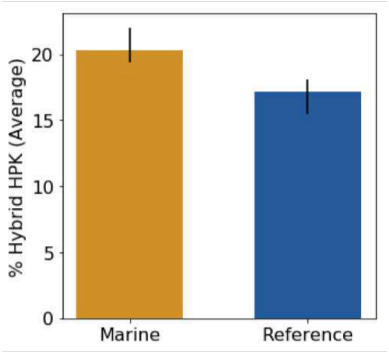


Figure 2.6 Proportion of histidine kinases that are hybrids in marine (orange) and reference (blue) bacteria. The marine bacteria have a greater percent of hybrid histidine kinases than the reference bacteria. Error bars represent a bootstrapped 95% confidence interval. The difference is statistically significant by a student's t-test ($p = 3e^{-10}$)

2.4.3 Patterns in Proteobacteria and cyanobacteria

We examined the TCS systems of Proteobacteria and cyanobacteria to explore differences among phylogenetically related organisms. Proteobacteria tend to have many histidine kinases (Fig. 7a). As before, we found that the Deltaproteobacteria devote a large portion of their genome to histidine kinases (see Fig. 3). In the cyanobacteria, we observed more variation, with *Prochlorococcus* and *Synechococcus* tending to have few histidine kinases and the nitrogen fixing diazotrophs having more (Fig. 7c).

Proteobacteria ted to have RR:HPK ratios greater than one (1.03 on average), suggesting the possibility of extra response regulators in the genomes. The RR:HPK ratios of the cyanobacteria are more variable than that of the Proteobacteria. *Prochlorococcus* and *Synechococcus*, for instance, have very high RR:HPK ratios (1.22 and 1.23 respectively) indicating that there are many extra response regulators in the genome, while *Trichodesmium* has a low RR:HPK ratio (0.75), indicating the possibility of extra histidine kinases.

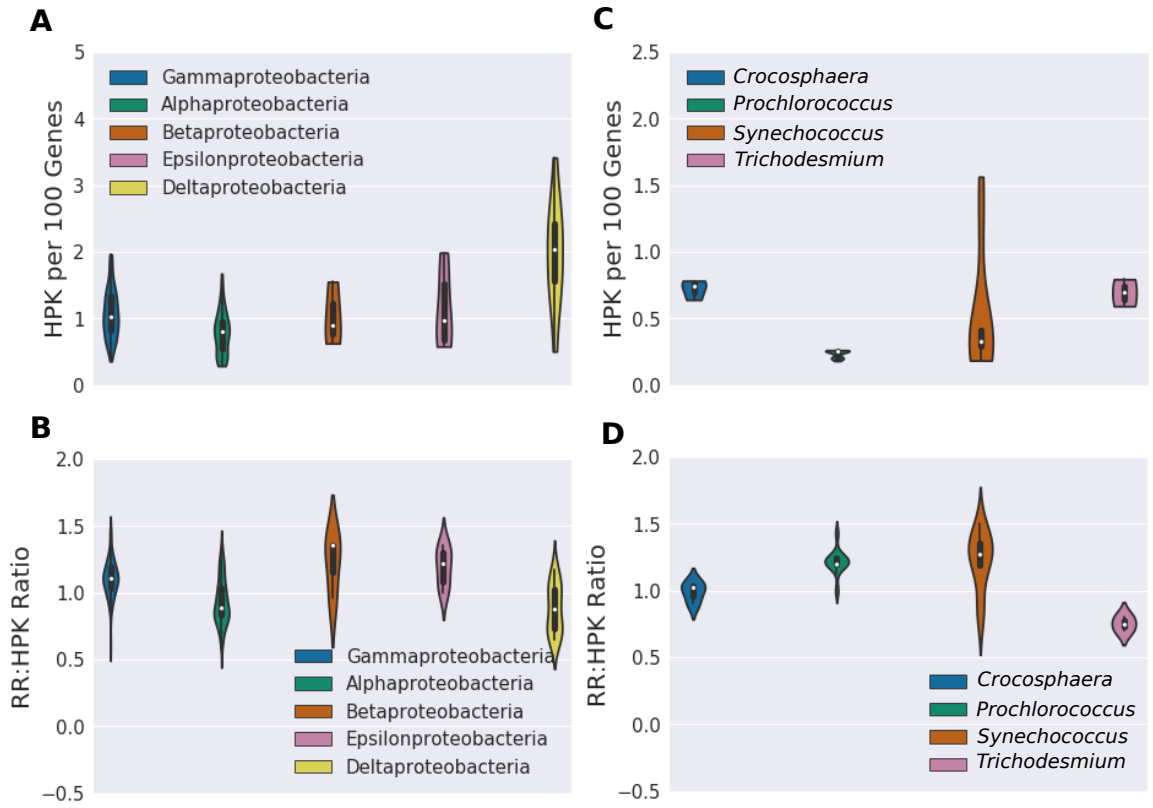


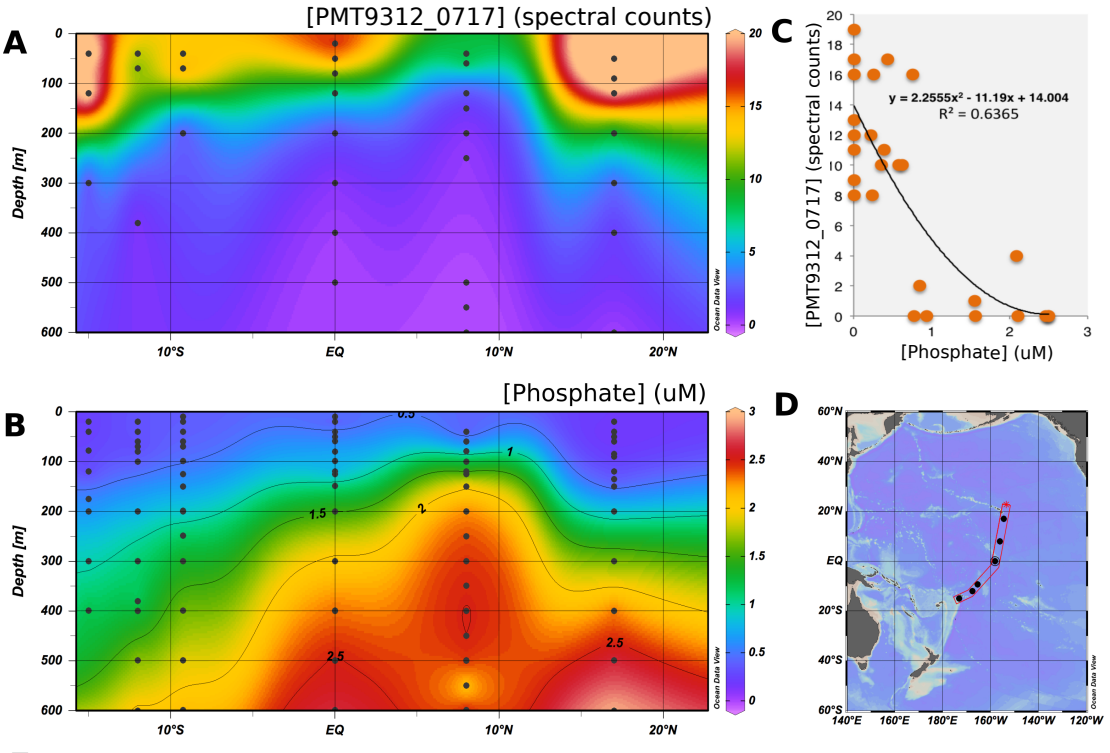
Figure 2.7 A) Number of histidine kinases per 100 protein encoding genes and B) RR:HPK ratio of Proteobacteria, and C) Number of histidine kinases per 100 protein encoding genes and D) RR:HPK ratio of Cyanobacteria. Proteobacteria, in particular the Deltaproteobacteria, have many histidine kinases per gene compared with the Cyanobacteria. The RR:HPK ratio of the Proteobacteria tends to be greater than 1. Note that *Crocosphaera*, *Synechococcus*, and *Trichodesmium* have more histidine kinases per 100 protein encoding genes than *Prochlorococcus*, which is adapted to highly oligotrophic conditions. The picocyanobacteria *Prochlorococcus* and *Synechococcus* have particularly high RR:HPK ratios.

2.4.4 Biogeochemical relevance of two component sensory systems

To investigate whether TCS proteins can be used as biomarkers of oceanographic processes, we examined the distribution of a putative *Prochlorococcus* MIT9312 phosphate sensing response regulator (PMT9312_0717) in metaproteomes of the tropical Pacific (Fig. 8). Data acquisition and analysis were previously described by Saito et. al in November 2011 (16). We identified two unique peptides from this protein. Using the open-source Metatryp software package (17), we determined that one peptide is specific to *Prochlorococcus*, and the other is specific to the order Synecoccales. Thus the distribution presented here can be thought of as a general picocyanobacteria signal (Fig. 8d).

PMT9312_0717 was confined to the upper euphotic zone and was less prevalent in phosphate rich waters near the Equator (Fig 8a). Notably, protein abundance was inversely correlated to phosphate concentration; the relationship can be modeled with a simple power

law ($r^2 = 0.64$) (Fig 8b, c). *Prochloroccocus* is highly abundant throughout the transect, while the protein is not, suggesting that PMT9312_0717 is specifically involved in phosphate regulation (Figure S2). However, like many orphan systems in marine bacteria, we were unable to identify a histidine kinase with corresponding relationships to phosphate concentration. Thus, the exact regulatory function and mechanism of PMT9312_0717 remains a mystery, despite its probable biogeochemical relevance.



E

Peptide	Least Common Ancestor
AESVVPIIFLTALEAISER	*Prochlorococcus*
VAGLDLGADDYLSKPFSPK	*Synechococcus*

Figure 2.8 A) Distribution of the response regulator PMT9312_0717 in metaproteomes of the South Pacific Ocean. B) Distribution of dissolved phosphate in the water column. C) The abundance of PMT9312_0717 (measured as spectral counts) is correlated to phosphate concentrations. The relationship can be modeled by a power law. The distribution of Prochlorococcus cells is high across the transect (Fig S2). D) Map of the METZYME transect where these samples were acquired. E) Taxonomic information for the identified peptides. Two peptides were identified, one of which was annotated separately to two different *Prochlorococcus* strains. METATRYP analysis suggests that both of these peptides are specific to the order Synechoccales.

2.5 DISCUSSION

Two component sensory systems allow bacteria to directly sense their internal and external surroundings, and therefore play key roles in bacterial "intelligence" (18-19). Because TCS systems are the most common regulatory systems in bacteria, studying them can provide insight into how individual cells might interact with their surroundings. As a result, TCS systems are of potentially significant ecological and biogeochemical importance, yet they have been little studied in this context (7, 16). In this study, we surveyed the TCS systems of 328 diverse marine bacteria and described patterns in marine two component sensing. We identified evidence of regulatory networking that may distinguish marine bacteria from other organisms, and demonstrated that the abundance of a TCS system protein can be linked to oceanographic patterns.

2.5.1 Lifestyle influences TCS gene abundance

As in prior surveys, we found that the number of histidine kinases in the genome was largely determined by genome size (2, 4). However, the proportion of the genome encoding histidine kinases varies and is related to lifestyle and niche. The concept of copiotrophy versus oligotrophy is a common theme in microbial oceanography due to pronounced patterns in nutrient abundance and scarcity, respectively, found in ocean environments (9-12). It is similar to K vs. r selection in that oligotrophs which are adapted to low nutrient conditions grow significantly slower than copiotrophs which are adapted to high nutrient conditions. Oligotrophs have smaller genomes and cell sizes that allow the organism to thrive in resource limited environments (Table 1).

The small genomes of oligotrophs are thought to be the result of genome streamlining processes that are in turn driven by the need to conserve energetic and elemental resources (nitrogen and phosphorus in particular) (9, 11, 19). Our analysis suggests that in oligotrophic environments the nutritional and energetic costs of maintaining TCS systems outweighs the regulatory benefits. Lack of TCS genes has previously been observed in extremely oligotrophic organisms such as *Pelagibacter*, which may rely on simpler regulatory structures such as one-component systems or riboswitches which require fewer energetic and nutritional resources (20-22). The extent to which other marine oligotrophs may utilize simplified regulatory systems instead of or in addition to two component systems is not yet known, but this survey found that few histidine kinases per gene is a hallmark of oligotrophy.

A previous effort to identify genomic signatures of oligotrophy found that histidine kinases were a weak indicator of copiotrophy versus oligotrophy (19). However, the study compared only two microbes, in contrast to the broader study across a larger number of diverse marine bacteria described here. The inability to adapt to rapid changes in nutrient availability and other environmental conditions may explain why oligotrophs are unable to survive in nutrient rich environments (23). A recent comparison of a coastal and open ocean *Synechococcus* cyanobacteria strains supports this notion, where a coastal strain was found to have a more dynamic proteome response to iron scarcity than an oligotrophic strain (24).

In contrast to oligotrophs, certain organisms devote a comparatively large portion of their genome to histidine kinase genes. These bacteria often have specific traits such as mat formation that may require coordinated physiological alterations (25-26). Many are sulfate or nitrate reducers from deep sea sediments or hydrothermal vents, which might represent

particularly dynamic environments. Two component systems may be particularly important for redox regulation in these organisms (27).

	Example genome size	Example growth rate (per day)	Lifestyle	HPK/100 genes	RR:HPK ratio	% Hybrid HPKs	Reference
SAR11	1200-1400	0.4-0.58	Oligotroph	0.387	0.83	typically 0 %	Rappe 2002
Prochlorococcus	1200-2000	0.51-0.83	Oligotroph	0.76	1.22	typically 0 %	Moore 1995
Synechococcus	1500-3000	1	Oligotroph	1.005	1.23	0-40%	Moore 1995
Trichodesmium	~5000	0.29	Oligotroph	0.694	0.753	15-35%	Hutchins 2007
Crocosphaera	~6000	0.5	Oligotroph	0.723	0.992	~35%	Webb 2009
Roseobacter	~5000	1.45	Varies/Copiotroph	0.755	0.991	10-40%	Teira 2007
Vibrio	~5000	up to 14.3	Copiotroph	1.25	1.07	25-50%	Mourio-Perez 2003
Alteromonas	4000-4500	6	Copiotroph	1.43	1.06	~40%	López-Pérez 2012
Pseudoalteromoas	3000-5000	~30	Copiotroph	1.5	1.1	40-50%	Pernthaler 2001

Table 2.1 Characteristics of copiotrophs vs. oligotrophs and their TCS sensory system genes.

2.5.2 Unique patterns in marine TCS systems: RR:HPK ratios, orphan genes, and hybrid systems

We identified unique trends in the TCS system genes of marine bacteria. Specifically, we found that in comparison with a reference dataset, marine bacteria tend to have higher RR:HPK ratios, 2) many orphan TCS genes, and 3) more hybrid histidine kinases. We discuss each of these observations in turn.

Marine bacteria have significantly higher RR:HPK ratios than bacteria as a whole, suggesting that they have "extra" response regulator genes. By considering the RR:HPK ratio, we found that approximately one in fifty response regulator genes lacks a histidine kinase partner (Fig. 4). The origin of the extra response regulators is puzzling. Some, lacking a histidine kinase partner, may have no regulatory function and could be the result of incomplete genetic innovations or horizontal gene transfers (HGT). However, this explanation is not consistent with the tendency towards genome streamlining in the nutrient limited ocean (9, 11). An intriguing alternative is that some of the extra response regulators participate in regulatory networks in which multiple response regulators interact with a single histidine kinase. Such networks are a topic of increasing study and have been implicated in coordinating nutrient acquisition (specifically phosphate and iron), sporulation processes, stress response, and circadian rhythms (30-33).

Colocalization of TCS genes is thought to provide for concerted transcription of the sensory genes and better success in HGT (15, 35). Marine bacteria seem to be an exception to this rule, having many orphan genes in their genomes (Fig. 5). Orphan TCS genes are present in even the most streamlined genomes (i.e. *Pelagibacter ubique*) suggesting that they play important biochemical roles. The non-modularity of TCS systems in marine bacteria suggests that the genes are not acquired through HGT, but instead through gene duplication and genetic remodeling (35). TCS systems created in this way are thought to be more likely to participate in regulatory crosstalk than systems that are acquired through horizontal gene transfer (3). Indeed, orphan genes are often involved in essential regulatory networks in model organisms (36-39). Alternatively, it is possible that in situations in which regulatory networks occur, the relationship between HPK and RR is not as specific as in normal two component systems. This lack of specificity could allow for non-modular TCS

genes to become fixed in the genome. Most of what we know about two component systems is based on studies of modular systems from model organisms; additional studies on non-model organisms may thus reveal new mechanisms of acquisition and action of TCS genes.

In hybrid histidine kinases, the phosphorylation signal is relayed by one or more histidine phosphotransfer (Hpt) protein(s) before it reaches a terminal response regulator (Fig. 1). The added complexity of the phosphorelay is thought to provide multiple points of regulation, allowing for fine-tuned physiological responses. For example, a hybrid histidine kinase regulates glycan utilization in *Bacteriodes thetaiotaomicron* by integrating both intracellular metabolism and extracellular substrate signals (28). Hybrid histidine kinases are associated with physiological and behavioral complexity, being especially prevalent in higher eukaryotes (29). Marine bacteria such as Proteobacteria and Bacteriodetes can have many hybrid TCS systems (Fig. 5 and Table 1), concurrent with a tendency towards metabolic complexity and particle association, which may drive multicellular behaviors (30).

Together, the presence of "extra" response regulators, non-modularity of TCS systems, and prevalence of hybrid TCS systems suggest increased regulatory complexity in marine bacteria. This may confer advantages in the ocean environment, where bacteria are often chronically nutrient limited. In the ocean, nutrient availability is determined by diffusion rates, resulting in co-variance in the distribution of organic nitrogen, carbon, phosphorus, and trace nutrients, especially at the microscale (30, 41). Nutrient co-limitation has been demonstrated in a number of marine environments and may be more prevalent than originally thought (40, 42-43). Marine organisms appear to have specific physiological responses to co-limitation, for example, proteome restructuring and cell size decreases in iron and phosphate co-limited *Trichodesmium* cells (44). Although the regulatory systems for co-limitation in marine microbes have yet to be elucidated, precedence for regulatory networking has been identified in non-marine organisms such as *Edwardsiella tarda*, in which the Pho and Fur systems interact with one another (33). The irregularities in marine bacterial TCS genes suggest that similar regulatory networks may underpin concerted responses to multiple environmental perturbations.

2.5.3 Comparison of Proteobacteria and cyanobacteria

Proteobacteria and cyanobacteria are perhaps the most abundant and well-studied cells in the ocean. Comparing them demonstrates the impact of lifestyle traits on the number of histidine kinases in the genome. With the exception of the oligotrophic *Pelagibacter* species, the Proteobacteria have a larger proportion of genes encoding histidine kinases (0.95 per 100 protein encoding genes on average) than the cyanobacteria (0.57 per 100 protein encoding genes on average) (Fig. 7A, C). This may be related to their tendency towards copiotrophy, which is associated with large numbers of TCS system genes. Within the cyanobacteria, nitrogen fixing organisms have more histidine kinases. Diazotrophs are not subjected to the same genome streamlining pressure as other marine cyanobacteria owing to the fact that they can access an unlimited supply of atmospheric nitrogen. Yet, the complexities of the diazotrophic lifestyle may necessitate a large number of regulatory genes. For instance, nitrogen fixation rates are known to respond to many environmental parameters such as iron nitrogen, phosphorus, dust, and light availability (45-49). Reflecting

this, marine diazotrophs have two component systems regulating nutrient availability, complex circadian rhythms, and redox state.

Evidence for regulatory networking is prevalent in both the Proteobacteria and cyanobacteria (Fig 7b, d). Most Proteobacteria and picocyanobacteria have elevated RR:HPK ratios suggesting branched regulatory networks in which one histidine kinase communicates with multiple response regulators. Branched regulatory networks could allow for multiple operons to be affected by a single sensory input, with each chemical sensor triggering multiple downstream operons, providing greater metabolic flexibility and dynamism. This could provide for fine tuned responses to multiple stimuli, such as nutrient co-limitation. In nutrient limited environments, regulatory networking may provide an advantage for cellular resource conservation. For instance, a single stimulus can trigger multiple physiological reactions without the need to express an entire two component system for each operon. Consistently, we find that the oligotrophic genomes from organisms such as *Pelagibacter* and *Prochloroccocus* have especially high RR:HPK ratios (Table 1).

Notably, the diazotrophic *Trichodesmium* species have low RR:HPK ratios suggesting regulatory networking in which multiple histidine kinases interact with a single response regulator. This may allow for integration of multiple environmental signals in regulating a single physiological response. This may underpin extensive proteomic changes such as coordination of carbon and nitrogen fixation processes over the course of the diel cycle (50). However, identifying these networks is challenging because the majority of two component system genes in *Trichodesmium* (and many marine bacteria) have not been characterized.

2.5.4 Two component systems as potential biogeochemical biomarkers

Having so far considered TCS genes, we next turned to the gene products and their relationship to oceanographic processes. We studied a *Prochlorococcus/Synechococcus* PhoB-like response regulator protein, PMT9312_0717, in the tropical Pacific Ocean. PMT9312_0717 is one of the most abundant TCS system proteins in this transect. It is an orphan, and its partner histidine kinase is not known. In the metaproteomics analysis, we found that the protein is inversely correlated with inorganic phosphate concentration, particularly below 1uM phosphate such as near the nutricline (Fig. 8). *Prochlorococcus* cells are abundant throughout the METZYME transect, while PMT9312_0717 is not and instead follows trends in phosphate concentrations (Fig S2). This implies that abundance of the protein is related to oceanographic processes, suggesting that the protein self-regulates its production. Increased abundance of TCS systems in response to low dissolved phosphorus concentrations has been previously observed for phosphate regulating systems (7, 16). Measurement of protein phosphorylation (i.e. the activity of the TCS system), while technically difficult, could provide additional information about the function of the protein.

In addition to this protein distribution, genomic evidence also suggests that PMT9312_0717 is involved in phosphate regulation. PMT9312_0717 is present in many *Prochlorococcus* species, both high and low light ecotypes, as well as in *Synechococcus* species. It is similar to both the *Prochlorococcus* sp. 9312 phosphate sensing histidine kinase PhoB (37% identity) and the *Synechococcus* sp. WH8102 PhoP protein

(WP_025362545.1, 37% identity) (16), but is a distinct protein. Because just a few point mutations can change the function of a response regulator, we hypothesized that this protein participates in phosphate regulation, but in different ways than PhoB/PhoP. Corroborating its possible role in phosphate sensing, PMT9312_0717 is located near a DedA family alkaline phosphatase related gene (e.g. PMT9312_0712) in multiple strains of *Prochlorococcus*. TCS systems have amino acids known as specificity residues that govern the histidine kinase-response regulator interaction (4). The specificity residues of the PMT9312_0717 and PhoB genes in *Prochloroccocus* sp. MIT 9312 are not shared, indicating that the PMT9312_0717 is unlikely to interact with the phosphate sensing histidine kinase PhoR.

TCS systems underpin many of the physiological changes observed in laboratory and field perturbations of marine bacteria. They are involved in nutrient acquisition, detoxification, quorum sensing, and other topical themes in marine microbiology. However, our knowledge of these systems in marine species is limited. Sequence based identification of TCS systems is difficult because histidine kinases and response regulators share highly conserved catalytic domains. Thus, while it is possible to identify TCS genes, identifying their physiological functions is tricky. Few two component systems have been experimentally verified in marine species, despite the fact that just a few amino acid substitutions can drastically change physiological function (4).

The environmental drivers behind gain/loss of two-component system genes and protein synthesis processes are not well understood. For instance, distribution of the phosphate sensing phoR-phoB two-component genes in *Prochlorococcus*, while initially hypothesized to be correlated to phosphate availability, cannot be consistently linked to large scale oceanographic patterns (7, 51). An important consideration is that two-component systems act on short time scales – seconds, minutes, or hours (52, 53). The presence of a TCS system thus suggests the need to continuously monitor the stimulus. For nutrient sensing regulators, it may be more accurate to suggest that distribution of the TCS system is related to varying, not necessarily chronically deplete, nutrient concentrations such as are found in surface waters. This is corroborated by our finding that the amount of PMT9312_0717 protein in the water column increases significantly near nutricline like concentrations. Given this circumstantial, yet consistent evidence, a detailed biochemical characterization of PMT9312_0717 is an intriguing topic for future study. However, based on this and previous work, it is clear that TCS system protein abundances can contain valuable biogeochemical information (15).

2.6 CONCLUSION

Two-component sensory systems reveal characteristics of both individual cells and the ecosystems in which they live. In this way, they represent a unique opportunity to link microbial physiology to the environment. We know relatively little about TCS systems in marine bacteria, but it is clear that the distribution of TCS genes and proteins is dependent both on the traits of the organism and its surrounding environment. For instance, we found that oligotrophs have significantly fewer histidine kinases per gene than copiotrophs and that diazotrophy is associated with greater numbers of TCS system genes. Importantly, we found that marine microbes may have adapted unique ways to sense their environments using complex regulatory networks. Additional characterization of these networks may

provide us with a greater appreciation for both the uniqueness of the ocean environment and the breadth of sensory systems used by prokaryotes. Detailed biochemical characterization of marine two-component systems is greatly needed, and has great potential to advance our understanding of microbial life and its connections to global biogeochemical cycles.

2.7 MATERIALS and METHODS

2.7.1 Marine and reference bacteria datasets

We acquired the genomes of 328 marine bacteria available in the JGI IMG data warehouse (Table S1). All of the genomes were high quality finished genomes or permanent drafts. We note that permanent draft genomes could be missing some RR or HPK genes due to incomplete assembly, but this is not expected to significantly affect results. The marine dataset is phylogenetically diverse; the best represented phyla are the Cyanobacteria, Proteobacteria, Fermicutes, Bacteriodetes, and Actinobacteria (Figure S1). The bacteria were isolated from varied habitats including coastal ecosystems, the open ocean, hydrothermal vent systems, marine sediments, and microbial mats. Most of the genomes we examined are from bacterial isolates in culture; as such, bacteria of particular oceanographic interest such as *Prochlorococcus*, as well as easily cultivated organisms such as *Alteromonas,* form a large fraction of the dataset. It is important to note that while we attempted to capture maximal phylogenetic and habitat level variability, this dataset does not necessarily represent actual bacterial diversity nor cell abundance in the ocean environment. Rather, the intention was to allow us to examine marine bacteria for which we have good quality genomic information at this time.

The reference bacteria dataset was largely derived from the GEBA-I initiative organisms, which provides a collection of bacterial genomes spanning the breadth of known phylogenetic diversity (Table S2) (8). The GEBA-I database includes both bacteria and archaea genomes; we used only the bacterial genomes here. We observed that Cyanobacteria are underrepresented in the GEBA-I dataset relative to the marine dataset described above. To facilitate comparisons between the marine and reference datasets, we included 180 high quality, non-marine Cyanobacteria genomes from the JGI IMG warehouse in the reference dataset.

2.7.2 Identification of two component system genes

We identified TCS system genes by protein family (pfam) domain annotations (54). The pfam domain annotation was performed for each genome as part of their initial ingestion into IMG. Comparison was thus facilitated by the fact that all genomes underwent similar annotation analysis pipelines in the JGI IMG portal.

Histidine kinases are composed of two domains – an HATPase and a phosphoacceptor domain, which are each separate entries in the pfam domain database. We tested both the HATPase and phosphoacceptor domains by comparing the number of histidine kinases identified with the results of an extensively curated previously published survey of TCS genes in bacteria (18). We found that the HATPase domain was more well-

conserved, identified more histidine kinases, and resulted in comparable histidine kinase identifications to the previous study. Thus, we used the HATPase domain for histidine kinase identification moving forward. The specific pfams used were pfam02518 (HATPase_c), pfam13581 (HATPase_c_2), pfam13589 (HATPase_c_3), pfam14501 (HATPase_c_5), pfam07536 (HWE_HK). The HATPase domain is also present in proteins such as DNA gyrase (pfam00204), HSP90 (pfam00183), and MutL (pfam13941); we removed genes containing these three pfams from the analysis.

Response regulators are composed of a phospho-receiver domain (sometimes known as a REC domain) and an output domain. The output domain can vary significantly and determines the biological function of the response regulator (53). We used the protein family domain for the response regulator receiver (pfm00072) to identify response regulators in our datasets. When possible, we compared the results of this pfam based identification of response regulators to that of a previously published survey and found that the pfam based analysis was comparable (18).

Hybrid histidine kinases occur when the response regulator phospho-acceptor domain is present on the same protein as the histidine kinase sensory and phosphoacceptor domains. We define them here as genes containing both a HATPase domain (pfam02518, pfam13581, pfam13589, pfam14501, or pfam07730) and the response regulator phospho-receiver domain (pfam00072). Thus, the hybrid histidine kinases are present both in the histidine kinase and response regulator data for a given genome.

2.7.3 Statistical and meta-analyses

All analyses were conducted in Python 3. The entire data analysis and visualization pipeline, including statistical analyses and machine learning algorithms, can be recreated by accessing the scripts at https://github.com/naheld/patterns_TCS_sensing_marine_bacteria. A fully executable cloud environment is provided courtesy of the Binder project (https://mybinder.org/). We used the Bokeh (https://bokeh.pydata.org/en/latest/) and seaborn (https://seaborn.pydata.org/) libraries to generate visualizations and perform statistical analyses. For ratio and statistical analyses, genomes containing 0 histidine kinases and/or 0 response regulators were excluded.

For K means clustering analyses, we counted the number of histidine kinases, including hybrid genes, and normalized the data by dividing by the number of protein encoding genes. For human readability, we express this value as the number of histidine kinases per 100 genes. To eliminate the effect of scaling, we unit normalized the data. We then performed the clustering analysis to identify groups of genomes with similar signaling repertoires in relation to genome size. We used the Silhouette method implemented in the Scikit Learn Cluster module to select the optimal number of clusters (4) by maximizing the average Silhouette coefficient values (56). Clustering analysis was performed with the SciKit Learn K-means clustering algorithm.

2.7.4 Locations of TCS genes

The locations of TCS genes in the genomes of marine bacteria were examined by plotting the gene starting locations on a linearized depiction of the circular bacterial genome

(a number line). For each genome examined, the average gene size was calculated. A gene was considered to be an orphan if its start point was farther than four average gene lengths from another TCS gene.

2.7.5 PhoB protein distribution analysis

We examined the distribution of response regulator PMT9312_0717 in marine metaproteomes from the METZYME expedition in the tropical Pacific. Analysis of these metaproteomes has been previously described (16). Briefly, microbial biomass was collected with *in situ* particle collection pumps (McLane Labs) on the KM1128 METZYME research expedition. Proteins from the 0.2-3uM size fraction were SDS detergent extracted and trypsin digested in-gel as previously described (16). Peptides were analyzed by LC-MS using 1-dimensional chromatography on a Thermo Fusion Orbitrap mass spectrometer (16). Spectral counts for the protein were generated by mapping against 6 metagenomes sampled from the METZYME expedition. These were sequenced at JGI and assembled using metaSPAdes (57). Genes were predicted and annotated using the pipeline described in Dupont et al. (2015) (58). Peptide to spectrum matches (PSMs) were identified by SEQUEST and restricted to a 10ppm peptide mass threshold and a 99.0% protein probability threshold (16). Two unique peptides for PMT9312_0717 were identified; we performed redundancy analysis using the openly available Metatryp software (17). Corresponding inorganic phosphate analyses were conducted by Joe Jennings at Oregon State University as previously described (59).

2.8 ACKNOWLEDGEMENTS

I thank my co-authors on this project, Matthew McIlvin, Mike Laub, and Mak Saito. We also Joe Jennings at Oregon State University and Chris Dupont at the J. Craig Venter Institute for providing nutrient and metagenomic analyses, respectively, for the KM1128 METZYME research expedition. This material is based on work supported by a National Science Foundation Graduate Research Fellowship under grant number 1122274 (N. Held). It is also supported by the Gordon and Betty Moore Foundation grant number 3782 (M. Saito) and National Science Foundation grant numbers OCE-1657766, EarthCube 1639714, OCE-1658030 and OCE-1260233.

2.9 REFERENCES

1. Fuhrman, J. A., Cram, J. A., & Needham, D. M. 2015. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews: Microbiology*, *13*(3), 133–146. https://doi.org/10.1038/nrmicro3417
2. Galperin, M. Y., Higdon, R., & Kolker, E. 2010. Interplay of heritage and habitat in the distribution of bacterial signal transduction systems. *Molecular bioSystems*, *6*(4), 721–728. https://doi.org/10.1039/b908047c
3. Heermann, R., & Jung, K. 2010. Stimulus Perception and Signaling in Histidine Kinases. *Bacterial Signaling*, *32*, 135–161. https://doi.org/10.1002/9783527629237.ch8

4. Capra, E. J., & Laub, M. T. 2012. Evolution of Two-Component Signal Transduction Systems. *Annual Reviews of Microbiology*, *66*(1), 325–347. https://doi.org/10.1146/annurev-micro-092611-150039
5. Gao, R., Mack, T. R., & Stock, A. M. 2007. Bacterial response regulators: versatile regulatory strategies from common domains. *Trends in Biochemical Sciences*, *32*(5), 225–234. https://doi.org/10.1016/j.tibs.2007.03.002
6. Yamada, M., Makino, K., Amemura, M., Shinagawa, H., & Nakata, A. 1989. Regulation of the phosphate regulon of Escherichia coli: Analysis of mutant phoB and phoR genes causing different phenotypes. *Journal of Bacteriology*, *171*(10), 5601–5606.
7. Martiny, A. C., Coleman, M. L., & Chisholm, S. W. 2006. Phosphate acquisition genes in Prochlorococcus ecotypes: Evidence for genome-wide adaptation. *Proceedings of the National Academy of Sciences*, *103*(33), 12552–12557. https://doi.org/10.1073/pnas.0601301103
8. Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-fadrosh, E. A., Meier-kolthoff, J. P., Göker, M., Cameron Coates, R., Hadjithomas, M. Pavlopoulos, G.A., Paez-Espino, D., Yoshikuni, Y., Visel, A., Whitman, W.B., Garrity, G. M., Eisen, J.A., Hugenholtz, P., Pati, Am., Ivanova, N. N., Woyke, T., Klenk, H., Kyrpides, N.C. 2017. 1003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nature*, *35*(7), 676–683. https://doi.org/10.1038/nbt.3886
9. Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M., González, J. M., Luo, H., Wright, J. J., Landry, Z. C., Hanson, N. W., Thompson, B. P., Poulton, N. J., Schwientek, P., Acinas, S.G., Giovannoni, S.J., Moran, M.A., Hallam, S. J., Cavicchioli, R., Woyke, T., Stepanauskas, R. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(28), 11463–8. https://doi.org/10.1073/pnas.1304246110
10. Kirchman, D. L. (2016). Growth Rates of Microbes in the Oceans. *Annual Review of Marine Science*, *8*(1), 285–309. https://doi.org/10.1146/annurev-marine-122414-033938
11. Grote, J., Thrash, J. C., Huggett, M. J., Landry, Z. C., Carini, P., & Giovannoni, S. J. (2012). Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade, *3*(5), 1–13. https://doi.org/10.1128/mBio.00252-12.Editor
12. Ivars-Martinez, E., Martin-Cuadrado, A.-B., D'Auria, G., Mira, A., Ferriera, S., Johnson, J., Friedman, R., Rodriguez-Valera, F. (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *The ISME Journal*, *2*(12), 1194–212. https://doi.org/10.1038/ismej.2008.74
13. Staley J. T., Konopka A. 1985. Measurement of in situ activities of non- photosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. 39:321–346
14. Barakat, M., Ortet, P., & Whitworth, D. E. 2011. P2CS: A database of prokaryotic two-component systems. *Nucleic Acids Research*, *39*, 771–776. https://doi.org/10.1093/nar/gkq1023
15. Mitrophanov, A. Y., & Groisman, E. A. 2008. Signal integration in bacterial two-component regulatory systems. *Gene Development*, *22*, 2601–2611. https://doi.org/10.1101/gad.1700308.response
16. Saito, M. A., McIlvin, M. R., Moran, D. M., Goepfert, T. J., DiTullio, G. R., Post, A. F., & Lamborg, C. H. 2014. Multiple nutrient stresses at intersecting Pacific Ocean biomes

detected by protein biomarkers. *Science*, *345*(6201), 1173–7. https://doi.org/10.1126/science.1256450
17. Saito, M. A., Dorsk, A., Post, A. F., McIlvin, M. R., Rappé, M. S., DiTullio, G. R., & Moran, D. M. 2015. Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* 00, 1-11. https://doi.org/10.1002/pmic.201400630
18. Galperin, M. Y. 2005. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiology*, *5*, 35. https://doi.org/10.1186/1471-2180-5-35\
19. Lyon, P. (2015). The cognitive cell: Bacterial behavior reconsidered. *Frontiers in Microbiology*, *6*(MAR), 1–18. https://doi.org/10.3389/fmicb.2015.00264
20. Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., DeMaere, M.Z., Tiny, L., Ertan, H., Johnson, J., Ferriera, S., Lapidus, A., Anderson, I., Kyrpides, N., Munk, A.C., Detter, C., Han, C.S., Brown, M.V., Robb, F.T., Kjellebert, S., Cavicchioli, R. 2009. The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(37), 15527–33. https://doi.org/10.1073/pnas.0903507106
21. Giovannoni, S.J., Tripp, H.J., Givan, S., Podar M., Vergin, K. L., Baptista D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., Rappé M.S., Short, J.M., Carrington, J.C., Mathur, E.J. 2005. Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309, 1242-1245.
22. Meyer, M. M., Ames, T. D., Smith, D. P., Weinberg, Z., Schwalbach, M. S., Giovannoni, S. J., & Breaker, R. R. 2009. Identification of candidate structured RNAs in the marine organism "Candidatus Pelagibacter ubique." *BMC Genomics*, *10*, 1–16. https://doi.org/10.1186/1471-2164-10-268
23. Koch, A. L. 2001. Oligotrophs versus copiotrophs. *BioEssays*, *23*(7), 657–661. https://doi.org/10.1002/bies.1091
24. Mackey, K. R. M., Post, A. F., McIlvin, M. R., Cutter, G. A, John, S. G., & Saito, M. A. 2015. Divergent responses of Atlantic coastal and oceanic Synechococcus to iron limitation. *Proceedings of the National Academy of Sciences*, *112*(32), 9944–9949. https://doi.org/10.1073/pnas.1509448112
25. Lehti, T. A., Heikkinen, J., Korhonen, T. K., & Westerlund-Wikström, B. 2012. The response regulator RcsB activates expression of Mat fimbriae in meningitic Escherichia coli. *Journal of Bacteriology*, *194*(13), 3475–3485. https://doi.org/10.1128/JB.06596-11
26. Reynolds, T. B., Jansen, A., Peng, X., & Fink, G. R. 2008. Mat formation in Saccharomyces cerevisiae requires nutrient and pH gradients. *Eukaryotic Cell*, *7*(1), 122–130. https://doi.org/10.1128/EC.00310-06
27. Kusian, B., & Bowien, B. 1997. Organization and regulation of *cbb* CO_2 assimilation genes in autotrophic bacteria. *FEMS Microbiology Reviews*, *21*(2), 135–155.
28. Sonnenburg, E. D., Sonnenburg, J. L., Manchester, J. K., Hansen, E. E., Chiang, H. C., & Gordon, J. I. 2006. A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing in vivo to carbohydrate metabolism. *Proceedings of the National Academy of Sciences*, *103*(23), 8834–8839. https://doi.org/10.1073/pnas.0603249103

29. Wuichet, K., Cantwell, B. J., & Zhulin, I. B. 2010. Evolution and phyletic distribution of two-component signal transduction systems. *Current Opinion in Microbiology*, *13*(2), 219–225. https://doi.org/10.1016/j.mib.2009.12.011
30. Lyons, N. A., & Kolter, R. 2016. On the evolution of bacterial multicellularity. *Curr. Opin. Microbiol*, 24, 21-28.
31. Espinosa, J., Boyd, J. S., Cantos, R., Salinas, P., Golden, S. S., & Contreras, A. 2015. Cross-talk and regulatory interactions between the essential response regulator RpaB and cyanobacterial circadian clock output. *Proceedings of the National Academy of Sciences*, 201424632. https://doi.org/10.1073/pnas.1424632112
32. Chakraborty, S., Sivaraman, J., Leung, K. Y., & Mok, Y. K. 2011. Two-component PhoB-PhoR regulatory system and ferric uptake regulator sense phosphate and iron to control virulence genes in type III and VI secretion systems of Edwardsiella tarda. *Journal of Biological Chemistry*, *286*(45), 39417–39430. https://doi.org/10.1074/jbc.M111.295188
33. Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R., & Ishihama, A. 2005. Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. *Journal of Biological Chemistry*, *280*(2), 1448–1456. https://doi.org/10.1074/jbc.M410104200
34. Fisher, S. L., Jiang, W., Wanner, B. L., & Walsh, C. T. 1995. Cross-talk between the Histidine Protein Kinase VanS and the Response Regulator PhoB. *Journal of Biological Chemistry*, *270*(39), 23143–23149. http://www.jbc.org/cgi/content/abstract/270/39/23143
35. Alm, E., Huang, K., & Arkin, A. 2006. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Computational Biology*, *2*(11), 1329–1342. https://doi.org/10.1371/journal.pcbi.0020143
36. Overgaard, M., Wegener-Feldbrügge, S., & Søgaard-Andersen, L. 2006. The orphan response regulator DigR is required for synthesis of extracellular matrix fibrils in Myxococcus xanthus. *Journal of Bacteriology*, *188*(12), 4384–4394. https://doi.org/10.1128/JB.00189-06
37. Idone, V., Brendtro, S., Gillespie, R., Kocaj, S., Peterson, E., Rendi, M., Warren, W., Michalek, S., Krastel, K., Cvitkovitch, D., Spatafora, G. 2003. Effect of an orphan response regulator on Streptococcus mutans sucrose-dependent adherence and cariogenesis. *Infection and Immunity*, *71*(8), 4351–4360. https://doi.org/10.1128/IAI.71.8.4351-4360.2003
38. Lu, Y., He, J., Zhu, H., Yu, Z., Wang, R., Chen, Y., Dang, F., Zhang, W., Yang, S., Jiang, W. 2011. An orphan histidine kinase, OhkA, regulates both secondary metabolism and morphological differentiation in Streptomyces coelicolor. *Journal of Bacteriology*, *193*(12), 3020–3032. https://doi.org/10.1128/JB.00017-11
39. Workentine, M. L., Chang, L., Ceri, H., & Turner, R. J. 2009. The GacS-GacA two-component regulatory system of Pseudomonas fluorescens: A bacterial two-hybrid analysis. *FEMS Microbiology Letters*, *292*(1), 50–56. https://doi.org/10.1111/j.1574-6968.2008.01445.x
40. Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., Galbraith, E.D., Geider, R.J., Guieu, C., Jaccard S.L., Jickells, T.D., La Roche, J., Lenton, T.M., Mahowald, N.M. Marañón, E. Marinov, I., Moore, J.K., Nakatsuka T., Oschlies, A., Saito, M.A., Thingstad, T.F., Tsuda, A., Ulloa, O. 2013. Processes and

patterns of oceanic nutrient limitation. *Nature Geosci*, *6*(9), 701–710. https://doi.org/10.1038/ngeo1765
41. Stocker, R. 2012. Marine Microbes See a Sea of Gradients. *Science*, *338*(6107), 628–633. https://doi.org/10.1126/science.1208929
42. Saito, M. A., Goepfert, T. J., & Ritt, J. T. 2008. Some thoughts on the concept of co-limitation: Three definitions and the importance of bioavailability. *Limnology and Oceanography*, *53*(1), 276–290. https://doi.org/10.4319/lo.2008.53.1.0276
43. Browning, T. J., Eric, P., Rapp, I., Engel, A., Bertrand, E. M., Tagliabue, A., & Moore, C. M. 2017. Nutrient co-limitation at the boundary of an oceanic gyre. *Nature Publishing Group*, *12*. https://doi.org/10.1038/nature24063
44. Walworth, N. G., Fu, F.X., Webb, E. A., Saito, M. A., Moran, D., Mcllvin, M. R., Lee, M.D., Hutchins, D. A. 2016. Mechanisms of increased Trichodesmium fitness under iron and phosphorus co-limitation in the present and future ocean. *Nat Commun*, *7*(May), 1–11. https://doi.org/10.1038/ncomms12081
45. Karl, D., Michaels, A., Bergman, B., Capone, D., Carpenter, E., Letelier, R., Fipschultz, L., Paerl, H., Sigman, D. Stal, L. 2002. Dinitrogen fixation in the world's oceans. *Biogeochemistry*, *57–58*, 47–98. https://doi.org/10.1023/A:1015798105851
46. Sanudo-Wilhelmy, S. A., A. B. Kustka, C. J. Gobler, D. A. Hutchins, M. Yang, K. Lwiza, J. Burns, D. G. Capone, J. A. Raven, and E. J. Carpenter. 2001. Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean, *Nature*, 411(6833), 66-69
47. Orchard, E. D., Webb, E. A., & Dyhrman, S. T. 2009. Molecular analysis of the phosphorus starvation response in trichodesmium spp. *Environmental Microbiology*, *11*(9), 2400–2411. https://doi.org/10.1111/j.1462-2920.2009.01968.
48. Chen, Y. B., Chen, Y. B., Dominic, B., Mellon, M. T., Zehr, J. P. 1998. Circadian rhythm of nitrogenase gene expression in the diazotrophic filamentous nonheterocystous cyanobacterium *Trichodesmium* sp strain IMS101. *J. Bacteriol.*, *180*(14), 3598–3605.
49. Rubin, M., Berman-Frank, I., and Shaked, Y. 2011. Dust- and mineral-iron utilization by the marine dinitrogen-fixer Trichodesmium. *Nat. Geosci.* 4, 529–534
50. Saito, M. A., Bertrand, E. M., Dutkiewicz, S., Bulygin, V. V, Moran, D. M., Monteiro, F. M., Follows, M.J., Vaolis, F.W., Waterbury, J. B. 2011. Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph Crocosphaera watsonii. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(6), 2184–9. https://doi.org/10.1073/pnas.1006943108
51. Luo, C., & Konstantinidis, K. T. 2011. Phosphorus-related gene content is similar in Prochlorococcus populations from the North Pacific and North Atlantic Oceans. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(16) https://doi.org/10.1073/pnas.1018662108
52. Mary, I., & Vaulot, D. 2003. Two-component systems in Prochlorococcus MED4: Genomic analysis and differential expression under stress. *FEMS Microbiology Letters*, *226*(1), 135–144. https://doi.org/10.1016/S0378-1097(03)00587-1
53. Ogawa, T., Bao, D. H., Katoh, H., Shibata, M., Pakrasi, H. B., & Bhattacharyya-Pakrasi, M. 2002. A two-component signal transduction pathway regulates manganese homeostasis in Synechocystis 6803, a photosynthetic organism. *Journal of Biological Chemistry*, *277*(32), 28981–28986. https://doi.org/10.1074/jbc.M204175200

54. Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., Punta, M. 2014. Pfam: The protein families database. *Nucleic Acids Research*, *42*(D1), 222–230. https://doi.org/10.1093/nar/gkt1223
55. Galperin, M. Y. 2006. Structural classification of bacterial response regulators: Diversity of output domains and domain combinations. *Journal of Bacteriology*, *188*(12), 4169–4182. https://doi.org/10.1128/JB.01887-05
56. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. 2012. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2
57. Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. 2017. metaSPAdes : a new versatile metagenomic assembler. *Genome Res* 27(5), 824–834. https://doi.org/10.1101/gr.213959.116.4
58. Dupont, C. L., McCrow, J. P., Valas, R., Moustafa, A., Walworth, N., Goodenough, U., Roth, R., Hogle, S.L., Bai, J., Johnson, Z.I., Mann, E., Palenik, B., Barbeau, K.A., Venter, J.C., Allen, A. E. 2015. Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *The ISME Journal*, *9*(5), 1076–92. https://doi.org/10.1038/ismej.2014.198
59. Noble, A.E. C. H. Lamborg, D. C. Ohnemus, P. J. Lam, T. J. Goepfert, C. I. Measures, C.H. Frame, K. L. Casciotti, G. R. DiTullio, J. Jennings, M. A. Saito. 2012. Basin-scale inputs of cobalt, iron, and manganese from the Benguela-Angola front into the South Atlantic Ocean. Limnol. Oceanogr. 57, 989–1010 2012. doi:10.4319/lo.2012.57.4.0989
60. Rappé, M. S., Connon, S. a, Vergin, K. L., & Giovannoni, S. J. 2002. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature*, *418*(6898), 630–633. https://doi.org/10.1038/nature00917
61. 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. 116:259-275.
62. Hutchins, D. A., Fu, F.-X., Zhang, Y., Warner, M. E., Feng, Y., Portune, K., Bernhard, P.W., Mulholland, M. R. 2007. CO2 control of Trichodesmium N2 fixation, photosynthesis, growth rates, and elemental ratios: Implications for past, present, and future ocean biogeochemistry. *Limnology and Oceanography*, *52*(4), 1293–1304. https://doi.org/10.4319/lo.2007.52.4.1293
63. Webb, E. A., Ehrenreich, I. M., Brown, S. L., Valois, F. W., & Waterbury, J. B. 2009. Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, Crocosphaera watsonii, isolated from the open ocean. *Environmental Microbiology*, *11*(2), 338–348. https://doi.org/10.1111/j.1462-2920.2008.01771.x
64. Teira, E., Gasol, J.M., Aranguren-Gassis, M., Fernandez, A., Gonzalez, J., Lekunberri, I., Alvarez-Salgado, X.A. 2008. Linkages between bacterioplankton community composition, heterotrophic carbon cycling and environmental conditions in a highly dynamic coastal ecosystem. *Environ Microbiol* 10: 906-917.
65. Mourin, R. R., Worden, A. Z., & Azam, F. 2003. Growth of Vibrio cholerae O1 in Red Tide Waters off California, *69*(11), 6923–6931. https://doi.org/10.1128/AEM.69.11.6923

66. López-Pérez, M., Gonzaga, A., Ivanova, E. P., & Rodriguez-Valera, F. 2014. Genomes of Alteromonas australica, a world apart. *BMC Genomics*, *15*(1), 483. https://doi.org/10.1186/1471-2164-15-483
67. Pernthaler, A., Pernthaler, J., Eilers, H., & Amann, R. 2001. Growth Patterns of Two Marine Isolates : Adaptations to Substrate Patchiness *67*(9), 4077–4083. https://doi.org/10.1128/AEM.67.9.4077

2.10 SUPPLEMENTARY FIGURES

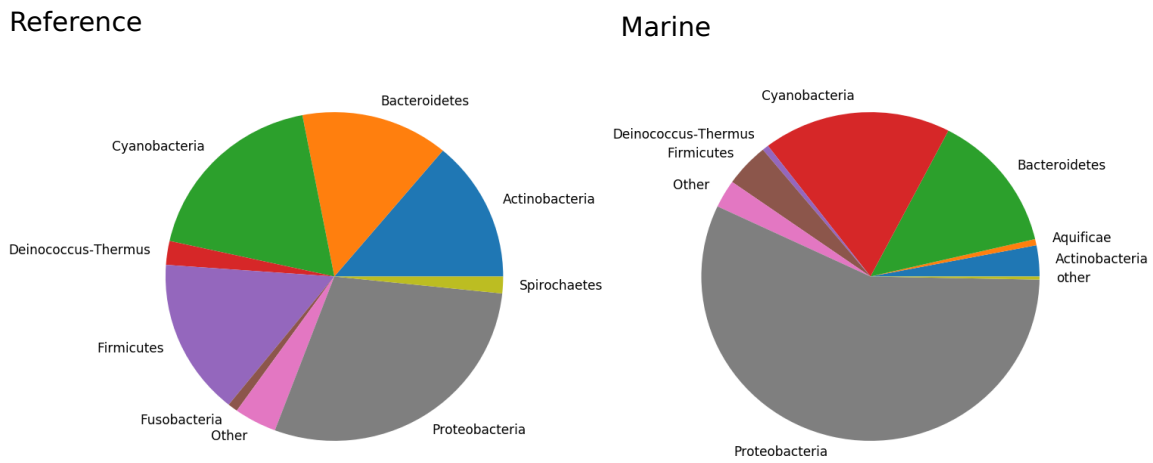


Figure S2.1 Phylogenetic breakdown of the marine and reference datasets, showing the phylogenetic diversity of the genomes used in this analysis.

