Prochlorococcus: life in light

by

Jessica Weidemier Thompson

A.B. Molecular Biology, Certificate in Visual Arts
Princeton University, 2008

Submitted to the Microbiology Graduate Program
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Microbiology
at the
Massachusetts Institute of Technology

June 2015

Signature of Author

Signature redacted

Microbiology Graduate Program and Department of Civil and Environmental Engineering
Massachusetts Institute of Technology, May 21, 2015

Certified by

Signature redacted

Sallie W. Chisholm
Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by

Signature redacted

Mike Laub
Professor of Biology
Chair, Committee for Graduate Students, Microbiology Graduate Program

Prochlorococcus: life in light

by

Jessica Weidemier Thompson

Submitted to the Microbiology Graduate Program on May 21, 2015
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Microbiology

Abstract

The marine cyanobacterium *Prochlorococcus*, a single-celled organism less than 1 μm in diameter, is highly abundant in the vast low-nutrient regions of the open oceans, and plays an important role in energy and nutrient flow in marine ecosystems. This thesis presents a body of work exploring several aspects of the role of light in the evolution of diversity within *Prochlorococcus*, combining approaches from genomics, field oceanography and laboratory cultures. Through isolation efforts targeted at low-light-adapted *Prochlorococcus*, clades that live deep in the water column where light is scarce, the representation of this group in cultures and genomic datasets has substantially expanded, leading to an improved picture of the deep diversity they contain. To explore the relationship between genomic variation and light physiology, cultures spanning the diversity of *Prochlorococcus* were screened for their ability to withstand severe, transient increases in light. Different clades of *Prochlorococcus* showed different