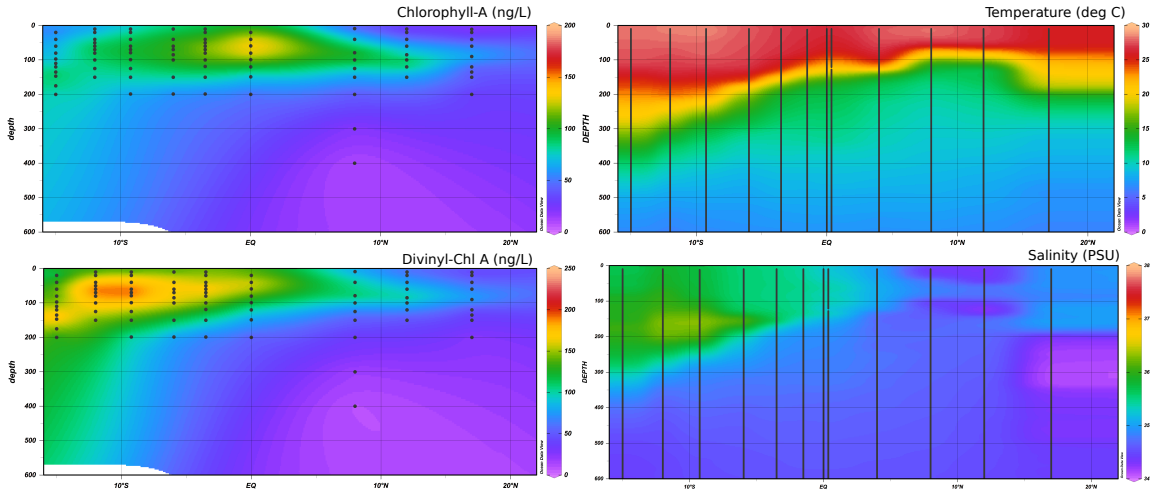


Figure S2.2 Hydrographic and pigment data from the METZYME cruise. The distribution of PMT9312_0717 is not well correlated with chlorophyll-a, divinyl chlorophyll-a, temperature, nor salinity in this region, but is well correlated with dissolved phosphate concentrations (see Fig 8).

2.11 SUPPLEMENTARY TABLES (list)

Table S2.1 TCS gene data for the 328 marine bacteria surveyed, including taxonomic information for each genome. Available at https://msystems.asm.org/content/4/1/e00317-18

Table S2.2 TCS gene data for the 1152 reference bacteria, including taxonomic information for each genome. Available at https://msystems.asm.org/content/4/1/e00317-18

CHAPTER 3. Co-occurrence of Fe and P stress in natural populations of the marine diazotroph *Trichodesmium*

3.1 ABSTRACT

Trichodesmium is a globally important marine microbe that provides fixed nitrogen to otherwise N limited ecosystems. In nature, nitrogen fixation is likely regulated by iron or phosphate availability, but the extent and interaction of these controls is unclear. From metaproteomics analyses using established protein biomarkers for iron and phosphate stress, we found that co-stress is the norm rather than the exception for field *Trichodesmium* colonies. Counter-intuitively, the nitrogenase enzyme was most abundant under co-stress, consistent with the idea that *Trichodesmium* has a specific physiological state under nutrient co-stress. Organic nitrogen uptake was observed to occur simultaneously with nitrogen fixation. Quantification of the phosphate ABC transporter PstC combined with a cellular model of nutrient uptake suggested that *Trichodesmium* is confronted by the biophysical limits of membrane space and diffusion rates for iron and phosphate acquisition. Colony formation may benefit nutrient acquisition from particulate and organic nutrient sources, alleviating these pressures. The results indicate that to predict the behavior of *Trichodesmium*, we must consider multiple nutrients simultaneously across biogeochemical contexts.

3.2 INTRODUCTION

The diazotrophic cyanobacterium *Trichodesmium* plays an important ecological and biogeochemical role in the tropical and subtropical oceans globally. By providing bioavailable nitrogen (N) to otherwise N-limited ecosystems, it supports basin-scale food webs, increasing primary productivity and carbon flux from the surface ocean.[1–5] Nitrogen fixation is energetically and nutritionally expensive, so it typically occurs when other sources of N are unavailable, i.e. in N-starved environments.[6] However, nitrogen availability is not the sole control on nitrogen fixation, which must be balanced against the cell's overall nutritional status. Because it can access a theoretically unlimited supply of atmospheric nitrogen, *Trichodesmium* often becomes phosphorus (P) limited.[7–11] It also has a tendency to drive itself towards iron (Fe) stress because the nitrogenase enzyme is iron-demanding.[12–16] Colony associated epibionts associated may also induce nutrient stress by competing with Trichodesmium for phosphorous compounds and iron.[17]

There is uncertainty about when and where *Trichodesmium* is Fe and P stressed and how this impacts nitrogen fixation in nature. Some reports suggest that *Trichodesmium* is primarily phosphate stressed in the North Atlantic, and primarily iron stressed in the Pacific, owing to relative Fe and P availability in these regions.[8–12,15] However, others have suggested that Fe and P can be co-limiting to *Trichodesmium*; one incubation study found two examples of Fe/P co-limitation in the field.[15] Even less clear is how Fe and/or P stress impacts nitrogen fixation. For instance, despite the intuitive suggestion that nitrogen fixation is limited by Fe or P availability, laboratory evidence indicated that *Trichodesmium* is specifically adapted to co-limited conditions, with higher growth and N_2-fixation rates under co-limitation than under single nutrient limitation[14,18].

There are several established protein biomarkers for Fe and P stress in *Trichodesmium,* all of which are periplasmic binding proteins involved in nutrient acquisition. For iron, this includes the IdiA and IsiB proteins and for phosphorus,

specifically phosphate, the PstS and SphX proteins (see Table S1). In *Trichodesmium* both IsiB, a flavodoxin, and IdiA, an ABC transport protein, are expressed under iron limiting conditions, and both are conserved across species with high sequence identity.[15,19] Transcriptomes and proteomes have shown that they are more abundant under iron stress conditions.[14,15,20] In this dataset, IsiB and IdiA were both highly abundant and correlated to one another (Figure S1). IdiA was used as the molecular biomarker of iron stress in the following discussion, but the same conclusions could be drawn from IsiB distributions. Like IdiA and IsiB, SphX and PstS are conserved across diverse *Trichodesmium* species.[15,21] SphX transcript and protein abundances increase more than two fold under phosphate limiting conditions.[10,14,22] PstS, a homologous protein located a few genes downstream of SphX, responds less clearly to phosphate stress. In *Trichodesmium*, the reason may be that PstS is not preceded by a Pho box, which is necessary for P based regulation.[22] Thus, in this study we focused on SphX as a measure of phosphate stress and IdiA as a marker of iron stress.

Here, we present evidence based on field metaproteomes that *Trichodesmium* colonies were simultaneously Fe and P stressed throughout the world's oceans, but particularly in the tropical and subtropical Atlantic. While Fe/P stress has been suggested before, this study provides molecular evidence for co-stress in a broad geographical and temporal survey. This co-stress occurred across significant gradients in Fe and P concentration, suggesting nutrient stress was driven not only by biogeochemical gradients but also by the inherent physiology of *Trichodesmium* itself. Fe and P stress were positively associated with nitrogen fixation and organic nitrogen uptake proteins, suggesting that *Trichodesmium's* Fe, P, and N statuses were closely linked.

3.3 RESULTS and DISCUSSION

3.3.1 Overview of the dataset

This study presents 36 field metaproteomes of colonial *Trichodesmium* populations collected at sixteen locations on four expeditions (Table S2). Most samples were from the subtropical and tropical Atlantic, and the majority were collected in the early morning hours to avoid changes occurring on the diel cycle (Figure 1 and Table S2). At each location, *Trichodesmium* colonies were hand picked from plankton net tows, rinsed in filtered seawater, collected onto filters, and immediately frozen. The metaproteomes were analyzed with a two-dimensional LC-MS/MS workflow that provided deep coverage of the proteome. Peptides were identified using a publicly available *Trichodesmium* metagenome (IMG ID 2821474806) and the CyanoGEBA project genomes as the sequence database.[23] This resulted in 4478 protein identifications, of which 2944 were *Trichodesmium* proteins. The remaining proteins were from colony-associated epibionts, which will be discussed in a future publication. Protein abundance is presented as precursor (MS1) intensity of the three most abundant peptides for each protein, normalized to total protein in the sample. Thus, changes in protein abundance were interpreted as changes in the fraction of the proteome devoted to that protein. The most abundant were GroEL, ribosome, and phycobilisome proteins.

A self-organizing map analysis identified groups of proteins with similar profiles, i.e. proteins whose abundances correlate cohesively, indicative of proteins that may be regulated similarly.[24] This revealed the central importance of nitrogen fixation to *Trichodesmium*. The nitrogenase proteins were among the most abundant in the proteome and were located in clusters 1 and 2 (Figure 2 and Table S3). Also in these clusters were nitrogen metabolism proteins including glutamine synthetase, glutamine hydrolyzing guanosine monophosphate (GMP) synthase and glutamate racemase. This is consistent with previous reports finding that N assimilation is synchronized with nitrogen fixation.[25]

Nitrogen fixation was closely linked to carbon fixation. Many photosystem proteins clustered with the nitrogenase proteins, including phycobilisome proteins, photosystem proteins, and the citric acid cycle protein 2-oxoglutarate dehydrogenase. Because these samples were collected pre-dawn to reduce diel effects, the relationships between C and N fixation proteins indicated that these processes are regulated directly in relation to the cell's physiological status. It seemed that the NtcA and P-II regulatory proteins, which link the carbon and nitrogen status of the cell by monitoring intracellular 2-oxoglutarate, were involved in this coordination.[26,27] 2-oxoglutarate is a key intermediate in the citric acid cycle but can also be converted into glutamate in the GS-GOGAT N assimilation pathway; therefore it is indicative of cellular C:N balance. In non-nitrogen fixing cyanobacteria, NtcA and P-II indicate nitrogen stress.[28,29] In diazotrophs, their role is unclear; for instance in *Trichodesmium* they do not respond to ammonium as they do in other cyanobacteria.[30] In this study, NtcA and P-II clustered with the N and C fixation proteins, suggesting that they may help to balance C and N fixation in the cell, though the details of this role have yet to be elucidated.

This self-organizing map analysis demonstrated that field populations of *Trichodesmium* invest heavily in macro- and micro-nutrient acquisition. There were clusters of proteins involved in trace metal acquisition and management, including iron, zinc, and metal transport proteins, with the latter including proteins likely involved in Ni and Mo uptake (TCCM_0270.00000020 & TCCM_0481.00000160). We also noted clusters of proteins involved in phosphate acquisition. Importantly, SphX and PstS appear in separate clusters, highlighting differential regulation of these functionally related proteins.

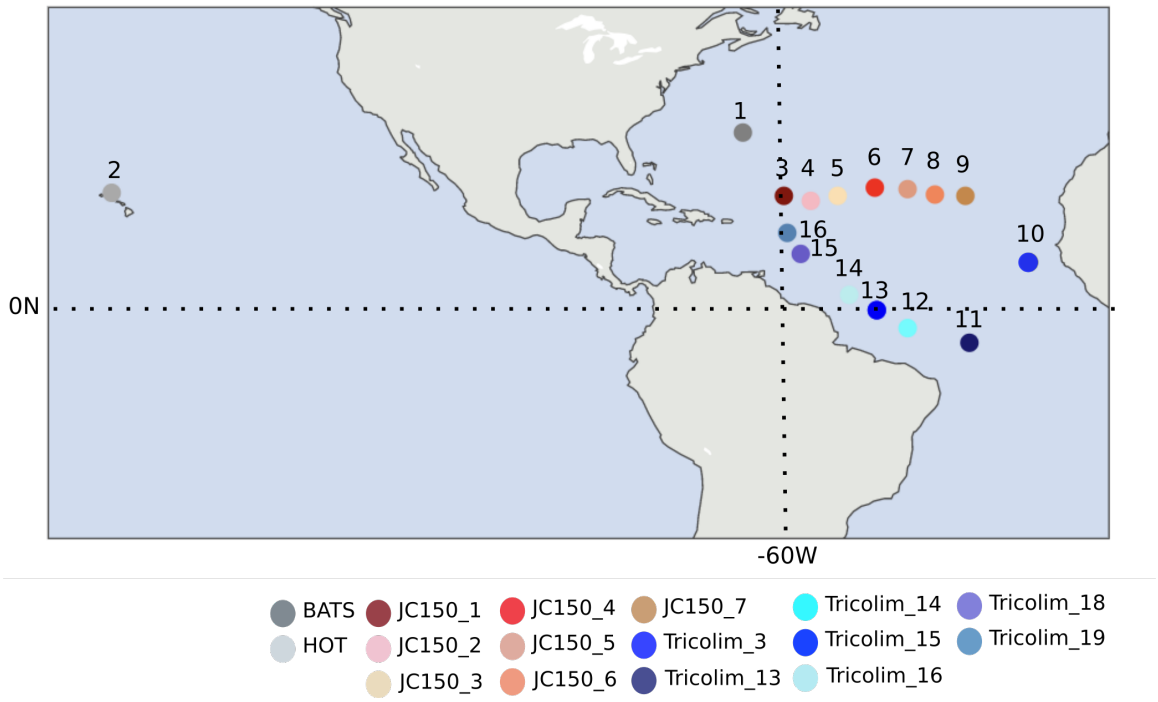


Figure 3.1. Sampling locations. Red/pink colors indicate JC150 stations; blue colors indicate Tricolim stations, dark grey indicates the Bermuda Atlantic Time Series (BATS) and light grey indicates Hawaii Ocean Time Series (HOT). Most samples exist in duplicate or triplicate; see Table S1 for detailed information.

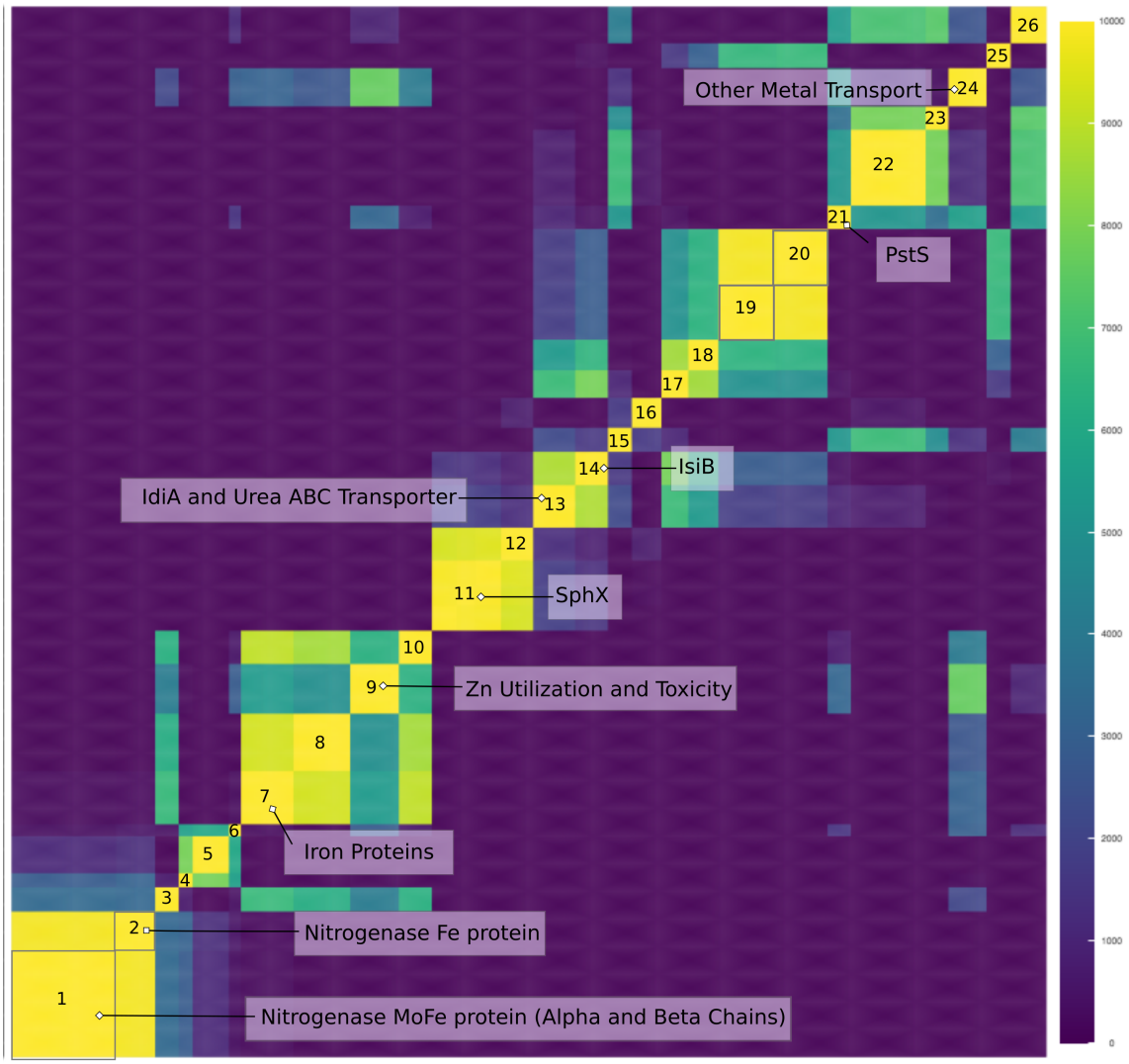


Figure 3.2. Heatmap displaying results of self-organizing map analysis. Each protein was mapped to a self-organizing map grid, and the grids subsequently clustered by a k-means clustering algorithm. The process was repeated 10,000 times and the results displayed here as a heatmap with warm colors representing proteins that appear in the same cluster. Dark yellow indicates proteins that appear in the same cluster 99.99% of the time. Only the top 500 most abundant proteins are displayed, with the same protein order on the x and y axes. Proteins group with themselves 100% of the time resulting in the diagonal line pattern. Clusters # 1 and 2 contain nitrogen fixation, carbon fixation, and nitrogen assimilation proteins as well as the regulatory systems NtcA and P-II. The cluster assignments for the proteins are available in Table S4.

3.3.2 Trichodesmium *is simultaneously iron and phosphate stressed throughout its habitat*

A surprising emergent observation from the *Trichodesmium* metaproteomes was the co-occurrence of multiple nutrient stress biomarkers in all samples. Biomarkers for iron (IdiA) and phosphate (SphX) stress were highly abundant and positively associated with surface Fe or P concentrations, following oceanographic trends (Figure 3). For instance, SphX varied up to 7.5 fold and was more abundant in populations from the North Atlantic gyre (JC150 expedition) compared with populations from near the Amazon river plume (Tricolim expedition) or at station ALOHA, where phosphate concentrations were greater (see Figure S3.2).[7,8,11] IdiA varied up to 8 fold, and increased moving West to East across the JC150 transect, consistent with an observed decrease in dFe concentrations.

The ubiquitous presence of both Fe and P stress biomarkers implied that co-stress was the norm rather than the exception for *Trichodesmium* colonies in the field. This interpretation is supported by prior laboratory studies demonstrating that IdiA and SphX protein abundances are greater under Fe and P depletion, and are less prevalent when nutrients are abundant.[10,14,15,19,20] This conclusion supports the growing acknowledgement that *Trichodesmium* experiences both iron and phosphate stress throughout its habitat.[31] If an implicit goal in understanding *Trichodesmium's* molecular physiology is to model the organism's behavior, these metaproteomes suggest that both iron and phosphorus conditions must be jointly considered.

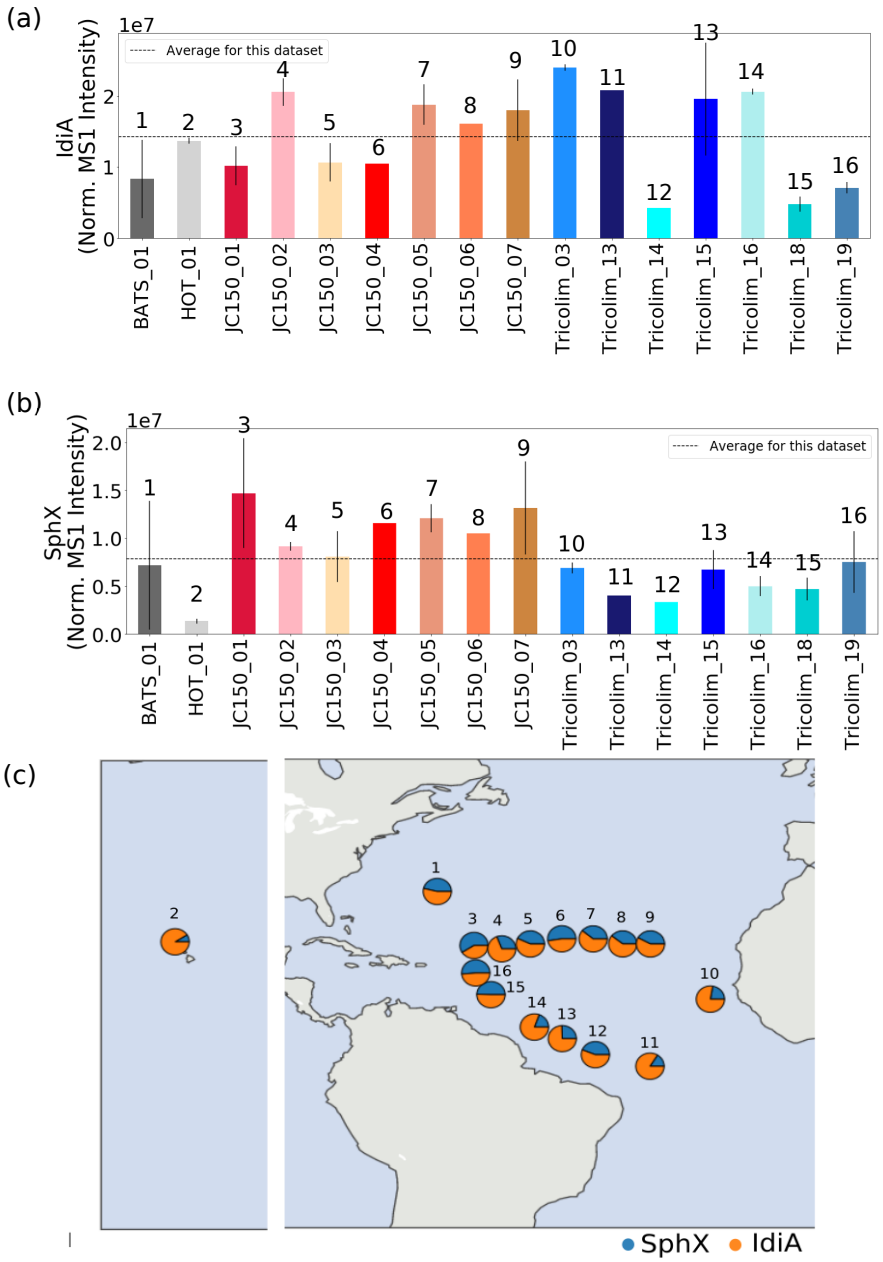


Figure 3.3. (A) Relative abundance of iron stress protein IdiA (A) and phosphate stress protein SphX (B). IdiA and SphX were among the most abundant proteins in the entire dataset. Error bars are one standard deviation on the mean when multiple samples were available. Dashed lines represent average values across the dataset. (C) Relative abundance of IdiA (orange) and SphX (blue) overlaid on the sampling locations.

3.3.3 The intersection of Fe, P and N stress

The metaproteomes enabled the relationship between Fe and P stress and overall cellular metabolism to be explored. Nitrogenase protein abundance was positively correlated with both IdiA and SphX, and was in fact highest at the intersection of high Fe and P stress (Figure 4). This observation contrasts with the current paradigm that *Trichodesmium* down

regulates nitrogen fixation when it is Fe or P stressed. Instead, it suggests that the nutritional demands of nitrogen fixation may drive *Trichodesmium* to Fe and P stress, thereby initiating an increase in Fe and P acquisition proteins. This indicates that cellular N, P and Fe status are linked, and perhaps by a regulatory network as is common in marine bacteria (Figure 5).[32] This network, which has yet to be fully characterized, responds in a specific manner to nutrient co-stress and affects the overall physiology of the cell. For instance, Fe and P co-limited *Trichodesmium* cells may reduce their cell size to optimize their surface area: volume quotient for nutrient uptake. However, a putative cell size biomarker Tery_1090, while abundant in co-limited cells in culture, was not identified in these metaproteomes despite bioinformatic efforts to target it, likely because it is a low abundance protein.[31]

Nitrogen fixation is not the only way that *Trichodesmium* can acquire fixed N.[8,31,33,34] In culture, *Trichodesmium* can be grown on multiple nitrogen sources including urea; in fact, it has been reported that nitrogen fixation provides less than 20% of the fixed N demand of cells, and a revised nitrogen fixation model requires Trichodesmium to take up nitrogen in the field.[35,36] In the current dataset, a urea ABC transporter was abundant, indicating that urea could be an important source of fixed nitrogen to colonies (Figure 6a). This urea transporter is unambiguously attributed to *Trichodesmium* rather than a member of the epibiont community. Of course, this does not rule out the possibility that urea or other organic nitrogen sources such as TMA are also utilized by epibionts, although no such epibiont transporters were identified in the metaproteomes.

Typically, elevated urea concentration decreases or eliminates nitrogen fixation in colonies.[37] However, in laboratory studies urea exposure must be unrealistically high (often over 20μM) for this to occur, compared with natural concentrations which are much lower.[37,38] In the field, urea utilization and nitrogen fixation seem to occur simultaneously, with a urea uptake protein positively correlated to nitrogenase abundance (Figure 6b). Urea and other organic nitrogen sources such as trimethylamine (TMA) could be sources of nitrogen for *Trichodesmium*, and the relationship with nitrogenase abundance could be a function of N stress driving organic nitrogen uptake and nitrogen fixation simultaneously.[39] Alternatively, urea uptake could be a colony-specific behavior, since colonies were sampled here as opposed to laboratory cultures that typically grow as single filaments. For instance, urea could be used for recycling of fixed N within the colony, or there could be heterogeneity in nitrogen fixation, with some cells taking up organic nitrogen and others fixing it. These unexpected observations of co-occurring nitrogen fixation and organic nitrogen transport show the value of exploratory metaproteomics, which does not require targeting of a specific protein based on a prior hypothesis.

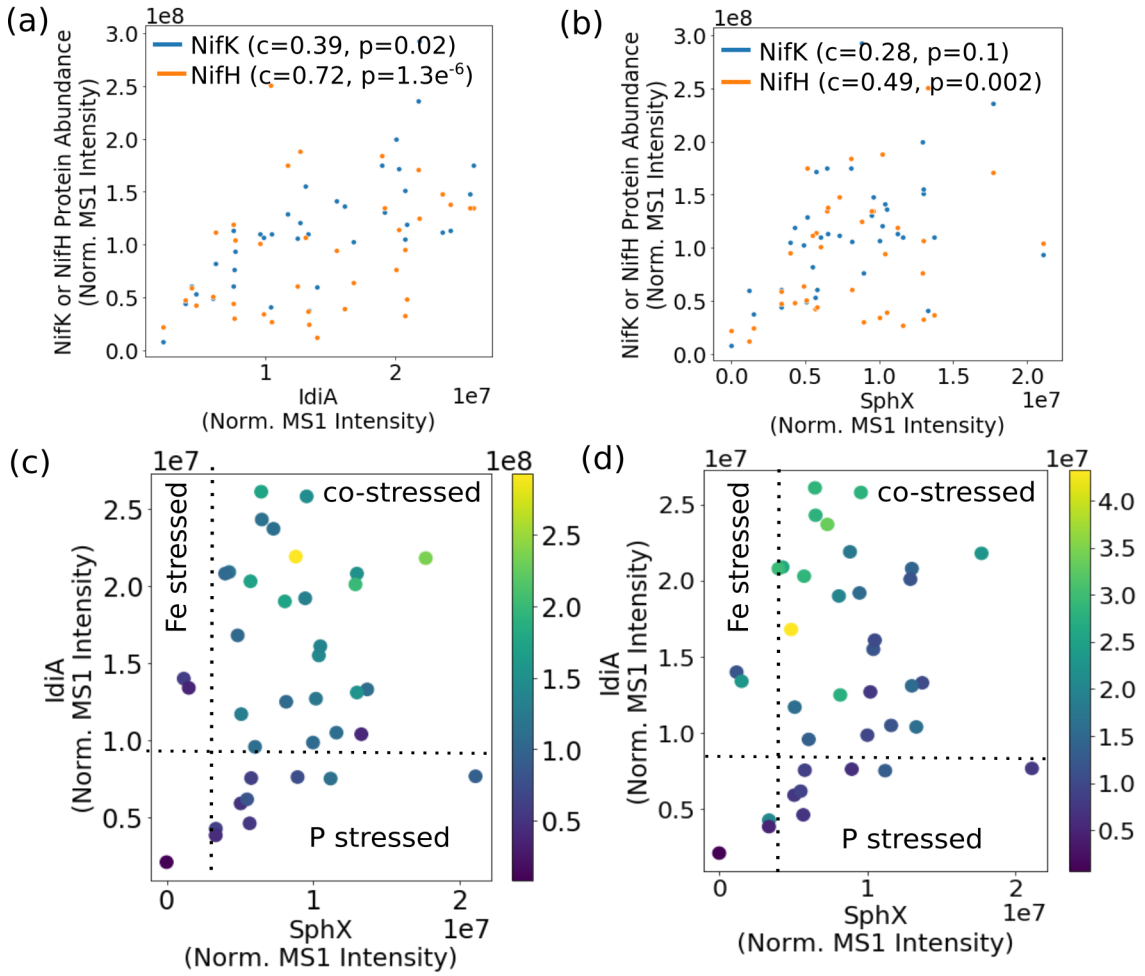


Figure 3.4. Nitrogenase abundance is highest at the intersection of high iron and phosphate stress. A) IdiA and B) SphX abundances are positively related to nitrogenase abundance (c=Spearman rank-order correlation coefficient, p = probability of Spearman correlation). Effects of combined iron and phosphate stress biomarkers on nitrogenase abundance. Marker colors represent abundance of NifK (panel C) and NifH (panel D).

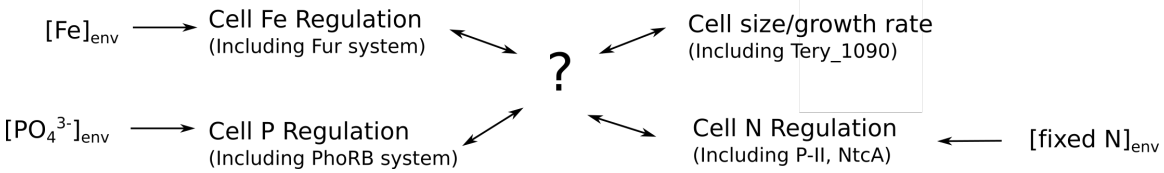


Figure 3.5. The metaproteomes suggest that there is a currently unknown regulatory link between cellular Fe, P, and N regulation. Key: Fur = ferric uptake regulator, PhoRB = phosphate two component sensory system, Tery_1090 = putative cell size regulator, P-II/NtcA = nitrogen regulatory proteins.

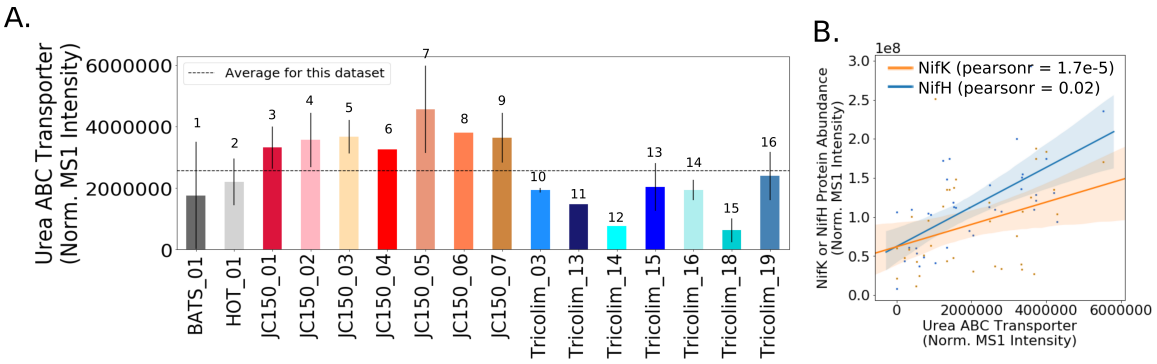


Figure 3.6. A) Relative abundance of the *Trichodesmium* urea ABC transporter. B) The abundance of the urea ABC transporter is positively correlated with NifH and NifK abundance. Statistical r values are pearson linear correlation coefficients (p value for NifK = $1.7e^{-5}$, NifH = 0.02)

3.3.4 Mechanisms of simultaneous iron and phosphate stress – membrane crowding

ABC transporters are multi-unit, trans-membrane protein complexes that use ATP to shuttle substrates across membranes. Specific ABC transporters are required for iron versus phosphate uptake. Nutrient transport rates can be modulated by changing the number of uptake ligands installed on the cell membrane or the efficiency of the uptake ligands through expression of assisting proteins such as IdiA and SphX, which bind Fe or P respectively in the periplasm and shuttle the elements to their respective membrane transport complexes. The high abundance of these proteins in all field populations suggested that nutrient transport rates could limit the amount of Fe and P *Trichodesmium* can acquire. Thus, we explored whether membrane crowding, i.e. lack of membrane space, can constrain nutrient acquisition by *Trichodesmium*.

To investigate this, we quantified the absolute concentration of the phosphate ABC transporter PstC, which interacts with the phosphate stress biomarkers SphX and PstS. This analysis is distinct from the above global metaproteomes, which allowed patterns to be identified but did not allow for absolute quantitation of the proteins themselves. The

analysis is performed similar to an isotope dilution experiment where labeled peptide standards are used to control for analytical biases. The analysis was performed for three Tricolim and six JC150 stations. Briefly, ^{15}N labeled peptide standards were prepared and spiked in known amounts into the samples prior to PRM LC-MS/MS analysis. Similar to an isotope dilution experiment, the concentration of the peptide in fmol μg^{-1} total protein was calculated using the ratio of product ion intensities for the heavy (spike) and light (sample) peptide and converted to PstC molecules per cell (Table 1 and see also Table S4). The peptide sequence used to quantify PstC is specific to the *Trichodesmium* genus. Based on these calculations, on average 19 to 36% of the membrane was occupied by the PstC transporter. In one population (JC150 expedition, Station 7), up to 83% of the membrane was occupied by PstC alone. While these are first estimates, it is clear that the majority of *Trichodesmium* cells devoted a large fraction of their membrane surface area to phosphate uptake.

To examine whether membrane crowding can indeed cause nutrient stress or limitation, we developed a model of cellular nutrient uptake in *Trichodesmium*. The model identifies the concentration at which free iron or phosphate limits the growth of *Trichodesmium* cells. This is distinct from nutrient stress, which changes the cell's physiological state but does not necessarily impact growth. In the model, nutrient limitation occurs when the daily cellular requirement is greater than the uptake rate, a function of the cell's growth rate and elemental quota. Following the example of Hudson and Morel (1992)[40], the model assumes that intake of nutrients once bound to the ABC transporter ligand is instantaneous, i.e. that nutrient uptake is limited by formation of the metal-ligand complex at the cell surface. This is an ideal scenario, because if intake is the slow step, for instance in a high affinity transport system, the uptake rate would be slower and nutrient limitation exacerbated (discussed below).

We considered two types of nutrient limitation in the model. First, we considered a diffusion-limited case, in which the rate of uptake is determined by diffusion of the nutrient to the cell's boundary layer ($\mu^*Q = {}^2/_3 k_D[\text{nutrient}]$, where μ = the cell growth rate, Q = the cell nutrient quota, and k_D = the diffusion rate constant, dependent on the surface area and diffusion coefficient of the nutrient in seawater). Based on empirical evidence provided by Hudson and Morel (1992), limitation occurs when the cell quota is greater than ${}^2/_3$ the diffusive-limited flux because beyond this, depletion of the nutrient in the boundary layer occurs[40]. In the second case, membrane crowding limitation, the rate of uptake is determined by the rate of ligand-metal complex formation ($\mu^*Q = k_f[\text{transport ligand}][\text{nutrient}]$, where k_f = the rate of ligand-nutrient complex formation). Here, up to 50% of the membrane can be occupied by the transport ligand following the example of Hudson and Morel (1992).[39] This is within the range of the above estimates of membrane occupation by phosphate transporter PstC. The model uses conservative estimates for diffusion coefficients, cell quotas, growth rates, and membrane space occupation to identify the lowest concentration threshold for nutrient limitation; as a result it is likely that *Trichodesmium* becomes limited at higher nutrient concentrations than the model suggests. At this time, the model can only consider labile dissolved iron and inorganic phosphate, though *Trichodesmium* can also acquire particulate iron, organic phosphorus, phosphite, and phosphonates.[9,34,41–43]

We first considered a spherical cell, where the surface area: volume quotient decreases as cell radius increases (Figure 7). As the cell grows in size, higher nutrient

concentrations are required to sustain growth. This is consistent with the general understanding that larger microbial cells with lower surface area: volume quotient are less competitive in nutrient uptake.[40,44] For a given surface area: volume quotient, we take the driver of nutrient limitation to be whichever model (membrane crowding or diffusion limitation) requires higher nutrient concentrations to sustain growth. For a spherical cell, iron limitation is driven by diffusion when the cell is large and the surface area: volume quotient is low (Figure 7a). However, when cells are smaller and the surface area: volume quotient is high, membrane crowding drives nutrient limitation, meaning that the number of ligands, and not diffusion from the surrounding environment, is the primary control on nutrient uptake. For phosphate, diffusion is almost always the driver of nutrient limitation owing to the higher rate of ligand-nutrient complex formation (k_f) for phosphate[45], which causes very fast membrane transport rates and relieves membrane-crowding pressures across all cell sizes (Figure 7b).

While this model may be directly applicable to some N_2-fixing cyanobacteria such as Groups B and C, which have roughly spherical cells, *Trichodesmium* cells are not spheres but rather roughly cylindrical. Thus, we repeated the model calculations for cylinders with varying radii (r) and heights (2r or 10r) based on previous estimates of *Trichodesmium* cell sizes.[12,46] Cylinders have lower surface area: volume quotient than spheres of similar sizes. In addition, the rate constant (k_D) for diffusion, which is a function of cell geometry, is greater. This increases the slope of the diffusion limitation line such that membrane crowding is important across a greater range of cell sizes (Figure 7c-d). *Trichodesmium* cell sizes vary in nature, for instance the cylinder height can be elongated, improving the surface area: volume quotient. However, the impact of cell elongation to radius r and height 10r on both diffusion limitation and membrane crowding is subtle (Figure 7e-f). Thus, we conclude that in certain scenarios, lack of membrane space could indeed limit Fe and P acquisition by *Trichodesmium*.

A key assumption of the model is that uptake rates are instantaneous. In the above calculations, we use the dissociation kinetics of iron from water and phosphorus with common seawater cations as the best case (i.e. fastest possible) kinetic scenario for nutrient acquisition. The model does not account for delays caused by internalization kinetics, which would exacerbate nutrient limitation. For instance it does not consider nutrient speciation, which could affect internalization rates particularly for iron.[40] Furthermore, the responsiveness of the periplasmic binding proteins IdiA and SphX/PstS to environmental abundance (Figure 3) suggests that uptake is not simultaneous; their involvement is likely associated with a kinetic rate of binding and dissociation from the periplasmic proteins in addition to any rate of ABC transport. Indeed, membrane crowding could explain why there is not a perfect dose-dependent response of IdiA and SphX in response to nutrient availability (see Figure 3). If nutrient acquisition is limited because the cell's uptake systems are saturated, increasing the number of periplasmic binding proteins will have limited effect.

Membrane crowding could produce real cellular challenges, leading to the observation of Fe and P co-stress across the field populations examined. The above model explicitly allows 50% of the cell surface area to be occupied by any one nutrient, consistent with our estimate of cell surface area occupied by the PstC transporter. If 50% of the membrane is occupied by phosphate transporters, and another 50% for iron transporters, this would leave no room for other essential membrane proteins and even the membrane lipids themselves. The problem is further exacerbated if the cell installs transporters for nitrogen

compounds such as urea, as the metaproteomes suggest. Thus, installation of transporters for any one nutrient must be balanced against transporters for other nutrients. This interpretation is inconsistent with Liebig's law of nutrient limitation, which assumes that nutrients are independent.[47,48] In an oligotrophic environment, membrane crowding could explicitly link cellular Fe, P, and N uptake status, driving the cell to be co-stressed for multiple nutrients.

Station	[Pst] in fmol/ug total protein (replicate average)	standard deviation replicates (if available)	Pst molecules per cell assuming 30% w/w protein content*	% surface area occupied assuming 30% w/w^	Pst molecules per cell assuming 55% w/w protein content**	% surface area occupied assuming 55% w/w^
Tricolim_18	13.0	1.8	3.8E+05	3.6	7.0E+05	6.6
Tricolim_15	11.2	3.4	3.3E+05	3.1	6.1E+05	5.7
Tricolim_16	89.1	123.1	2.6E+06	24.5	4.8E+06	45.0
JC150_3	38.7	63.3	1.1E+06	10.7	2.1E+06	19.5
JC150_4	89.6	14.7	2.6E+06	24.7	4.9E+06	45.2
JC150_5	74.2	36.4	2.2E+06	20.4	4.0E+06	37.5
JC150_6	61.6	40.1	1.8E+06	17.0	3.3E+06	31.1
JC150_7	165.7		4.9E+06	45.6	9.0E+06	83.6
JC150_1	106.1		3.1E+06	29.2	5.7E+06	53.5
average				19.9		36.4
stdev				13.4		24.6

* calculated using Trichodesmium cell volume of 3000um3 (Berman-Frank et al., 2001), cell volume to carbon conversion logC = 0.716log(V)-0.314 (Strathman, 1967), protein content of a cyanobacterium 30% w/w (Gonzalez Lopez et al., 2010), carbon to total protein conversion 0.53 g C/ g total protein (Rowenhoerst et al., 1991). **calculated as in (*) but with protein content of a cyanobacterium 50% w/w (Gonzalez Lopez et al., 2010). ^calculated using cross sectional area of an Ca ATPase of 0.0000167 um2 (Hudson and Morel 1992)

Table 3.1. Quantification of the Pst ABC transporter and estimation the percent of the membrane space occupied by this protein.

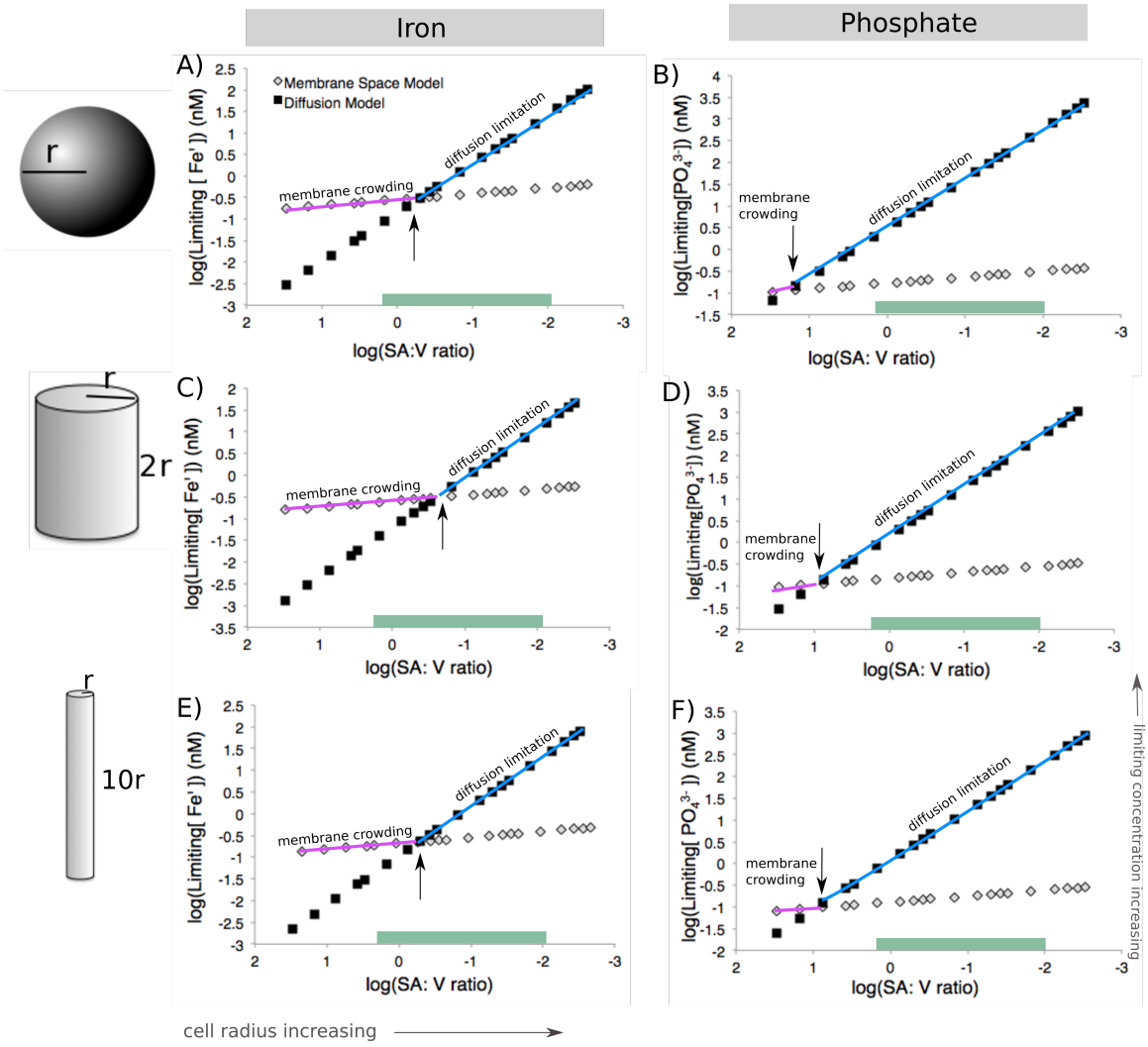


Figure 3.7. Model calculations for membrane space and diffusion based nutrient limitation reveal that membrane crowding could drive *Trichodesmium* to iron or phosphate stress, particularly when cells are small. Two cell morphologies (sphere and cylinder) were modeled for both iron and phosphate limitation. As the cell radius increases and the surface area: volume quotient decreases, the limiting concentration increases. This is concurrent with the current understanding that the low surface area: volume quotient of large cells leads to limitation. Green bars represent common SA: V ratios for *T. Theibautii*.[46] (A-B). Membrane crowding (purple) occurs if the limiting nutrient concentration is greater than in the diffusion limitation model (blue). Membrane crowding is more significant for cylindrical cells in particular (C-D); altering the length of the cylinder only minimally affects the model (E-F).

3.3.5 Advantages of the colonial form

Living in a colony has specific advantages and disadvantages for a *Trichodesmium* cell. Colonies may be able to access nutrient sources that would be infeasible for use by single cells or filaments. For instance, *Trichodesmium* colonies have a remarkable ability to entrain dust particles and can move these particles into the center of the colony.[42,43,49,50] In this study, which focused on *Trichodesmium* colonies, the chemotaxis response regulator CheY was very abundant, particularly in populations sampled near the Amazon and Orinoco river plumes. CheY was positively correlated with iron stress biomarker IdiA, but not with phosphate stress biomarker SphX (Figure 8). This suggests that chemotactic movement is involved in entrainment of mineral particles in colonies.

The metaproteomes and nutrient uptake modeling presented in this paper support the growing understanding that *Trichodesmium* must be able to access particulate and organic matter. Living in a colony can be advantageous because such substrates can be concentrated, improving the viability of extracellular nutrient uptake systems. *Trichodesmium's* epibiont community produces siderophores, which assist in iron uptake, particularly from particulate sources.[51,52] Siderophore production is energetically and nutritionally expensive, so it is most advantageous when resource concentrations are high and loss is low, as would occur in the center of a colony.[53] Colonies may similarly enjoy advantages for phosphate acquisition, particularly when the excreted enzyme alkaline phosphatase is utilized to access organic sources[9,10,54–56]. Additionally, concentration of cells in a colony means that the products of nitrogen fixation, including urea, can be recycled and are less likely to be lost to the environment. By increasing the effective size and concentrating deterrent toxins, colony formation may also protect against grazing.[57]

A key hallmark of *Trichodesmium* colony formation is production of mucus, which can capture particulate matter and concentrate it within the colony. In addition to particle entrainment, the mucus layer can benefit cells by protecting them from oxygen, facilitating epibiont associations[58–60], regulating buoyancy, defending against grazers and helping to "stick" trichomes together. However, these benefits come at a cost because the mucus layer hinders diffusion to the cell surface (Figure 9), reducing contact with the surrounding seawater. Despite this, the benefits of colony formation seem to outweigh the costs, since *Trichodesmium* forms colonies in the field, particularly under stress.[12,46,61]

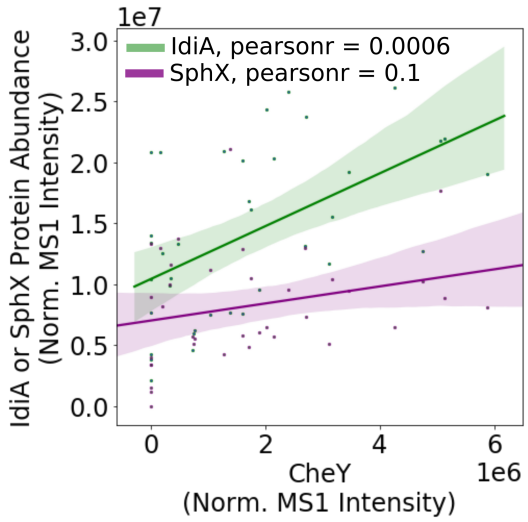


Figure 3.8. CheY is positively correlated with the iron stress biomarker IdiA, but has a weaker association with phosphate stress biomarker SphX. This suggests that it might be involved in iron acquisition, for instance by helping colonies to move dust particles to the colony center. Pearson linear correlation coefficients (r values) are provided (p value for IdiA = $6e^{-4}$, SphX = 0.1)

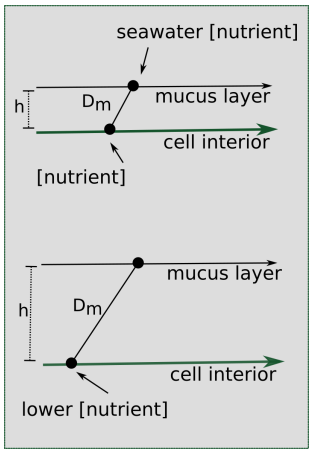


Figure 3.9. Scheme for the effect of a mucus layer on nutrient diffusion. h = height of the mucus membrane, D_m = diffusion coefficient of the mucus. Assuming some diffusion constant for the nutrient through the mucus and the same starting seawater nutrient concentration, a thicker layer of mucus surrounding a cell in a colony would result in a lower concentration of nutrient experienced at the cell surface.

3.4 CONCLUSION

Trichodesmium's colonial lifestyle likely produces challenges for dissolved Fe and P acquisition, which must be compensated for by production of multiple nutrient transport systems, such as for particulate iron and organic phosphorous, at a considerable cost. While laboratory studies have largely focused on single nutrient stresses in free filaments, these metaproteomic observations and accompanying cellular modeling demonstrate that Fe and P co-stress is the norm rather than the exception. This means that the emphasis on single limiting nutrients in culture studies and biological models may not capture the complexities of *Trichodesmium's* physiology in situ. Thus, biogeochemical models should consider incorporating Fe and P co-stress. Specifically, in this study and in others there is evidence that nitrogen fixation is optimal under co-limited or co-stressed conditions, implying that an input of either Fe or P could counter-intuitively decrease N_2 driven new production.[14,18]

These data demonstrate that *Trichodesmium* cells are confronted by the biophysical limits of membrane space and diffusion rates for their Fe, P, and possibly urea, acquisition systems. This means that there is little room available for systems that interact with other resources such as light, CO_2, Ni, and other trace metals, providing a mechanism by which nutrient stress could compromise acquisition of other supplies. The cell membrane could be a key link allowing *Trichodesmium* to optimize its physiology in response to multiple environmental stimuli. This is particularly important in an ocean where nutrient availability is sporadic and unpredictable. Future studies should aim to characterize the specific regulatory systems, chemical species and phases, and symbiotic interactions that underlie *Trichodesmium's* unique behavior and lifestyle.

3.5 MATERIALS and METHODS

3.5.1 Sample Acquisition

A total of 37 samples were examined in this study. Samples were acquired by the authors on various research expeditions and most exist in biological duplicate or triplicate. *Trichodesmium* colonies were hand-picked from 200μm surface plankton net tows, rinsed thrice in 0.2μm filtered surface seawater, decanted onto 0.2-5μm filters, and frozen until protein extraction. The samples were of mixed puff and tuff morphology types, depending on the natural diversity present at the sampling location. The majority of samples considered in this study were taken in the early morning pre-dawn hours. Details such as filter size, morphology, location, cruise, date, and time of sampling are provided in Table S1.

3.5.2 Protein extraction and digestion

Proteins were extracted by a previously described detergent based method.[27,62] To reduce protein loss and contamination, all tubes were ethanol rinsed and dried prior to use and all water and organic solvents used were LC/MS grade. Sample filters were placed in a tube with 1-2mL 1% sodium dodecyl sulfate (SDS) extraction buffer (1% SDS, 0.1M Tris/HCL pH 7.5, 10mM EDTA) and incubated for 10min at 95°C with shaking, then for one hour at room temperature with shaking. The protein extract was decanted and clarified

by centrifugation (14100 x g) at room temperature. The crude protein extracts were then quantified with the colormetric BCA protein concentration assay with bovine serum albumin as a standard. Extracts were concentrated by 5 kD membrane centrifugation (Vivaspin spin columns). The protein extracts were purified by organic precipitation in 0.5mM HCl made in 50% methanol and 50% acetone at -20°C for at least one week, then collected by centrifugation at 14100xg for 30min at 4°C, decanted and dried by vacuum concentration for 10min. The protein pellets were resuspended in a minimum amount of 1% SDS extraction buffer, and re-quantified by BCA protein concentration assay to assess extraction efficiency.

Protein extracts were digested in a polyacrylamide tube gel as previously described.[27,63] The gel was cut up into 1mm pieces to maximize surface area, then rinsed in 50:50 acetonitrile: 25mM ammonium bicarbonate overnight. The 50:50 rinse was then replaced and the rinse repeated for 1 hour. Gels were dehydrated thrice in acetonitrile, dried by vacuum centrifugation, and rehydrated in 10mM dithiothrietol (DTT) in 25mM ammonium bicarbonate, then incubated for one hour at 56°C with shaking. Unabsorbed DTT solution was removed and the volume recorded, allowing for calculation of the total gel volume. Gels were washed in 25mM ammonium bicarbonate, then incubated in 55mM iodacetamide for one hour at room temperature in the dark. Gels were again dehydrated thrice in acetonitrile. Trypsin (Promega Gold) was added at a ratio of 1:20 total protein in 25mM ammonium bicarbonate in a volume sufficient to barely cover the gel cubes. Proteins were digested overnight at 37°C with shaking. Any unabsorbed solution was then removed to a new tube and 50µL of peptide extraction buffer (50% acetonitrile, 5% formic acid in water) was added and incubated for 20 min at room temperature. The supernatant as then decanted and combined with the unabsorbed solution, and the step then repeated. The resulting peptide mixture was then concentrated by vacuum centrifugation to 1µg/µL concentration. Finally, the peptides were clarified by centrifugation at room temperature, taking the top 90% of the volume to reduce the carry over of gel debris.

3.5.3 LC x LC-MS Global Proteome Analysis

The global proteomes were analyzed by online comprehensive active-modulation two-dimensional liquid chromatography (LC x LC-MS) using high and low pH reverse phase chromatography. 10ug of protein was injected per run directly onto the first column using a Thermo Dionex Ultimate3000 RSLCnano system, and an additional RSLCnano pump was used for the second dimension gradient. The samples were then analyzed on a Thermo Orbitrap Fusion mass spectrometer with a Thermo Flex ion source.

3.5.4 Relative quantitation of peptides and proteins

Raw spectra were searched with SequestHT using a custom built genomic database. The genomic database consisted of a publically available *Trichodesmium* community metagenome available on the JGI IMG platform (IMG ID 2821474806), as well as the entire contents of the CyanoGEBA project genomes [23]. Protein annotations were derived from the original metagenome. SequestHT mass tolerances were set at +/- 10ppm (parent) and +/- 0.8 Dalton (fragment). Cysteine modification of +57.022 and methionine modification of +16

were included. Protein identifications were made with Peptide Prophet in Scaffold (Proteome Software) at the 95% protein and peptide identification levels. Relative abundance is measured by normalized top 3 precursor (MS1) intensities. Normalization and false discovery rate (FDR) calculations, which were 0.1% peptide and 1.2% protein, were performed in Scaffold (Proteome Software). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD016225 and 10.6019/PXD016225.[64]

3.5.5 Absolute quantitation of peptides

A small number of peptides were selected for absolute quantitative analysis using a modified heterologous expression system (McIlvin and Saito in prep). A custom plasmid was designed which contained the *Escherichia coli* K12 optimized reverse translation for peptides of interest separated by tryptic spacers (protein sequence = TPELFR). To avoid repetition of the spacer nucleotide sequence, twelve different codons were utilized to encode the spacer. Six equine apomyoglobin and three peptides from the commercially available Pierce peptide retention time calibration mixture (product number 88320) were also included. The gene was inserted into a pet(30a)+ plasmid using the BAMH1 5' restriction site and XhoI 3' restriction site such that one his tag would be present on the overexpressed protein for purification purposes.

The plasmid was transformed into competent tuner(DE3)pLys *E.coli* cells and grown on kanamycin amended LB agar plates to ensure plasmid incorporation. A single colony was used to inoculate a small amount of kanamycin containing ^{15}N labeled S.O.C. media (Cambridge Isotope Laboratories) as a starter culture. These cells were grown overnight and then used to inoculate 10mL of ^{15}N labeled, kanamycin-containing SOC media. Cells were grown to approximately OD600 0.6, then induced with 1mM isopropyl β-D-1-thiogalactopyranoside (IPTG), incubated in the overexpression phase overnight at room temperature and harvested by centrifugation.

Cells were lysed using BugBuster detergent with added benzonase nuclease. The extracts were centrifuged and a large pellet of insoluble cellular material remained. Because the plasmid protein is large, this pellet contained a large number of inclusion bodies containing nearly pure protein. The inclusion bodies were solubilized in 6M urea at 4°C overnight. The protein was reduced, alkylated, and trypsin digested in solution to generate a standard peptide mixture.

The standard peptide mixture was calibrated to establish the exact concentration of peptides. A known amount (10fmol/μL) of the commercially available Pierce standard peptide mixture and an apomyoglobin digest was spiked into the standard. The ratio of Pierce (isotopically labeled according to JPT standards) or apomyoglobin (light) to heavy standard peptide MS2 peak area was calculated and used to establish the final concentration. MS2 peak area is the gold standard for absolute protein quantitation.[27] Multiple peptides were used for this calibration and the standard deviation among them was approximately 10%. Finally, the linearity of the peptide standard was tested by generating a dilution curve and ensuring that the concentration of each peptide versus MS2 peak area was linear between 0.001 and 20fmol/μL concentration.

The sample was prepared at 0.2μg/μL concentration, with 10μL injected to give a total of 2μg total protein analyzed. The heavy labeled standard peptide mixture was spiked

into each sample at a concentration of 10fmol/µL.The concentration of the light peptide was calculated as the ratio of the MS2 area of the light:heavy peptide multiplied by 10µg/µL. A correction was applied for protein recovery before and after precipitation, and the result was the absolute concentration of the peptide in fmol/µg total protein.

The percent of the membrane occupied by the ABC transporter was calculated by converting the absolute concentration to molecules per *Trichodesmium* cell, using average values for *Trichodesmium* cell volume [12], carbon content per volume[65], protein content per g carbon[66], and the cross sectional area of a calcium ATPase[40] (see Table S4).

3.5.6 Self-organizing map analyses

Self-organizing maps were used to reduce the dimensionality of the data and explore relationships among co-varying proteins of interest. Only *Trichodesmium* proteins were considered. Analyses were conducted in Python 3.0 and fully reproducible code is available at https://github.com/naheld/self_organizing_map_tricho_metaP.

The input data consisted of a table of protein names (rows) and samples (columns) such that the input vectors contained 2818 features. To eliminate effects of scaling, the data was unit normalized with the Scikit-learn preprocessing package. The input vectors were used to initialize a 100 output node (10x10) self-organizing map using the SOMPY Python library (https://github.com/sevamoo/SOMPY). The output nodes were then clustered using a k-means clustering algorithm (k = 10) implemented in scikit learn. The input nodes (proteins) assigned to each map node were then retrieved and the entire process repeated 10,000 times. Proteins were considered in the same cluster if they appeared in the same cluster of output nodes more than 99.99% of the time.

3.6 ACKNOWLEDGEMENTS

I thank my coauthors on this project, Eric Webb, Matthew McIlvin, Dave Hutchins, Natalie Cohen, Dawn Moran, Korrina Kunde, Maeve Lohan, Claire Mahaffey, Malcolm Woodword, and Mak Saito. I also thank Ben Van Mooy for insightful discussions early in this work. We also acknowledge Petroc Shelley, Elena Cerdan Garcia, Despo Polyviou, Asa Conover and Joanna Harley for assistance with *Trichodesmium* sampling and phosphate measurements. We thank Joe Jennings (Oregon State) for performing nutrient measurements for the Tricolim cruise. This work was supported by the Gordon and Betty Moore Foundation (grant number 3782 [M.Saito]), by the National Science Foundation (grant numbers 1657766, 1639714, 1260233 [M. Saito]) and by a National Science Foundation Graduate Research Fellowship # 1122274 [N.Held]).

3.7 REFERENCES

1. Capone, D. G. *Trichodesmium*, a Globally Significant Marine Cyanobacterium. *Science (80-.).* **276**, 1221–1229 (1997).
2. Sohm, J. A., Webb, E. A. & Capone, D. G. Emerging patterns of marine nitrogen fixation. *Nat. Rev. Microbiol.* **9**, 499–508 (2011).
3. Deutsch, C., Sarmiento, J. L., Sigman, D. M., Gruber, N. & Dunne, J. P. Spatial coupling of nitrogen inputs and losses in the ocean. *Nature* **445**, 163–167 (2007).
4. Carpenter, E. J. & Romans, K. Major Role of the Cyanobacterium *Trichodesmium* in Nutrient Cycling in the North Atlantic Ocean. *Science* **254**, 1989–1992 (1991).
5. Coles, V. J., Hood, R. R., Pascual, M. & Capone, D. G. Modeling the impact of *Trichodesmium* and nitrogen fixation in the Atlantic ocean. *J. Geophys. Res. C Ocean.* **109**, 1–17 (2004).
6. Karl, D. *et al*. Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57–58**, 47–98 (2002).
7. Wu, J. F., Sunda, W., Boyle, E. A. & Karl, D. M. Phosphate depletion in the western North Atlantic Ocean. *Science (80-.).* **289**, 759–762 (2000).
8. Sañudo-Wilhelmy, S. A. *et al*. Phosphorus limitation of nitrogen fixation by *Trichodesmium* in the central Atlantic Ocean. *Nature* **411**, 66–69 (2001).
9. Frischkorn, K. R., Krupke, A., Guieu, C., Louis, J. & Rouco, M. *Trichodesmium* physiological ecology and phosphate reduction in the western tropical South Pacific. *Biogeosciences* 15, 5761–5778 (2018).
10. Orchard, E. D. Phosphorus physiology of the marine Cyanobacterium Trichodesmium. *Massachusetts Inst. Technol.* 130 (2010). doi:10.1575/1912/3366
11. Hynes, A. M., Chappell, P. D., Dyhrman, S. T., Doney, S. C. & Webb, E. A. Cross-basin comparison of phosphorus stress and nitrogen fixation in *Trichodesmium*. *Limnol. Oceanogr.* **54**, 1438–1448 (2009).
12. Bergman, B., Sandh, G., Lin, S., Larsson, J. & Carpenter, E. J. *Trichodesmium*--a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol. Rev.* **37**, 286–302 (2013).
13. Sunda, W. G. Feedback interactions between trace metal nutrients and phytoplankton in the ocean. *Front. Microbiol.* **3**, 1–22 (2012).
14. Walworth, N. G. *et al*. Mechanisms of increased *Trichodesmium* fitness under iron and phosphorus co-limitation in the present and future ocean. *Nat Commun* **7**, 1–11 (2016).
15. Chappell, P. D., Moffett, J. W., Hynes, A. M. & Webb, E. A. Molecular evidence of iron limitation and availability in the global diazotroph *Trichodesmium*. *ISME J.* **6**, 1728–1739 (2012).
16. Rouco, M., Frischkorn, K. R., Haley, S. T. & Alexander, H. Transcriptional patterns identify resource controls on the diazotroph *Trichodesmium* in the Atlantic and Pacific oceans. *ISME J.* 1486–1495 (2018). doi:10.1038/s41396-018-0087-z
17. Frischkorn, K. R., Rouco, M., Mooy, B. A. S. Van, & Dyhrman, S. T. Epibionts dominate metabolic functional potential of *Trichodesmium* colonies from the oligotrophic ocean, 2090–2101 (2017)
18. Garcia, N. S., Fu, F., Sedwick, P. N. & Hutchins, D. A. Iron deficiency increases growth and nitrogen-fixation rates of phosphorus-deficient marine cyanobacteria.

ISME J. **9**, 238–45 (2015).
19. Webb, E. A., Jakuba, R. W., Moffett, J. W. & Dyhrman, S. T. Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol*. *Oceanogr*. **52**, 2221–2232 (2007).
20. Snow, J. T., Polyviou, D., Skipp, P., Chrismas, N. A. M., Hitchcock, A., Geider, R., Bibby, T. S. (2015). Quantifying integrated proteomic responses to iron stress in the globally important marine diazotroph *Trichodesmium*. *PLoS ONE*, *10*(11), 1–24. (2015)
21. Walworth, N. G., Lee, M. D., Fu, F.-X., Hutchins, D. A. & Webb, E. A. Molecular and physiological evidence of genetic assimilation to high CO_2 in the marine nitrogen fixer *Trichodesmium*. *Proc*. *Natl*. *Acad*. *Sci*. 201605202 (2016). doi:10.1073/pnas.1605202113
22. Orchard, E. D., Webb, E. A. & Dyhrman, S. T. Molecular analysis of the phosphorus starvation response in *Trichodesmium* spp. *Environ*. *Microbiol*. **11**, 2400–2411 (2009).
23. Shih, P. M. *et al*. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc*. *Natl*. *Acad*. *Sci*. *U*. *S*. *A*. **110**, 1053–1058 (2013).
24. Reddy, R. J. *et al*. Early signaling dynamics of the epidermal growth factor receptor. *Proc*. *Natl*. *Acad*. *Sci*. *U*. *S*. *A*. **113**, 201521288 (2016).
25. Carpenter, E. J. *et al*. Glutamine synthetase and nitrogen cycling in colonies of the marine diazotrophic cyanobacteria *Trichodesmium* spp. *Appl*. *Environ*. *Microbiol*. **58**, 3122–9 (1992).
26. Flores, E. & Herrero, A. Nitrogen assimilation and nitrogen control in cyanobacteria: Figure 1. *Biochem. Soc. Trans.* **33**, 164–167 (2005).
27. Saito, M. A. *et al.* Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **345**, 1173–7 (2014).
28. Mohr, W., Intermaggio, M. P. & LaRoche, J. Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium Crocosphaera watsonii WH8501. *Environ. Microbiol.* **12**, 412–421 (2010).
29. Saito, M. A. *et al*. Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph *Crocosphaera watsonii*. *Proc*. *Natl*. *Acad*. *Sci*. *U*. *S*. *A*. **108**, 2184–9 (2011).
30. Küpper, H. *et al*. Traffic Lights in *Trichodesmium* . Regulation of Photosynthesis for Nitrogen Fixation Studied by Chlorophyll Fluorescence Kinetic Microscopy Published by : American Society of Plant Biologists (ASPB) Linked references are available on JSTOR for this arti. **135**, 2120–2133 (2019).
31. Flores, E. & Herrero, A. Nitrogen assimilation and nitrogen control in cyanobacteria: Figure 1. *Biochem*. *Soc*. *Trans*. **33**, 164–167 (2005).
32. Saito, M. A. *et al*. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **345**, 1173–7 (2014).
33. Mohr, W., Intermaggio, M. P. & LaRoche, J. Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium Crocosphaera watsonii WH8501. *Environ. Microbiol.* **12**, 412–421 (2010).
29. Saito, M. A. *et al.* Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph Crocosphaera watsonii. *Proc. Natl. Acad. Sci. U. S. A.* **108**,

2184–9 (2011).
30. Forchhammer, K. & De Marsac, N. T. The P(II) protein in the cyanobacterium *Synechococcus* sp. strain PCC 7942 is modified by serine phosphorylation and signals the cellular N-status. *J. Bacteriol.* **176**, 84–91 (1994).
31. Mills, M. M., Ridame, C., Davey, M., La Roche, J. & Geider, R. J. Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic. *Nature* **429**, 292–294 (2004).
32. Held, N. A., Mcilvin, M. R., Moran, D. M., Laub, M. T. & Saito, A. Unique Patterns and Biogeochemical Relevance of Two-Component Sensing in Marine Bacteria. *mSystems* 1–16 (2019).
33. Küpper, H., Setlik, I., Siebert, S., Prasil, O., Setlikova, E., Stritmatter, M., Levitan, O., Lohscheider, J., Adamska, I., Berman-Frank, I. Iron limitation in the marine cyanobacterium *Trichodesmium* reveals new insights into regulation of photosynthesis and nitrogen fixation. New Phytologist **179**: 784-798 (2008).
34. Dyhrman, S. T. *et al.* Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*. *Nature* **439**, 68–71 (2006).
35. Mulholland, M. R. & Capone, D. G. Nitrogen utilization and metabolism relative to patterns of N_2 fixation in populations of *Trichodesmium* from the North Atlantic Ocean and Caribbean Sea. *Mar. Ecol. Prog. Ser.* **188**, 33–49 (1999).
36. McGillicuddy Jr., D. J. Do *Trichodesmium* spp. populations in the North Atlantic export most of the nitrogen they fix? *Global Biogeochem. Cycles* **28**, 103–114 (2014).
37. Ohki, K., Zehr, J. P., Falkowski, P. G. & Fujita, Y. Regulation of nitrogen-fixation by different nitrogen sources in the marine non-heterocystous cyanobacterium *Trichodesmium* sp. NIBB1067. *Arch. Microbiol.* **156**, 335–337 (1991).
38. Wang, Q., Li, H. & Post, A. F. Nitrate assimilation genes of the marine diazotrophic, filamentous cyanobacterium *Trichodesmium* sp. strain WH9601. *J. Bacteriol.* **182**, 1764–1767 (2000).
39. Walworth, N. G. *et al.* Nutrient-colimited *Trichodesmium* as a nitrogen source or sink in a future ocean. *Appl. Environ. Microbiol.* **84**, 1–14 (2018).
40. Hudson, Robert J Morel, Morel, F. M. M. Trace metal transport by marine microorganisms: implications of metal coordination kinetics. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **40**, 129–150 (1992).
41. Polyviou, D., Hitchcock, A., Baylay, A. J., Moore, C. M. & Bibby, T. S. Phosphite utilization by the globally important marine diazotroph *Trichodesmium*. *Environ. Microbiol. Rep.* **7**, 824–830 (2015).
42. Rubin, M., Berman-Frank, I. & Shaked, Y. Dust-and mineral-iron utilization by the marine dinitrogen-fixer *Trichodesmium*. *Nat. Geosci.* **4**, 529–534 (2011).
43. Poorvin, L., Rinta-kanto, J. M., Hutchins, D. A. & Wilhelm, S. W. Viral release of iron and its bioavailability to marine plankton. **49**, 1734–1741 (2004).
44. Chisholm, S. W. Phytoplankton Size. in *Primary Productivity and Biogeochemical Cycles in the Sea* (eds. Falkowski, P. G., Woodhead, A. D. & Vivirito, K.) 213–237 (Springer US, 1992). doi:10.1007/978-1-4899-0762-2_12
45. P.N. Froelich, M.L. Bender, N.A. Luedtke, G.R. Heath, T. D. The Marine Phosphorus Cycle. *Am. J. Sci.* **282**, 464–511 (1982).
46. Hynes, A. M., Webb, E. A., Doney, S. C. & Waterbury, J. B. Comparison of cultured

Trichodesmium (cyanophyceae) with species characterized from the field. *J. Phycol.* 48 , 196–210 (2012).
47. Liebig, J. V. Principles of agricultural chemistry with special reference to the late researches made in England. (Dowden, Hutchinson, & Ross, 1855).
48. Saito, M. a., Goepfert, T. J. & Ritt, J. T. Some thoughts on the concept of colimitation: Three definitions and the importance of bioavailability. *Limnol. Oceanogr.* **53**, 276–290 (2008).
49. Basu, S. & Shaked, Y. Mineral iron utilization by natural and cultured *Trichodesmium* and associated bacteria. *Limnol. Oceanogr.* **63**, 2307–2320 (2018).
50. Basu, S., Gledhill, M., de Beer, D., Prabhu Matondkar, S. G. & Shaked, Y. Colonies of marine cyanobacteria *Trichodesmium* interact with associated bacteria to acquire iron from dust. *Commun. Biol.* **2**, 1–8 (2019).
51. Lee, M. D. *et al.* Transcriptional activities of the microbial consortium living with the marine nitrogenfixing cyanobacterium *Trichodesmium* reveal potential roles in community-level nitrogen cycling. *Appl. Environ. Microbiol.* **84**, (2018).
52. Chappell, P. D. & Webb, E. A. A molecular assessment of the iron stress response in the two phylogenetic clades of *Trichodesmium*. *Environ. Microbiol.* **12**, 13–27 (2010).
53. Leventhal, G. E., Ackermann, M. & Schiessl, K. T. Why microbes secrete molecules to modify their environment: The case of iron-chelating siderophores. *J. R. Soc. Interface* **16**, (2019).
54. Yamaguchi, T., Furuya, K., Mitsihide Sato & Kazutaka Takahasi. Phosphate release due to excess alkaline phosphatase activity in *Trichodesmium* erythraeum. *Plankt. Benthos Res* **11**, 29–36 (2016).
55. Orcutt, K. M., Gundersen, K. & Ammerman, J. W. Intense ectoenzyme activities associated with *Trichodesmium* colonies in the Sargasso Sea. *Mar. Ecol. Prog. Ser.* **478**, 101–113 (2013).
56. Yentsch, C.M. Yentsch, C.S. and Perras, J.P. Alkaline phosphatase activity in the tropical marine blue-green alga *Trichodesmium*. *Limnol. Oceanog.* 17,772–774 (1970).
57. Hawser, S. P., O'Neil, J. M., Roman, M. R. & Codd, G. A. Toxicity of blooms of the cyanobacterium *Trichodesmium* to zooplankton. *J. Appl. Phycol.* **4**, 79–86 (1992).
58. Sheridan, C. C. The microbial and metazoan community associated with colonies of *Trichodesmium* spp.: a quantitative survey. *J. Plankton Res.* **24**, 913–922 (2002).
59. Eichner, M. *et al.* N_2 fixation in free-floating filaments of *Trichodesmium* is higher than in transiently suboxic colony microenvironments. *New Phytol.* **222**, 852–863 (2019).
60. Lee, M. D. *et al.* Transcriptional activities of the microbial consortium living with the marine nitrogen-fixing cyanobacterium *Trichodesmium* reveal potential roles in community-level nitrogen cycling. *Appl. Environ. Microbiol.* **84**, AEM.02026-17 (2017).
61. Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B. & Carpenter, E. J. *Trichodesmium*, a globally significant marine cyanobacterium. *Science (80-.).* **276**, 1221–1229 (1997).

62. Saito, M. Dorsk, A., Post, A., McIlvin, M., Rappe, M., DiTullio, G., Moran, D.Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome. *Proteomics* (2015).
63. Lu, X. & Zhu, H. Tube-Gel Digestion: A Novel Proteomic Approach for High Throughput Analysis of Membrane Proteins. *Mol Cell Proteomics* **4**, 1948–1958 (2005).
64. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
65. Strathman, R.R., Estimating Organic Carbon Content of Phytoplankton from Cell Volume or Plasma Volume. *Limnol. Oceanogr.* **12**, 411- (1967).
66. López, C. V. G. *et al.* Protein measurements of microalgal and cyanobacterial biomass. *Bioresour. Technol.* **101**, 7587–7591 (2010).
67. Martiny, A. C., Lomas, M. W., Fu, W., Boyd, P. W., Chen, Y. L., Cutter, G. A., Moore, J. K. Biogeochemical controls of surface ocean phosphate. *Science Advances*, *5* (2019).

3.8 SUPPLEMENTARY FIGURES

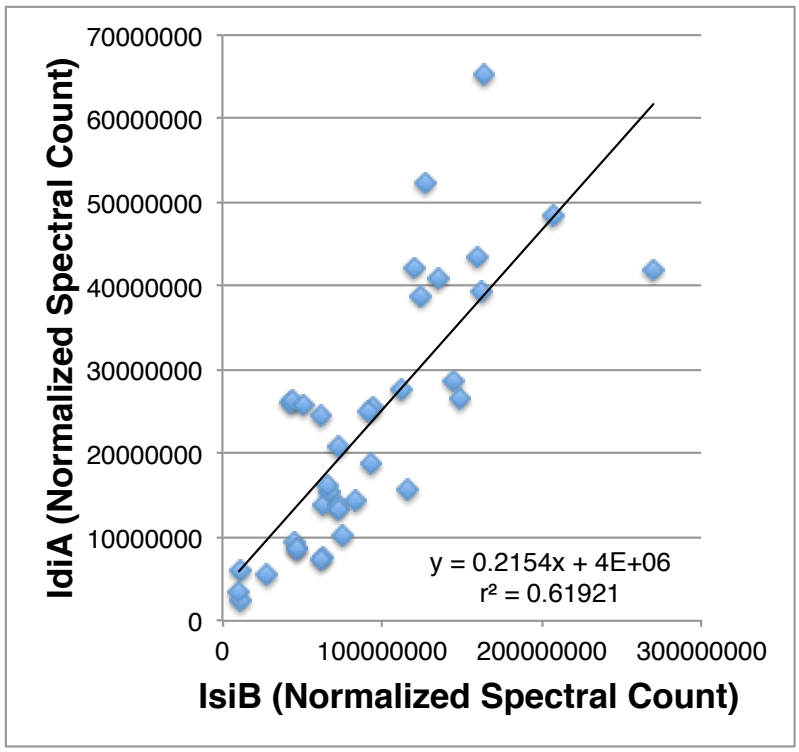


Figure S3.1. Relative abundance of IdiA and IsiB. Both are biomarkers of iron stress in *Trichodesmium*, and behave similarly in this dataset.

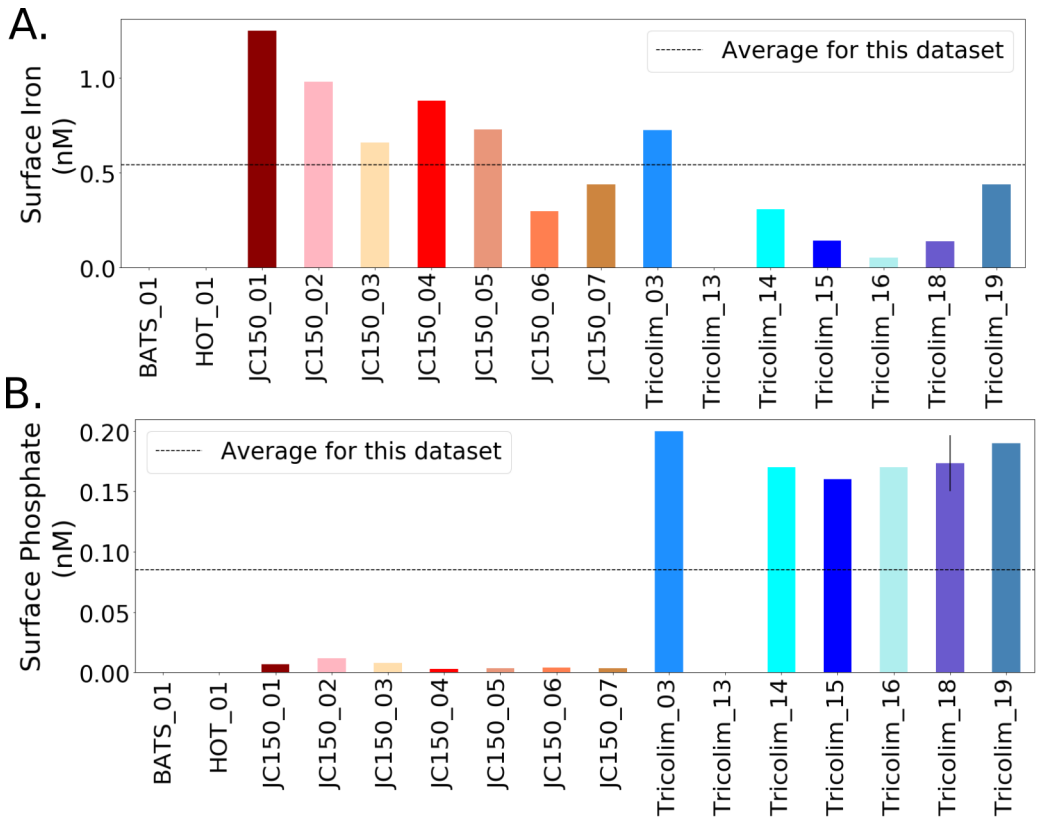


Figure S3.2. (A) Surface iron and (B) surface phosphate distributions at each location, where measured. Note nutrient data is lacking for BATS, HOT, and Tricolim_13. Dashed lines represent average values across the dataset. The phosphate data for the JC150 cruise has been previously published and was measured at the nm scale. Note that the phosphate concentrations from the Tricolim cruise were not measured at the nm scale and were below the detection limit, likely explaining the differences in phosphate concentrations.[67]

3.9 SUPPLEMENTARY TABLES (list)

TABLE S3.1. Important Fe and P stress biomarkers in *Trichodesmium*
TABLE S3.2. Sample provenance
TABLE S3.3. Relative protein abundance data (available at BCO-DMO)
TABLE S3.4. SOM cluster assignments for most abundant *Trichodesmium* proteins (available at BCO-DMO)
TABLE S3.5 Quantitative proteomics data for PstC transporter

Table S3.1. Important Fe and P stress biomarkers in *Trichodesmium*

Name	Nutrient sensed	T. erythraeum sp. IMS101 gene name	Gene name in metagenome used in this study (IMG ID 2821474806)
IdiA	Fe	Tery_3377	TCCM_0877.00000020
PstS	Phosphate	Tery_3537	TCCM_0018.00000050
SphX	Phosphate	Tery_3434	TCCM_0018.00000040

Table S3.2. Sample provenance information

Name	Replication	Cruise	Cruise station #	Time sampled	Date sampled	Net size (uM)	Filter size	Latitude	Longitude
JC150_03	triplicate	JC150	3	pre-dawn	7/10/17	200	0.2uM	22	-50
JC150_04	singlicate	JC150	4	pre-dawn	7/14/17	200	0.4uM	23.22	-44.47
JC150_05	triplicate	JC150	5	pre-dawn	7/19/17	200	0.4uM	23	-39.6
JC150_06	singlicate	JC150	6	pre-dawn	7/24/17	200	0.4uM	22.19	-35.53
JC150_07	triplicate	JC150	7	pre-dawn	7/29/17	200	0.4uM	22	-31
JC150_01	triplicate	JC150	1	pre-dawn	7/3/17	200	0.2uM	22	-58
JC150_02	duplicate	JC150	2	pre-dawn	7/4/17	200	0.2uM	21.26	-54
Tricolim_18	triplicate	Tricolim	18	pre-dawn	3/9/18	130	0.2uM	13.38	-55.5
Tricolim_15	triplicate	Tricolim	15	pre-dawn	3/5/18	130	0.2uM	5.0239	-44.1836
Tricolim_19	triplicate	Tricolim	19	pre-dawn	3/11/18	130	0.2uM	16.5	-57.5
Tricolim_14	duplicate	Tricolim	14	afternoon	3/1/18	130	0.2uM	2.36061	-39.591
Tricolim_15	singlicate	Tricolim	15	9	3/3/18	130	0.2uM	5.0239	-44.1836
Tricolim_03	duplicate	Tricolim	3	11:30	2/15/18	130	0.2uM	12.13	-21.59
Tricolim_16	duplicate	Tricolim	16	14:45	3/4/18	130	0.2uM	7.3	-48.29
Tricolim_13	singlicate	Tricolim	13	15:00	2/27/18	130	0.2uM	0.17976	-30.3984
HOT_01	duplicate	HOT 117	n/a	13:30	7/27/00	130	5uM	22.45	-158
BATS_01	singlicate	BATS	n/a	morning	2/4/15	200	0.2uM	31.4	-64.1

Table S3.5. Absolute abundance of the Pst protein and calculation of surface area occupied

Assuming 30% w/w cyanobacteria protein content (Gonzalez Lopez et al., 2010) - lower bound

Assume a cylindrical cell of height 15um, width 11um (Bergmann et al., 2013)

cell volume (um3)	1424.8
cell surface area (um	708.07

Assume average protein contains 0.53 g C / g protein (Rouwenhorst et al)

Assume ATP transporter has cross sectional area of 1.66e-5 um2 (Hudson and Morel 1992)

pg C per cell	pg C in protein	pg protein/cell	ug protein/cell
86.65	26.00	49.05	0.00005

from cell volume per Strathman 1967

Station	Avg [Pst] fmol/ug total protein	St Dev	fmol protein per cell	Pst molecules per cell	S.A. occupied per cell (um2)	% surface area occupied
Tricolim_18	13.00	1.80	0.0006	383994.82	25.34	3.58
Tricolim_15	11.22	3.45	0.0006	331540.40	21.88	3.09
Tricolim_16	89.06	123.06	0.0044	2630552.93	173.62	24.52
JC150_3	38.73	63.34	0.0019	1143789.85	75.49	10.66
JC150_4	89.58	14.74	0.0044	2645858.35	174.63	24.66
JC150_5	74.24	36.42	0.0036	2192888.10	144.73	20.44
JC150_6	61.64	40.07	0.0030	1820590.66	120.16	16.97
JC150_7	165.72		0.0081	4894655.41	323.05	45.62
JC150_1	106.08		0.0052	3133303.68	206.80	29.21
					average	19.86

Assuming 55% w/w cyanobacteria protein content (Gonzalez Lopez et al., 2010) - upper bound

Assume a cylindrical cell of height 15um, width 11um

pg C per cell	pg C in protein	pg protein/cell	ug protein/cell
86.65	47.66	89.92	0.00009

Station	Avg [Pst] fmol/ug total protein	St Dev	fmol protein per cell	Pst molecules per cell	S.A. occupied per cell (um2)	% surface area occupied
Tricolim_18	13.00	1.80	0.0012	703990.50	46.46	6.56
Tricolim_15	11.22	3.45	0.0010	607824.06	40.12	5.67
Tricolim_16	89.06	123.06	0.0080	4822680.37	318.30	44.95
JC150_3	38.73	63.34	0.0035	2096948.06	138.40	19.55
JC150_4	89.58	14.74	0.0081	4850740.31	320.15	45.21
JC150_5	74.24	36.42	0.0067	4020294.85	265.34	37.47
JC150_6	61.64	40.07	0.0055	3337749.54	220.29	31.11
JC150_7	165.72		0.0149	8973534.91	592.25	83.64
JC150_1	106.08		0.0095	5744390.09	379.13	53.54
					average	36.41

CHAPTER 4. Active processing of mineral particles by the colonial cyanobacterium *Trichodesmium*

4.1 ABSTRACT

The marine cyanobacterium *Trichodesmium* impacts global biogeochemistry by supplying fixed nitrogen to the oligotrophic ocean. This study describes heterogeneity in *Trichodesmium* colonies collected from a single plankton net, including associations with metal-enriched particles consistent with iron oxide and iron clay minerals. Metaproteome analysis of individual colonies revealed that iron, nickel, copper, zinc, and chemotaxis proteins were enriched when particles were present. The differential responses of the iron stress protein ferritin versus the iron transport protein IdiA indicated that *Trichodesmium* has a specific physiological response to particulate versus dissolved iron. This demonstrates that *Trichodesmium* modulates its proteome in response to minerals, and implies that particle interaction is an important ecological niche for colonies.

4.2 INTRODUCTION

Nitrogen fixation by the diazotrophic cyanobacterium *Trichodesmium* directly impacts primary production in the global ocean, increasing the net export of carbon that fuels marine food-webs at depth.[1,2] *Trichodesmium* is a mysterious organism exhibiting complex, multicellular-like behaviors.[3–5] In the ocean, *Trichodesmium* forms distinct colony morphologies including "puffs and "tufts" which do not necessarily correspond to species differences.[6] Specific epibiont populations, distinct from the surrounding seawater, are associated with these colonies.[7–9] There is uncertainty as to why *Trichodesmium* forms colonies in nature, but hypotheses include sequestration of the oxygen-sensitive nitrogenase enzyme, buoyancy regulation, and iron acquisition.[5,10,11] In particular, *Trichodesmium* colonies have a unique ability to access iron from dust particles that it captures and entrains into the colony center.[5,12–14] The exact mechanisms and impacts of particle utilization are not well understood, particularly in natural populations.

Small-scale "-omics" sampling provides key insights into the lives of marine microbes. Typically, molecular investigations on *Trichodesmium* have integrated signals from 50-100 colonies, or thousands of individual cells/filaments, masking heterogeneity in the physiology of individual colonies.[15–18] In the field, we know that such heterogeneity exists because of differences in colony morphology. Individual colonies likely experience unique local conditions owing to small-scale gradients in the aquatic environment. Illustrating this, single cell genomics studies have revealed significant micro-diversity within marine phytoplankton populations.[19,20–23]

In this study, we investigated morphological and physiological differences in *Trichodesmium* colonies at a single Caribbean Sea location. Under the microscope, we observed striking variation in colony morphology, including associations with auto-fluorescent particles hypothesized to be of mineral origin. Using a combination of μ-XRF based element mapping, μ-XANES spectroscopy, and single colony metaproteome measurements, we demonstrate that some *Trichodesmium* colonies actively process iron oxide and iron clay minerals.

4.3 RESULTS

4.3.1 Sampling and visualization of colonies

Trichodesmium colonies were collected from a single phytoplankton net sampled at 17:00 local time on March 11, 2018 at 65.22°W 17.02°N. This is a Caribbean Sea location near Puerto Rico and in the vicinity of the Orinoco river plume.[24] Both puff and tuft morphologies were present, though puffs dominated (see Figure 4.1). Surface phosphate concentrations were low (~0.2μM) as is typical in an oligotrophic ecosystem, while surface iron concentrations were relatively high (0.44nM), suggesting coastal or atmospheric inputs. The most abundant species at this location were *Trichodesmium theibautii* sp. V-I and *Trichodesmium tenue* sp. H94 (Eric Webb, personal communication). Thirty individual *Trichodesmium* colonies of mixed morphology were hand-picked and immediately examined by fluorescent microscopy with a DAPI (4',6-diamidino-2-phenylindole) long-pass filter (> 420nm). No reagent was added, so observed fluorescence was due to inherent properties of the colonies. From the microscopy images it was apparent that some colonies were associated with particulate matter (Figure 4.1 and Figure S4.1). The particles fluoresced in the visual light range, often appearing as yellow, red, or blue dots. Strikingly, colonies either had many such particles or none at all. In general the particles were concentrated in the center of puff type colonies, though they were also present in tufts.

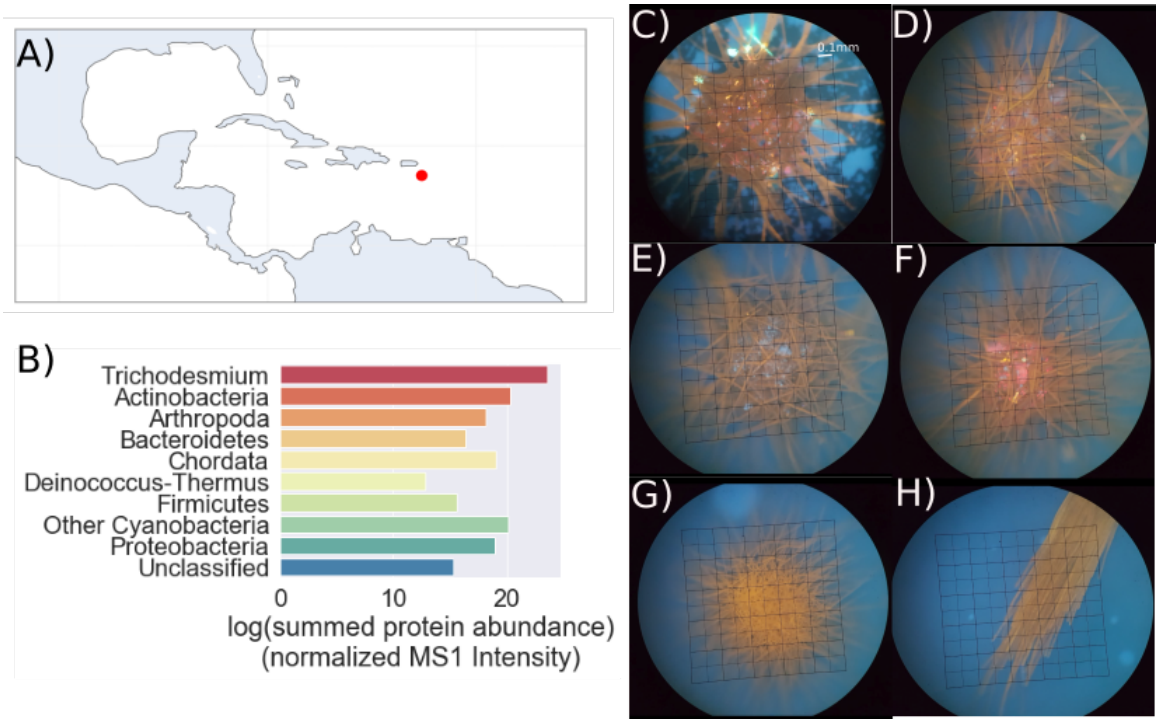


Figure 4.1. (A) Map of the location at which all of the following images and data were collected. (B) Taxonomic distribution of the proteins identified on the log scale (C-F) Representative images of puff colonies with particles, (G) puff without particles and (H) a tuft colony. Images were collected in epifluorescent mode using a DAPI long pass filter, but no dyes were used so all fluorescence was due to the inherent properties of the material. The particles are of different shapes and colors, which may indicate variance in character (i.e. minerals versus organic matter). The scale bar in panel (C) applies to all images.

4.3.2 Identification of metal rich mineral particles

Synchrotron based micro X-ray fluorescence (μ-XRF) element mapping of individual colonies demonstrated the presence of metal enriched particles associated with colony biomass. The particles were enriched in iron, copper, zinc, titanium/barium (which cannot be distinguished by this method), and manganese and cobalt, though the data was nosier for the latter two elements (Figure 4.2). Iron concentrations were particularly high in the particles, suggesting mineral origin such as atmospheric dust, which *Trichodesmium* is known to utilize.[5,12–14] Micro X-ray absorption near-edge structure (μ-XANES) spectra for Fe was conducted on six particles from two puff type colonies (Figure 4.2 and Figure S4.3). The particles contained mineral bound iron with iron oxidation states of 2.6, 2.7, two of oxidation state 2.9, and two of oxidation state 3.0. While the mineralogy of these particles could not be definitively resolved using μ-XANES, the post-edge region provided insight into broad mineral groups (Table S2). Both Fe(III) oxides and mixed valence Fe-bearing clays were present, suggesting

heterogeneity in mineral origin. Iron oxides are known to interact with *Trichodesmium* colonies, and this is the first evidence of iron clay interactions.[5,12,14,25] In this region, iron oxides could be sourced from atmospheric dust, and clays from the Orinoco river and/or island of Puerto Rico.

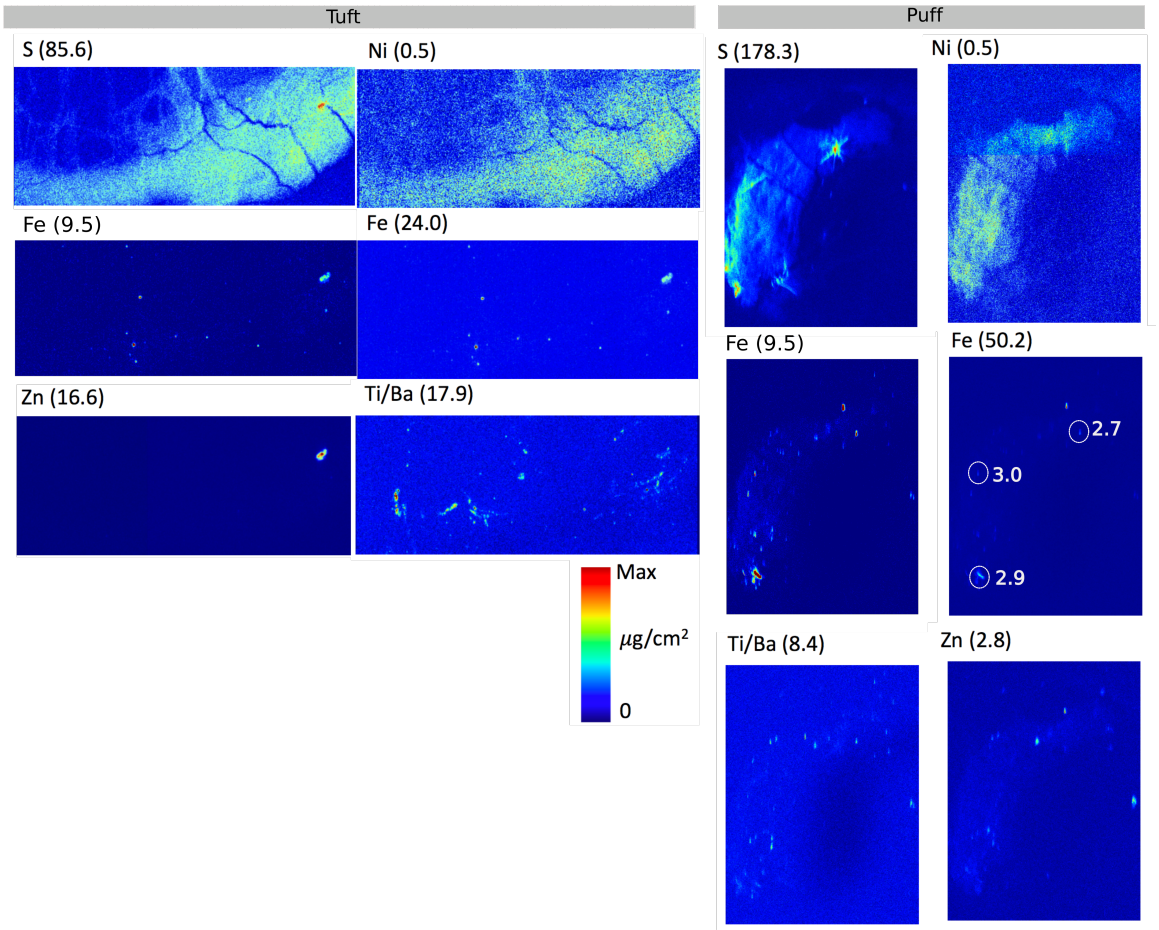


Figure 4.2. μXRF based element mapping of a *Trichodesmium* tuft (left) and puff (right) colony. Sulfur and nickel are associated with colony biomass, as are metal-enriched particles. The maximum of the color scale is provided next to each element, and iron is displayed twice with different scales to demonstrate the extremity of iron concentration in the biomass associated particles. Iron oxidation states were determined via μ-XANES for three particles in the puff colony, and their oxidation state is annotated.

4.3.3 Metaproteome analyses of individual colonies

To investigate the impact of the particles on *Trichodesmium* physiology, individual colonies were isolated for metaproteomic analysis. Despite starting with 10-100x less starting material, the complexity of the single colony metaproteomes were comparable to that of a more typical bulk metaproteome integrating 50-100 colonies (Figure S4.2). In total, 1107 *Trichodesmium* and 485 epibiont proteins were identified

across the 27 individual metaproteomes versus 2944 *Trichodesmium* and 1534 epibiont proteins in the bulk metaproteome (Tables S4.1 and S4.4).

Particle presence significantly impacted the proteomes of individual *Trichodesmium* colonies. In particular, puffs with particles were enriched for proteins involved in phosphate and amino acid transport, peptide and nickel transport, and mineral and organic ion transport, suggesting enhanced transport of elements and compounds across the cell membrane (Figure S4.5). Cell signaling proteins and two component sensory systems were also enriched when particles were present, and these likely regulate the proteome expression changes described above. Particle presence did not impact nitrogenase abundance, however the colonies were sampled at dusk when nitrogenase concentrations decrease, so if there was an effect, it may have been missed (Figure 3).

The data supported the hypothesis that tufts and puffs harbor distinct epibiont populations, but also suggested that the chemical environment, i.e. particle presence, is important.[9] Puffs with particle associations were associated with cyanobacteria including *Lyngbya*, *Cyanothece*, *Microcoeleus*, and *Pontibacter* (Figure 4.1B and Figure S4.6). *Trichodesmium* is known to associate with other cyanobacteria, for instance a puff-specific *Trichodesmium-Calothrix* association has been previously identified.[26] In such cases, the colony may provide a buoyant environment for other light-seeking organisms. Consistent with greater levels of cyanobacteria associations, puffs with particles were also enriched in epibiont proteins for carbon fixation and carbohydrate metabolism (Figure S4.6). A key function of *Trichodesmium* epibionts is to produce siderophores to assist in iron uptake from particulate sources.[12] For the most part, siderophore synthesis and uptake proteins were not identified in this dataset, likely because siderophore metabolism is poorly annotated, however there was some evidence of increased siderophore production by epibionts in the presence of particles, for instance a *Rhodococcus* arylsulfatase protein, which may sulfonate siderophores such as petrobactin, was significantly enriched (Figure 4.4D).[27,28]

All of the identified epibionts were known members of the *Trichodesmium* epibiont community, though we did not identify some commonly associated species such as *Alteromonas*.[29,7,8] The sequence database included *Alteromonas* species, but sample complexity and sequence heterogeneity may have resulted in fewer epibiont protein identifications. Importantly, however, a rarefaction curve of the epibiont species identified suggested that saturation of the population was reached based on the analytical workflow and sequence database used here (Figure S4.7). In general, puffs had more epibiont protein diversity at the phylogenetic and functional level than tufts, and in particular had more eukaryotic proteins including actins, tubulins, ATP synthases, and histone proteins (Figure 4.1B). Taxonomic attributions of these proteins appeared to be non-marine, however, the proteins more probably relate to zooplankton not present in the genomic database. Many of the eukaryote proteins we observed have homologs in the copepod model *Calanus finmarchicus* such as histone protein TCCM_0779.00001180 (BLAST hit to *C. finmarchus* histone protein, $E = 1e^{-4}$) and tubulin protein TCCM_0148.00002430 (BLAST hit to *C. finmarchus* alpha-tubulin, $E = 4e^{-23}$). We observed many copepods at this sampling location, which were associated particularly with puffs, even after seawater rinsing (e.g. Figure S4.6 inset). Copepods are known predators of *Trichodesmium*; certain species may even have specialized hooks for grabbing filaments and one species, *Macrosetella gracilis,* houses its eggs in

Trichodesmium filaments.[30,31] Copepods are phylogenetically diverse, and few example genomes exist, explaining the lack of coverage in metagenome annotations.[32] Based on these data, copepods may preferentially associate with puff type colonies.

The metaproteomes revealed heterogeneity in metal containing proteins, suggesting that the mineral particles were being actively processed by the colonies. These included nickel superoxide dismutase, an enzyme that may be involved in protecting nitrogenase from damage by molecular oxygen (Figure 4.4F). Nickel has been known to limit cell growth and nitrogen fixation in *Trichodesmium*, so provision of this metal may be a key benefit of the mineral particles.[33,34] Superoxide may be used to reduce Fe(III) to Fe(II) for uptake, therefore superoxide dismutase may be enriched to help balance the redox state of colonies when particles are present.[35] Other enriched metalloproteins included a zinc peptidase, used to break down proteins, and a copper chaperone (Figure 4.4G-H). The copper chaperone may be involved in metal detoxification when particles are present because *Trichodesmium* is very susceptible to elevated metal concentrations (J. Waterbury, personal communication).

Particle presence directly impacted the colony's iron status. Consistent with the high concentration of iron in the mineral particles, multiple iron proteins including the electron transport protein ferredoxin and the iron storage protein ferritin were significantly enriched when particles were associated with the colony (Figure 4.4A-B). The high abundance of ferritin implied that in addition to directly using the iron accessed from particulate sources, *Trichodesmium* was storing it for future use. Despite strong signals in ferritin and other proteins, the iron transport protein IdiA, which is responsive to dissolved iron concentrations and is therefore often used as a biomarker for iron stress, was unaffected by particle presence (Figure 4.4C).[36,37] This result does not conflict with current interpretations of IdiA activity, which focus on the protein's response to dissolved iron sources. It may in fact be advantageous for colonies to maintain high levels of IdiA when particles are present because the protein can assist in iron uptake in both the Fe^{3+} and Fe^{2+} states, which would be provided by iron oxide and iron clay minerals, respectively.[38] In addition to uptake via IdiA, *Trichodesmium* may also be able to access mineral derived iron via siderophores produced by the epibiont community, the production of which was enriched when particles were present as described earlier (see Figure S4).[36]

The chemotaxis protein CheY also responded positively to particle presence, implying that it might be involved in the entrainment of mineral particles. Microscopic observations have shown colonies have the ability to shuttle particles to the colony center on the order of minutes, and that this involves complex rotation, bending, stretching, and flipping movements of trichomes and whole colonies.[5] Consistent with this observations, the chemotaxis response regulator CheY, which regulates flagellar motor direction, was abundant only in puffs with particles, and not identified at all in puffs without particles or in tufts (Figure 4.4E). This implied that *Trichodesmium's* chemotaxis machinery is involved in particle entrainment, either by translocating particles along the trichome or through collective movement of the trichomes to push the particles to the colony center.

This study indicated that *Trichodesmium* behaves differently in response to dissolved versus particulate iron, and challenges the current model of iron homeostasis in *Trichodesmium*, where IdiA and ferritin abundance exist on an iron continuum, with negative and positive relationships to iron availability, respectively. Instead, it seems that

while dissolved iron is taken up by IdiA and used directly, particle-derived iron is taken up by IdiA and/or siderophore transport proteins, and is preferentially stored. The ability to store particulate iron when rich particulate matter is available could provide colonies with an advantage in ecosystems with patchy nutrient distributions. The ferric iron regulator (Fur) protein regulates cellular iron homeostasis and is thought to control both IdiA and ferritin.[39,40] However, since both IdiA and ferritin were high with particulate iron, this suggests that a different regulatory system is involved (Figure 4.4I). This system may interact with chemotaxis machinery to enhance particle entrainment. Marine bacteria, particularly oligotrophs such as *Trichodesmium*, are known to "network" their sensory systems in this way.[41]

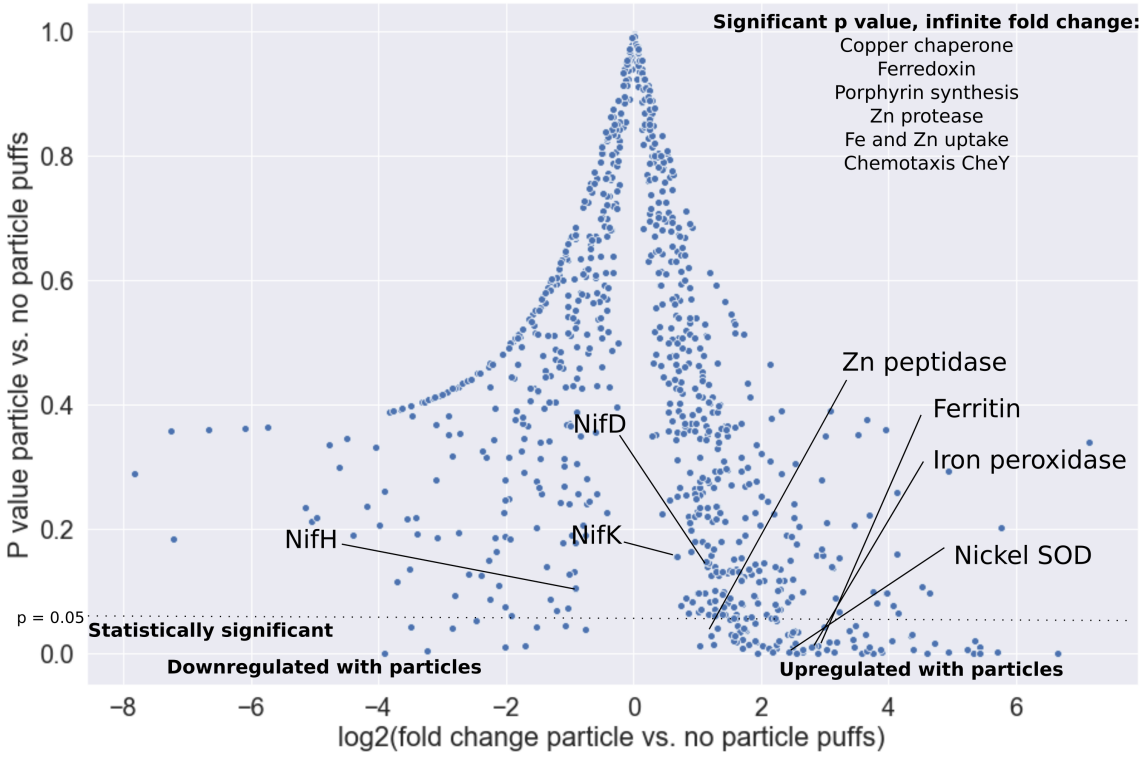


Figure 4.3. Volcano plot of the p value (two tailed t-test) versus the fold change of the average protein abundance for particle versus non-particle containing puffs, including both *Trichodesmium* and epibiont proteins. All proteins were identified at the 1% protein and peptide FDR levels. The dash line indicates the statistically significant ($p < 0.05$) value. Many metal containing proteins were significantly enriched when particles were present.

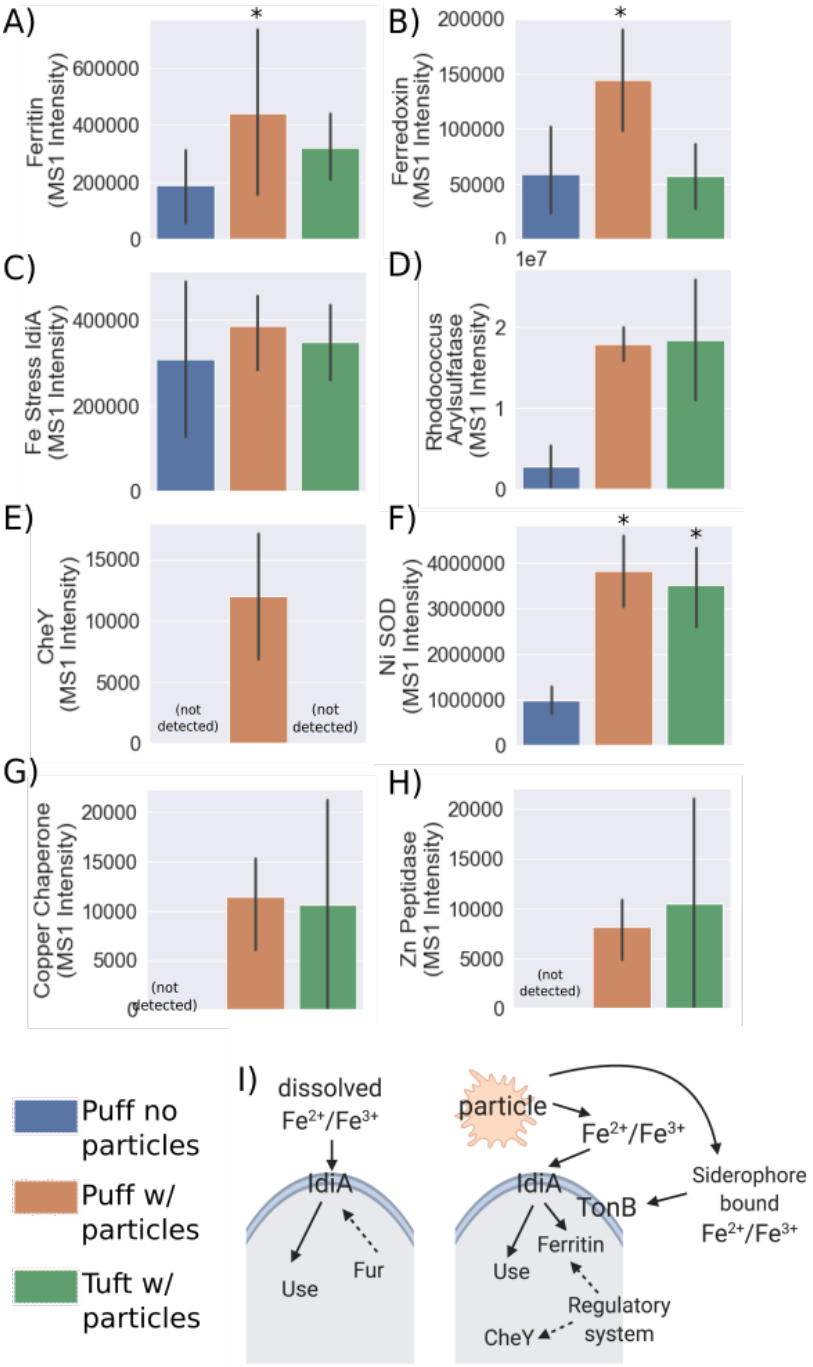


Figure 4.4. (A-H) Average relative abundance of metal containing proteins for the different morphology types. *Indicates statistically significant difference to the puffs without particles by a two-tailed t-test, $p < 0.05$. Error bars are one standard deviation of the mean. Note that even within these broad morphological classes, there is significant variation in protein abundance. (I) Model of dissolved and particulate iron uptake and its regulation based on results of this study.

4.4 CONCLUSION

Trichodesmium colonies clearly interact with mineral particles, including iron oxide and clay minerals. However, only some of the colonies at this location interacted with particles, perhaps due to stochastics of the particle – colony encounter. Another explanation could be species or sub-species level differences among the colonies such that some can access metals from the particles and others cannot. It will be possible to investigate this in the future as more *Trichodesmium* genomes are sequenced. Future studies can also focus on identifying the exact mechanisms by which *Trichodesmium* acquires and metabolizes mineral metals, a process that likely also involves the epibiont community.[13]

This study demonstrates that metal-enriched particles from both oxide and clay minerals are actively processed by *Trichodesmium* colonies, suggesting that *Trichodesmium* has many mechanisms for acquiring iron and other metals from the environment. Examining the colonies individually allowed observation of mineral-utilizing behavior that was masked in corresponding bulk samples, revealing differential responses to dissolved versus particulate iron that is likely relevant for modeling nitrogen fixation by *Trichodesmium* at the biogeochemical scale. This study reveals the unique and surprisingly complex physiology of this globally important organism, pointing to a new factor to be considered in future biogeochemical models.

4.5 MATERIALS AND METHODS

4.5.1 Sampling and microscopy

All of the samples in this study were taken from a single plankton net conducted at 65.22 W 17.02 N at 17:00 local time on March 11, 2018 on the Tricolim expedition (R.V. Atlantic Explorer, Chief Scientist D. Hutchins). A 200μm net was released to approximately 20m, then pulled back to surface and the process repeated five times. Colonies were hand picked, rinsed twice in 0.2μm filtered seawater, and stored in filtered seawater until imaging. Colonies were imaged with a Zeiss epifluorescent microscope using transmitted light and/or a DAPI long-pass fluorescent filter set. At the time of imaging they were labeled as "particle containing" or not and classified as puffs or tufts. They were then decanted individually onto a 0.2 μm Supor filter and flash frozen in liquid nitrogen until analysis. Images were captured with a Samsung Galaxy Note 4 using a SnapZoom universal digiscoping adapter.

4.5.2 Sample handling/small scale proteomics optimization

Upon return to the lab, the colonies were carefully cut out of the filter to reduce the volume of liquid needed for protein extraction. The filters were treated in PBS buffer with 10% sodium dodecyl sulfate (SDS), 1mM magnesium chloride, 2M urea and benzonase nuclease, heated at 95°C for 10min, then shaken at room temperature for one

hour. Proteins were quantified by the BCA assay. The proteins were digested with a modified tube gel protocol following Saito et al., 2014, but instead of the typical 200µL final volume only 50µL final volume was used.[42,43] Additionally, the protein precipitation/purification step was eliminated because this is another source of total protein loss. Instead, the samples were treated with benzonase nuclease to solubilize any DNA/RNA components, allowing the purification step to be skipped.The resulting peptide mixtures were concentrated to 0.2µg total protein/µL final concentration. While the tube gel method was used for samples presented here, magnetic bead and soluble protein digestion methods were also tested. Total protein recovery was lower with these methodologies, perhaps because these methods do not use SDS, which in our hands is a good lysing agent for *Trichodesmium*.

4.5.3 LC-MS/MS analysis

Metaproteome analyses were conducted by tandem mass spectrometry on a Thermo Orbitrap Fusion using 0.5µg total protein injections and a one-dimensional 120min non-linear gradient on a 15cm C18 (packed in house with 3µm beads with a picofrit tip) column. LC lines were shortened when possible to reduce sample loss. Blanks were run between each sample to avoid carryover effects. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD016330 and 10.6019/PXD016330.[46]

4.5.4 Bioinformatics analyses

The spectra were searched using the SEQUEST algorithm with a trimmed *Trichodesmium* sequence database. To generate the sequence database, triplicate bulk metaproteomes from the same location, each integrating ~50 colonies, were analyzed using a publically available metagenome from BATS (IMG ID 2821474806). Then, the sequence database was trimmed to include only the proteins identified at a 1% protein and peptide FDR level calculated with the Scaffold program (Proteome Software, Inc). This was used to search the single colony metaproteomes using the SEQUEST search engine. The results were statistically validated at the 1% FDR level using the Scaffold program. This resulted in 1592 protein identifications across the individual colonies. When the whole metagenome was used, only 800 proteins were identified at the 1% protein and peptide FDR level, so reducing the search space significantly improved data quality.

4.5.5 Micro-X-ray fluorescence and Micro-X-ray absorption spectroscopy

Micro-X-ray Fluoresence (µ-XRF) and micro-X-ray-absorption-spectroscopy (µ-XAS) were conducted at the Stanford Synchrotron Radiation Lightsource (SSRL) on beamline 2-3 with a 3 µm raster and a 50 ms dwell time on each pixel. µ-XRF data were analyzed using MicroAnalysis Toolkit.[44] Elemental concentrations were determined using standard foils containing each element of interest. The relative proportions of Fe(II) and

Fe(III) were determined by fitting the edge position of the background subtracted, normalized XANES spectra. Fe XANES spectra were fit using the SIXPACK Software package[45], and redox state was estimated by fitting the absorption edge (7115-7140 eV) with linear combination fitting of standard spectra using the model compounds siderite and 2-line ferrihydrite as endmember representatives of Fe(II) and Fe(III), respectively. Further, these values were confirmed through deconvolution of the edge shape using Gaussian peaks at two fixed energies corresponding to primary Fe(II) (7122 eV) and Fe(III) (7126 eV) contributions (PeakFit software, SeaSolve Inc.)[47]

Although mineral identity cannot be conclusively determined with XANES, visual comparison of the edge features are indicative of broad Fe-bearing mineral groups including many common oxides and silicate minerals. Thus, to get a general sense of mineral groups, LCF fitting was also conducted using 2-line ferrihydrite and goethite (as Fe oxide phases), ferrosmectite and nontronite (as Fe-bearing secondary clays), biotite (as a primary silicate), and siderite (as an Fe(II)-bearing carbonate). Linear combinations of the empirical model spectra were optimized where the only adjustable parameters were the fractions of each model comound contributing to the fit. The goodness of fit was established by minimization of the R-factor.[48,49]

4.6 ACKOWLEDGEMENTS

I thank my co-authors on this project, Kevin Sutherland, Matthew McIlvin, Eric Webb, Dave Hutchins, Colleen Hansel, and Mak Saito. I also thank the science and crew of the Tricolim cruise, especially Natalie Cohen, Michael Mazzota, Jaclyn Saunders and Asa Conover. This work was supported by NSF Graduate Research Fellowship grant # 1122274 [N.Held], the Gordon and Betty Moore Foundation (grant number 3782 [M.Saito]) the National Science Foundation (grant number 1657766 [M.Saito, E.Webb]), and the Woods Hole Oceanographic Institution Ocean Ventures Fund. Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

4.7 REFERENCES

1. Capone, D. G. *Trichodesmium*, a Globally Significant Marine Cyanobacterium. *Science (80-.).* **276**, 1221–1229 (1997).
2. Buchanan, P.J., Chase, Z., Matear, R. J., Phipps, S. J. & Bindoff, N.L. Marine nitrogen fixers mediate a low latitude pathway for atmospheric CO_2 drawdown. *Nat. Commun.* 1–10 (2019).
3. Walworth, N. G., Lee, M.D., Suffirdge, C., Qu, P., Fu, F., Saito, M.A., Webb, E.A., Sañudo-Wilhelmy, S.A., Hutchins, D.A. Functional Genomics and Phylogenetic Evidence Suggest Genus-Wide Cobalamin Production by the Globally Distributed Marine Nitrogen Fixer *Trichodesmium*. *Front. Microbiol.* **9**, 1–12 (2018).
4. Frischkorn, K. R., Haley, S. T. & Dyhrman, S. T. Coordinated gene expression between *Trichodesmium* and its microbiome over day–night cycles in the North Pacific Subtropical Gyre. *ISME J.* 1–11 (2018). doi:10.1038/s41396-017-0041-5
5. Rubin, M., Berman-Frank, I. & Shaked, Y. Dust-and mineral-iron utilization by

the marine dinitrogen-fixer *Trichodesmium*. *Nat. Geosci.* **4**, 529–534 (2011).
6. Orcutt, K. M. , Rasmussen, U., Webb, E.A., Waterbury, J.B., Gundersen K., Bergman, B. Characterization of *Trichodesium* spp. by Genetic Techniques. *App. Env. Microbiology* **68**, 2236–2245 (2002).
7. Frischkorn, K. R., Rouco, M., Mooy, B. A. S. Van & Dyhrman, S. T. Epibionts dominate metabolic functional potential of *Trichodesmium* colonies from the oligotrophic ocean. *ISME-J* 2090–2101 (2017).
8. Hmelo, L. R., Van Mooy, B. A. S. & Mincer, T. J. Characterization of bacterial epibionts on the cyanobacterium *Trichodesmium*. *Aquat. Microb. Ecol.* **67**, 1–14 (2012).
9. Rouco, M., Haley, S. T. & Dyhrman, S. T. Microbial diversity within the *Trichodesmium* holobiont. *Environ. Microbiol.* **18**, 5151–5160 (2016).
10. Saino, T. & Hattori, A. Aerobic nitrogen fixation by the marine non-heterocystous cyanobacterium *Trichodesmium* (Oscillatoria) spp.: Its protective mechanism against oxygen. *Mar. Biol.* **70**, 251–254 (1982).
11. Walsby, A. E. The properties and buoyancyproviding role of gas vacuoles in *Trichodesmium* ehrenberg. *Br. Phycol. J.* **13**, 103–116 (1978).
12. Basu, S., Gledhill, M., de Beer, D., Prabhu Matondkar, S. G. & Shaked, Y. Colonies of marine cyanobacteria *Trichodesmium* interact with associated bacteria to acquire iron from dust. *Commun. Biol.* **2**, 1–8 (2019).
13. Basu, S. & Shaked, Y. Mineral iron utilization by natural and cultured *Trichodesmium* and associated bacteria. *Limnol. Oceanogr.* **63**, 2307–2320 (2018).
14. Rueter, J.G., Hutchins, D.A., Smith, R.W., Unsworth, N. L. Iron nutrition of *Trichodesmium*. in *Marine Pelagic Cyanobacteria: Trichodesmium and other Diazotrophs* (ed. Carpenter, E.J., Capone, D.G., Rueter, J. G.) 289–306 (Kluwer Academic Publishers, 1992).
15. Frischkorn, K. R., Krupke, A., Guieu, C., Louis, J. & Rouco, M. *Trichodesmium* physiological ecology and phosphate reduction in the western tropical South Pacific. *Biogeosciences* **15**, 5761–5778 (2018).
16. Rouco, M., Frischkorn, K. R., Haley, S. T. & Alexander, H. Transcriptional patterns identify resource controls on the diazotroph *Trichodesmium* in the Atlantic and Pacific oceans. *ISME J*. 1486–1495 (2018).
17. Pfreundt, U., Kopf, M., Belkin, N., Berman-Frank, I. & Hess, W. R. The primary transcriptome of the marine diazotroph *Trichodesmium* erythraeum IMS101. *Sci. Rep.* **4**, 6187 (2014).
18. Lee, M. D. *et al*. Transcriptional activities of the microbial consortium living with the marine nitrogenfixing cyanobacterium *Trichodesmium* reveal potential roles in community-level nitrogen cycling. *Appl. Environ. Microbiol.* **84**, (2018).
19. Stocker, R. Marine Microbes See a Sea of Gradients. *Science (80-.).* **338**, 628–633 (2012).
20. Berube, P. M. *et al*. Data descriptor: Single cell genomes of *Prochlorococcus*, *Synechococcus*, and sympatric microbes from diverse marine environments. *Sci. Data* **5**, 1–11 (2018).
21. Stepanauskas, R. Single cell genomics: An individual look at microbes. *Curr. Opin. Microbiol.* **15**, 613–620 (2012).
22. Malmstrom, R. R. *et al*. Ecology of uncultured *Prochlorococcus* clades revealed

through single-cell genomics and biogeographic analysis*ISME J*. **7**, 184–98 (2013).
23. Kashtan, N. *et al*. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science (80-.)*. **344**, 416–420 (2014).
24. Corredor, J. E., Morell, J. M., Lopez, J. M., Capella, J. E. & Armstrong, R. A. Cyclonic eddy entrains orinoco river plume in eastern caribbean. *Eos (Washington. DC)*. **85**, 0–4 (2004).
25. Wu, J. F., Sunda, W., Boyle, E. A. & Karl, D. M. Phosphate depletion in the western North Atlantic Ocean. *Science (80-.)*. **289**, 759–762 (2000).
26. Momper, L. M., Reese, B. K., Carvalho, G., Lee, P. & Webb, E. A. A novel cohabitation between two diazotrophic cyanobacteria in the oligotrophic ocean. *ISME J*. **9**, 882–893 (2015).
27. Neilands, J. B. Siderophores: Structure and function of microbial iron transport compounds. *J. Biol. Chem*. **270**, 26723–26726 (1995).
28. Bosello, M. *et al*. Structural characterization of the heterobactin siderophores from *rhodococcus erythropolis* PR4 and elucidation of their biosynthetic machinery. *J. Nat. Prod*. **76**, 2282–2290 (2013).
29. Lee, M. D. *et al*. Transcriptional activities of the microbial consortium living with the marine nitrogen-fixing cyanobacterium *Trichodesmium* reveal potential roles in community-level nitrogen cycling. *Appl. Environ. Microbiol*. **84**, AEM.02026-17 (2017).
30. O'Neil, J. M. & Roman, M. R. Grazers and Associated Organisms of *Trichodesmium*. in *Marine Pelagic Cyanobacteria: Trichodesmium and other Diazotrophs* (eds. Carpenter, E. J., Capone, D. G. & Rueter, J. G.) 61–73 (Springer Netherlands, 1992).
31. Bron, J. E. *et al*. Observing copepods through a genomic lens. *Front. Zool*. **8**, 22 (2011).
32. Ho, T.-Y. Nickel limitation of nitrogen fixation in *Trichodesmium*. *Limnol. Oceanogr*. **58**, 112–120 (2013).
33. Mackey, K. R. M. *et al*. Phytoplankton responses to atmospheric metal deposition in the coastal and open-ocean Sargasso Sea. *Front. Microbiol*. **3**, 1–15 (2012).
34. Roe, K. L. & Barbeau, K. A. Uptake mechanisms for inorganic iron and ferric citrate in *Trichodesmium* erythraeum IMS101. *Metallomics* **6**, 2042–2051 (2014).
35. Webb, E. A., Jakuba, R. W., Moffett, J. W. & Dyhrman, S. T. Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol. Oceanogr*. **52**, 2221–2232 (2007).
36. Chappell, P. D. & Webb, E. A. A molecular assessment of the iron stress response in the two phylogenetic clades of *Trichodesmium*. *Environ. Microbiol*. **12**, 13–27 (2010).
37. Polyviou, D. *et al*. Structural and functional characterization of IdiA/FutA (Tery_3377), an iron-binding protein from the ocean diazotroph *Trichodesmium* erythraeum. *J. Biol. Chem*. **293**, 18099–18109 (2018).
38. Morrissey, J. & Bowler, C. Iron utilization in marine cyanobacteria and eukaryotic algae. *Front. Microbiol*. **3**, 1–13 (2012).
39. Webb, E. A., Moffett, J. W. & Waterbury, J. B. Iron Stress in Open-Ocean

Cyanobacteria *App. Envi. Microbiology*. **67**, 5444–5452 (2001).
40. Held, N. A., Mcilvin, M. R., Moran, D. M., Laub, M. T. & Saito, A. Unique Patterns and Biogeochemical Relevance of Two-Component Sensing in Marine Bacteria. *mSystems* 1–16 (2019).
41. Saito, M. A. *et al*. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **345**, 1173–7 (2014).
42. Lu, X. & Zhu, H. Tube-Gel Digestion: A Novel Proteomic Approach for High Throughput Analysis of Membrane Proteins. *Mol Cell Proteomics* **4**, 1948–1958 (2005).
43. Webb, S. M. The microAnalysis toolkit: X-ray fluorescence image processing software. *AIP Conf. Proc*. **1365**, 196–199 (2010).
44. Webb, S. M. SIXPack a Graphical User Interface for XAS Analysis Using IFEFFIT. *Phys. Scr*. 1011 (2005).
45. Saito, M. A. *et al*. Progress and Challenges in Ocean Metaproteomics and Proposed Best Practices for Data Sharing. *J. Proteome Res*. **18**, 1461–1476 (2019).
46. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, Pérez E, Uszkoreit J, Pfeuffer J, Sachsenberg T, Yilmaz S, Tiwary S, Cox J, Audain E, Walzer M, Jarnuczak AF, Ternent T, Brazma A, Vizcaíno JA The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* 47 (2019)
47. Grabb, K.C., C. Buchwald, C.M. Hansel, S.D. Wankel. 2017. A dual nitrite isotopic investigation of chemodenitrification by mineral associated Fe(II) and its production of nitrous oxide. *Geochimica et Cosmochimica Acta* 196:388-402.
48. Newville M. (2001) EXAFS analysis using FEFF and FEFFIT. J. Synchr. Rad. 8, 96–100.
49. Hansel, C.M., S.G. Benner, J. Neiss, A. Dohnalkova, R.K. Kukkadapu, and S. Fendorf. 2003. Secondary mineralization pathways induced by dissimilatory iron reduction of ferrihydrite under advective flow. *Geochimica et Cosmochimica Acta* 67, 2977-2992.

4.8 SUPPLEMENTAL FIGURES

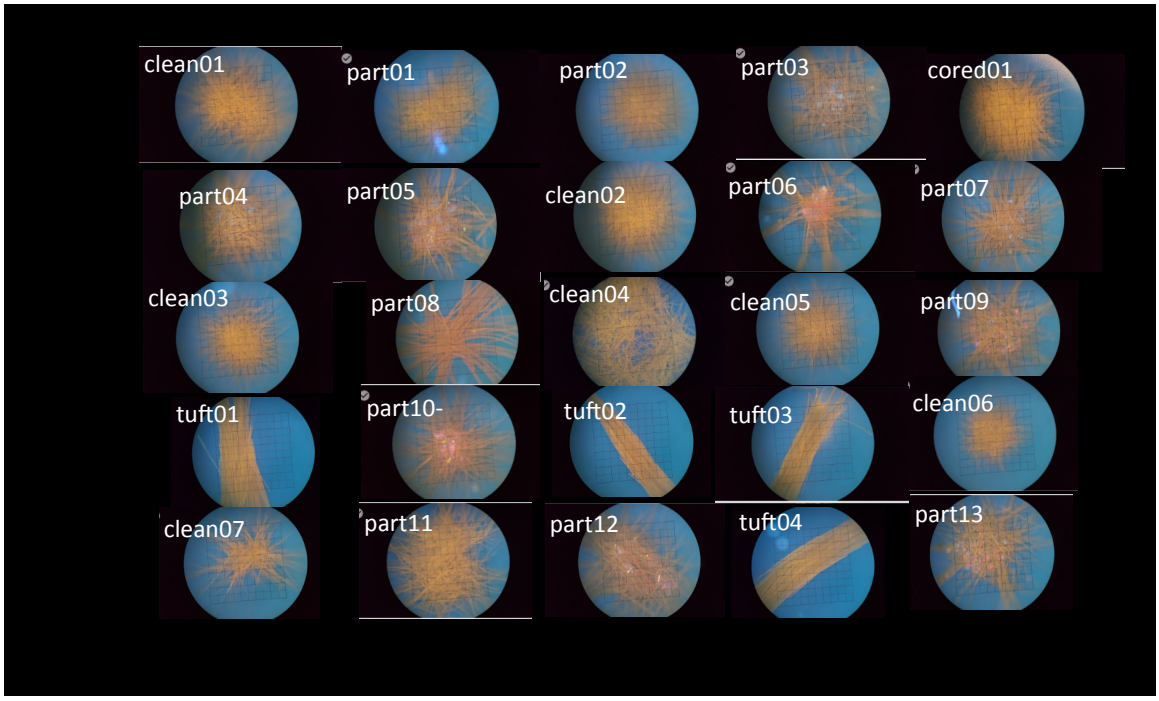


Figure S4.1. DAPI long pass filter images for all of the colonies examined in this study.

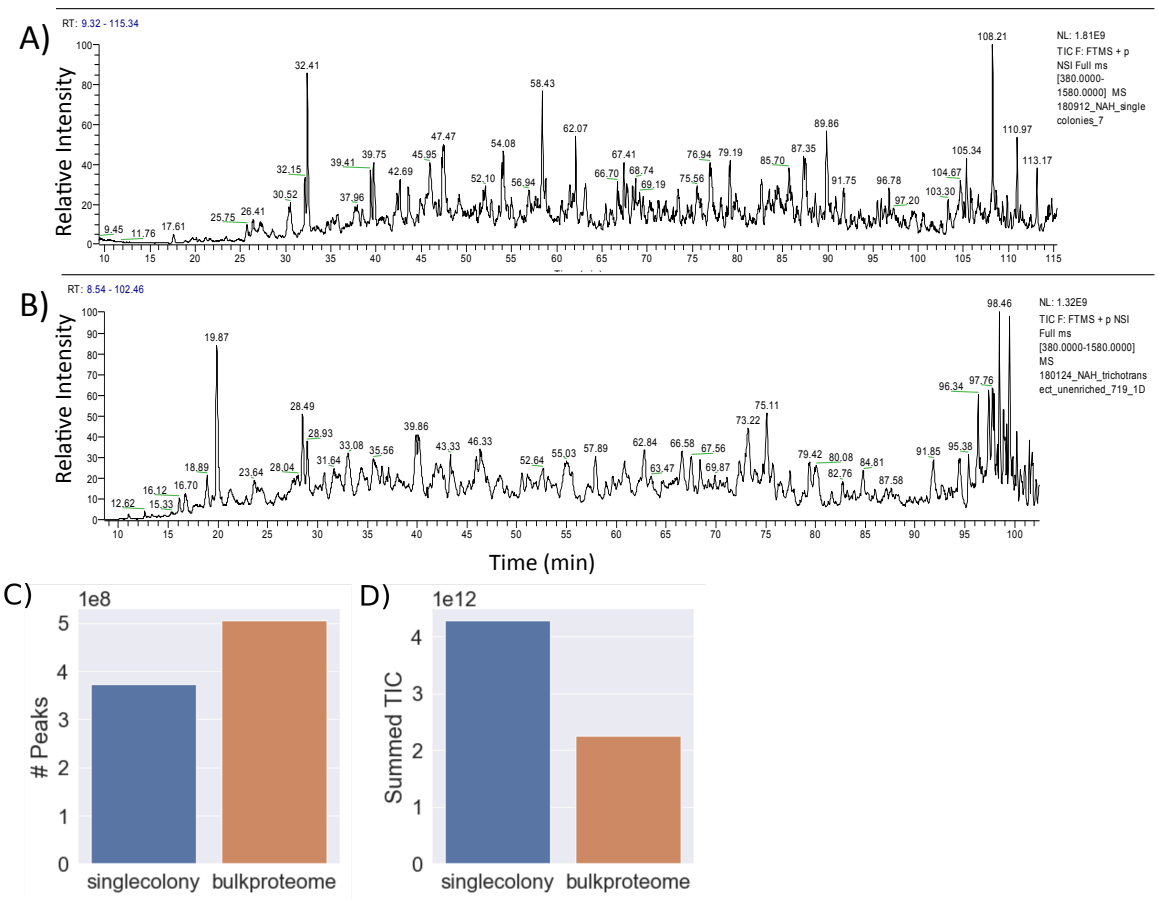


Figure S4.2. Chromatographic traces for (A) a single colony and (B) bulk *Trichodesmium* metaproteome from the same station. (C) number of peaks identified in the spectrum and (D) summed total ion current (TIC) for the two samples. The single colony metaproteome has higher total intensity and fewer identified peaks. Peak counting was performed following Saito et al., 2018.[44] However, it is still very complex illustrating the challenge in acquiring high quality data from a small amount of complex sample.

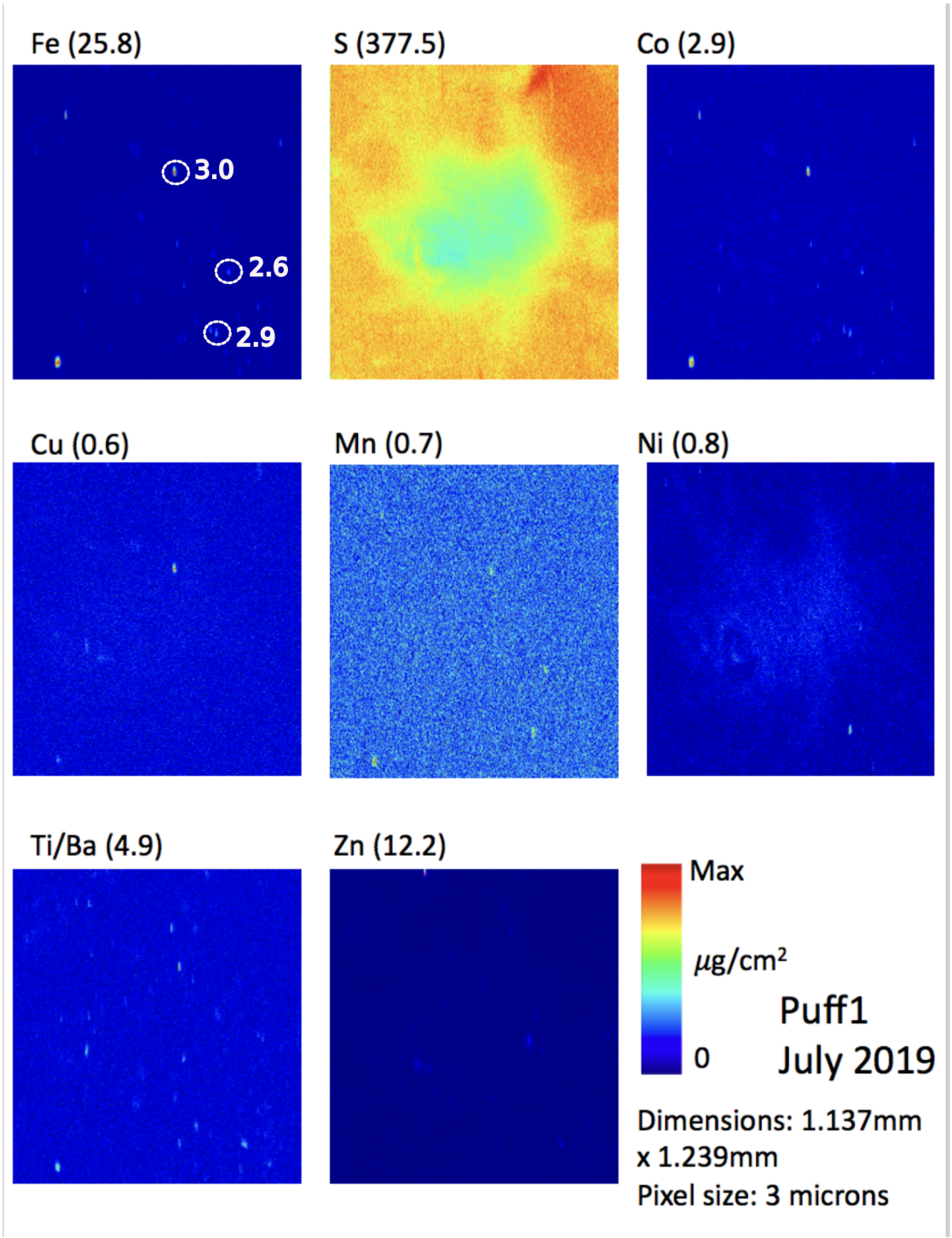


Figure S4.3. uXRF based element mapping of a second *Trichodesmium* puff colony. Iron oxidation states were determined for three particles and are annotated in the iron panel.

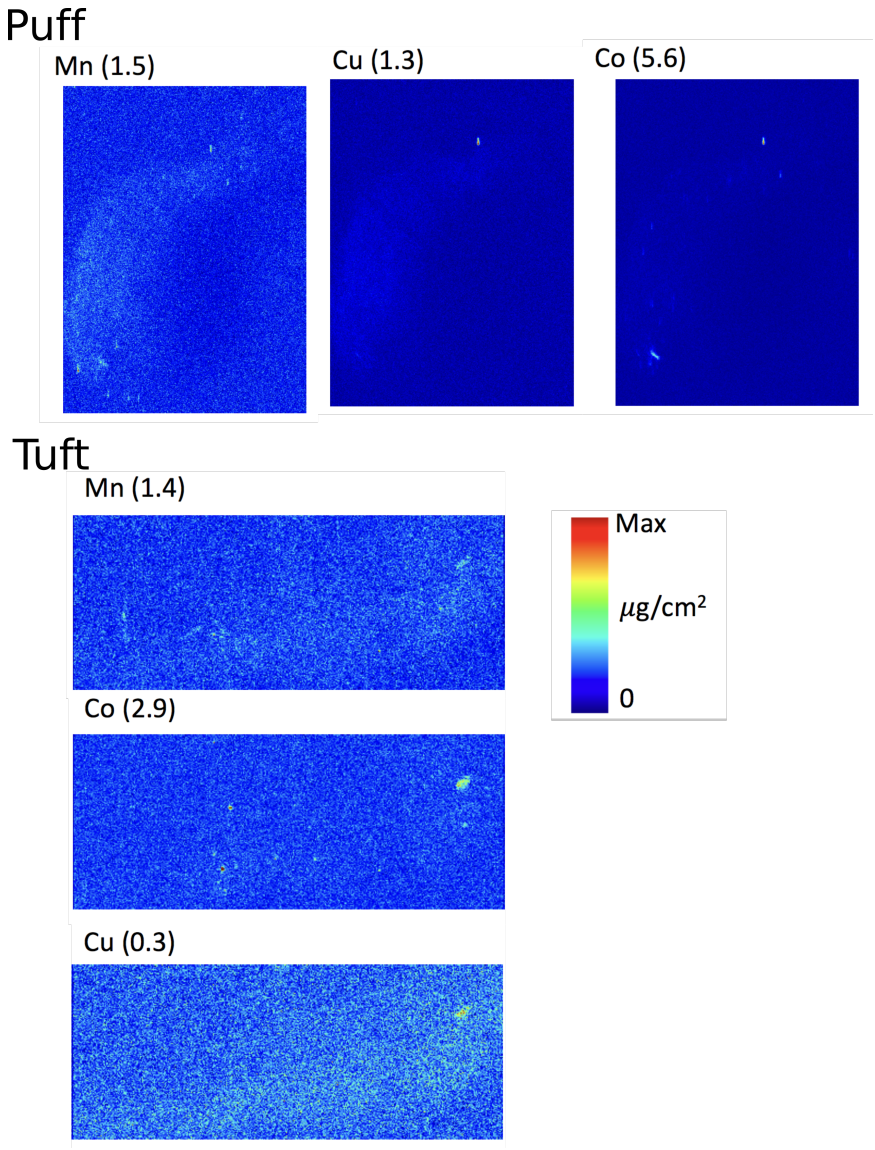


Figure S4.4. Additional uXRF element maps of the *Trichodesmium* tuft and puff colony featured in Figures 8 and 9, respectively. The maximum of the color scale is given next to each element.

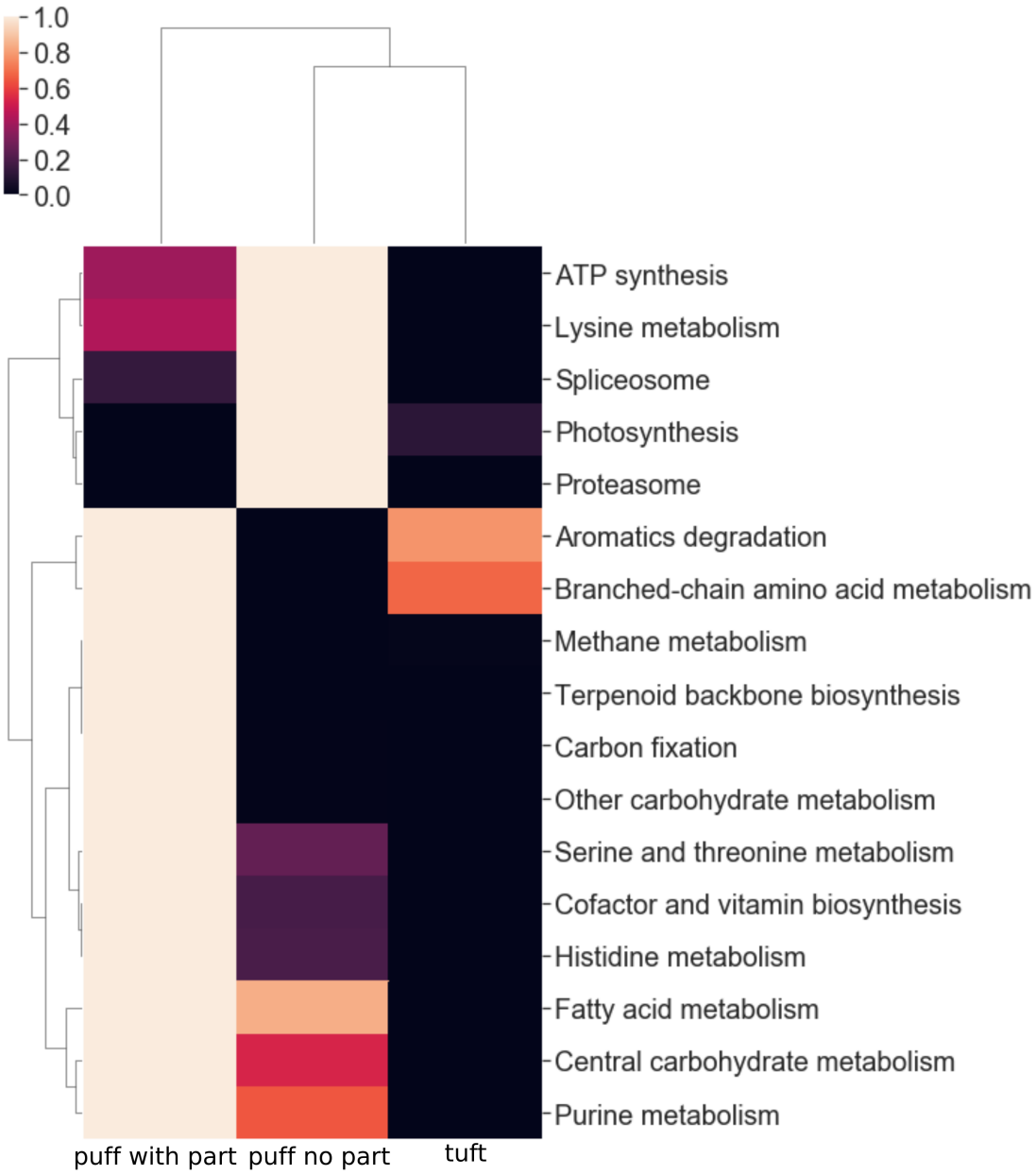


Figure S4.5. Clustered heat map of the abundance of each Kegg ontology (KO) module for the different morphology types. Abundances are the sum of the normalized protein abundances assigned to the KO module and are unit-normalized across the row. The different morphologies had similar proteomes overall.

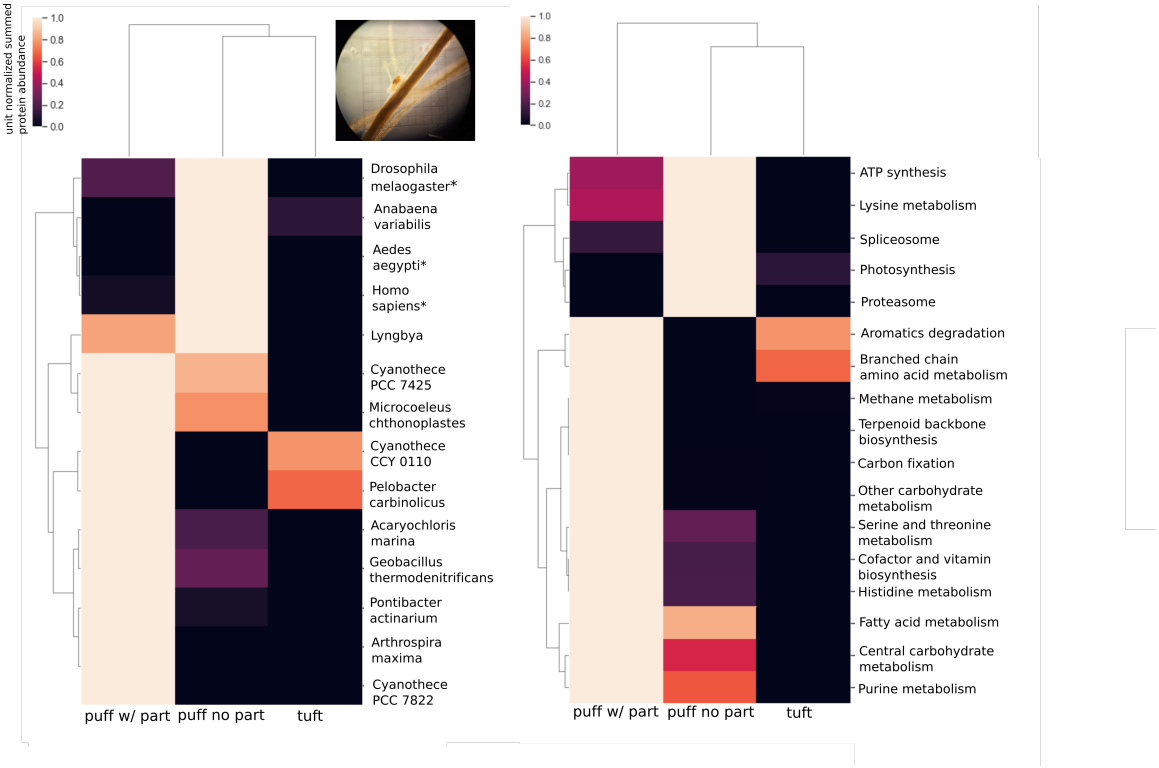


Figure S4.6. Clustered heatmap of summed relative protein abundances of the phylogenetic and Kegg ontology (KO) functional modules of epibiont proteins, classified by colony morphology. Summed protein abundances have been unit normalized across the row.

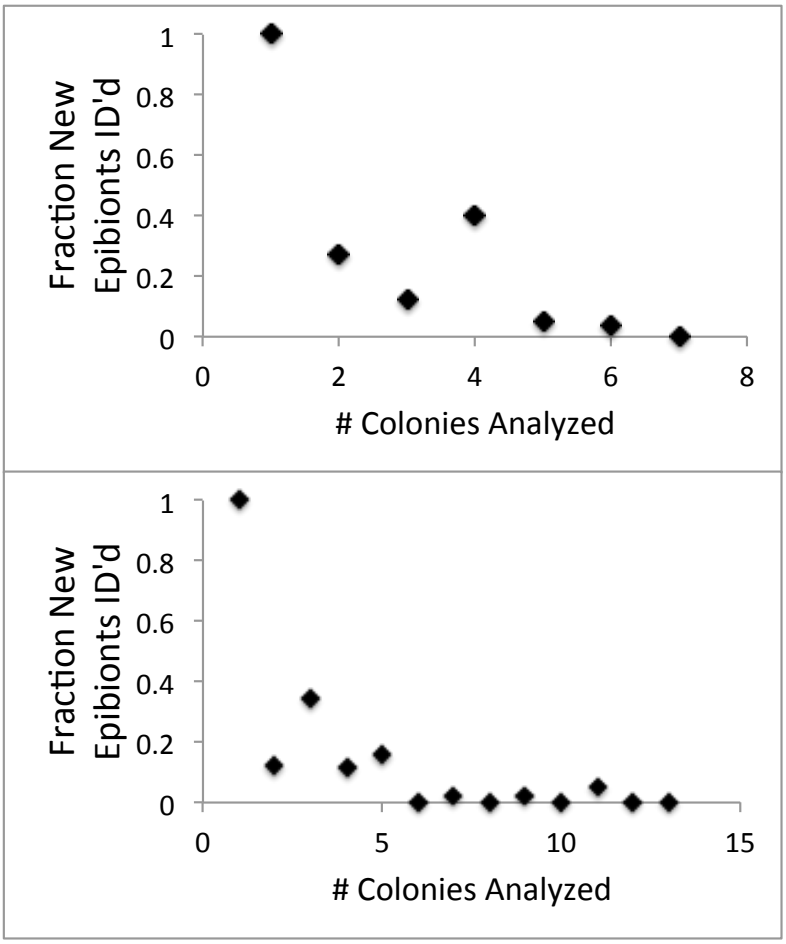


Figure S4.7. Rarefaction curve of the epibiont species identified in the individual colony metaproteomes for colonies without (top) and with (bottom) particle associations. In both cases saturation of the epibiont community was reached after just a few colonies were analyzed, indicating complete coverage of the species diversity that the analytical workflow and sequence database allows.

4.9 SUPPLEMENTAL TABLES (LIST)

Table S4.1. Protein identifications and relative quantitation data (available at BCO DMO)
Table S4.2. XANES based iron oxidation state and mineralogy data for selected particles
Table S4.3. Significant p-values for puffs with and without particles (available at BCO DMO)
Table S4.4. Metaproteome data from the bulk population at this location (available at BCO DMO)

Table S4.2 XANES iron oxidation state and mineralogy data for selected particles

puff 1	siderite_fit (Fe(II))	ferrihydrite_fit (Fe(III))	calc ox. state	comments
Fe_NH_puff1_1	0	1.011	3	
Fe_NH_puff1_2	0.127	0.8626	2.871665319	
Fe_NH_puff1_3	0.4104	0.5406	2.568454259	iron clay/magnetite like
puff 5				
Fe_puff5_1	0.0833	0.8687	2.9125	likely smectite
Fe_puff5_2	0	0.9414	3	ferrosmectite and oxide mixture
Fe_puff5_3	0.2869	0.6331	2.688152174	illite clay and oxide mixture

CHAPTER 5. Regulation of daytime nitrogen fixation, and associated buoyancy benefits, in *Trichodesmium erythraeum* sp. IMS101

5.1 ABSTRACT

Trichodesmium is a globally important marine nitrogen fixer, providing substantial amounts of fixed nitrogen to the oligotrophic ocean. *Trichodesmium* is enigmatic because it fixes both carbon and nitrogen simultaneously during the photoperiod despite the incompatibility of the nitrogenase enzyme with oxygen produced in photosynthesis. Its ecological success suggests there is a physiological benefit to this strategy. In this study, we investigate the diel proteome of *Trichodesmium erythraeum* sp. IMS101 to identify why and how it fixes nitrogen during the day. The proteome varies significantly within one hour increments, indicating tight coordination. Specific observations include changes in photosynthetic efficiency due to cyclic degradation of the phycobilisome proteins, which may protect nitrogenase from molecular oxygen. Regulatory proteins including P-II and RpaA are likely involved in regulating these processes. Evidence suggests that daytime nitrogen fixation occurs in support of *Trichodesmium's* unique vertical migration patterns, a key nutrient acquisition strategy. For much of the day, solar energy is directly supplied to the nitrogenase enzyme, instead of being stored as glycogen for later use. This minimizes cellular ballast, helping large filaments/colonies to remain at the surface. In the evening, perturbations in the cellular C:N ratio stimulate a spike in photosystem protein abundance which is concurrent with a spike in glycogen production one hour later, providing ballast for sinking motility at night. Buoyancy is regained as these storage compounds are consumed, assisted by gas vesicle production. Reflecting these processes, cellular POC and PON content can be predicted from just a handful of protein biomarkers, indicating the potential utility of global proteomic data for obtaining biogeochemical parameters. Together, this study highlights how a temporally dynamic proteome contributes to the complex lifestyle of this key marine cyanobacterium.

5.2 INTRODUCTION

The abundant marine cyanobacterium *Trichodesmium* sp. fixes atmospheric nitrogen, thereby providing a substantial source of N to otherwise limited ecosystems.[1–4] This input stimulates primary production and therefore impacts global carbon and nitrogen cycles. Understanding controls on nitrogen fixation by *Trichodesmium* is thus crucial for marine biogeochemical models.[2,5–8] *Trichodesmium* resides in tropical and subtropical surface waters, where it experiences changes in light, temperature, and nutrient availability over the course of the day. Like other cyanobacteria, *Trichodesmium* coordinates its cellular processes to make best use of these natural rhythms.[9,10]

One unique pattern in *Trichodesmium* is that it fixes both carbon and nitrogen during the light period. This is perplexing because the nitrogenase enzyme is susceptible to damage by molecular oxygen, which is produced during photosynthesis. Diazotrophs have evolved two strategies for solving these problems – some, like *Crocosphaera*, separate the processes temporally, fixing carbon during the day and nitrogen at night.[11,12] Others, like *Anabaena*, they separate the processes spatially, forming differentiated heterocyst cells that lack the oxygen evolving photosystem II complex.[13,14]

Trichodesmium is the exception, fixing carbon and nitrogen during the photoperiod within the same cell.

It has been speculated based on photosynthetic efficiency measurements that fine-tuned separation of nitrogen and carbon fixation occurs during the photoperiod, and that this may protect nitrogenase from molecular oxygen.[15,16] Spatial segregation, specifically formation of partially differentiated, nitrogen fixing "diazocytes" has also been suggested, but the existence and importance of this is uncertain. The first suggestion of spatial separation arose from observations of low-pigment cells at the trichome center. It was hypothesized that centrally located cells may experience lower oxygen concentrations and could therefore supply the filament/colony with fixed nitrogen.[17] However, it was later found that colony formation is not a prerequisite for nitrogen fixation.[18,19] Following this were observations of concentrated nitrogenase enzyme in certain cells, coined diazocytes, however these observations suffered from high background autofluorescence of the *Trichodesmium* cells, and an attempt to reproduce the experiment with a protocol optimized to reduce autofluorescence found nitrogenase to be equally distributed among the cells, contradicting the prior interpretation. [15,20–22]

Additional evidence against the spatial separation theory includes the lack of differential ^{13}C and ^{15}N uptake along the trichome[23], and lack of a plausible mechanism for transporting the fixed nitrogen from the "diazocytes" to neighboring cells. Specifically, in true heterocyst forming diazotrophs nitrogen is stored as cyanophycin which is concentrated at the heterocyst poles; actual nitrogen transfer occurs by exchange of glutamate-glutamine.[14,24,25] No corollory has been observed in *Trichodesmium*; instead, it has been suggested that "diazocytes" release fixed nitrogen to the environment.[3,26] This would be a decidedely risky strategy and has never been directly demonstrated. Based upon this evidence, this study works from the assumption that spatial segregation is not the main strategy for simultaneous N and C fixation in *Trichodesmium*.

This study seeks to identify the reasons for and mechanisms behind simultaneous N and C fixation in *Trichodesmium* cells. Given that *Trichodesmium* is ecologically successful and widely distributed in the marine enviornment, it assumes that the benefits of daytime nitrogen fixation outweigh the costs, and seeks to identify not only how but also why daytime nitrogen fixation occurs. Towards this end, we present a proteomic dataset of a model species, *Trichodesmium erythraeum* IMS101 with high temporal resolution over the diel cycle. Sampling occurred approximately every 1-2 hours, concentrated around dawn and dusk. We examine the overall dynamics of the proteome and focus on regulatory mechanisms coordinating nitrogen and carbon fixation on the diel cycle, finally concluding that that a key benefit of daytime nitrogen fixation is maintenance of cellular buoyancy.

5.3 RESULTS and DISCUSSION

5.3.1 The Trichodesmium proteome is dynamic over the diel cycle

The proteome changed dramatically over the course of the diel cycle. Fourteen samples were analyzed with 1-3 hour resolution between each sampling time point. Between them 2389 proteins were identified representing approximately 46% of the protein encoding genome of *T. erythraeum*. 562 proteins had significant periodicity based on a RAIN rhythmicity test ($p < 0.05$). Protein abundances varied up to 14 fold, and each hour hundreds of distinct proteins reached their maxima (Figure 1 and Figure S1). These fluctuations were surprising and extreme, suggested that *T. erythraeum* extensively coordinates its physiology on the time scale of one hour or less, and may explain why *T. erythraeum* has more regulatory proteins than other marine cyanobacteria.[27]

There were obvious distinctions in major processes occurring in the day versus the night. A weighted correlation network (WGCNA) analysis revealed groups of proteins that had similar abundance proteins throughout the diel cycle. During the photoperiod, *T. erythraeum* devoted much of its proteome to carbon and nitrogen fixation, as expected. Daytime activity was preceded in the early morning hours by proteins involved in vitamin, co-factor, nitrogenase, and photosystem synthesis. At night, *T. erythraeum* invested in reproduction as evidenced by increases in ribosomal proteins, cell division protein FtsZ, and peptidoglycan synthesis, corroborating prior reports that division occurs at night (see Figure S2).[28] Dawn and dusk were marked periods of transition, characterized by enrichment in regulatory proteins such as the light sensing NblS-SasA two component system, the nitrogen regulator NtcA, and chemotaxis regulators.

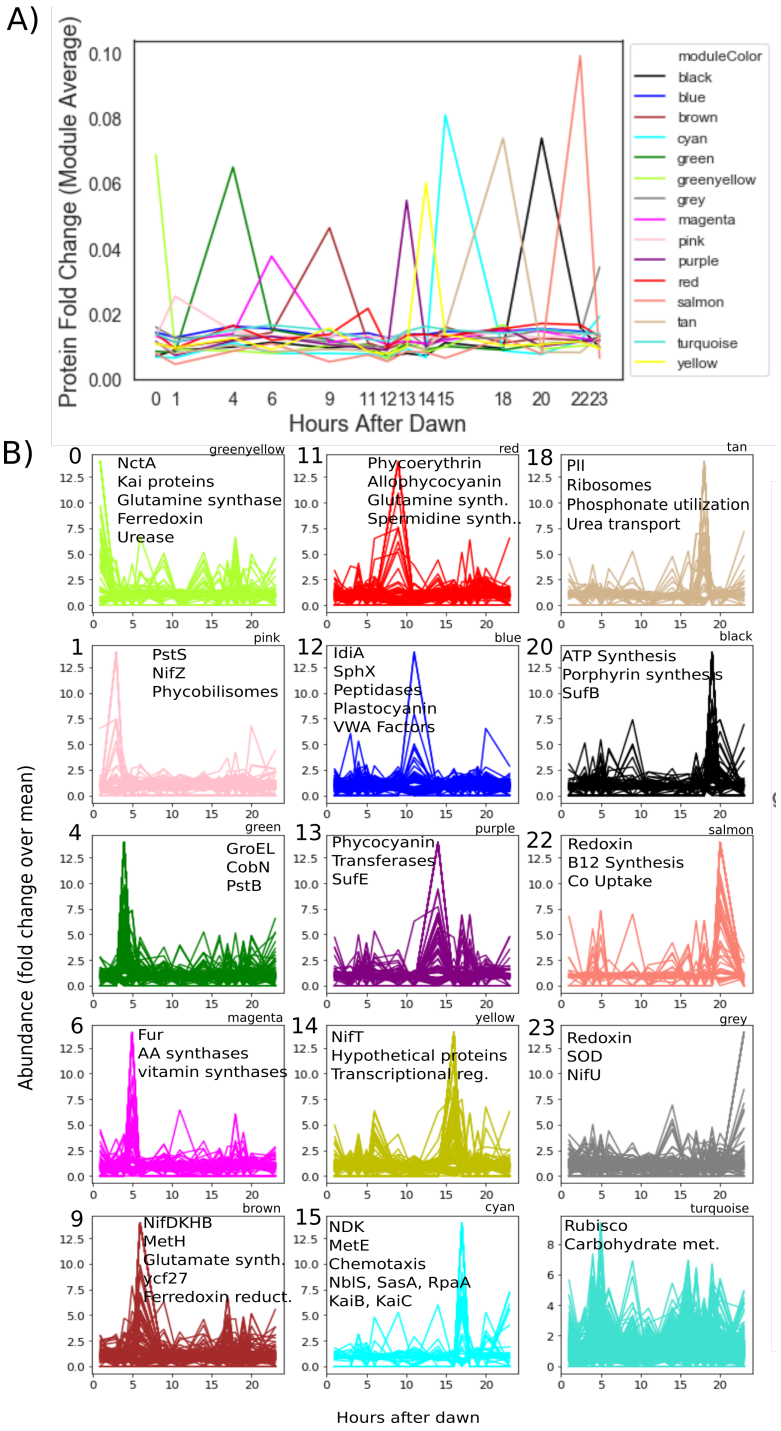


Figure 5.1. (A) WGCNA based overview of the *Trichodesmium* proteome. Fifteen modules were identified, all but one corresponding to a specific sampling time. (B) Each module displayed individually. Proteins of particular interest have been annotated in their respective plots, and assignments for every identified protein can be found in Table S2.

5.3.2 Oscillations in phycobilisome and nitrogenase protein abundance

Phycobilisome protein abundance cycled multiple times during the day. During sample preparation there were obvious changes in cell lysate coloring, which varied from reddish orange to pink indicative of changes in phycoerythrin, the phycobilisome pigment protein that gives *Trichodesmium erythraeum* sp. IMS101 its characteristic red color (Figure 2A). Confirming this, phycobilisome protein abundance oscillated over the course of the day, peaking in the early morning and afternoon (Figure 2B). Oscillations continued into the night, suggesting that they were under the control of a regulator with an inherent circadian rhythm as opposed to one that is responsive specifically to light. Indeed, the master circadian regulator RpaA oscillated in tandem with the phycobilisome rod proteins, and was positively correlated with phycobilisome abundance. This oscillating pattern appears to be unique to *Trichodesmium*, as in other cyanobacteria RpaA is activated only once during the diel cycle.[9,10] Degradation of the phycobilisome rods was likely mediated by abundant proteases such as Tery_1247, which displayed a similar circadian pattern to RpaA and the phycobilisome rod proteins (Figure S3). The proteomic analysis thus provided molecular evidence that changes in photosynthetic efficiency brought on by oscillations in the phycobilisome rod structures protect nitrogenase from destruction by molecular oxygen.

In contrast to the phycobilisome proteins, nitrogenase abundance cycled just once during the day, increasing in the morning and decreasing at dusk (Figure 2E). Importantly, nitrogenase abundance never reached zero even at night when nitrogen fixation does not occur, suggesting there was a second control on nitrogen fixation independent of protein concentration. This could include a post-translational modification of the nitrogenase enzyme as has been previously suggested.[29,30] The temporal persistence of nitrogenase implies that iron involved in N_2 fixation remains bound throughout the day, assuming apo-nitrogenase (the protein without the metal cofactor) does not occur. This differs from the marine diazotroph *Crocosphaera*, which shares its iron between its photosynthesis and nitrogenase proteins, thus significantly reducing its iron demand.[31]

In diazotrophs and non-diazotrophs alike, the regulatory proteins P-II and NtcA monitor the C:N ratio by sensing 2-oxoglutatate, a key intermediate in the TCA cycle and an ingredient for glutamate production in the GS-GOGAT pathway.[32–34] This suggests that nitrogen regulation is generally pointed inward, i.e. is focused on sensing nitrogen demand rather than environmental concentrations. In non-diazotrophs such as *Synechocystis*, P-II and NtcA respond to changes in external nitrogen concentrations, however this does not occur in *Trichodesmium*.[33,35] In the diel proteomes, P-II was negatively correlated with nitrogenase abundance (Figure 2 F, G), while NtcA was not correlated at all. This indicated that P-II either controls or is controlled by nitrogen fixation. This contrasts with recent field metaproteomes where both P-II and NtcA were positively correlated with nitrogenase abundance (see Chapter 3). It indicates that it is likely that a second, currently unidentified, control on nitrogen fixation occurs over the diel cycle, which may be related either to circadian or light rhythms. Such regulation would be consistent with the single oscillation in nitrogenase abundance as opposed to multiple oscillations in phycobilisome protein abundance.

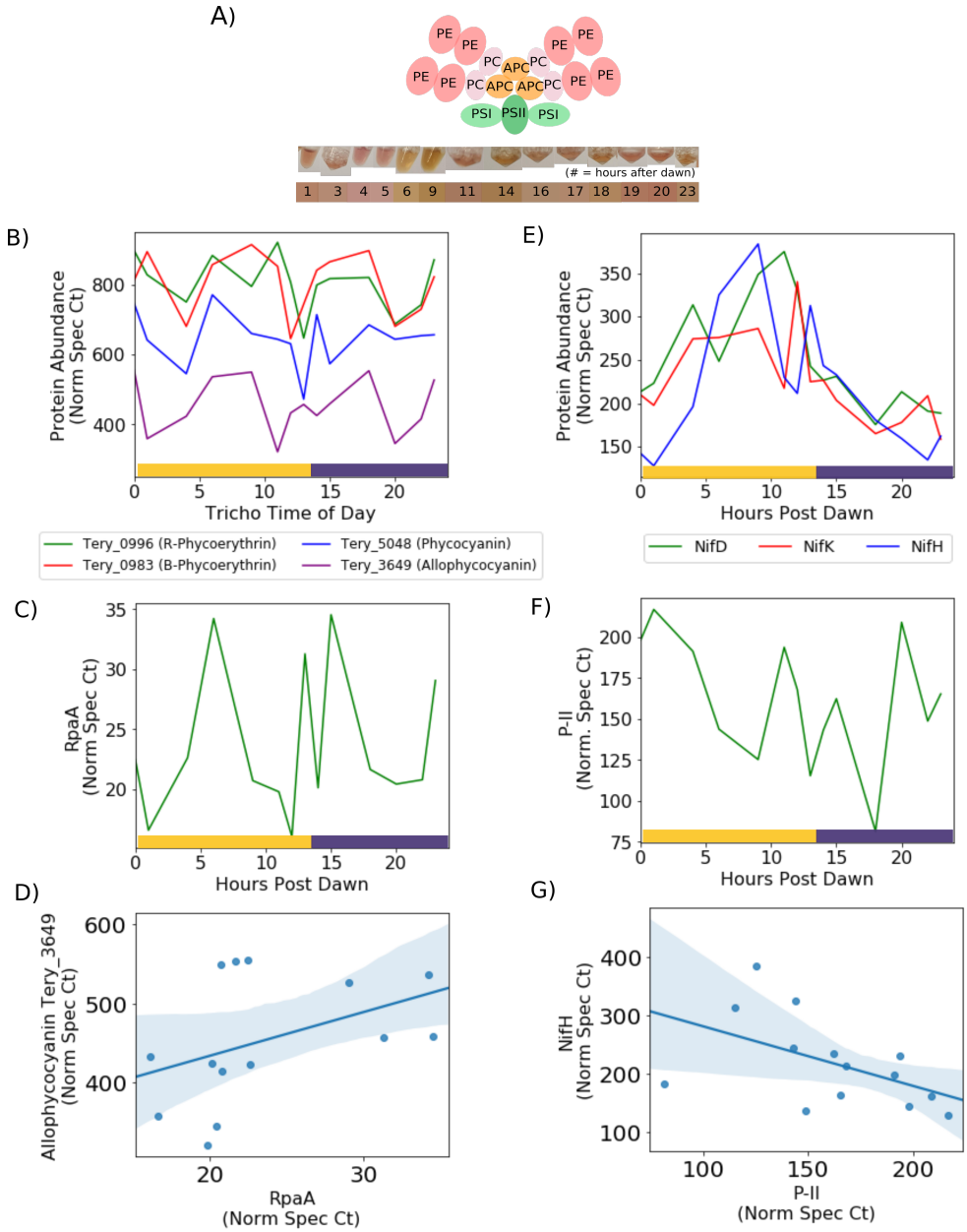


Figure 5.2. (A) Schematic of the phycobilisome rod proteins. PC = phycocyanin, PE = phycoerythrin, APC = allophycocyanin, PSI = photosystem I, PSII = photosystem II. Colors of the protein extracts are displayed for each time point measured. (B) Abundance of the phycobilisome rod proteins and (C) the RpaA regulatory protein throughout the day. (D) RpaA is positively correlated with phycobilisome protein abundance. (E) Abundance of the nitrogenase proteins and (F) the P-II nitrogen regulatory protein. (G) Nitrogenase abundance is negatively correlated with P-II abundance.

5.3.3 Regulation of electron transport to carbon versus nitrogen fixation and associated buoyancy benefits

Oscillations in nitrogen and carbon fixation proteins were associated with changes in the flow of electrons to these processes. The electrons produced by photosynthesis are used for three main processes – carbon fixation/carbohydrate metabolism, nitrogen fixation, and respiration.[36] While the energy-harvesting photosystem machinery remained relatively constant, proteins involved in carbon fixation (including rubisco, uridine kinase, phosphoenolpyruvate carboxylase, and others, see Table S3) and carbohydrate metabolism (including transaldolases, isocitrate dehydrogenase, citrate synthase, and others) were less abundant when nitrogenase was high (Figure 3A and B). This suggested that at the height of nitrogen fixation, solar derived electrons were directly diverted to nitrogenase as opposed to carbon fixation.

In cyanobacteria such as *Trichodesmium*, electrons are transported via iron ferredoxin or non-iron flavodoxin proteins. *T. erythraeum* has at least thirteen ferredoxin genes, ten of which were identified in this experiment; both of *Trichodesmium's* two flavodoxins were identified. A specific ferredoxin, Tery_4145, was positively correlated with nitrogenase abundance and negatively correlated with carbon fixation (Figure 3D, F), suggesting that it mediated the flow of electrons to nitrogenase. Corroborating its role, the protein is located close to, and possibly within, the nitrogenase operon (genes Tery_4133 to at least Tery_4143). By contrast, the second most abundant ferredoxin, Tery_0454, was positively correlated with carbon fixation and negatively associated with nitrogenase (Figure 3 C, E). Together, this suggested that the relative abundance of ferredoxins Tery_4145 versus Tery_0454 contribute to the trafficking of electrons throughout the photoperiod.

This brings us to a major downside of daytime nitrogen fixation, which is that it increases cellular iron demand. In *Crocosphaera*, iron is shuttled from the photosystem to the nitrogenase proteins as during the diel cycle, significantly reducing cellular iron quotas.[12] The need to maintain nitrogenase and the photosystems simultaneously means that *Trichodesmium* cannot take advantage of this strategy. Evidence indicated that *Trichodesmium* is less iron-optimized in general. Specifically, unlike *Crocospheara* which uses a non-iron flavodoxin protein to supply electrons to nitrogenase, *Trichodesmium* appeared to use an iron containing ferredoxin protein, and maintained a large pool of iron-demanding nitrogenase protein throughout the diel cycle. *Trichodesmium's* unique ability to access particulate iron may facilitate this, since it reduces pressure to conserve this resource. Additionally, it should be noted that nitrogen fixation may be further assisted by photoxidation of Fe^{3+} sources during the photoperiod.[37,38]

Direct diversion of energy to nitrogenase during the photoperiod may be crucial for maintaining the buoyancy of large *Trichodesmium* cells. To illustrate this, it is helpful to consider by analogy a nighttime N_2 fixing diazotroph such as *Crocosphaera*. To support the energy demands of the nitrogenase enzyme, *Crocosphaera* produces large stores of glycogen during the day.[11] Glycogen is heavy and acts as sinking ballast, but because *Crocosphaera* cells are small, they remain suspended in the water column.[39] Compared with *Crocosphaera*, *Trichodesmium* filaments and colonies are much larger and therefore are subject to a serious buoyancy problem.[40] They invest heavily in

maintaining their surface buoyancy, for instance by producing gas vesicles that can occupy 70% or more of the cell volume.[41] The buoyancy problem seems to be significant enough to drive the organism towards daytime nitrogen fixation. Assuming that *Trichodesmium* fixes 5-10nmol N colony^{-1} day^{-1} (Mulholland 2006), that 16 ATPs are required per reaction, and that each glucose supplies 8 ATP, each colony would need to store an additional 1-2μg mass of glucose per day to fuel nighttime nitrogen fixation. This would double the typical daytime carbohydrate content, which is approximately 2μg mass per colony, or about 10% of the colony's total mass.[42] In other words, night-time nitrogen fixation would require *Trichodesmium* to be twice as heavy as it actually is, making buoyancy impossible to maintain. Thus, daytime nitrogen fixation provides a key benefit to the cells because it minimizes the amount of glycogen ballast the cell must produce each day.

While energy/carbon is directly channeled to nitrogen fixation during the day, *Trichodesmium* cells must eventually produce glycogen in order to fuel metabolic processes at night. A progression of activities was observed that explains how glycogen production may be stimulated. First, while the cellular POC:PON ratio was relatively constant for much of the day, it did begin to decrease in the late afternoon due to nitrogen fixation (Figure 4B). This stimulated a spike in photosystem protein abundance (Figure 4A) which was followed one hour later but a spike in the POC:PON ratio concurrent with glycogen synthesis (Figure 4C). This late afternoon peak in glycogen production, which occurs after the nitrogen fixation period, may be crucial for the survival of *Trichodesmium* during the night.

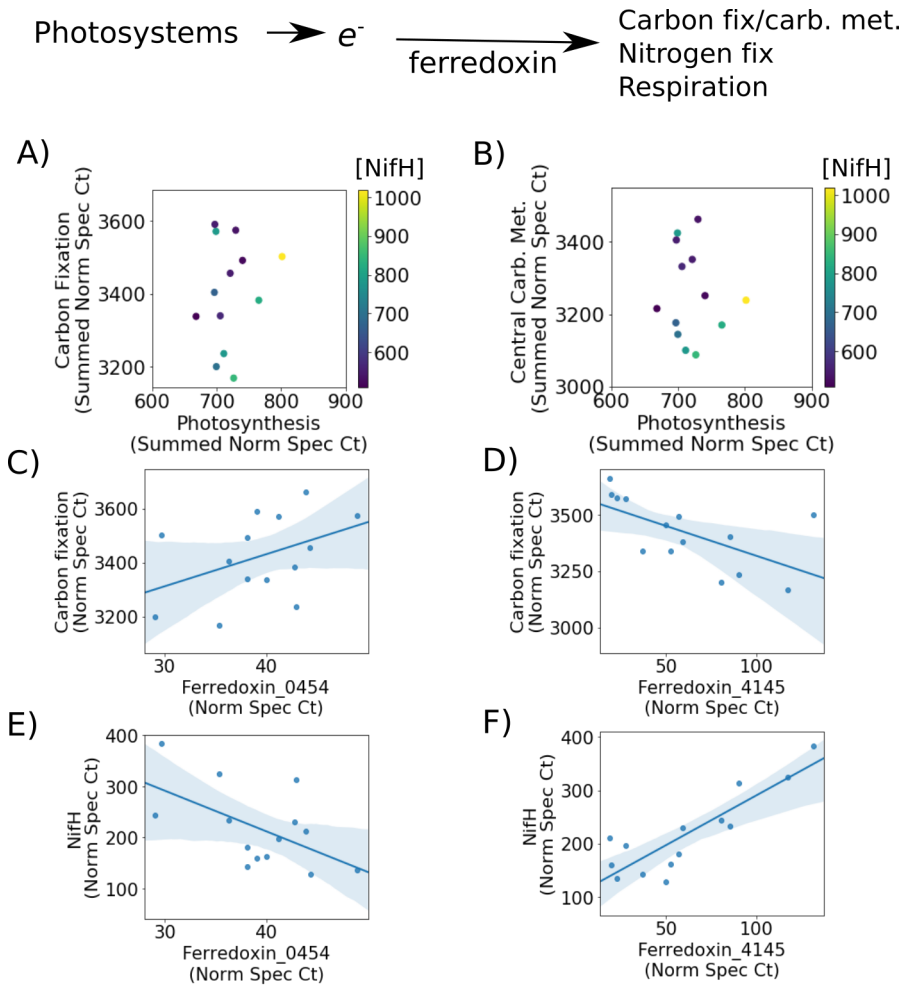


Figure 5.3. (A) and (B) when nitrogenase abundance is high, the abundance of proteins involved in carbon fixation and central carbohydrate metabolism (see Table S1 for Kegg Ontology assignments) are lower relative to photosystem abundance, indicating that electrons are being directly shuttled to the nitrogenase enzyme. (C – F) Ferredoxin Tery_0545 is positively correlated with carbon fixation rate (C) but not nitrogenase abundance (D), while ferredoxin Tery_4145 is positively correlated with nitrogenase (F) but not carbon fixation protein abundance (E).

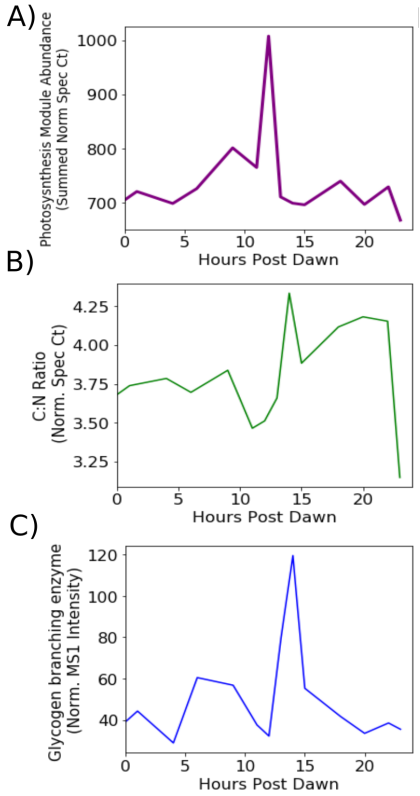


Figure 5.4. (A) Photosystem abundance is relatively constant throughout the day except for a significant peak at dusk. This peak corresponds with a spike in the cellular C:N ratio one hour later (B), which coincides with maximum glycogen synthesis (C).

5.3.5 Modeling cellular POC and PON content from proteomic data

Consistent with concurrent changes in the cell's proteome and POC:PON ratio (see Figure 4), a simple box model allowed POC and PON content to be modeled from proteins in the dataset. In the box model, POC and PON inputs are photosynthesis and nitrogen fixation, respectively (Figure 5). This assumption is valid because no nitrogen or carbon source besides the gaseous forms were provided to the cultures. The POC and PON outputs were carbon respiration and nitrogen excretion, respectively. The offset between the measured POC/PON values and those modeled from photosynthesis/nitrogenase protein content therefore revealed patterns in these processes over the diel cycle. Respiration and nitrogen excretion are cyclic with a period of approximately six hours, peaking just before increases in nitrogenase abundance similar to the phycobilisome proteins described earlier. This supported the growing consensus that respiratory oxygen consumption is a significant source of carbon/energy in *Trichodesmium* cells.[36] Two proteins - the glutamate producing enzyme 5-oxoprolinase and the antioxidant protein thioredoxin were used to estimate the respiration and excretion offsets, respectively. The resulting model captured POC and PON as a function of a few proteins and linear r^2 values of 0.6 or more, demonstrating the utility of molecular biomarkers for modeling biogeochemical parameters.

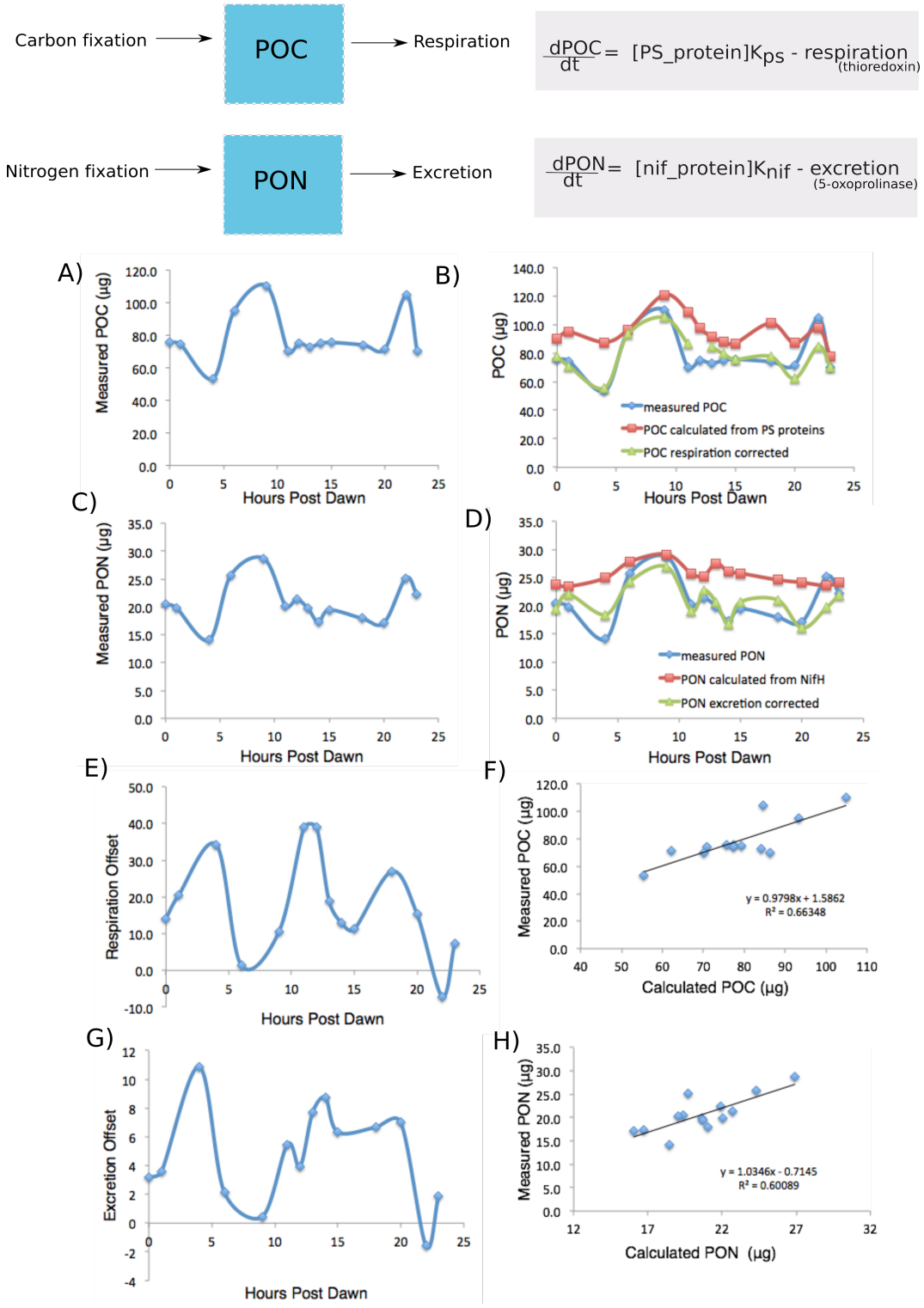


Figure 5.5. (A and B) Box model schema for model of POC and PON abundances in the cultures. (C –H) Measured and modeled POC and PON abundances, as well as offsets due to respiration and excretion demonstrating their cyclic nature. (I and J) Correlation of the modeled versus measured POC and PON abundances. Accurate estimations of POC/PON content can be made from protein abundance data.

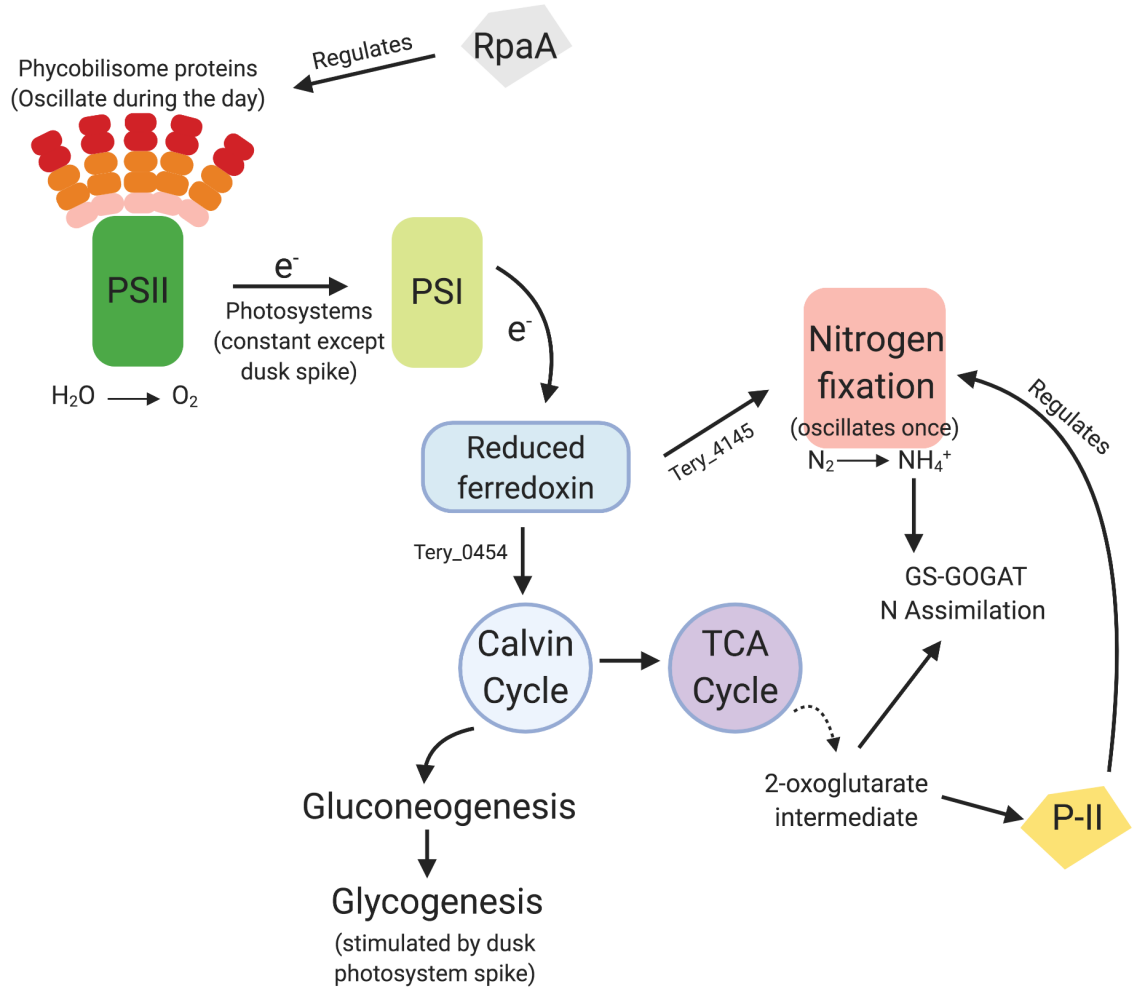
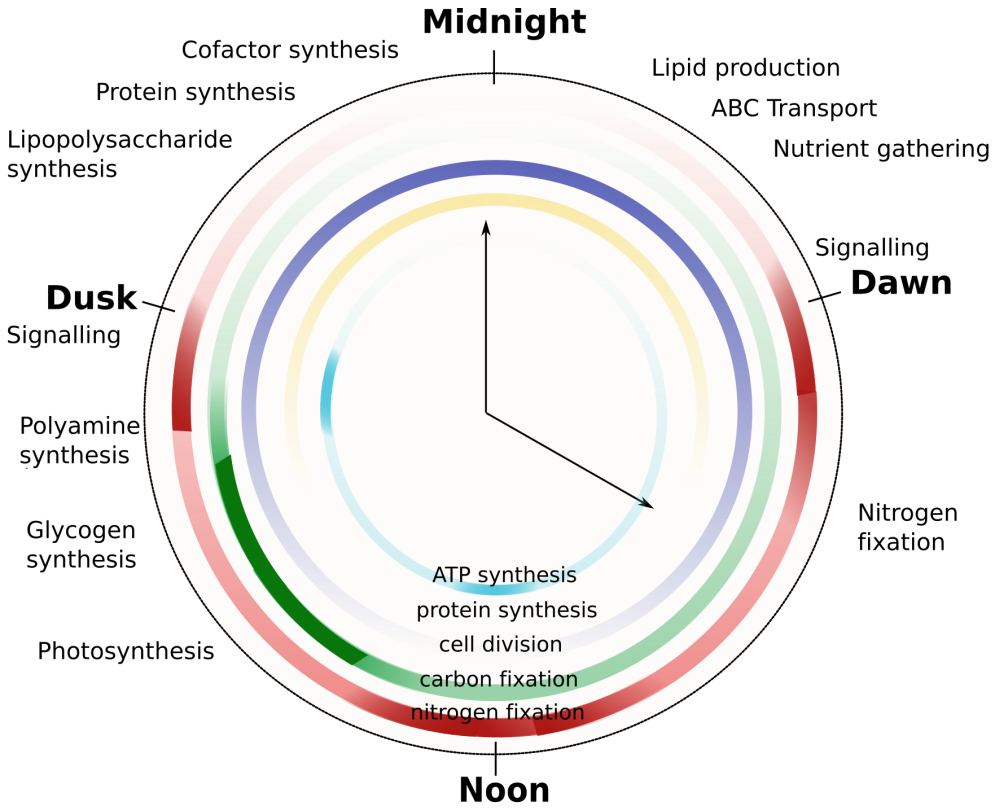


Figure 5.6. Summary of nitrogen and carbon fixation regulation in *Trichodesmium*.



Daytime N_2 fixation

Why?	How?
Energy availability	Temporal separation
Buoyancy regulation	Respiratory protection
Regulation of C:N Ratio	Bet-Hedging

Figure 5.7. Summary of the findings presented here, including how and why daytime nitrogen fixation occurs.

5.4 CONCLUSION

Trichodesmium has long been an enigmatic and mysterious marine microbe. Better understanding how it regulates nitrogen fixation will enhance our ability to predict its effect on global nitrogen and carbon cycles. Daytime nitrogen fixation is crucial for *Trichodesmium's* buoyancy and vertical migration patterns. It requires considerable investment in regulating and implementing changes in phycobilisome, nitrogenase, respiration, and glycogen production throughout the day (Figure 6). This is reflected in the extreme dynamicity of the proteome on short time scales. Multiple mechanisms allow daytime nitrogen fixation to occur, such as temporal separation during the photoperiod, increased respiratory protections such as the Mehler reaction (see Figure S5 and

supplementary text), and "bet-hedging," that is compensating for potential damage by constantly maintaining a pool of nitrogenase enzyme (Figure 7). We note that anything we have identified evidence against the theory that *Trichodesmium* forms "diazocytes", finding that the abundance of a putative differentiation protein HetR was most abundant at night and not during the nitrogen fixation period as has been previously suggested (Figure S6).[43]

Many of the advantages and strategies for daytime nitrogen fixation are based on the peculiar physiology of *Trichodesmium* (Figure 7). Specifically, the large size of *Trichodesmium* filaments/colonies presents a significant problem for maintaining surface buoyancy, but also facilitates nutrient acquisition, reducing a key biogeochemical pressure on the cell. These trade offs provide *Trichodesmium* with a unique and competitive niche in the oligotrophic ocean. The need to coordinate carbon and nitrogen fixation over the diel cycle requires a large number of genes and proteins, particularly those involved in cell sensing and regulation.[27] It certainly seems to underlie the extreme dynamics of *T. erythraeum's* proteome over the course of the day. Future work can focus on clarifying the benefits of *Trichodesmium's* dynamic lifestyle, and can also continue calibrating the protein biomarkers identified here as markers of biogeochemical processes.

5.5 MATERIALS AND METHODS

5.5.1 Cell culturing and sampling

Trichodesmium erythraeum sp IMS101 was grown in RMP growth media prepared with oligotrophic Sargasso seawater (Webb et al., 2001). The cultures were not axenic but had been partially purified by serial sterile transfer for hundreds of generations into RMP media. The cultures were maintained for months at 26.9°C in a 14:10 day:night light cycle, in which light ramps up and down mimicking light intensity over the diel cycle at Station ALOHA. The experiment was conducted in the same incubator.

Approximately 200mL of dense, healthy filaments were inoculated into 1 L RMP media for a final volume of 1200mL in a 2.5 L sampling flask with constant, gentle stirring. An aquarium pump with a sparger and sterile air filter provided gentle oxygenation at the surface of the culture. The cultures were allowed to acclimate and grow for 5 days prior to the beginning of sampling, which occurred in exponential growth. Samples were taken by sterile pipetting 80mL of the culture into a separate culture flask, and immediately filtering by gentle vacuum onto 0.2μm supor filters. Filters were immediately frozen at -80°C until analysis. Sampling was spaced approximately every two hours, with sampling concentrated around the dawn and dusk hours. Care was taken during sampling to prevent exposure to light/darkness during the night/day respectively.

5.5.2 Protein extraction and digestion

Proteins were extracted with a detergent based method as described in detail in Chapter 3. Briefly, proteins were extracted in SDS detergent buffer and purified by precipitation in organic solvent. The proteins were quantified by BCA protein concentration assay (Thermo Fisher). Protein extracts were trypsin digested in-gel using a 1μg:20μg trypsin:total protein ratio as described in Chapter 3. The resulting peptide mixtures were concentrated to 1μg protein μL^{-1} solution for LC-MS/MS analysis.

5.5.3 LC-MS/MS analysis

The global proteomes were analyzed by LCaXmlLC-MS/MS. 10ug of protein was injected per run. To maximize coverage of the proteome, two orthogonal steps of chromatography were performed (PLRP-S column followed by a C18 column), both in-line on a Thermo Dionex Ultimate3000. The samples were then analyzed on a Thermo Orbitrap Fusion mass spectrometer. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD016332 and 10.6019/PXD016332.[44]

5.5.4 Relative quantitation of peptides and proteins

Raw spectra were searched using SequestHT using the *Trichodesmium erythraeum* sp. IMS101 genome plus non-*Trichodesmium* sequences identified in a recent metatranscriptome analysis of epibiont organisms associated with IMS101 cultures.[45] SequestHT mass tolerances were set at +/- 10ppm (parent) and +/- 0.8 Dalton (fragment). Cysteine modification of +57.022 and methionine modification of +16 were included. Protein identifications were made with Peptide Prophet in Scaffold (Proteome Software) at the 95% protein and 99% peptide identification levels. Relative abundance is measured by normalized spectral count. Normalization and FDR calculations were performed in Scaffold (Protein Metrics). The FDRs were 0.01% peptide and 0.3% protein. In total, 2799 proteins (2396 specific to *Trichodesmium*) were identified representing 50.5% of the IMS101 genome. Because epibiont protein identifications were sparse, only *Trichodesmium* proteins were considered moving forward. Protein identification data are provided in Table S1, and peptide identification data in Table S2.

5.5.5 Meta-analysis of the dataset using WGCNA

A weighted correlation network analysis was conducted to identify major trends in protein abundance over the diel cycle. The analysis was performed with the WGCNA library in R on log2 normalized protein fold change data. *Trichodesmium* proteins were considered. Fifteen modules given color labels were identified in the analysis and each was displayed in Figure 2. The groups had similar expression patterns demonstrating the success of the WGCNA analysis in picking up similar proteins. Color module assignments for each *Trichodesmium* protein are provided in Table S2.

5.5.6 Assessing periodicity with RAIN

The periodicity of protein expression over the diel cycle was tested using the RAIN algorithm.[46] The algorithm fits the data to a sinusoidal curve and calculates amplitude, lag, and period, as well as a p value for the model. Proteins were considered periodic at $p < 0.05$.

5.6 ACKNOWLEDGEMENTS

I thank my co-authors on this project, Matthew McIlvin for help with proteomics analysis and John Waterbury, and Mak Saito for insightful discussions. This work was supported by an NSF Graduate Research Fellowship grant # 1122274 [N.Held], the Gordon and Betty Moore Foundation (grant number 3782 [M.Saito]) the National Science Foundation (grant number 1657766 [M.Saito]).

5.7 REFERENCES

1. Karl, D. M., Letelier, R., Hebel, D. V, Bird, D. F. & Winn, C. D. *Trichodesmium* Blooms and New Nitrogen in the North Pacific Gyre. in *Marine Pelagic Cyanobacteria: Trichodesmium and other Diazotrophs* (eds. Carpenter, E. J., Capone, D. G. & Rueter, J. G.) 219–237 (Springer Netherlands, 1992).
2. Karl, D. *et al*. Dinitrogen fixation in the world's oceans. *Biogeochemistry* 57–58, 47–98 (2002).
3. Bergman, B., Sandh, G., Lin, S., Larsson, J. & Carpenter, E. J. *Trichodesmium*--a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol. Rev.* 37, 286–302 (2013).
4. Capone, D. G. *Trichodesmium*, a Globally Significant Marine Cyanobacterium. *Science (80-.).* 276, 1221–1229 (1997).
5. Walworth, N. G. *et al*. Nutrient-colimited *Trichodesmium* as a nitrogen source or sink in a future ocean. *Appl. Environ. Microbiol.* 84, 1–14 (2018).
6. McGillicuddy Jr., D. J. Do *Trichodesmium* spp. populations in the North Atlantic export most of the nitrogen they fix? *Global Biogeochem. Cycles* 28, 103–114 (2014).
7. Coles, V. J., Hood, R. R., Pascual, M. & Capone, D. G. Modeling the impact of *Trichodesmium* and nitrogen fixation in the Atlantic ocean. *J. Geophys. Res. C Ocean.* 109, 1–17 (2004).
8. Dutheil, C. *et al*. Modelling N_2 fixation related to *Trichodesmium* sp.: Driving processes and impacts on primary production in the tropical Pacific Ocean. *Biogeosciences* 15, 4333–4352 (2018).
9. Cohen, S. E. & Golden, S. S. Circadian Rhythms in Cyanobacteria. *Microbiol. Mol. Biol. Rev.* 79, 373–385 (2015).
10. Welkie, D. G. *et al*. A Hard Day's Night: Cyanobacteria in Diel Cycles. *Trends Microbiol.* 27, 231–242 (2019).
11. Mohr, W., Intermaggio, M. P. & LaRoche, J. Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium Crocosphaera watsonii WH8501. *Environ. Microbiol.* 12, 412–421 (2010).
12. Saito, M. A. *et al*. Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph Crocosphaera watsonii. *Proc. Natl. Acad. Sci. U. S. A.* 108, 2184–9 (2011).
13. Golden, J. W. & Yoon, H. S. Heterocyst formation in Anabaena. *Curr. Opin. Microbiol.* 1, 623–629 (1998).
14. Sherman, D. M., Tucker, D. & Sherman, L. A. Heterocyst development and localization of cyanophycin in N_2 fixing cultures of *Anabaena* sp. PCC 7120. *J. Phycol* **36,** 932–941 (2000).
15. Berman-Frank, I. *et al*. Segregation of nitrogen fixation and oxygenic photosynthesis in the marine cyanobacterium *Trichodesmium*. *Science (80-.).* 294, 1534–1537 (2001).
16. Küpper, H. *et al*. Traffic Lights in *Trichodesmium* . Regulation of Photosynthesis for Nitrogen Fixation Studied by Chlorophyll Fluorescence Kinetic Microscopy *Plant Physiology* **135**, 2120–2133 (2019).
17. Carpenter, E. J. & Price IV, C. C. Marine Oscillatoria (*Trichodesmium*):

Explanation for aerobic nitrogen fixation without heterocysts. *Science (80-.).* 191, 1278–1280 (1976).
18. Ohki, K. & Fujita, Y. Aerobic nitrogenase activity measured as acetylene reduction in the marine non-heterocystous cyanobacterium *Trichodesmium* spp. grown under artificial conditions. *Mar. Biol.* 98, 111–114 (1988).
19. Eichner, M. *et al.* N 2 fixation in free-floating filaments of *Trichodesmium* is higher than in transiently suboxic colony microenvironments. *New Phytol.* 222, 852–863 (2019).
20. Bergman, B. & Carpenter, E. J. Nitrogenase confined to randomy distributed trichomes in the marine cyanobacterium *Trichodesmium theibautii*. *J. Phycol.* **27**, 158–165 (1991).21. Ohki, K. Intercellular localization of nitrogenase in a non-heterocystous cyanobacterium (cyanophyte), *Trichodesmium* sp. NIBB1067. *J. Oceanogr.* 64, 211–216 (2008).
22. Ohki, K. & Taniuchi, Y. Detection of nitrogenase in individual cells of a natural population of *Trichodesmium* using immunocytochemical methods for fluorescent cells. *J. Oceanogr.* 65, 427–432 (2009).
23. Finzi-Hart, J. A. *et al.* Fixation and fate of C and N in the cyanobacterium *Trichodesmium* using nanometer-scale secondary ion mass spectrometry *Proc. Natl. Acad. Sci. U. S. A.* 106, 9931 (2009).
24. Flores, E. & Herrero, A. Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nat. Rev. Microbiol.* 8, 39–50 (2010).
25. Lamont, H. C., Silvester, W. B. & Torrey, J. G. Nile red fluorescence demonstrates lipid in the envelope of vesicles from N 2 -fixing cultures of *Frankia Can. J. Microbiol.* 34, 656–660 (1988).
26. Boatman, T. G., Davey, P. A., Lawson, T. & Geider, R. J. The physiological cost of diazotrophy for *Trichodesmium* erythraeum IMS101. *PLoS One* 13, 1–24 (2018).
27. Held, N. A., Mcilvin, M. R., Moran, D. M., Laub, M. T. & Saito, M.A. Unique Patterns and Biogeochemical Relevance of Two-Component Sensing in Marine Bacteria. *mSystems* 1–16 (2019).
28. Sandh, G., El-Shehawy, R., Díez, B. & Bergman, B. Temporal separation of cell division and diazotrophy in the marine diazotrophic cyanobacterium *Trichodesmium* erythraeum IMS101. *FEMS Microbiol. Lett.* 295, 281–288 (2009).
29. Ohki, K., Zehr, J. P., Falkowski, P. G. & Fujita, Y. Regulation of nitrogen-fixation by different nitrogen sources in the marine non-heterocystous cyanobacterium *Trichodesmium* sp. NIBB1067. *Arch. Microbiol.* 156, 335–337 (1991).
30. Zehr, J. P., Wyman, M., Miller, V., Capone, D. G. & Duguay, L. Modification of the Fe Protein of Nitrogenase in Natural Populations of *Trichodesmium* thiebautii Modification of the Fe Protein of Nitrogenase in Natural Populations of *Trichodesmium* thiebautii. (1993).
31. Saito, M. A. *et al.* Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph Crocosphaera watsonii. *Proc. Natl. Acad. Sci. U. S. A.* 108, 2184–9 (2011).
32. Lee, H. M., Vásquez-Bermúdez, M. F. & De Marsac, N. T. The global nitrogen regulator NtcA regulates transcription of the signal transducer P(II) (GlnB) and influences its phosphorylation level in response to nitrogen and carbon supplies in

the cyanobacterium *Synechococcus* sp. strain PCC 7942. *J. Bacteriol.* 181, 2697–2702 (1999).
33. Muro-Pastor, M. I., Reyes, J. C. & Florencio, F. J. Ammonium assimilation in cyanobacteria. *Photosynth. Res.* 83, 135–150 (2005).
34. Muro-Pastor, M. I., Reyes, J. C. & Florencio, F. J. Cyanobacteria Perceive Nitrogen Status by Sensing Intracellular 2-Oxoglutarate Levels. *J. Biol. Chem.* 276, 38320–38328 (2001).
35. Post, A. F., Rihtman, B. & Wang, Q. Decoupling of ammonium regulation and ntcA transcription in the diazotrophic marine cyanobacterium *Trichodesmium* sp. IMS101. *ISME J.* 6, 629–637 (2012).
36. Science, E., Inomura, K., Wilson, S. T. & Deutsch, C. and Photosynthesis in Marine *Trichodesmium*. 4, 1–13 (2019).
37. Tagliabue, A. & Arrigo, K. R. Processes governing the supply of iron to phytoplankton in stratified seas. *J. Geophys. Res. Ocean.* 111, 1–14 (2006).
38. Voelker, B. M., Morel, F. M. M. & Sulzberger, B. Iron redox cycling in surface waters: Effects of humic substances and light. *Environ. Sci. Technol.* 31, 1004–1011 (1997).
39. Berthelot, H., Bonnet, S., Grosso, O., Cornet, V. & Barani, A. Transfer of diazotroph-derived nitrogen towards non-diazotrophic planktonic communities: A comparative study between *Trichodesmium erythraeum, Crocosphaera watsonii* and *Cyanothece* sp. *Biogeosciences* 13, 4005–4021 (2016).
40. Hynes, A. M., Webb, E. A., Doney, S. C. & Waterbury, J. B. Comparison of cultured *Trichodesmium* (cyanophyceae) with species characterized from the field. *J. Phycol.* 48 , 196–210 (2012).
41. Walsby, A. E. The properties and buoyancyproviding role of gas vacuoles in *Trichodesmium* ehrenberg. *Br. Phycol. J.* 13, 103–116 (1978).
42. Villareal, T. A. & Carpenter, E. J. Buoyancy regulation and the potential for vertical migration in the oceanic cyanobacterium *Trichodesmium*. *Microb. Ecol.* 45, 1–10 (2003).
43. Sandh, G., Xu, L. & Bergman, B. Diazocyte development in the marine diazotrophic cyanobacterium *Trichodesmium*. *Microbiology* 158, 345–352 (2012).
44. Perez-Riverol, Y. *et al*. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* 47, D442–D450 (2019).
45. Lee, M. D. *et al*. Transcriptional activities of the microbial consortium living with the marine nitrogen-fixing cyanobacterium *Trichodesmium* reveal potential roles in community-level nitrogen cycling. *Appl. Environ. Microbiol.* 84, AEM.02026-17 (2017).
46. Thaben, P. F. & Westermark, P. O. Detecting rhythms in time series with rain. *J. Biol. Rhythms* 29, 391–400 (2014).
47. Ho, T.-Y. Nickel limitation of nitrogen fixation in *Trichodesmium*. *Limnol. Oceanogr.* 58, 112–120 (2013).

5.8 SUPPLEMENTAL TEXT

This data provided preliminary evidence that nickel cycling may occur over the diel cycle in *Trichodesmium* (see Figure S5). Superoxide dismutase was most abundant during the day, suggesting it may be involved in regulating the oxidative products of photosynthesis. The next most abundant nickel containing protein, urease, had a nearly opposite pattern, being most abundant at nighttime. Thus it is possible that nickel is recycled from superoxide dismutase to urease throughout the day. Nickel is known to limit *Trichodesmium* in certain situations, and increasing nickel availability can extend the length of the nitrogen fixation period.[47] This indicates the need for further investigations on trace metal utilization and cycling in *Trichodesmium*.

5.9 SUPPLEMENTAL FIGURES

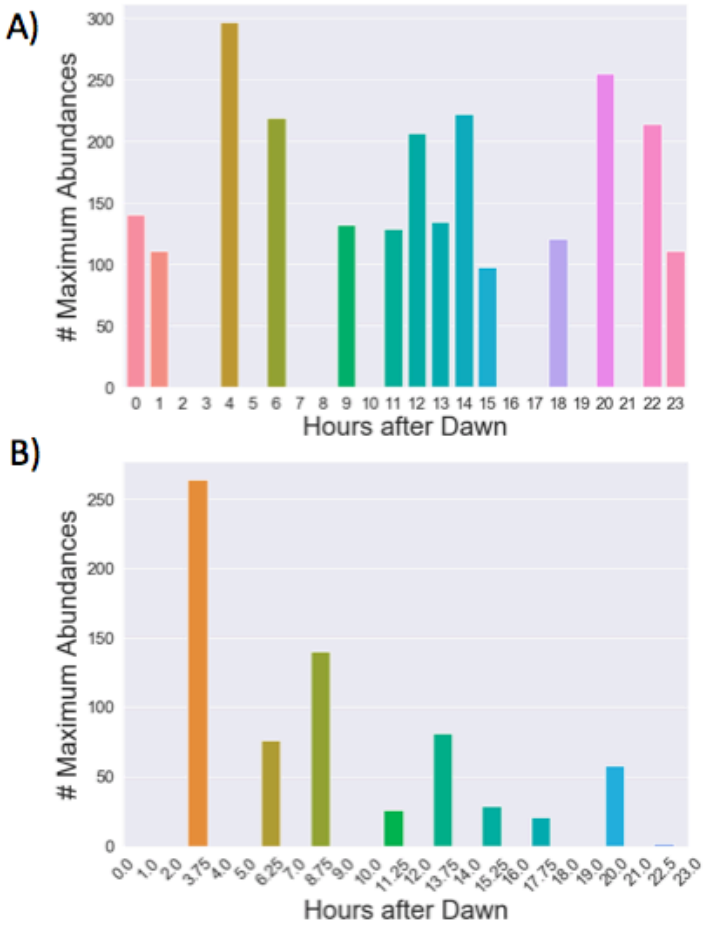


Figure S5.1. Histogram of the number of proteins reaching their maximum abundances at each point of sampling for (A) the *Trichodesmium* dataset presented here and (B) the *Crocosphaera* dataset provided by Saito et al., 2013.

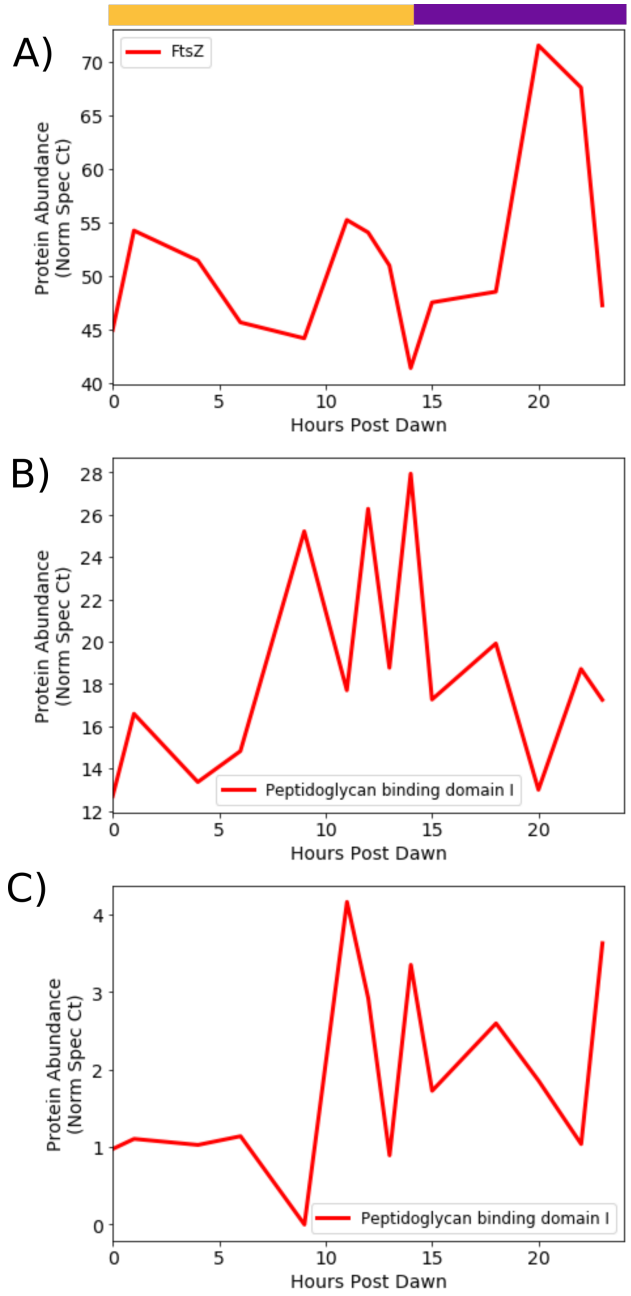


Figure S5.2. Abundance of the cell division proteins FtsZ and peptidoglycan synthesis proteins, which are more abundant in the late afternoon and night. Day (yellow) and night (purple) are indicated.

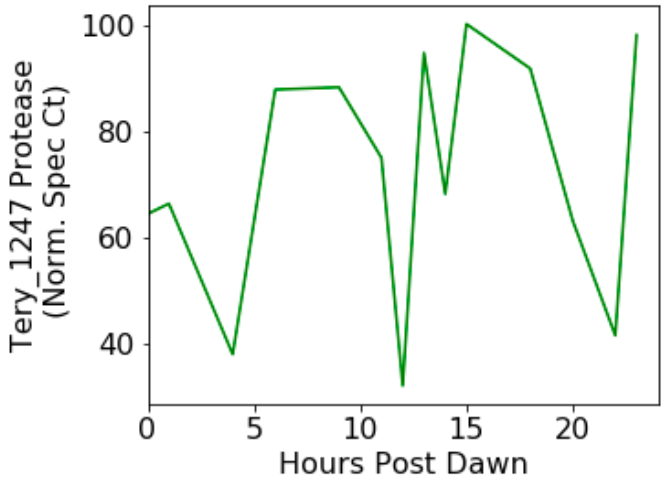


Figure S5.3. Abundance of the Tery_1247 protease, which cycles similarly to RpaA and the phycobilisome proteins and may be involved in degradation of their degradation.

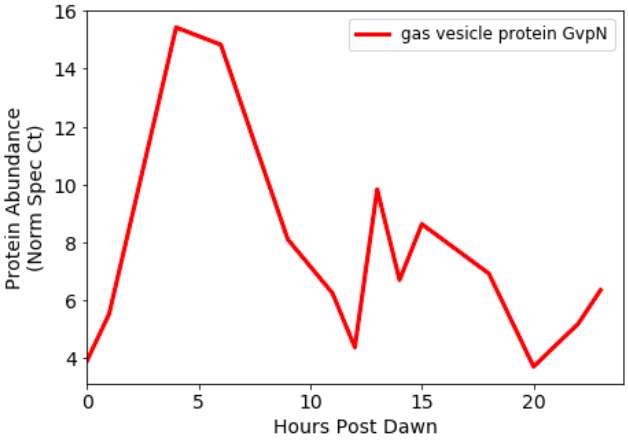


Figure S5.4. Abundance of gas vesicle protein GvpN, which is most abundant during the photoperiod.

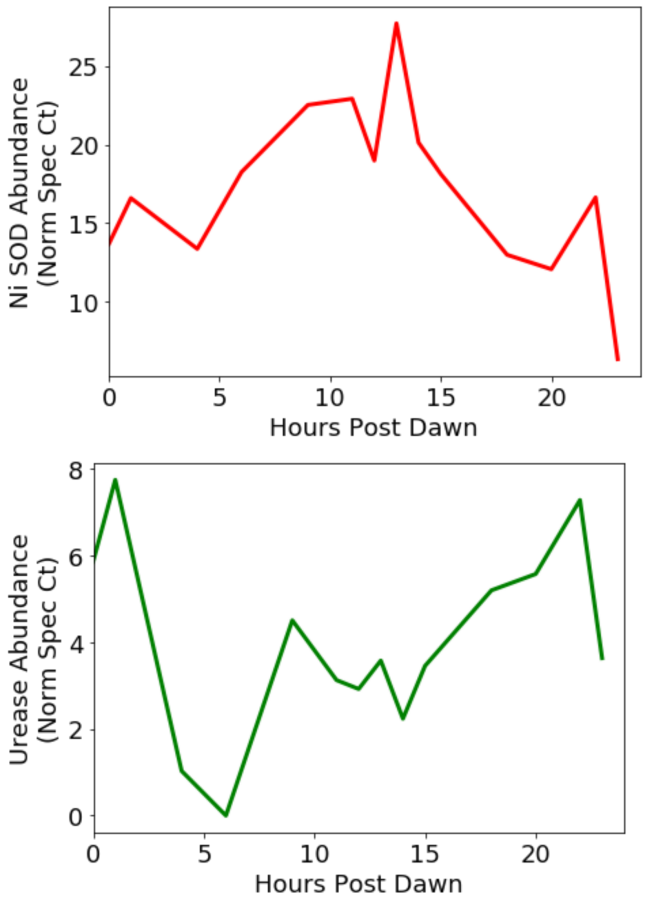


Figure S5.5. Abundance of the Ni superoxide dismutase and urease proteins over the diel cycle. Superoxide dismutase is involved in the Mehler reaction and is most abundant at the height of the nitrogen fixation period. The abundance patterns are nearly opposite to each other, suggesting that nickel may be recycled among them, though more work is needed to clarify this.

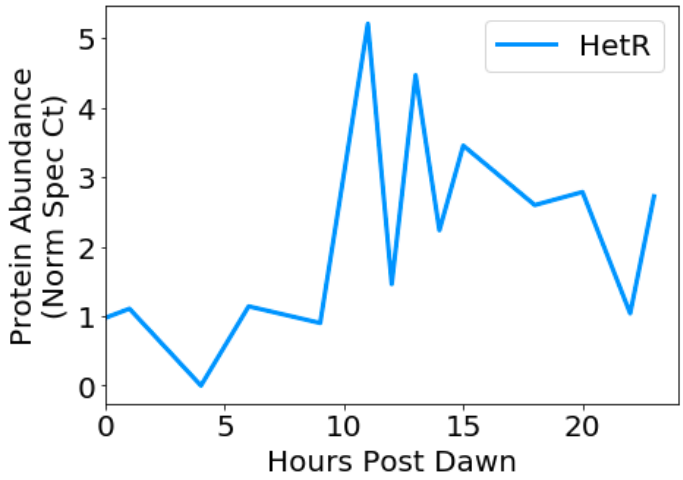


Figure S5.6. Abundance of the putative heterocyst differentiation protein HetR over the course of the day. It was more abundant in the late afternoon towards thee end of nitrogen fixation, contrasting with previous reports.

5.10 SUPPLEMENTAL TABLES (LIST)

Table S5.1 Protein abundance data with functional annotations (available at BCO DMO)
Table S5.2 WGCNA module assignments (available at BCO DMO)

CHAPTER 6. Conclusions

In this thesis, I used molecular biomarkers to obtain information about microbe-chemistry interactions, with an emphasis on the nitrogen fixer *Trichodesmium*. Molecular biomarkers are useful because they allow us to both observe and interpret the behavior of marine microbes. First, they can be used to determine when and where an organism or community of organisms is experiencing an environmental stressor such as nutrient starvation. For instance, in Chapter 3 we determined that iron and phosphate stress co-occurs in *Trichodesmium* populations throughout the tropical and subtropical North Atlantic. This approach can be expanded to gain information about a biogeochemical process of interest, such as in Chapter 5 where we used nitrogenase abundance to understand nitrogen fixation patterns throughout *Trichodesmium's* diel cycle, in Chapter 6 where a modification of the NifH nitrogenase protein was used to model nitrogen fixation rates, and in Chapter 2 where we identified the sensitivity and relevance of two component sensory systems as biomarkers of enviornmental stress.

Second, and most powerfully, molecular biomarkers allow the causes and effects of these processes to be examined. An example of this is provided in Chapter 3, where protein abundance data demonstrated that *Trichodesmium* becomes iron and phosphate starved because of the biophysical limits of membrane space and diffusion. Another example is provided in Chapter 4, where the different responses of ferritin vs. IdiA indicated that *Trichodesmium* has a specific response to particulate iron sources, and in Chapter 2 where we identified unique patterns of cell regulation that may represent a specific adaptation to the marine environment. This thesis thus provides key examples of the power of molecular biomarkers for gaining biogeochemically relevant information with predictive power.

This work develops a detailed understanding of *Trichodesmium* physiology, specifically how the organism interacts with nutrients in its environment. The overarching goal was to understand *Trichodesmium's* physiology well enough that we can start to use molecular biomarkers to monitor and make predictions about populations in the field. This is important because *Trichodesmium* is a key player in marine nitrogen and carbon biogeochemical cycles. This thesis makes significant headway towards this goal, specifically by spelling out causes of simultaneous iron and phosphate stress in the environment (Chapter 3), by identifying a post-translational modification involved in regulation of nitrogenase (Chapter 6), by demonstrating heterogeneity in iron-*Trichodesmium* associations in the field (Chapter 4), and by possibly identifying some reasons why *Trichodesmium* fixes nitrogen during the day (Chapter 5). This thesis thus makes a significant contribution to our understanding of how *Trichodesmium* is "wired" to respond to its environment, information can be directly applied to ongoing field studies and modeling efforts.

While this thesis makes headway in understanding *Trichodesmium*, it also highlights the complexity of its physiology. Systems biology approaches are particularly useful in such cases, since they take a holistic attitude towards the organism. For instance, in Chapter 3 we identified the importance of considering multiple nutrients, including how the acquisition of one nutrient affects the acquisition of another. Similarly, in Chapter 4 we demonstrate that *Trichodesmium* has multiple responses to iron availability depending on whether it comes from a dissolved or particulate source. Key to the systems biology approach is an appreciation for the complexity of decision-making processes even in "simple" organisms such as a cyanobacterium. This complexity arises

as the result of multiple evolutionary and adaptive pressures occurring at the same time. This thesis demonstrates that systems biology approaches are crucial for continued understanding of ocean chemistry, since they can identify molecular "switches" and key cellular decisions that govern the rates of biogeochemical reactions.

This thesis benefited from access to cutting edge instrumentation, and future work in this field will depend on continuing to increase access to mass spectrometry and bioinformatics technology. Most specifically, there is significant work to be done in improving the throughout and cost of post-translational modification proteomics, particularly for complex environmental samples. In addition to technological access, this thesis benefitted significantly from access to the sea allowing large-scale molecular surveys. Continuing to increase this access is crucial if we hope to someday map the distributions of proteins in the ocean. This highlights a key benefit of proteomics, however, which is that thousands of biomarkers for different processes of interest can be obtained from a single sample. Proteomics spectra are records of an environment at a given time, and thus can continue to be mined for new targets as they become of interest.

There is growing acknowledgement that inter-laboratory calibrations of "omics" type measurements is needed if the goal is to obtain global coverage of molecular biomarker distributions and process rates in the ocean. Some such efforts, such as BioGeoSCAPES, are already underway, and will set the groundwork for large surveys of ocean molecular ecology. In particular, molecular biomarkers will need to be calibrated to reaction rates, which in turn will require further laboratory experimentation to benchmark microbial behavior under different controlled conditions. These are significant needs but are made tractable by increasing international interest and collaboration.

An overarching theme throughout this thesis is that microbe-chemistry interactions are often more complex than they appear. Honing in on processes involved in cellular decision-making, specifically cell sensory and regulatory proteins, provided a useful framework for beginning to tease apart this complexity. A poignant finding of this thesis is that microbes begin linking elemental cycles at the sensory/biochemical level. This has implications for our understandigns of biogeochemical cycles, because microbes mediate key chemical transformations of these elements. Molecular biomarkers such as proteins and protein post-translational modifications are powerful because they provide information about these links at multiple temporal and spatial time scales. When combined with a systems biology approach and with continued development, these biomarkers can provide crucial information at the organismal, community, ecosystem, and global level.

APPENDIX 1. The primary phosphoproteome of *Prochlorococcus* and *Alteromonas*

A1.1 SUMMARY

This appendix describes efforts to analyze the phosphoproteomes of two marine organisms – *Prochlorococcus* sp. 9601 and *Alteromonas* sp. BB2AT2. The phosphoproteomes were analyzed in exponential and stationary phase growth, and over time after exposure to low-nutrient media. Analysis was conducted by an iterative method, where the sample was re-analyzed multiple times with exclusion lists included for all peptides and phosphopeptides that had already been found. This provided good coverage of the phosphoproteome and may be complementary to chemical enrichment methods. Many of the phosphoproteins identified were involved in basic metabolisms such as gluconeogenesis. When measured over time, the phosphoproteomems changed on the time scale of just hours. While in Prochlorococcus the number of phosphopeptides increased after starvation, in Alteromonas it decreased. This may be related to the lifestyles of these organisms, with the oligotroph tuned to respond to low nutrient conditions and the copiotroph tuned to respond to high. This work suggests that phosphoproteomics could have potential importance for marine microbiology, but also highlights the need for basic studies to identify regulatory events and make such analyses more tractable.

A1.2 INTRODUCTION

Posttranslational modification is the dynamic addition of chemical groups to specific amino acids, resulting in a change in the function and activity of the protein. Posttranslational modifications (PTMs) regulate protein activity, flag proteins for degradation, modulate bi-functional proteins, assist in recruitment of proteins to complexes, and transmit cellular signals.[1,2] Recent advancements in instrumentation and data analysis algorithms make it possible to study PTMs in proteins using mass spectrometry; this presents a unique opportunity to study the biochemistry of microbes in unprecedented depth.[3]

Protein phosphorylation, the dynamic addition of phosphate to amino acids (typically serine, threonine, tyrosine, histidine or aspartate), is a key regulator of protein activity in both prokaryotes and eukaryotes. It is the most well-characterized protein modification to date. Recent developments in analytical chemistry and liquid chromatography mass spectrometry (LC-MS/MS) make it possible to study the collection of phosphorylation events in the proteome (the "phosphoproteome"). However, the most popular strategies require large amounts of sample and introduce biases into the analysis. For instance, treating samples with TiO_2 improves recovery of phosphorylated residues, but selectively enriches for phosphorylated serine and threonine residues. This hinders identification of phosphorylation on tyrosine, histidine and aspartate residues, which are important sites for prokaryote cell signaling. A wide body of knowledge exists concerning the identification, localization, and quantification of phosphorylation events, particularly for model organisms such as *Escherichia coli*.[4] This allows us to validate results, making phosphorylation a good place to start the exploration of PTMs in marine systems.

Protein phosphorylation represents a small yet essential sink of phosphate in cyanobacteria. A back-of the envelope calculation suggests that protein phosphorylation composes 0.2-0.5% of the intracellular phosphorous pool (Box 1). This is a small amount but may be enough to affect the relative fitness of the organisms. Cellular C:P ratios are relatively plastic; for instance, phytoplankton can reduce their phosphorous demand by making substitutions to phospholipids under phosphate stress.[5] It is unclear whether similar substitutions can be made for post-translational modifications/signaling mechanisms. We might ask the extent to which organisms prioritize the use of phosphate for cell signaling vs. use in other biological molecules such as DNA/RNA, particularly when under phosphate stress. This makes phosphorylation an intriguing PTM to study in marine systems, where nutrient availability varies significantly.

This appendix describes efforts to identify and evaluate a reproducible method for global phosphoproteome analysis in marine microbes. Specifically it explores whether iterative re-analysis of the sample with targeted exclusion of previously identified peptides can generate high quality datasets without the need for chemical phosphopeptide enrichment (Figure 1). The iterative method is applied to two marine organisms – *Prochlorococcus* and *Alteromonas* – representing the oligotrophic vs. copiotrophic lifestyle. The phosphoproteomes of these organisms were measured in the logarithmic and stationary phases, as well as in response to nutrient starvation. This demonstrates the potential for phosphoproteomics for understanding microbial physiology, and indicates that copiotrophs and oligotrophs may have different regulatory strategies in response to nutrient availability.

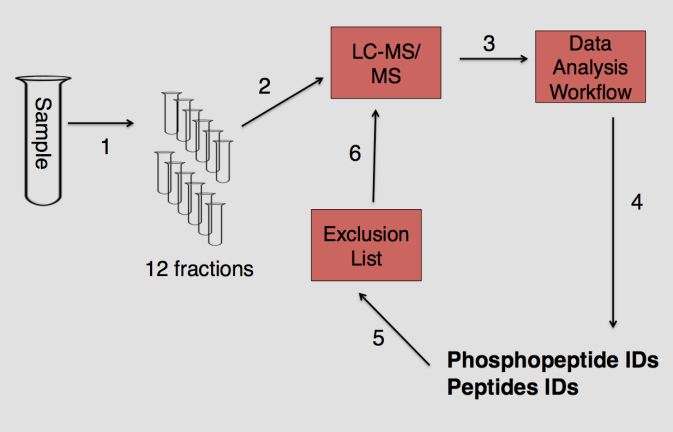


Figure A1.1. Phosphoproteome analysis by the iterative method. (1) the sample is fractionated off-line, (2) each fraction is analyzed by LC-MS/MS, (3) data enters the bioinformatics workflow and peptides are identified (4). These peptides are used to generate an exclusion list (5) that is used for subsequent analyses of the same fractions (6). Steps 3-6 may be repeated.

Box 1. How of the phosphate pool is used for protein modification?
We estimated the amount of intracellular phosphate used in protein modification as follows:

mass P/cell * molar mass P * Avogadro's number = atoms P/cell

pro cell volume * protein concentration per volume * % phosphorylated = phosphate bound to proteins/cell

phosphate bound to proteins/atoms P = percent phosphate pool used for PTM

Values used:
mass P/cell: 0.98fg P/cell for P replete cells, 0.34 P/cell for P deplete cells (Bertilsson 2003)
molar mass P: 30.97g/mol
average protein concentration in a bacterium: $3e^6$ proteins/uM^3 (Milo 2011)
pro cell volume: 0.6 uM diameter (Partensky 1999) calculated as a sphere
percent phosphorylated: 10% (this study)

A1.3 MATERIALS AND METHODS

A1.3.1 Prochloroccus *Experiments and Protein Digestion*

Prochlorococcus sp. 9601 (Pacific ocean high light ecotype) was grown in eight 250mL volumes in trace metal clean LDPE bottles containing Pro 99 media under constant light.[6] Growth was monitored daily by fluorescent absorbance at 700nm. Two bottles were harvested by centrifugation in mid-log growth. Four bottles were harvested at the same time by centrifugation, then washed twice in phosphate deplete media (Pro99 media with no added phosphate), and finally re-suspended in 200mL phosphate deplete media. These bottles were returned to the incubator and harvested 0, 4, 12, or 24hrs after treatment. The remaining two bottles were allowed to grow undisturbed and were harvested in the stationary phase. Cells were lysed by sonication in 100mM ammonium bicarbonate with phosphatase inhibitors (50mM each sodium fluoride, sodium pyrophosphatase, and sodium orthovanadate). Phosphatase inhibitors are necessary to prevent removal of phosphate groups by native enzymes in the cells. Cellular debris was removed by centrifugation. Proteins in the resulting crude extract were purified by 1:4 sample:solvent organic precipitation in 100% acetone overnight at -20°C. The precipitated proteins were then collected by centrifugation.

Proteins were digested in a gel matrix to improve recovery of membrane bound proteins such as cell sensory systems.[7,8] Purified extracts were resuspended in 100mM ammonium bicarbonate buffer containing 6M urea to denature and solubilize proteins. Resuspended protein extracts were embedded in the gel matrix and incubated with dithiothreitol (to reduce disulfide bonds) followed by iodacetemide (to alkylate cysteine residues, preventing disulfide bonds from reforming), and finally a second dithiothreitol incubation (to prevent subsequent alkylation of the digestive enzyme, trypsin, by excess iodacetemide). Proteins were quantified with a colorimetric DC Protein Assay (BioRad) and digested at ratio of 50:1 protein: trypsin (Promega Gold) at 37°C overnight with

shaking. The following morning, the samples (now peptides) were removed from the incubator and analyzed by the LC-MS/MS method described below.

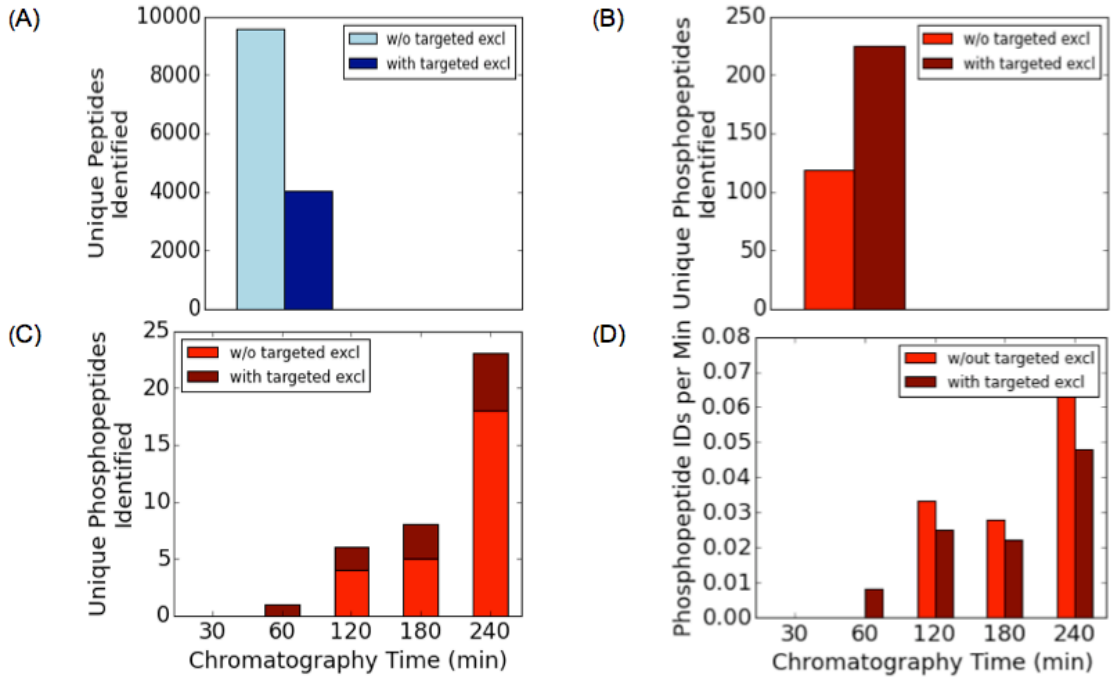


Figure A1.2. (A) Targeted exclusion of peptides using the iterative method results in fewer peptide identifications but more (B) phosphopeptide identifications. (C) the number of phosphopeptides identified by the iterative method increases with longer chromatography/analysis time, but (D) the efficiency baesd on identifications per minute decreases because the iterative method doubles the machine time required.

A1.3.2 Alteromonas BB2AT2 *(diatom associated heterotroph) phosphoproteome*

Alteromonas BB2AT2 isolates were obtained and grown in 250mL volumes of coastal seawater supplemented with peptone and yeast extract at room temperature under vigorous shaking.[9] Cells were harvested in mid-log growth and proteins purified and digested in solution as previously described. *Alteromonas BB2AT2* peptides were used for the majority of method testing and iterative analysis method development.

We examined the phosphoproteomes of *Alteromonas BB2AT2* subject to nutrient deplete media. Cells were grown for sixteen hours (~ 2 doublings) in sterile filtered coastal seawater with added peptone and yeast extracts. Cells were collected by centrifugation, rinsed twice and re-suspended in the experimental media (sterile filtered coastal seawater (coastalSW), sterile filtered coastal seawater amended with 5mM glucose and 3mM ammonium chloride (amendedSW), or the typical media (control)). At various time points (0h, 1h, 2h and 24h) after the experiment was initiated, cells were harvested by centrifugation and digested for proteomic analysis as in Saito 2014. Because of the large number of samples resulting from this experiment, the digestion was

performed in a liquid matrix instead of a gel (a less time and resource consuming method). Samples were analyzed by the iterative method (one iteration).

A1.3.3 LC-MS/MS Analyses

We made considerable efforts to optimize the data collection procedures for phosphoproteome analysis. All analyses were conducted on the Thermo Orbitrap Fusion mass spectrometer. Conditions tested included ion activation method, dynamic exclusion time, and on-line liquid chromatography gradients for best efficiency and separation of the samples (data not shown). Based on this testing, optimal instrument parameters were selected (activation method: higher-energy collision dissociation (HCD), dynamic exclusion time: 30s, chromatographic gradient: 180mins, ~500nL/min on a 50cm column). Parent ions were measured in the Orbitrap (high accuracy), and fragment ions were measured in the ion trap (lower accuracy, but faster scan time). Unless otherwise noted, these parameters were used for all data collection in this study.

The iterative phosphoproteome data collection method can be visualized in Figure 1 and is described as follows: the sample (typically 200μg of protein) was separated off-line in a high pH (ammonium formate pH ~11) reverse phase separation procedure using a 70min acetonitrile gradient on a Michrom Paradigm MS4 instrument.[10] The elevated pH helps to protect acid labile phospho-peptide bonds. Fractions were collected on 5min intervals for a total of 12 fractions. Fractions were acidified to pH 4 with glacial acetic acid just before analysis.

Each fraction was analyzed individually as described above. The collected spectra were then analyzed using Sequest HT against the genome for the target organism. Peptide identifications were filtered to a 5% false discovery rate (FDR); the resulting peptides were used to generate an "exclusion list" for the mass spectrometer. The exclusion list consists of a list of mass : charge (m/z) ratios, charge states, and retention times for each identified peptide. Each fraction was then injected a second time into the mass spectrometer, this time with the instrument programmed to ignore precursor ions with characteristics matching an entry on the exclusion list. The spectra from the initial analysis and the analysis with targeted exclusion were combined and entered into the bioinformatics workflow.

A1.3.4 Bioinformatics workflow

Spectra were searched against strain-specific genomes acquired from collaborators (*Alteromonas* sp. BB2AT2) or Uniprot (*Prochlorococcus* sp. 9601). Raw files containing the spectra for each fraction were compiled and searched together in Sequest in Proteome Discoverer 1.6. When the iterative method was used, the raw files from all of the runs were combined in this initial Sequest search. Variable modifications were set for phosphorylation (+90Da) on S, T, Y, H and D residues. Results were filtered to a 5% FDR threshold with Percolator.[11] Phosphorylation site scoring and localization was performed with PhosphoRS at the 80% probability level.[12,13] Normalization of the spectral counts and FDR calculations were performed in Scaffold (Proteome Software).

A1.4 RESULTS AND DISCUSSION

A1.4.1 Phosphoproteome analysis by the iterative method

Targeted exclusion of previously identified peptides was effective in increasing coverage of the phosphoproteome. However, as with all LC-MS/MS experiments there was a trade off between depth of analysis and resource allocation, i.e. machine time. In each successive re-analysis of a sample, the number of new peptide identifications decreased while the number of new phosphopeptide identifications increased (Figure 2 A-B). This was consistent with preferential exclusion of high abundance peptides, freeing up analysis time for low level phosphopeptides that would otherwise be missed. However, because the iterative method doubled the analysis time, the efficiency was decreased (Figure 2 C-D). Thus phosphopeptide profiling by the iterative method is not particularly useful for large collections of samples. Practicality and efficiency dictate that iterations be limited to two or three per run at most, but deep coverage of an entire organism's phosphoproteome might take much longer. Moving forward, I suggest deep phosphoproteome analysis to identify interesting phosphopeptides in an organism. This can be followed by targeted analysis of these phosphopeptides in experimental conditions.

A1.4.2 The Prochlorococcus *Phosphoproteome*

The global phosphoproteome of *Prochlorococcus* sp. 9601 in the exponential and stationary phases is presented in Figure 3. The phosphoproteome was quite dynamic, with differences in phosphorylation status ranging from 0 to 99% of the total measured peptide abundance when both the phosphorylated and unphosphorylated peptide were identified. On average, 6.3% of the peptides and 10.3% of the proteins were phosphorylated in the exponential phase, and 5.1% of peptides and 18.3% of the proteins were phosphorylated in the stationary phase.

The phosphoproteome differed significantly in exponential versus stationary phase, providing clues about the metabolic activity of the organism. In theexponential phase, phosphoproteins included transporters, including transporters for phosphorous copounds, and proteins involved in the pentose phosphate pathway. *Prochlorococcus* responds quickly to changes in extracellular phosphate, and tightly regulates its P metabolism.[14] Knowing this, identification of regulatory phosphorylation events in proteins involved in phosphorous metabolism suggests that the differences we see are authentic. One question is whether phosphorylation of proteins for phosphate metabolism occurs purposefully (as a feedback mechanism) or simply because phosphorylation is common in cell signaling.

The strongest phosphorylation signal in the exponential phase was for the 50S ribosomal protein L11 (Fig 4A). The signal was so high, in fact, that it concealed other data and was removed from the visual data representation. Protein L11 is implicated in translation termination. Artificial phosphorylation of purified E.coli 50s ribosomal proteins by rabbit protein kinases found that L11 is phosphorylated on a serine or threonine residue.[15] The phosphorylated 50s subunit was less active, and the relative amount of protein phosphorylation was often lower in the intact subunit vs. isolated

proteins. Based on this information, ribosome activity seemed to be greater in the exponential phase relative to the stationary phase; this makes sense given the fast growth rate of the cells. However, the mechanism of ribosome inhibition by protein phosphorylation has not been elucidated; thus, it is too soon to know whether this phosphorylation is a true marker of ribosome activity.

In stationary growth, many amino acid and vitamin biosynthesis proteins were preferentially phosphorylated. This suggested tight regulation of biosynthetic pathways in cells that are growing more slowly. Again, a strong signal swamped out the remaining signals and was removed for the purposes of visualization. This signal corresponded to protein cbbA, which encodes a fructose-bisphosphate aldolase involved in the pentose-phosphate pathway for glycolysis/gluconeogenesis, and the Calvin cycle. This protein is also phosphorylated in *Escherichia coli K12*.[16] The aldolase also responds to phosphate stress in diatoms and *Synechococcus* (P limited cells have more of the protein than cells in P replete media).[17] Studies in protozoans suggest that phosphorylation may help localize the protein in the glycosome (an organelle used for glycolysis). While cyanobacteria do not have glycosomes, they do have carboxysomes where the first step of carbon fixation is concentrated. Phosphorylation may thus assist in recruiting the aldolase enzyme to its proper location in the cell.[18] However, there is no evidence to suggest that the aldolase resides in the carboxysome itself (canonically, the carboxysome is composed of tightly packed rubisco proteins), and this explanation does not readily explain why the phosphorylation status changes in the log vs. stationary phase. Alternatively, phosphorylation status may modulate the function of the protein in some way, perhaps acting as a biochemical switch for gluconeogenesis/carbon fixation vs. glycolysis. This data suggests that phosphorylation is a "turn on" signal for glycolysis, since it was elevated in stationary phase cells (Fig 4B). Detailed biochemical studies would be needed to substantiate either hypothesis, but evidence of this protein's importance in phosphate deplete conditions in numerous organisms suggest its importance in cellular regulation.

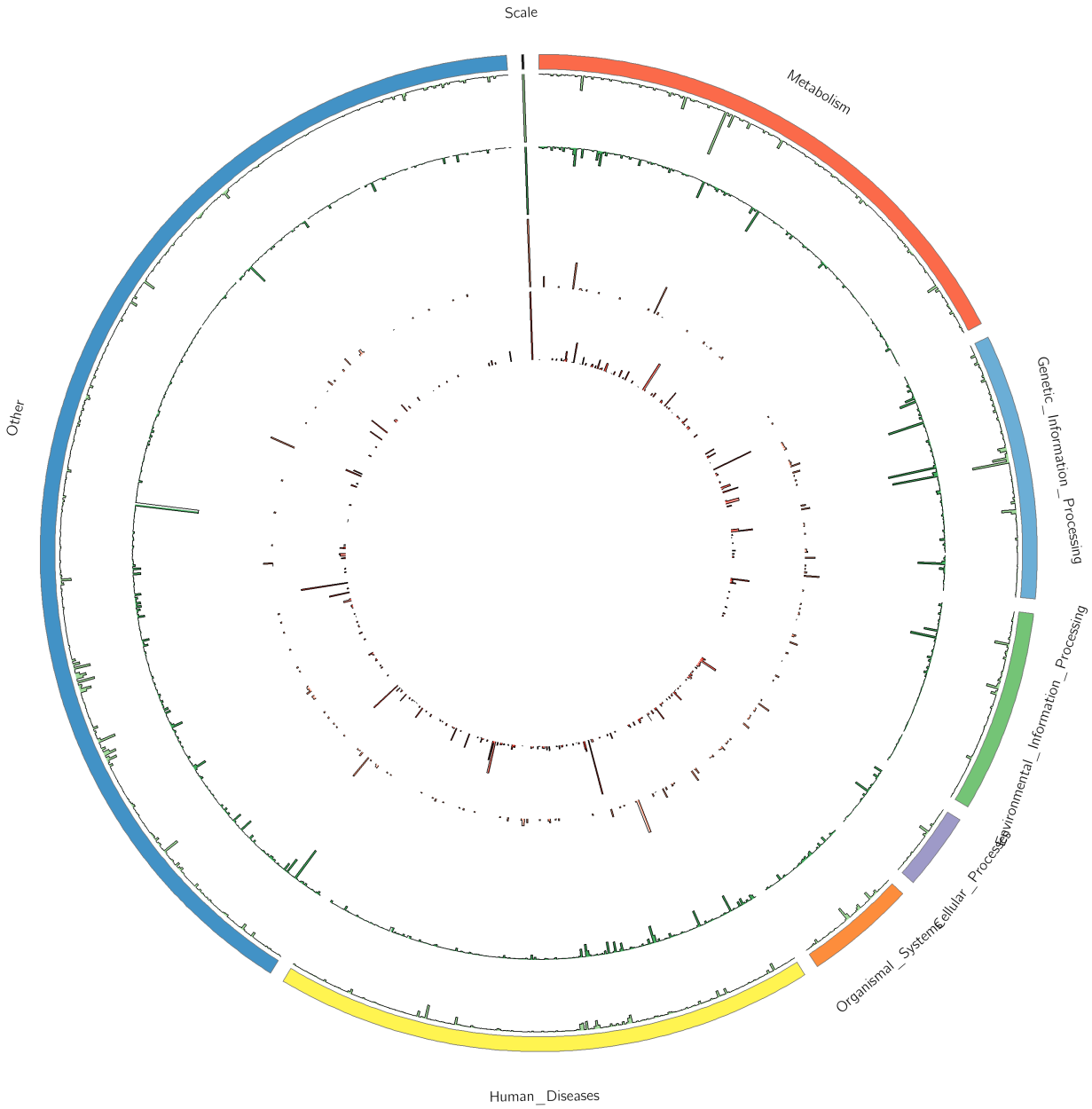


Figure A1.3. The global proteome (green) and phosphoproteome (red) of *Prochlorococcus* sp. 9601 in exponential (outer, lighter colors) and stationary (inner, darker colors) growth. Intensities are represented as normalized MS1 peak intensities. The scale represents intensity $6e^6$. For simplicity in the representation, the most abundant confidently localized phosphorylation site is displayed in cases where there are multiple phosphosites per protein. Not shown: ribosomal protein L11 and fructose-bisphosphate aldolase (see text).

A1.4.3 The Alteromonas *Phosphoproteome*

The phosphoproteome of Alteromonas sp. BB2AT2 is presented in Figure 4. Fewer proteins and protein modifications were identified here than in the *Prochlorococcus* experiment. This could reflect true physiological differences but is more likely an analytical issue. The phosphoproteome included a number of sensory histidine kinases, signal transduction proteins, porins, efflus transporters, and ribosomal proteins. Phosphoenolpyruvate (PEP) synthase is involved in gluconeogenesis protein and is controlled by a regulatory phosphorylation on a histidine residue (active form) or threonine residue (unactive form). The threonine phosphorylated form was more abundant in the stationary phase, suggesting lower metabolic activity consistent with growth state. This interpretation is different from what would be found if protein abundance was considered alone, since PEP synthase abundance was relatively constant in the two growth conditions. This illustrates the potential importance of post-translational modification measurements in understanding the physiology of marine organisms, particularly if the goal is to extrapolate metabolic function from protein abundance data.

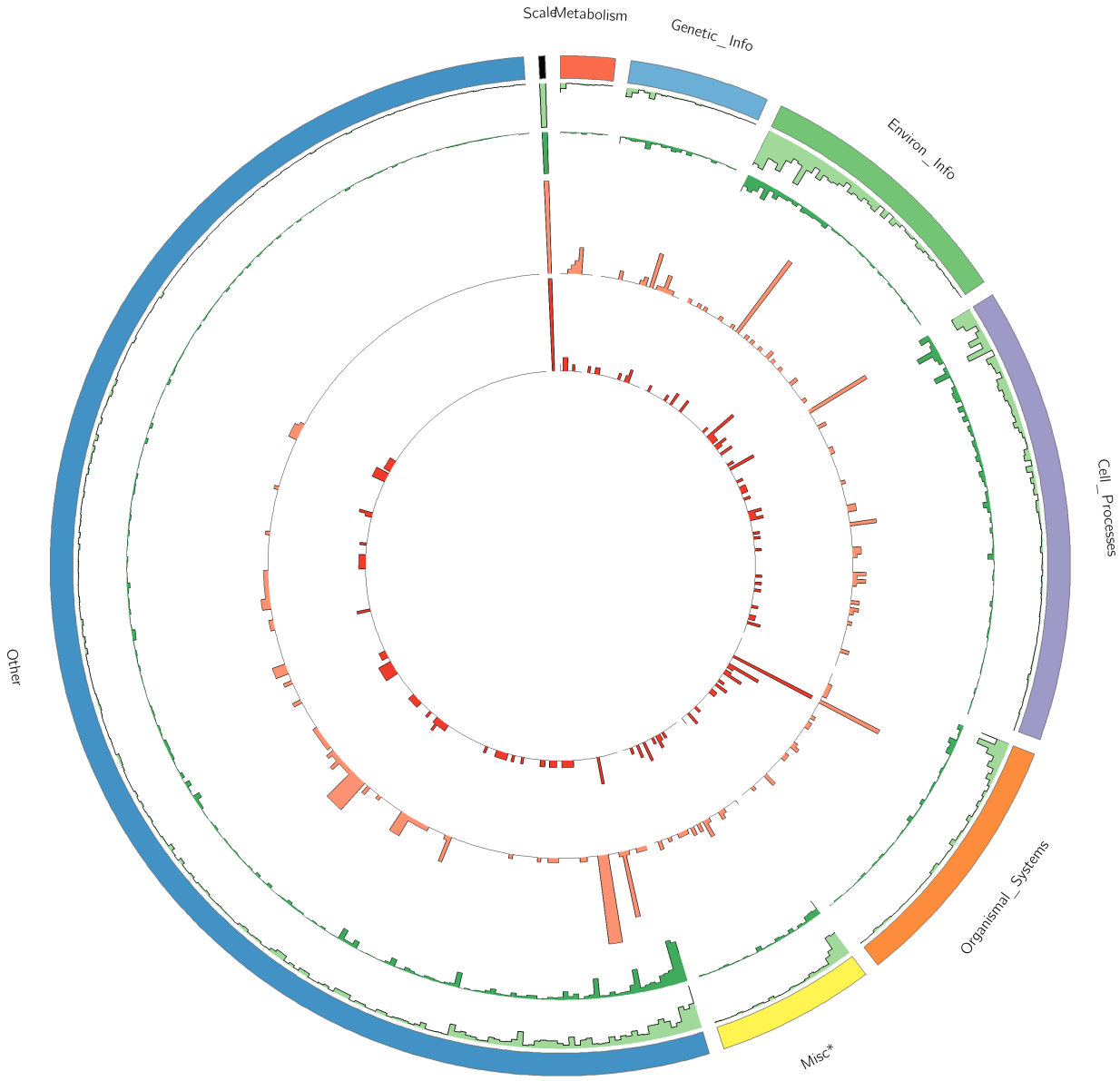


Figure A1.4. The global proteome (green) and phosphoproteome (red) of Alteromonas BB2AT2 in exponential (outer, lighter colors) and stationary (inner, darker colors) growth. Intensities are represented as normalized spectral counts. The scale for represents 17 spectral counts. For simplicity in the representation, the most abundant confidently localized phosphorylation site is displayed in cases where there are multiple phosphosites per protein.

A1.4.4 Phosphoproteome dynamics over time

The phosphoproteome of *Prochlorococcus* and *Alteromonas* changed rapidly after exposure to low nutrient media. Because biomass yields were low during nutrient starvation, these experiments suffered from low phosphopeptide recovery, making comparison of phosphopeptide abundances among the treatments difficult. However they did provide proof of concept for rapid protein regulation occurring in these marine bacteria. This is unsurprising since the phosphoproteome changes rapidly in other organisms, such as humans.[19,20] In *Prochlorococcus*, the number of phosphorylated peptides more than doubled twelve hours after exposure to low nutrient media, then decreased 24 hours after the exposure (Figure 5). This suggested dynamic regulation of the proteome shortly after the initial transfer, followed by a quieting of this regulation. *Alteromonas* had the opposite reaction to exposure to low nutrient media, specifically a decrease in the percent peptides phosphorylated under nutrient starvation relative to the control (Figure 6). The response was very rapid, occurring on the time scale of just one hour, indicating that protein phosphorylation events can change dramatically on short time scales.

The differences in the response of *Prochlorococcus* versus *Alteromonas* may have something to do with their respective lifestyles. *Prochlorococcus* is adapted to low-nutrient oligotrophic environments, and indeed seems to be "tuned" to ramp up regulation in response to starvation. By contrast, *Alteromonas* is adapted to make use of transient patches of concentrated nutrients, and therefore may ramp down its regulatory events in response to starvation. Further work will be needed to clarify this pattern and also to identify specific phosphopeptide biomarkers of nutrient stress in these organisms.

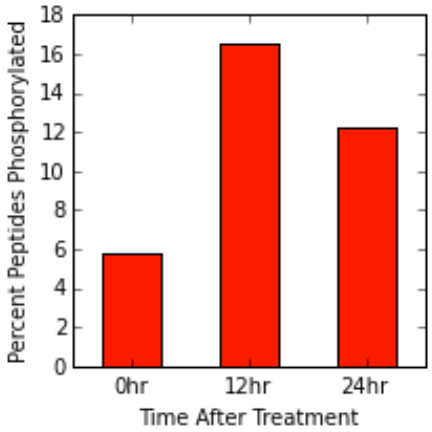


Figure A1.5. Percent peptides phosphorylated over time for *Prochlorococcus* sp. 9601 after exposure to nutrient deplete media.

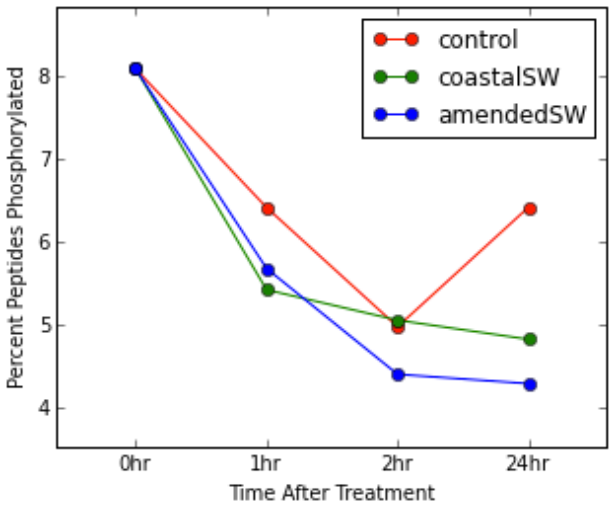


Figure A1.6. Percent peptides phosphorylated over time for *Alteromonas* sp. BB2AT2 after exposure to three different media.

A1.5 BRIEF CONCLUSIONS

This appendix provides a summary of our attempts to explore the phosphoptoeomes of two marine bacteria – *Prochlorococcus* and *Alteromonas*. Iterative re-analysis of the samples provided deeper coverage of the phosphoproteome, but at an analytical cost. Because it does not introduce chemical biases into the analysis, the iterative method described here may complement existing enrichment methods. It is particularly valuable for deep profiling of cultured organisms in order to identify phoshopeptides of interest that could then be profiled in more high-throughput analyses.

Phosphoproteomics is a valuable tool for understanding complex regulatory systems. Both *Prochlorococcus 9601* and *Alteromonas BB2AT2* have dynamic phosphoproteomes, supporting the idea that these species respond rapidly to environmental change. In particular, these phosphoproteomes suggest a fundamental difference in the regulatory responses of copiotrophs versus oligotrophs in response to nutrient starvation, with oligotrophs increasing regulatory events and copiotrophs decreasing them. This intriguing hypothesis should be clarified by profiling the phosphoproteomes of more marine organisms.

In some cases, the answers obtained from an experiment differ in the phosphoproteomic and proteomic analyses. This is an important consideration when identifying metabolic biomarkers, particularly in tightly regulated systems. As data collection and expertise grows, identification new regulatory systems or metabolic "cross-talk" in important marine microbes will be facilitated. Phosphoproteomics has great potential to be a hypothesis-generating tool; this paper acknowledges at least five unanswered questions about the identity or regulation of specific proteins. The ultimate goal is to bring this method to the field for better understanding of environmental function and identification of microbial biomarkers. However, before this can happen we will need to develop better understandings of these regulatory events in non-model marine organisms.

A1.6 LIST OF SUPPLEMENTARY TABLES

SA1.1 *Prochlorococcus* log and exponential phosphopeptides and phospho sites
SA1.2 *Alteromonas* log and exponential phosphopeptides and phospho sites
SA1.3 *Prochlorococcus* nutrient experiment phosphopeptides and phospho sites
SA1.4 *Alteromonas* nutrient experiment phosphopeptides and phospho sites

All are available at:
https://drive.google.com/drive/folders/1zic9tKx9gCvQzz6n5vFsxsjVe8Gl6asR?usp=sharing

A1.7 ACKNOWLEDGEMENTS

Special thanks to Matthew McIlvin for mass spectrometry training and maintenance. This work was supported by an NSF Graduate Research Fellowship grant (N. Held), the Moore Foundation and by the Gordon and Betty Moore Foundation grant number 3782 (M. Saito).

A1.8 REFERENCES

1. Stolarczyk, E. I., Reiling, C. J. & Paumi, C. M. Regulation of ABC transporter function via phosphorylation by protein kinases. *Curr. Pharm. Biotechnol.* **12**, 621–35 (2011).
2. Karve, T. M. & Cheema, A. K. Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease. *J. Amino Acids* **2011**, Article ID: 207691 (2011).
3. Lee, D. C. H., Jones, A. R. & Hubbard, S. J. Computational phosphoproteomics: From identification to localization. *Proteomics* 1–14 (2014). doi:10.1002/pmic.201400372
4. Macek, B., Mann, M. & Olsen, J. V. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol. Toxicol.* **49**, 199–221 (2009).
5. Van Mooy, B. a S. *et al*. Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. *Nature* **458**, 69–72 (2009).
6. Moore, L. R. *et al*. Culturing the marine cyanobacterium *Prochlorococcus*. *Limnol. Oceanogr. Methods* **5**, 353–362 (2007).
7. Lu, X. & Zhu, H. Tube-Gel Digestion: A Novel Proteomic Approach for High Throughput Analysis of Membrane Proteins. *Mol Cell Proteomics* **4**, 1948–1958 (2005).
8. Saito, M. A. *et al*. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **345**, 1173–7 (2014).
9. Bidle, K. D. & Azam, F. Bacterial control of silicon regeneration from diatom detritus: Significance of bacterial ectohydrolases and species identity. *Limnol. Oceanogr.* **46**, 1606–1623 (2001).

10. Batth, T. S., Francavilla, C. & Olsen, J. V. O ff -Line High-pH Reversed-Phase Fractionation for In-Depth Phosphoproteomics. *J. Proteome Res.* **13**, 6176-6186 (2014).
11. Spivak, M., Weston, J., Bottou, L., Kall, L., & Noble, W. S. Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. *J. Proteome Res.* **8**, 3737–3745 (2009).
12. Chalkley, R. J. & Clauser, K. R. Modification site localization scoring: strategies and performance. *Mol. Cell. Proteomics* **11**, 3–14 (2012).
13. Taus, T. *et al.* Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **10**, 5354–5362 (2011).
14. Martiny, A. C., Coleman, M. L. & Chisholm, S. W. Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc. Natl. Acad. Sci.* **103**, 12552–12557 (2006).
15. Traugh, J. & Traut, R. Phosphorylation of ribosomal proteins of *Escherichia coli* by protein kinase from rabbit skeletal muscle. *Biochemistry* 2503–2509 (1972).
16. Macek, B. *et al.* Phosphoproteome Analysis of *E . coli* Reveals Evolutionary Conservation of Bacterial Ser / Thr / Tyr Phosphorylation * *Annu. Rev. Pharmacol. Toxicol.* **49** 299–307 (2009).
17. Cox, A. D. & Saito, M. A. Proteomic responses of oceanic *Synechococcus* WH8102 to phosphate and zinc scarcity and cadmium additions. *Front. Microbiol.* **4**, 1–17 (2013).
18. Clayton, C. E. & Fox, J. Phosphorylation of fructose bisphosphate aldonale in *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **33**, 73–80 (1989).
19. Reddy, R. J. *et al.* Early signaling dynamics of the epidermal growth factor receptor. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 201521288 (2016).
20. Schmelzle, K., Kane, S., Gridley, S., Lienhard, G. E. & White, F. M. Temporal dynamics of tyrosine phosphorylation in insulin signaling. *Diabetes* **55**, 2171–2179 (2006).

APPENDIX 2. Phosphoproteomes of the North Atlantic surface ocean microbiome

A2.1 SUMMARY

The goal of this work was to identify phosphopeptides in a environmental metaproteome samples, focusing on serine/threonine phosphorylation. The motivation was to discover potential biomarkers for nutrient stress and biogeochemical processes in the ocean. Because protein phosphorylation is a dynamic process, phosphopeptides may be particularly sensitive biomarkers, and may be relevant on short time scales. In terms of the number of phosphopeptides identified, two-dimensional active modulation chromatography methods outperformed titanium dioxide phosphopeptide enrichment. Phosphopeptide identification was further enhanced by combining multiple database search engines and missed cleaveage cutoffs in the bioinformatics analysis. Phosphopeptides were clearly present in the ocean, however often the measurements were too patchy to allow for comparison across oceanic regimes. This indicated that we did not approach saturation of the phosphopeptide diversity in the surface ocean. This may reflect true biological diversity both geographically and temporally. It also highlights the need for further development to facilitate the consistent identification of phosphopeptides in marine metaproteome samples.

A2.2 INTRODUCTION

Phosphorylation is a key regulator of protein activity. As such, phosphopeptides may be good biomarkers for major biogeochemical processes of interest. Phosphoproteomics technology has matured rapidly in recent years, but it is rarely applied to complex samples such as metaprotomes that include signals from multiple organisms. This appendix describes some progress towards understanding the benefits and limitations of such phosphoproteome analyses in field ocean metaproteomes.

Phosphoproteomics analyses are difficult, particularly in the marine environment. First, access to sampling locations is limited and costly, making true replication nearly impossible to achieve.[1] Matrix effects complicate protein extraction and preservation and may be incompatible with certain phosphopeptide enrichment methods. Another key challenge is the large dynamic range of peptide intensities in marine metaproteome spectra, in particular because phosphopeptides tend to be low in abundance relative to other peptides. Even when a phosphorylation is identified, localization may be difficult owing to chimeric peaks in metaproteomics MS2 spectra.[1] In order to make consistent and comparable measurements across samples, both the modified and unmodified peptide must be consistently identified. Finally, while many amino acids can be phosphorylated, most existing technology focuses on serine/threonine phosphorylation. These are common in eukaryotes but may outshadowed by other types of phosphopeptides in bacteria. All of the above challenges mean that analysis of post-translational phosphorylation in metaproteome samples is uncommon at this time.

A2.3 MATERIALS AND METHODS

This appendix presents phosphoproteome profiles of the North Atlantic gyre sampled in 2017 on the JC150 cruise (see Chapter 3 and Figure 1). The expedition traversed major macro and micronutrient regimes, as well as different phytoplankton communities. Cells were collected via in situ pumps (McLane Corp) at 20m depth for seven stations. The data presented here is from the 0.2-3um filter fraction, largely consisting of marine phytoplankton such as *Prochlorococcus* and *Synechococcus*.

The proteins were extracted with SDS detergent and trypsin digested in-gel in the presence of phosphatase and protease inhibitors. A fraction of each sample was enriched for phosphopeptides by titanium dioxide using TiO_2 microcolumns (Glygen) in the presence of dihydroxybenzoic acid.[2,3] The spectra were analyzed on a Thermo Orbitrap Fusion with a Thermo Flex ion source using either a one-dimensinoal 140min gradient on a C18 column or a two-dimensional active-modulation method using in-line PLRP-S and C18 columns.

Database searches were conducted with different algorithms (Sequest and MSAmanda) using custom metagenomes from the Bermuda Atlantic Time Series (BATS). Dynamic phosphorylation of serine and threonine and up to 4 missed cleavages were permitted. Validation was performed at the 2% protein and peptide FDR level, which included the phosphopeptides. The actual FDRs were calculated in house using a custom Python script. Plotting and visualization was performed with the SciPy libraries Bokeh, Matplotlib, BioPython, and Seaborn.

Figure A2.1. Sampling locations and examples of the filters and collection techniques used.

A2.4 RESULTS and DISCUSSION

A2.4.1 Utility of orthogonal chromatography for phosphoproteome analyses

Environmental metaproteomes are complex owing to high sequence diversity coupled with high dynamic range in peak intensity.[1] This is a problem especially for phosphopeptide analyses since these are generaly low abundance. As expected, the surface seawater samples analyzed in this study retained significant complexity in a standard 120min LC-MS/MS run (Figure 2A). Enrichment for serine

and threonine phosphopeptides reduced this complexity somewhat, indicating that the enrichment did remove many high-abundance non phosphorylated peptides from the sample mixture (Figure 2B). To further reduce the complexity of the mixture reaching the mass spectrometer, we extended the chromatographic separation of the sample by using two dimensions of chromatography, a PLRP-S and a C18 column. Both separate on hydrophobicity but have slightly chemical characters. In this method, the eluent from the PLRP-S column was collected in 30min increments over 8 hours and then introduced onto the C18 column (Figure 2C).

In our hands, the two-dimensional chromatography method resulted in more peptide and phosphopeptide identifications than the typical 1D 120min LC-MS/MS runs, including after titanium dioxide phosphopeptide enrichment (Figure 3). Titanium dioxide enrichment reduces the number of non-phosphorylated peptides in the mixture (Figure 3A), but did not increase the number of phosphopeptides identified (Figure 3B). The possible reasons for this are multiple. First, titanium dioxide is biased towards specific types of phosphopeptides, particularly multiply phosphorylated peptides.[4] It is only selective towards serine/threonine phosphorylation, which are common in eukaryotes but are less important in These cyanobacteria, which are the primary type of organism in these size fractions. Additional types of phosphopeptide enrichment may therefore be needed to capture the diversity in these field samples. Additionally, the complexity of the sample may reduce the efficiency of binding and elution to the titanium dioxide beads, which are most often used for collections of one organism only.

Compared with the titanium dioxide enrichment method, introducing a second dimension of chromatography nearly doubled the number of peptides and phosphopeptides identified (Figure 3A and B). The main benefit of this long chromatography method was that it reduced the complexity of the sample mixture reaching the mass analyzer, thus allowing for more MS2 spectra to be collected on low abundance phosphopeptides. The high performance of this method for phosphopeptide identification indicated that it may be fruitful to search for phosphopeptides in existing spectra collected using this method, which is becoming standard for field metaproteome analysis in the Saito lab group.

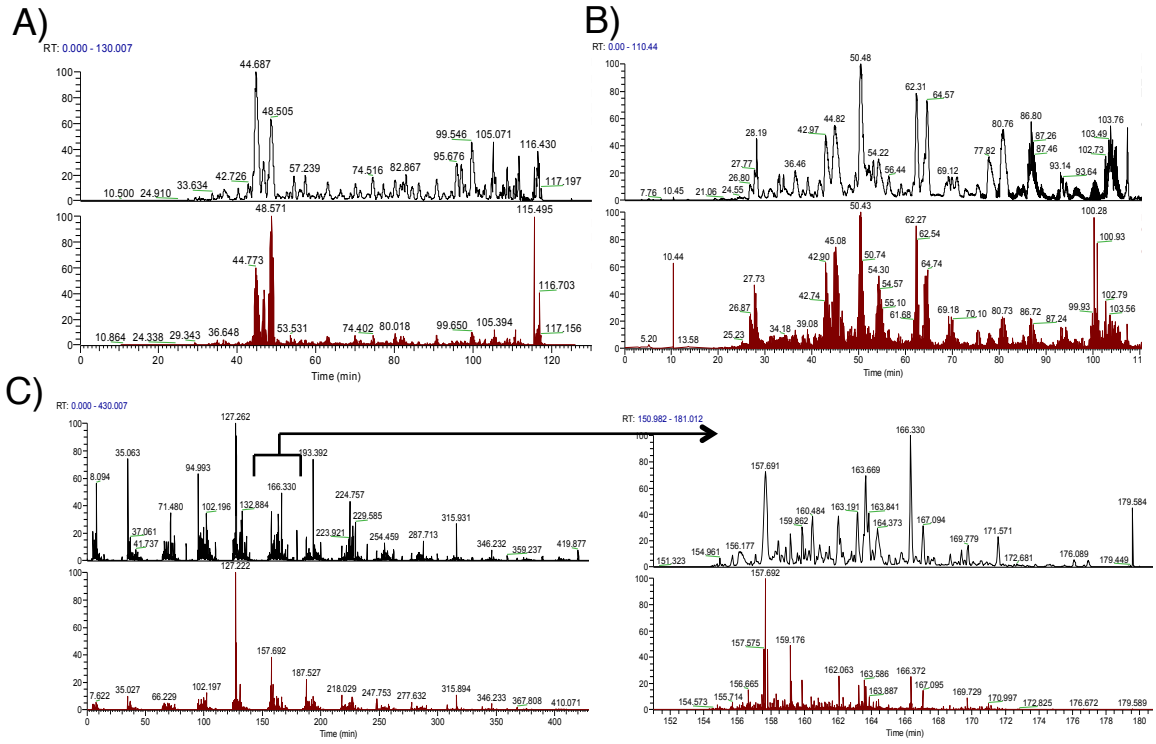


Figure A2.2. Chromatography trace of A) 120min 1D LC-MS/MS analytical run, B) 120min 1D LC-MS/MS analytical run after titanium dioxide enrichment, and C) 480min LCaXmLC/MS/MS run, with a zoomed in panel on the right. The complexity of the metaproteome sample is quite high even after one dimension of chromatography, as exemplified by panel C. However, two dimensions of orthogonal chromatography are able to significantly reduce complexity.

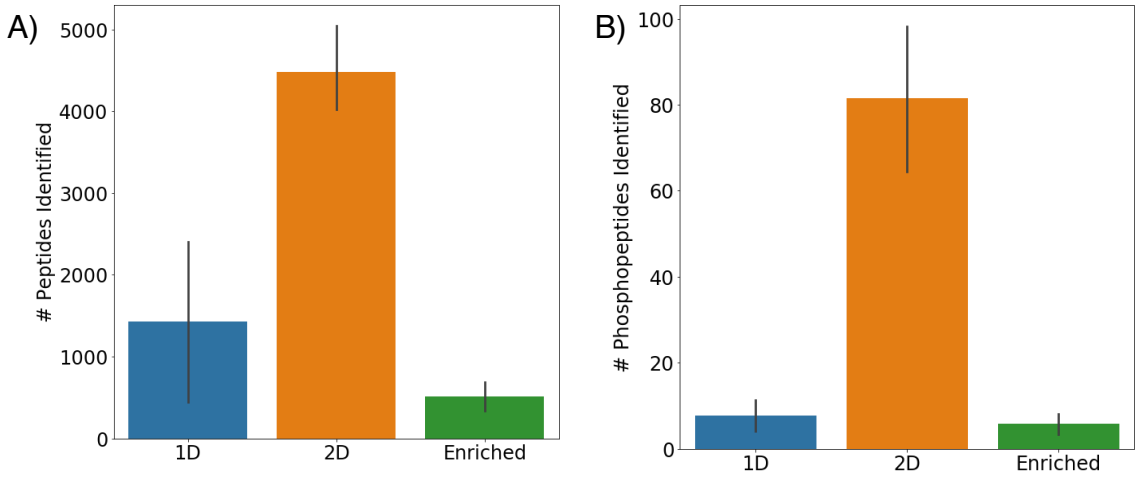


Figure A2.3. Number of peptide and phosphopeptide identifications made across all samples in 1D, 2D, and TiO_2 enriched samples. For both peptides and phosphopeptides, the number of identifications was significantly improved by using two dimensions of chromatography. This is the average number of peptides/phosphopeptides identified across the samples, and the error bar indicates one standard deviation.

A2.4.2 Search algorithm comparisons

Each peptide-to-spectrum matching algorithm has its own biases and thus results in different sets of peptides identified.[5] Indeed, the peptides and phosphopeptides identified at the 1% protein and peptide FDR levels were different for two algorithms commonly used in the Saito lab, SEQUEST and MSAmanda. This indicated that it was beneficial to use multiple algorithms to capture complementary sets of peptides nd phosphopeptides (Figure 4A and B). In addition, because phosphorylation tends to reduce the efficiency of trypsin digestion, increasing the number of missed cleaveages in the peptide-to-spectrum matching analysis resulted in more identifications (Figure 4B).

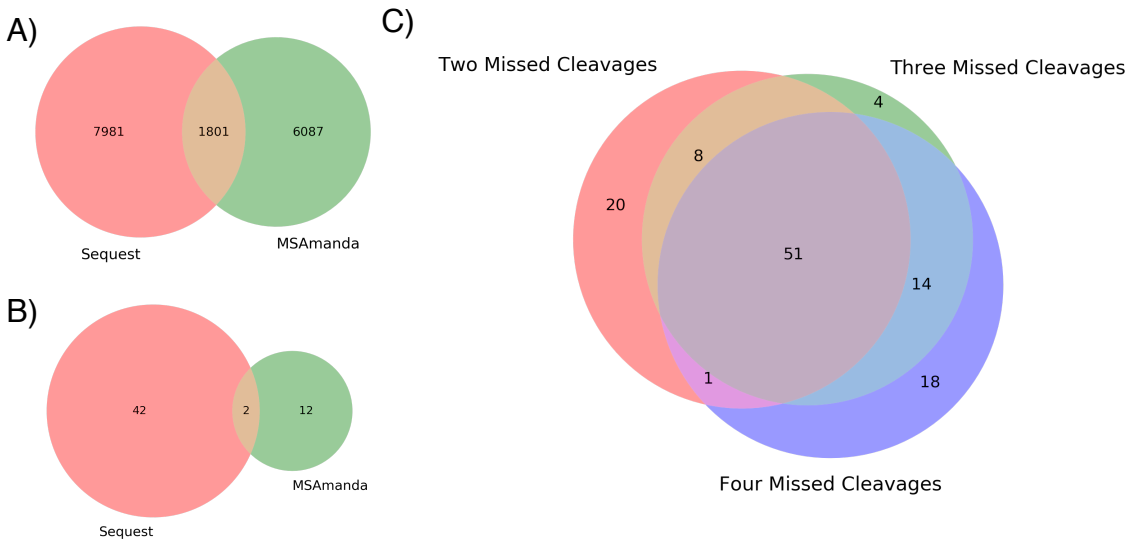


Figure A2.4. Effect of search engine on (A) peptide and (b) phosphopeptide identifications. The search engines capture complementary sets of peptides and phosphopeptides in the metaproteomes. In (C), the effect of modulating the number of allowed missed cleaveages in a SEQUEST search of the data. Because phosphorylation may reduce digestion efficiency, increasing the number of possible missed cleavages results in additional phosphopeptide identifications.

A2.4.3 Assessing completeness of the phosphopeptide analysis

To assess the completeness of the phosphoproteome profiles we generated, we borrowed the idea of rarefaction from the field of ecology. Rarefaction is typically used to assess species richness in an environment, and is performed by sub-sampling multiple times and identifying the number of new identifications made after each sampling event. Sampling saturation is reached when the number of new identifications plateaus. We note that we were not the first to apply this technique to peptide and phosphopeptide analyses.[6]

Keeping in mind the goal of profiling peptide and phosphopeptide diversity across North Atlantic subtropical surface waters, it was clear that saturation was not achieved in this study. With each additional sample analyzed, the number of unique peptides and phosphopeptides increased linearly (Figure 5A and C). However, the number of identifications that represented new peptides did decrease with each successive sample (Figure 5B), indicating that saturation was beginning to be reached. With each new sample analyzed, the % new peptide identifications dropped by 20%, so following logically from this saturation may be reached with analysis of just a few more samples. This occurred in spite of the fact that population shifts occurred across these samples, specifically a shift from a *Prochlorococcus* to *Synechococcus* dominated community moving West to East. It indicates that the diversity of peptides in the surface North Atlantic is relatively small, reflecting phylogenetic similarity among these communities.

The story for the phosphopeptide analysis was different. Unlike for the peptides, the phosphopeptides identified in each new sample were nearly all "new" analytes (Figure 5D). The number of new identifications did not decrease as more samples were analyzed. This indicated incomplete profiling of phosphopeptide diversity in the surface ocean and likely reflects the large amount of phosphopeptide diversity. Because protein phosphorylations are short lived and dynamic, different communities sampled at different locations, times of day, and experiencing different environmental conditions likely have completely different sets of phosphopeptides, despite similar peptide diversity. [7–9] Because most phosphopeptides were identified only in one sample across the dataset, it was impossible to compare abundances across oceanic regimes.

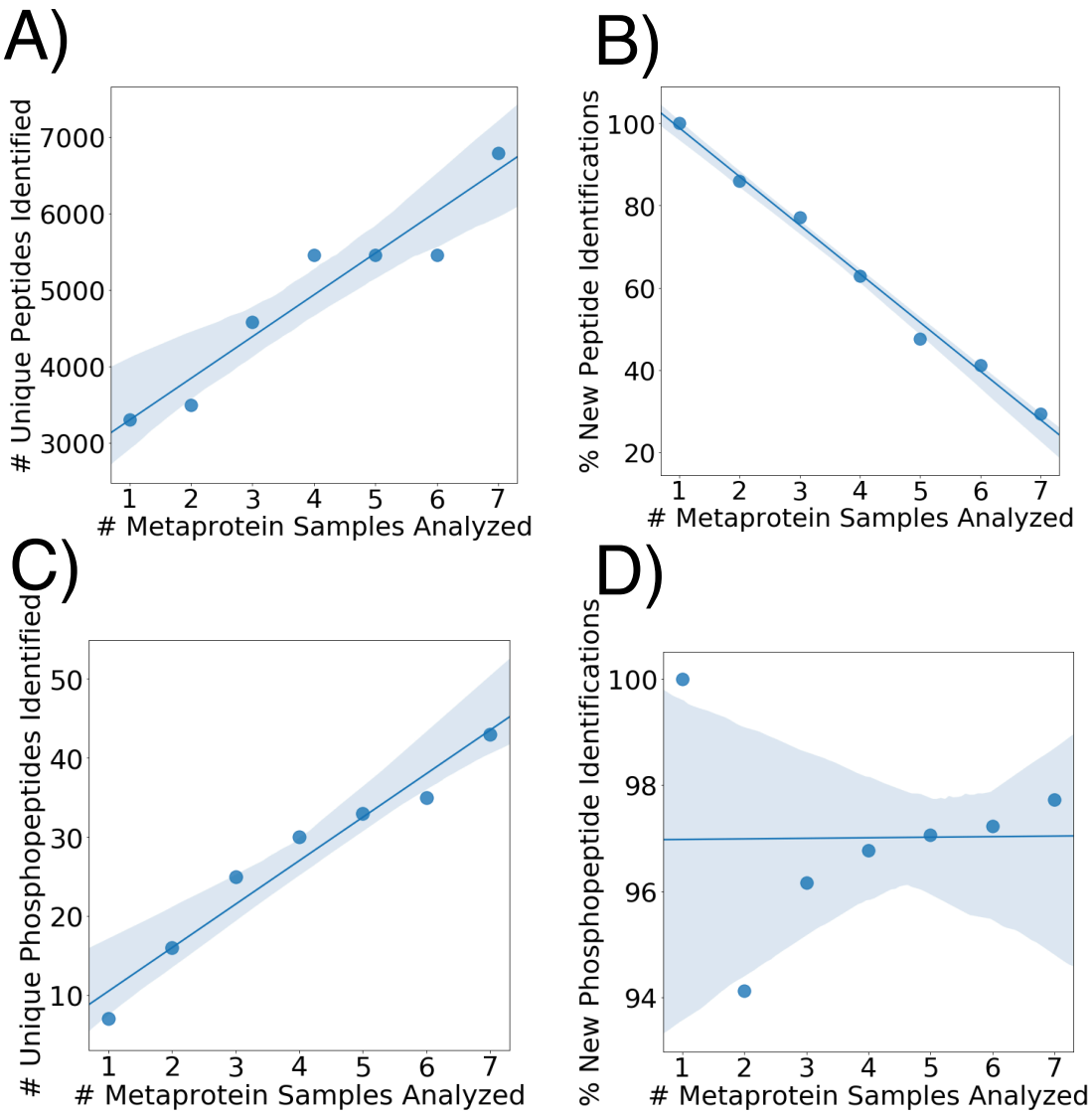


Figure A2.5. Rarefaction curves. A) The number of unique peptides identified across the entire dataset increases as more samples are analyzed, however the percent of new identifications decreases (B). For phosphopeptides, more peptides are identified s more samples are measured and nearly all of the phosphopeptides identified are new to the dataset (D). This indicates a high level of diversity in the phosphopeptide pool versus the peptide pool, and demonstrates that we only scratch the surface of phosphopeptide identifications in the surface ocean.

A2.5 BRIEF CONCLUSIONS

There is still a long way to go before phosphoproteomics measurements become commonplace in oceanography. This appendix highlights some major challenges in this work moving forward. A primary direction of future work will be continued testing of phosphoproteomics enrichment methods and other ways to reduce sample complexity. It is intriguing that the long, two-dimensional chromatography method surpassed traditional titanium dioxide enrichment for analysis of the serine/threonine phosphoproteome, because this means that it may be possible to extract phosphopeptide information from typical field metaproteome spectra. However, significant work is needed to improve coverage of the phosphoproteome in metaproteomics samples. The most viable direction for future work will be to target specific phosphopeptides of interest, which can be identified in culture studies. This cutting edge research will be further facilitated by improvements in database searching algorithms and mass spectrometry analysis allowing for faster peak identification and spectrum analysis such that more low-abundance ions are selected for MS2s.

Phosphopeptides are clearly important to marine microbes and likely have much to tell us about marine microbial physiology. This work, while unable to make oceanographic conclusions, clearly demonstrates the presence and diversity of phosphopeptides in the ocean environment. In particular, this study highlights the need for continued profiling of phosphopeptides in cultured marine microbes to help us to develop databases and understandings from which field studies can benefit.

A2.6 ACKNOWLEDGEMENTS

The authors thank Chris Dupont and at JCVI for the custom North Atlantic metagenome database. This work was supported by an NSF Graduate Research Fellowship grant (N. Held), the Moore Foundation and by the Gordon and Betty Moore Foundation grant number 3782 (M. Saito). Visualization development supported by National Science Foundation grant EarthCube 1639714.

A2.7 SUPPLEMENTARY TABLES (LIST)

Table SA2.1 Phosphopeptides identified across the sampling sites
Table SA2.2 Quantitative data for phosphopeptides and peptides across the sampling sites

All are available at:
https://drive.google.com/drive/folders/1zic9tKx9gCvQzz6n5vFsxsjVe8Gl6asR?usp=sharing

A2.8 REFERENCES

1. Saito, M. A. *et al*. Progress and Challenges in Ocean Metaproteomics and Proposed Best Practices for Data Sharing. *J. Proteome Res*. **18**, 1461–1476 (2019).
2. Aryal, U. K. & Ross, A. R. S. Enrichment and analysis of phosphopeptides under different experimental conditions using titanium dioxide affinity chromatography and mass spectrometry. *Rapid Commum. Mass Spectrom*. **24**, 219–231 (2010). doi:10.1002/rcm
3. Thingholm, T. E., Jørgensen, T. J. D., Jensen, O. N. & Larsen, M. R. Highly selective enrichment of phosphorylated peptides using titanium dioxide. *Nat. Protoc*. **1**, 1929–35 (2006).
4. Kweon, H. K. & Håkansson, K. Selective zirconium dioxide-based enrichment of phosphorylated peptides for mass spectrometric analysis. *Anal. Chem*. **78**, 1743–9 (2006).
5. Dorfer, V. *et al*. MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J. Proteome Res*. **13**, 3679–3684 (2014).
6. Boekhorst, J., Boersema, P. J., Tops, B. B. J., Breukelen, B. Van & Heck, A. J. R. Evaluating Experimental Bias and Completeness in Comparative Phosphoproteomics Analysis. **6**, 1–8 (2011).
7. Reddy, R. J. *et al*. Early signaling dynamics of the epidermal growth factor receptor. *Proc. Natl. Acad. Sci*. **113**, 201521288 (2016).
8. Schmelzle, K., Kane, S., Gridley, S., Lienhard, G. E. & White, F. M. Temporal dynamics of tyrosine phosphorylation in insulin signaling. *Diabetes* **55**, 2171–2179 (2006).
9. Nita-Lazar, A., Saito-Benz, H. & White, F. M. Quantitative phosphoproteomics by mass spectrometry: Past, present, and future. *Proteomics* **8**, 4433–4443 (2008).

APPENDIX 3. Progress towards phospho-histidine profiling of *Trichodesmium*

A3.1 SUMMARY

In this appendix I describe progress made towards identifying phosphohistidine (pHis) containing proteins in the marine cyanobacterium *Trichodesmium*. Histidine phosphorylation is a common protein modification in bacteria but is understudied owing to difficulties in preserving the modification throughout LC-MS/MS analyses. Western blot experiments using antibodies for the 1-pHis and 3-pHis isomers confirmed the presence of these moieties in *Trichodesmium*. An immunoprecipitation experiment identified some proteins that may contain phosphohistidine sites. This work is preliminary but provides a glimpse into the knowledge that can be gained from further development of pHis proteomics methods, particularly in the study of prokaryotes.

A3.2 INTRODUCTION and GOALS

The goal of this work was to identifying phosphohistidine (pHis) containing proteins in *Trichodesmium erythraeum* sp. IMS101. The motivation was to begin to characterize some of the 35 two-component sensory systems in the *Trichodesmium* genome, most of which have no functional annotation. Two-component sensory systems are activated by histidine phosphorylation. Previous work has demonstrated their potential importance as biomarkers of biogeochemical processes (see Chapter 2).

Historically it has been difficult to observe histidine phosphorylation in LC-MS/MS experiments. This is because phosphorylation of histidine occurs via a phosphoamidate bond, which is weaker than the phosphoester bonds that occur when serine and threonine are phosphorylated. Phosphoamidate bonds rapid hydrolysis under the acidic conditions typically used in LC-MS/MS proteomic analyses.[1,2] Additionally, phosphohistidine exists in two isomeric forms (1pHis and 3pHis). These challenges mean that current understandings of phosphohistidine is very limited, despite their importance in bacterial species.

In this appendix I summarize preliminary attempts at identifying phosphohistidine containing proteins in *Trichodesmium*. Using antibodies provided by Drs. Tony Hunter and Kevin Adam (Salk Institute), I confirmed the presence of phosphohistidine in *Trichodesmium* and identified some pHis containing proteins of interest [3]. While I was not able to reach the ability to quantify and compare pHis proteins across Trichodesmium samples, I was able to make some early progress towards this end. Most significantly, I demonstrate the presence and thus potential importance of pHis analysis for understanding the physiology of *Trichodesmium* and other marine cyanobacteria, and highlight the need for continued development of pHis measurement technology.

A3.3 MATERIALS AND METHODS

A3.3.1 Culturing/sampling conditions

Trichodesmium was either obtained from the field or the laboratory. Field samples were taken by hand-picking colonies from plankton nets conducted in the North Atlantic

surface ocean on 7/14/17 at 2:30am local time at 22°N, -50 °W. Cultured *Trichodesmium erythraeum* sp. IMS101 was grown following the procedures outlined in Chapter 5 and was sampled in the early morning hours, shortly after the incubator lights turned on. In all cases cells were collected on 0.2μm Supor filters and flash frozen at -80°C until analysis.

A3.3.2 Western Blot experiments

All steps were performed in the cold room to minimize loss of the pHis signal, which is acid and heat labile.[3] Proteins were extracted by soaking sample filters in sample buffer (10mM tris HCL, 30% glycerol, 50mM EDTA, 2% SDS, pH 8.8) with vigorous shaking for one hour at room temperature. The filters were then removed and sample clarified by centrifugation at 4°C. The supernatant was then sonicated with a tip-sonicator for 30s on ice, then allowed to rest and the process repeated five times. Care was taken to avoid heating up the sample during sonication. Negative controls were prepared by acidifying an aliquot of each sample to pH 4 with 10% trifluoroacetic acid, heating for 10 minutes at 95°C, cooling the sample, and finally neutralizing to pH 7 with 100mM ammonium bicarbonate. Protein concentrations were measured with a DC assay kit (BioRad). Finally, bromophenol blue was added to assist in sample loading (it was added last because it interferes with the DC assay).

Electrophoresis gels were run in the cold room using cold buffer solutions. The gels were poured in house and consisted of a 4% acrylamide stacking gel, pH 8.8 and a 12% acrylamide resolving gel, also pH 8.8. 10μL corresponding to approximately 20μg of each sample including negative controls were loaded. The ladder was the PageRule PLUS protein ladder (Fisher). The samples were resolved at 100V for 2-3 hours using a modified running buffer (1g SDS, 3g trizma, 14.4 g glycine per L, pH 8). Proteins were transferred to PVDF membranes in modified transfer buffer (1g SDS, 15g trizma, 14.1g glycine, 200mL methanol per L, pH 8) at 200A for 3 hours at 4°C. Membranes were either stained for total protein content using the REVERT protein stain kit (LiCor), or proceeded for pHis experiments. For the latter, membranes were blocked in Odyssey blocking buffer (LiCor) for two hours with gentle shaking at 4°C, then washed twice in TBST buffer (tris-buffered saline + tween20) and incubated with the 1pHis, 3pHis, or pNDK antibodies overnight. The next day, membranes were washed again and then incubated with the fluorescent secondary antibody (Alexafluor anti-rabbit 700nM) for two hours room temperature before imaging with an Odyssey Imaging System (LiCor).

A3.3.3 Immunoprecipitation and LC-MS/MS workflows

Proteins were extracted in a similar way as in the Western blots. Cells were lysed in sample buffer containing 8M urea, 50mm tris HCL pH 8.8, 1% SDS. They were shaken overnight in this solution at 4°C, then the filter removed and supernatant clarified by centrifugation. Sonication was performed using a tip-sonicator for 30s on ice, then the samples rested and the process repeated five times. The proteins were then precipitated overnight in ice-cold methanol, collected by centrifugation, and suspended in sample buffer. Next, the proteins were treated with dithiothreitol (DTT) at 10mM final concentration for one hour at room temperature with shaking. This was followed by one

hour treatment with 10mM final concentration iodacetamide (IODA) at room temperature with shaking. Finally, the proteins were digested by incubating the samples with trypsin (1μg trypsin/100μg total protein) for at least 18 hours at room temperature with shaking. A dot blot was performed, similar to the above Western blot experiments, to confirm that pHis residues were preserved throughout this process (Figure 3).

The immunocprecipitation was conducted using antibody conjugated IgG beads provided by Kevin Adam (Salk Institute). The column was washed in ammonium bicarbonate pH 8.8. The peptides were introduced and flow-through collected; this was repeated three times. The column was washed three times with ammonium bicarbonate pH 8.8. Then, the pHis containing peptides were eluted with 100mM triethylamine. Two elution fractions were collected and combined for LC-MS/MS analysis. The eluent was dried to completion in a speedvac and then reconstituted in LC-MS/MS buffer B, pH 4. Columns were washed at least four timex with phosphate buffered saline and stored in the same at 4C.

LC-MS/MS analyses were conducted with a 120min standard chromatography run on a Thermo Orbitrap fusion in DDA mode. Peptides/proteins were identified by a SEQUEST search against the *Trichodesmium erythraeum* sp. IMS101 genome.

A3.4 RESULTS and CHALLENGES

A3.4.1 Confirming the presence/maintenance of pHis proteins in Trichodesmium

This work confirmed the presence of pHis protein modifications in *Trichodesmium erythrauem* sp. IMS101, as well as in field populations collected 7/14/17 at 2:30am local time at 22°N, -50 °W. Significant efforts were made to maintain the pHis moieties throughout the sample preparation and analysis stages. The most significant was the use of sonication instead of heat to extract proteins from the cell lysate and disrupt their tertiary and quaternary structure. In addition to sonication, using in-house poured gels with neutral pH stacking gels was also key for maintaining the pHis signal. These steps were crucial for maintaining the pHis modification but did result in sub-optimal separation of proteins during the gel electrophoresis stage.

Phosphohistidine signals were very consistent across samples and experiments (see Figures A1 and A2). Encouragingly, signals were not identified in negative controls in which the pHis signals were deliberately destroyed. This indicated that Trichodesmium does indeed modify some of its histidine residues.

In addition to 1 and 3pHis antibodies, we also tested antibodies for phosphorylated nucleoside diphosphate kinase (pNDK). Nucleoside diphosphate kinase is a conserved in prokaryotes and eukaryotes alike. The antibody tested was developed for human cells, but the protein region that it recognizes is similar in *Trichodesmium*. Indeed, we were able to identify both the monomeric and hexameric forms of pNDK in both field and laboratory samples (see Figures 1 and 2). In marine diatoms, nucleoside diphosphate kinase is a biomarker for cell growth.[4] The phosphorylated form represents the reactive intermediate of the catalytic cycle. The ability to measure nucleoside diphosphate activity

"in action" may therefore provide a target for growth rate measurements from "omics" data in *Trichodesmium*.

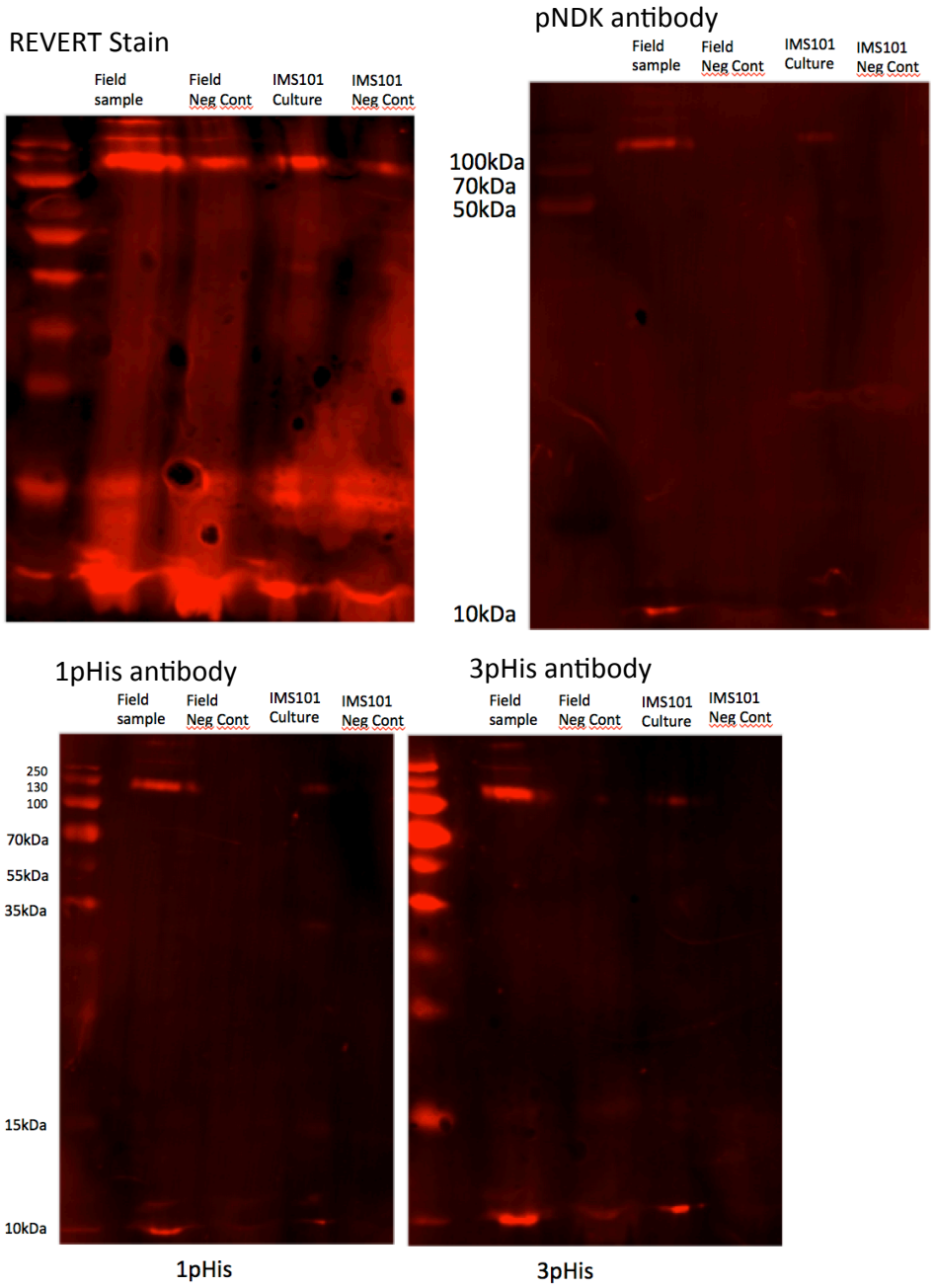


Figure A3.1. An early Western blot experiment on field and cultured Trichodesmium, looking for pNDK, 1pHis, and 3pHis containing proteins.

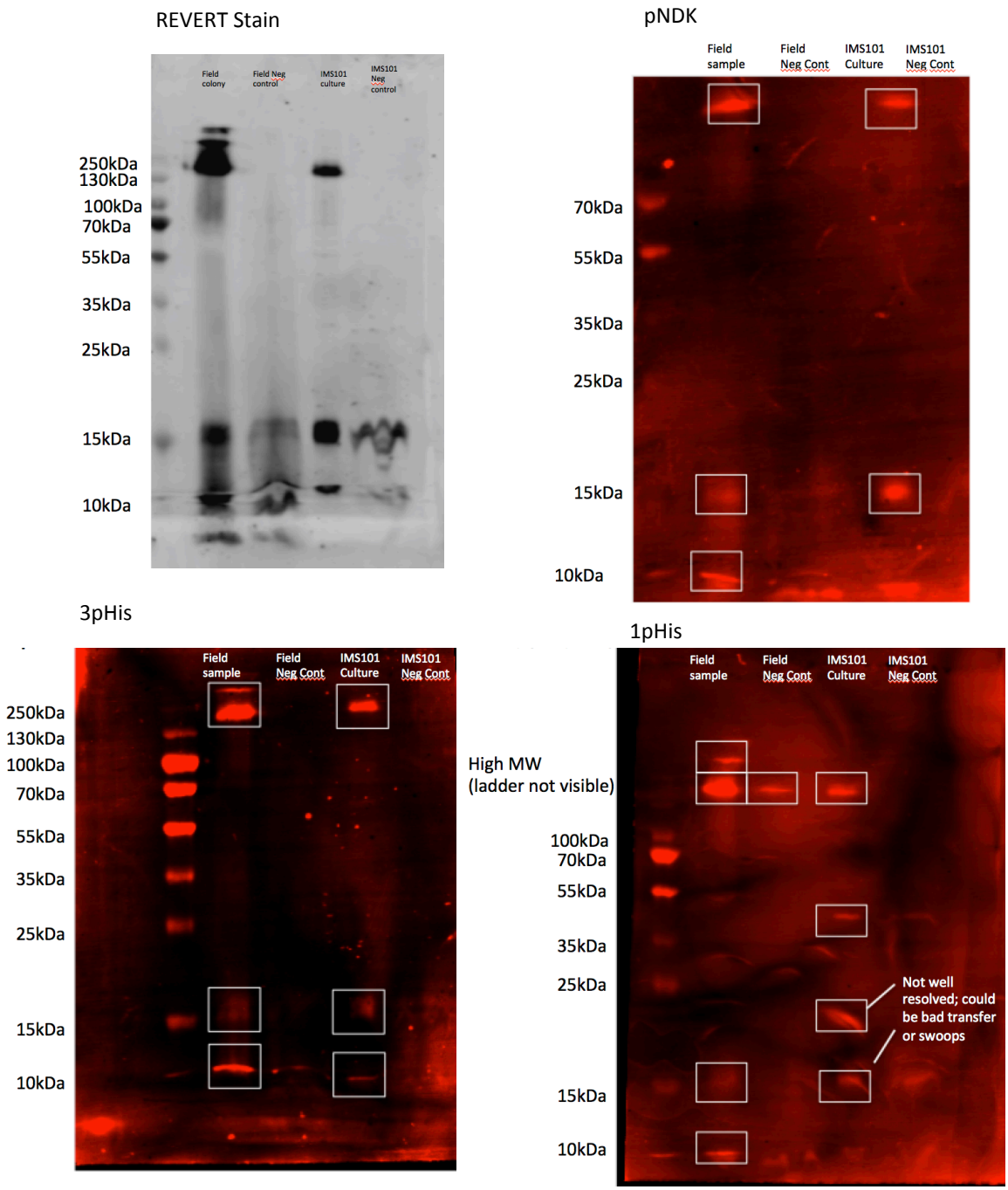


Figure A3.2. A later Western blot experiment on field and cultured *Trichodesmium*, looking for pNDK, 1pHis, and 3pHis containing proteins. The results were very consistent with the previous Western blot experiments and more examples exist.

A3.4.2 Identifying pHis containing proteins in Trichodesmium

The next step was pHis immunoprecipitation of pHis containing peptides to facilitate their identification by LC-MS/MS. The enrichment was performed separately

for the 1 and 3 pHis isomers on a field sample collected 7/14/17 at 2:30am local time at 22°N, -50 °W, and on cultured *Trichodesmium erythraeum* sp. IMS101. This latter wa not an axenic culture but it had been maintained/transferred for years in sterile conditions. The basic workflow is that the proteins are extracted and gently denatured. Dot blots were performed with the antibodies to ensure that the phosphohistidine signals were retained. Then they were enriched by immunoprecipitation on a small column containing the antibody conjugated beads. Elution occurred by competition with concentrated triethylamine. The resulting eluent was then analyzed by LC-MS/MS in DDA mode. This allowed us to identify phosphohistidine containing proteins but not necessarily the site of the histidine phosphorylation. The list of proteins is presented in Table 1.

Phosphohistidine containing peptides were identified in both the 1 and 3pHis immunoprecipitations. However, most peptides were identified using the 3pHis antibody suggesting possible biological bias towards this isomer in *Trichodesmium*. This is consistent with the Western blot experiments where the 3pHis signal was generally stronger than the 1pHis signal. An additional factor may be that the 1pHis isomer may be more labile than the 3pHis isomer (Kevin Adam, personal communication). Nucleoside diphosphate kinase appeared in the 3pHis fraction despite being in the 1pHis fraction in human cells. More identifications were made for the cultured *Trichodesmium* sample as opposed to the field sample, likely because more biomass was available and because the sequence database was more representative.

Many of the putative pHis containing peptides came from highly abundant proteins such as GroEL and phycobilisome proteins. This is suspicious and may reflect imperfect separation of pHis and non-pHis containing peptides. Many protein kinases were also identified, however most were serine/threonine protein kinases. Some such kinases have been demonstrated to contain pHis, which is used to regulate kinase activity.[5,6] Only one two component system protein, the histidine kinase Tery_2216, was observed in the pHis fractions.

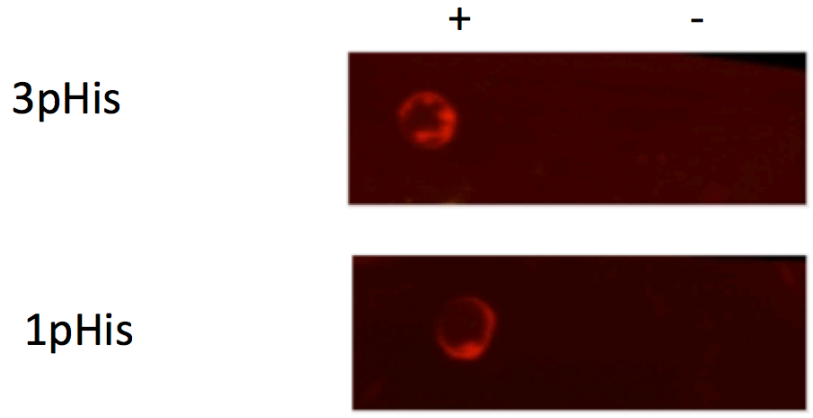


Figure A3.3. Dot blots of the *Trichodesmium erythraeum* sp. IMS101 peptides (positive and negative controls) demonstrating that the pHis signals were conserved throughout the protein extraction and digestion process.

A3.5 BRIEF CONCLUSIONS and FUTURE DIRECTIONS

It is difficult to benchmark the success of this work because it is the first investigation of phosphohistidine in *Trichodesmium,* and to my knowledge in any cyanobacterium. It is clear that we were successful in demonstrating the presence of phosphohistidine in *Trichodesmium*, both in the laboratory and in the field. The data is too patchy to facilitate biological interpretation of the pHis proteins identified, but this appendix does provide a list of potential proteins of interest, many of them of biogeochemical relevance. One particularly interesting opportunity lies in the potential of phosphorylated nucleoside diphosphate kinase (pNDK) as a possible indicator of growth rate, though significant work will be needed to benchmark this.

New technology continues to emerge for phosphohistidine measurements. Recent work suggests that slightly increasing the pH at which the sample is ionized can allow for phosphohistidine sites to be observed, even without immunoprecipitation enrichment.[1,7] It will be prudent to keep an eye on this exciting field moving forward so that the biogeochemical relevance of these regulatory phosphorylations can be investigated.

A3.6 ACKNOWLEDGEMENTS

I sincerely thank Dr. Kevin Adam and Dr. Tony Hunter (Salk Institute) for their assistance and willingness to allow us to test their antibodies in *Trichodesmium*. Thanks also to Alexander Devaux for help with the Western blot experiments. This work supported by an NSF Graduate Research Fellowship grant (N. Held), the Moore Foundation and by the Gordon and Betty Moore Foundation grant number 3782 (M. Saito).

A3.7 REFERENCES

1. Oslund, R. C. *et al*. A phosphohistidine proteomics strategy based on elucidation of a unique gas-phase phosphopeptide fragmentation mechanism. *J. Am. Chem. Soc*. **136**, 12899–12911 (2014).
2. Kee, J. M., Villani, B., Carpenter, L. R. & Muir, T. W. Development of stable phosphohistidine analogues. *J. Am. Chem. Soc*. **132**, 14327–14329 (2010).
3. Fuhs, S. R. *et al*. Monoclonal 1- and 3-Phosphohistidine Antibodies: New Tools to Study Histidine Phosphorylation. *Cell* **162**, 198–210 (2015).
4. Bacillariophyceae, T. P., Berges, J. A. & Harrison, P. J. Light-Limited Growth Rate in the Marine Diatom. *Cell* **53**, 45–53 (1993).
5. Schramm, A., Lee, B. & Higgs, P. I. Intra- and interprotein phosphorylation between two-hybrid histidine kinases controls myxococcus xanthus developmental progression. *J. Biol. Chem*. **287**, 25060–25072 (2012).
6. Ross, a. R. S. Identification of Histidine Phosphorylations in Proteins Using Mass Spectrometry and Affinity-Based Techniques. *Methods Enzymol*. **423**, 549–572 (2007).
7. Gonzalez-Sanchez, M.-B., Lanucara, F., Helm, M. & Eyers, C. E. Attempting to rewrite History: challenges with the analysis of histidine-phosphorylated peptides.

Biochem. Soc. Trans. **41**, 1089–95 (2013).

A3.8 TABLES

Table A3.1. Results of immunoprecipitation experiments – proteins that may contain pHis sites.

Key
Identified in 1 and 3phis cultures
Identified in 3pHis in field and in cultures
Identified in 1pHis in field and in cultures
1pHis field only
3pHis field only
1pHis culture only
3pHis culture only

Description
Tery_0476 tuf translation elongation factor 1A (EF-1A/EF-Tu)
Tery_0984 Phycobilisome protein
Tery_4327 groL2 chaperonin GroEL
Tery_0270 groL1 chaperonin GroEL
Tery_3385 atpD ATP synthase F1 subcomplex beta subunit
Tery_5049 phycocyanin, beta subunit
Tery_0162 Redoxin
Tery_0983 Phycobilisome protein
Tery_2362 peptidase S8 and S53, subtilisin, kexin, sedolisin
Tery_3650 allophycocyanin alpha subunit apoprotein
Tery_2199 atpA ATP synthase F1 subcomplex alpha subunit
Tery_5048 phycocyanin, alpha subunit
Tery_4030 glyceraldehyde 3-phosphate dehydrogenase (NADP+) (EC 1.2.1.13)
Tery_3834 L-glutamine synthetase (EC 6.3.1.2)
Tery_2623 ahcY adenosylhomocysteinase (EC 3.3.1.1)
Tery_2376 hypothetical protein
Tery_3286 putative transposase, IS891/IS1136/IS1341
Tery_3649 allophycocyanin beta subunit apoprotein
Tery_0454 psaC 4Fe-4S ferredoxin, iron-sulfur binding
Tery_1926 aldehyde dehydrogenase
Tery_0247 serine/threonine protein kinase
Tery_3498 Cna B-type
Tery_1486 hypothetical protein
Tery_3670 DNA binding domain, excisionase family
Tery_3727 RDD domain containing protein

Tery_0565 Na+/solute symporter
Tery_3284 PfkB
Tery_0702 hypothetical protein
Tery_2609 serine/threonine protein kinase
Tery_4249 AAA ATPase, central region
Tery_4012 dnaK chaperone protein DnaK
Tery_2688 lipopolysaccharide biosynthesis
Tery_2486 Phycobilisome linker polypeptide
Tery_0387 Cadherin
Tery_2677 hypothetical protein
Tery_2039 rimK SSU ribosomal protein S6P modification protein
Tery_1538 glycosyltransferase
Tery_1462 primary replicative DNA helicase (EC 3.6.1.-)
Tery_0600 hypothetical protein
Tery_4027 conserved hypothetical protein 103
Tery_0998 Phycobilisome protein
Tery_4410 cbbL ribulose 1,5-bisphosphate carboxylase large subunit (EC 4.1.1.39)
Tery_3791 photosystem I protein PsaD
Tery_4099 fructose-bisphosphate aldolase (EC 4.1.2.13)
Tery_1749 grpE GrpE protein
Tery_4326 groS chaperonin Cpn10
Tery_3959 Redoxin
Tery_2563 plastocyanin
Tery_4136 nifH Mo-nitrogenase iron protein subunit NifH (EC 1.18.6.1)
Tery_0948 sedoheptulose 1,7-bisphosphatase (EC 3.1.3.37)/D-fructose 1,6-bisphosphatase (EC 3.1.3.11)
Tery_4137 nifD Mo-nitrogenase MoFe protein subunit NifD precursor (EC 1.18.6.1)
Tery_0996 Phycobilisome protein
Tery_3376 pgk phosphoglycerate kinase (EC 2.7.2.3)
Tery_2325 gvpA gas vesicle protein GVPa
Tery_0262 rplL LSU ribosomal protein L12P
Tery_1666 fld flavodoxin
Tery_3314 Tetratricopeptide TPR_2
Tery_3654 phosphoribulokinase (EC 2.7.1.19)
Tery_4104 Phycobilisome linker polypeptide
Tery_1610 bacterial nucleoid protein Hbs
Tery_1306 dnaK chaperone protein DnaK
Tery_4702 RNA-binding region RNP-1
Tery_3849 Carbonate dehydratase
Tery_3312 thioredoxin
Tery_4335 ilvC ketol-acid reductoisomerase (EC 1.1.1.86)
Tery_5038 1-Cys peroxiredoxin (EC 1.11.1.15)
Tery_0847 metE methionine synthase (B12-independent) (EC 2.1.1.14)

Tery_1891 peptidase-like
Tery_1892 peptidase-like
Tery_4773 hypothetical protein
Tery_1460 rplI LSU ribosomal protein L9P
Tery_3372 beta-Ig-H3/fasciclin
Tery_3386 atpC ATP synthase F1 subcomplex epsilon subunit
Tery_3659 hypothetical protein
Tery_4408 ribulose 1,5-bisphosphate carboxylase small subunit (EC 4.1.1.39)
Tery_2536 rpsP SSU ribosomal protein S16P
Tery_4509 fusA2 translation elongation factor 2 (EF-2/EF-G)
Tery_2324 gvpA gas vesicle protein GVPa
Tery_4796 psbU photosystem II oxygen evolving complex protein PsbU
Tery_4661 metK methionine adenosyltransferase (EC 2.5.1.6)
Tery_4005 RNA-binding region RNP-1
Tery_4138 nifK Mo-nitrogenase MoFe protein subunit NifK (EC 1.18.6.1)
Tery_3658 oxidoreductase FAD/NAD(P)-binding
Tery_3835 allophycocyanin beta-18 subunit apoprotein
Tery_5019 ThiJ/PfpI
Tery_3544 sulfite reductase (ferredoxin) (EC 1.8.7.1)
Tery_3909 Phycobilisome linker polypeptide
Tery_2209 phycobilisome core-membrane linker protein
Tery_4142 nifW nitrogen fixation protein NifW
Tery_4663 SSU ribosomal protein S1P
Tery_3311 thioredoxin
Tery_4105 Phycobilisome linker polypeptide
Tery_1234 cyanobacterial porin (TC 1.B.23)
Tery_5062 stress protein
Tery_2593 Periplasmic protein TonB links inner and outer membranes-like
Tery_4024 pgi glucose-6-phosphate isomerase (EC 5.3.1.9)
Tery_0450 transketolase (EC 2.2.1.1)
Tery_4466 cyanobacterial porin (TC 1.B.23)
Tery_0301 Peptidylprolyl isomerase
Tery_0265 rplK LSU ribosomal protein L11P
Tery_2198 atpG ATP synthase F1 subcomplex gamma subunit
Tery_2686 psbV cytochrome c, class I
Tery_3647 apcC phycobilisome core linker protein
Tery_0999 Phycobilisome linker polypeptide
Tery_1809 hypothetical protein
Tery_4465 cyanobacterial porin (TC 1.B.23)
Tery_0493 hypothetical protein
Tery_3852 ccmK2 microcompartments protein
Tery_3293 dapL LL-diaminopimelate aminotransferase apoenzyme (EC 2.6.1.83)
Tery_3901 hypothetical protein

Tery_3816 hypothetical protein
Tery_3004 rpmC LSU ribosomal protein L29P
Tery_4701 RNA-binding region RNP-1
Tery_1519 ppa Inorganic diphosphatase
Tery_0596 GatB/Yqey
Tery_4081 phage shock protein A (PspA) family protein
Tery_1161 rpsB SSU ribosomal protein S2P
Tery_1798 petA cytochrome f
Tery_1634 photosystem I reaction center protein PsaF, subunit III
Tery_2202 atpG ATP synthase F0 subcomplex B' subunit
Tery_1863 RNA-binding region RNP-1
Tery_2437 clpC1 ATPase AAA-2
Tery_2960 hypothetical protein
Tery_3917 photosystem II manganese-stabilizing protein PsbO
Tery_1014 psaE photosystem I reaction centre subunit IV/PsaE
Tery_1687 fructose-bisphosphate aldolase (EC 4.1.2.13)
Tery_0513 psbC photosystem II 44 kDa subunit reaction center protein
Tery_3377 extracellular solute-binding protein, family 1
Tery_1834 6-phosphogluconate dehydrogenase (decarboxylating) (EC 1.1.1.44)
Tery_1799 petC Rieske (2Fe-2S) region
Tery_4666 psbB photosystem antenna protein-like
Tery_3895 hypothetical protein
Tery_2201 atpF ATP synthase F0, B subunit
Tery_1160 tsf translation elongation factor Ts (EF-Ts)
Tery_1998 neutral amino acid-binding protein/L-glutamate-binding protein/L-aspartate-binding protein
Tery_2613 hypothetical protein
Tery_2993 adk Adenylate kinase (EC 2.7.4.3)
Tery_3537 phosphate ABC transporter substrate-binding protein, PhoT family (TC 3.A.1.7.1)
Tery_3633 hypothetical protein
Tery_3887 rpsF SSU ribosomal protein S6P
Tery_0586 tig trigger factor
Tery_3003 rpsQ SSU ribosomal protein S17P
Tery_1141 ndk nucleoside diphosphate kinase (EC 2.7.4.6)
Tery_0149 Rho termination factor-like
Tery_2793 eno enolase (EC 4.2.1.11)
Tery_3183 SSU ribosomal protein S30P/sigma 54 modulation protein
Tery_1108 UBA/THIF-type NAD/FAD binding fold
Tery_4349 molybdopterin synthase subunit MoaD (EC 2.8.1.12)
Tery_2842 nitrogen regulatory protein P-II family
Tery_0986 CpcD phycobilisome linker-like
Tery_4163 hypothetical protein
Tery_4356 Iron-regulated ABC transporter ATPase subunit SufC

Tery_3734 OmpA/MotB
Tery_3167 Serine--glyoxylate transaminase
Tery_1537 glucose-1-phosphate thymidyltransferase
Tery_3541 hypothetical protein
Tery_0738 ftsY signal recognition particle-docking protein FtsY
Tery_1229 hypothetical protein
Tery_4108 CopG-like DNA-binding
Tery_0535 6-phosphogluconolactonase (EC 3.1.1.31)
Tery_1312 peptidyl-prolyl cis-trans isomerase, cyclophilin type
Tery_1467 Peptidylprolyl isomerase
Tery_2657 nitroreductase
Tery_0283 tal transaldolase (EC 2.2.1.2)
Tery_4082 phage shock protein A (PspA) family protein
Tery_0263 rplJ LSU ribosomal protein L10P
Tery_1810 hypothetical protein
Tery_1989 hypothetical protein
Tery_2561 petJ cytochrome c, class I
Tery_2129 pentapeptide repeat
Tery_5020 hypothetical protein
Tery_4503 UspA
Tery_4668 psaB photosystem I core protein PsaB
Tery_2339 Gas vesicle synthesis GvpLGvpF
Tery_0930 infC bacterial translation initiation factor 3 (bIF-3)
Tery_2710 Hemolysin-type calcium-binding region
Tery_1169 hypothetical protein
Tery_3766 ftsZ cell division protein FtsZ
Tery_0682 fbp D-fructose 1,6-bisphosphatase (EC 3.1.3.11)
Tery_2883 ndhM NADH dehydrogenase I subunit M
Tery_4514 amino acid ABC transporter substrate-binding protein, PAAT family
Tery_0451 3-oxoacyl-
Tery_2900 hypothetical protein
Tery_4799 allophycocyanin alpha-B subunit apoprotein
Tery_1830 hypothetical protein
Tery_0882 argG argininosuccinate synthase (EC 6.3.4.5)
Tery_3001 rplX LSU ribosomal protein L24P
Tery_0545 hypothetical protein
Tery_1765 hypothetical protein
Tery_0373 hypothetical protein
Tery_4135 Fe-S cluster assembly protein NifU
Tery_0511 peptidyl-prolyl cis-trans isomerase, cyclophilin type
Tery_2992 infA bacterial translation initiation factor 1 (bIF-1)
Tery_3009 rplB LSU ribosomal protein L2P
Tery_4788 Peptidoglycan-binding domain 1

Tery_2743 protein of unknown function DUF520
Tery_4747 NADPH-glutathione reductase (EC 1.8.1.7)
Tery_1859 transcriptional regulator AbrB
Tery_1091 phage Tail Collar
Tery_1137 petB cytochrome b/b6-like
Tery_3622 protoporphyrin IX magnesium-chelatase (EC 6.6.1.1)
Tery_1740 psb27 photosystem II 11 kD protein
Tery_1497 Chromosome segregation ATPase-like protein
Tery_1164 hypothetical protein
Tery_2990 rpsM SSU ribosomal protein S13P
Tery_2995 rplO LSU ribosomal protein L15P
Tery_3443 rpoZ DNA-directed RNA polymerase, omega subunit
Tery_2200 atpH ATP synthase F1 subcomplex delta subunit
Tery_2741 hemL glutamate-1-semialdehyde 2,1-aminomutase (EC 5.4.3.8)
Tery_1641 Tic22-like
Tery_3729 rpsR SSU ribosomal protein S18P
Tery_0985 Phycobilisome linker polypeptide
Tery_3006 rpsC SSU ribosomal protein S3P
Tery_3976 infB bacterial translation initiation factor 2 (bIF-2)
Tery_2941 rpsT SSU ribosomal protein S20P
Tery_3010 rplW LSU ribosomal protein L23P
Tery_4669 psaA photosystem I core protein PsaA
Tery_0830 hypothetical protein
Tery_4593 plasmid segregation actin-type ATPase ParM
Tery_0394 msrA peptide methionine sulfoxide reductase
Tery_0891 putative nickel-containing superoxide dismutase precursor (NISOD)
Tery_1092 hypothetical protein
Tery_2528 rpiA ribose-5-phosphate isomerase (EC 5.3.1.6)
Tery_2984 rpsI SSU ribosomal protein S9P
Tery_0585 clpP ATP-dependent Clp protease proteolytic subunit ClpP (EC 3.4.21.92)
Tery_3467 Hemolysin-type calcium-binding region
Tery_1313 efp translation elongation factor P (EF-P)
Tery_3508 hypothetical protein
Tery_1016 ndhO hypothetical protein
Tery_0033 protein of unknown function DUF477
Tery_2363 hypothetical protein
Tery_0420 Glyoxalase/bleomycin resistance protein/dioxygenase
Tery_2918 hypothetical protein
Tery_4517
Tery_0466 glutamate synthase (ferredoxin) (EC 1.4.7.1)
Tery_3319 fabZ 3-hydroxyacyl-
Tery_1084 phosphoglucomutase/phosphomannomutase alpha/beta/alpha domain I
Tery_0068 hypothetical protein

Tery_1987 pentapeptide repeat
Tery_2747 DRTGG
Tery_0534 FHA domain containing protein
Tery_4723 TPR repeat
Tery_0313 band 7 protein
Tery_1246 Endopeptidase Clp
Tery_2660 glyA serine hydroxymethyltransferase (EC 2.1.2.1)
Tery_2371 acsA acetyl-coenzyme A synthetase (EC 6.2.1.1)
Tery_0954 ycf27 two component transcriptional regulator, winged helix family
Tery_1259 ychF GTP-binding protein YchF
Tery_1873 SPFH domain, Band 7 family protein
Tery_1547 hypothetical protein
Tery_3548 hypothetical protein
Tery_2607 DSBA oxidoreductase
Tery_3778 frr ribosome recycling factor
Tery_2381 hypothetical protein
Tery_4323 two component transcriptional regulator, LuxR family
Tery_4128 NifT/FixU
Tery_2982 prfA bacterial peptide chain release factor 1 (bRF-1)
Tery_0264 rplA LSU ribosomal protein L1P
Tery_4659 hypothetical protein
Tery_3817 hypothetical protein
Tery_2055 Hemolysin-type calcium-binding region
Tery_2988 rpoA DNA-directed RNA polymerase subunit alpha (EC 2.7.7.6)
Tery_1093 hypothetical protein
Tery_2416 putative endoribonuclease L-PSP
Tery_3832 hypothetical protein
Tery_0275 glucose-methanol-choline oxidoreductase
Tery_4109 glycine betaine/L-proline ABC transporter, ATPase subunit
Tery_2728 homoaconitate hydratase family protein
Tery_1124 NUDIX hydrolase
Tery_5017 anti-sigma-factor antagonist
Tery_2240 putative signal transduction protein with CBS domains
Tery_2937 rpoC2 DNA-directed RNA polymerase subunit beta' (EC 2.7.7.6)
Tery_2355 protein of unknown function DUF181
Tery_3408 Na-Ca exchanger/integrin-beta4
Tery_0043 argJ N-acetylglutamate synthase (EC 2.3.1.1)/glutamate N-acetyltransferase (EC 2.3.1.35)
Tery_4196 D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95)
Tery_4045 hypothetical protein
Tery_0018 Rho termination factor-like
Tery_4120 hypothetical protein
Tery_3218 periplasmic solute binding protein
Tery_2024 fabI Enoyl-

Tery_1344 dTDP-4-dehydrorhamnose reductase (EC 1.1.1.133)
Tery_4395 glucose-1-phosphate adenylyltransferase
Tery_2997 rplR LSU ribosomal protein L18P
Tery_5053 stress protein
Tery_1380 tpiA triosephosphate isomerase (EC 5.3.1.1)
Tery_1541 pnp Polyribonucleotide nucleotidyltransferase
Tery_2188 hypothetical protein
Tery_2881 glyceraldehyde-3-phosphate dehydrogenase (NAD+) (EC 1.2.1.12)
Tery_5013 putative anti-sigma regulatory factor, serine/threonine protein kinase
Tery_1954 purS phosphoribosylformylglycinamidine synthase, purS
Tery_4127 ATP-binding region, ATPase-like
Tery_3499 ndhJ NADH dehydrogenase subunit C (EC 1.6.5.3)
Tery_3375 UspA
Tery_0169 rpsO SSU ribosomal protein S15P
Tery_2238 hypothetical protein
Tery_2341 hypothetical protein
Tery_3483 hypothetical protein
Tery_3171 dihydrolipoamide dehydrogenase (EC 1.8.1.4)
Tery_1458 Substrate-binding region of ABC-type glycine betaine transport system
Tery_3500 ndhK NADH dehydrogenase subunit B (EC 1.6.5.3)
Tery_0683 tal transaldolase (EC 2.2.1.2)
Tery_0419 Hemolysin-type calcium-binding region
Tery_1166 hemF coproporphyrinogen oxidase (EC 1.3.3.3)
Tery_0477 rpsJ SSU ribosomal protein S10P
Tery_1748 type II secretion system protein E
Tery_3012 rplC LSU ribosomal protein L3P
Tery_3152 hypothetical protein
Tery_2318 hypothetical protein
Tery_3743 hypothetical protein
Tery_4516 hypothetical protein
Tery_1699 Thioredoxin-disulfide reductase
Tery_5022 4-carboxymuconolactone decarboxylase (EC 4.1.1.44)
Tery_4809 calcium-translocating P-type ATPase, PMCA-type
Tery_3755 ftsH membrane protease FtsH catalytic subunit (EC 3.4.24.-)
Tery_2216 periplasmic sensor signal transduction histidine kinase
Tery_1632 Nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase
Tery_2469 hypothetical protein
Tery_3008 rpsS SSU ribosomal protein S19P
Tery_2062 FAD-dependent pyridine nucleotide-disulphide oxidoreductase
Tery_1824 protein of unknown function DUF1499
Tery_3842 Redoxin
Tery_2625 sat sulfate adenylyltransferase (EC 2.7.7.4)
Tery_4894 glutamyl-tRNA synthetase

Tery_2266 Tetratricopeptide TPR_2
Tery_2326 hypothetical protein
Tery_4653 surface antigen (D15)
Tery_2156 Methyltransferase type 12
Tery_1314 biotin carboxyl carrier protein
Tery_1204 psaL photosystem I reaction centre, subunit XI PsaL
Tery_3709 hypothetical protein
Tery_1794 hypothetical protein
Tery_2063 pheT phenylalanyl-tRNA synthetase beta subunit (EC 6.1.1.20)
Tery_2408 psb28 photosystem II protein PsbW, class I
Tery_1230 psbDII photosystem II D2 protein (photosystem q(a) protein)
Tery_3494 response regulator receiver protein
Tery_1831 catalytic domain of components of various dehydrogenase complexes
Tery_4385 protein of unknown function UPF0182

APPENDIX 4. Additional samples collected during the course of this work

Table A4.1 Expedition – JC150 (ZIPLOc)

Location: North Atlantic subtropical gyre

Experiment or type of sample	**Date(s) or stations collected**
0.2-3µm metaproteomes (multiple filter fractions – whole community)	Stations 1-7
3-51µm metaproteomes (multiple filter fractions – whole community)	Stations 1-7
>51µm metaproteomes (multiple filter fractions – whole community)- contain Tricho in abundance	Stations 1-7
Trichodesmium colony Mn, Ni, Fe matrix incubation proteomes	Stations 1-7
Picked colonies for diel proteomes - surface	Stations 1-7
Picked colonies for diurnal migration proteomes – 15m and 200m depth throughout the day	Stations 1-7

Table A4.2 Expedition – AT39-05 (TriCoLim)

Location: Tropical and subtropical Atlantic

Experiment or type of sample	**Date(s) or stations collected**
0.2-3µm metaproteomes (multiple filter fractions – whole community)	Stations 1-19
3-51µm metaproteomes (multiple filter fractions – whole community)	Stations 1-19
>51µm metaproteomes (multiple filter fractions – whole community)- contain Tricho in abundance	Stations 1-19
Picked Tricho colonies for diel proteomes - surface	Stations 1-19
Picked Tricho colonies trace metal clean for metalloproteomes	Dates: 2.24, 3.5, 3.10, 3.11
Trichodesmium colony Mn, Ni, Fe matrix incubation proteomes	Dates: 2.13, 2.15, 2.20, 2.24, 3.1, 3.2,.3.4, 3.5, 3.9
Nickel addition experiments to whole seawater	Dates: 2.17, 2.25, 3.7, 3.11

Table A4.3 Expedition – FK160115 (ProteOMZ)

Location: Tropical Pacific

Experiment or type of sample	Date(s) or stations collected
Whole seawater N, Co, Zn, Fe, B12 co-limitation incubation experiments (proteome samples)	St 8, 12, 13
Whole seawater increased temperature incubation experiments	St 6, 10, 13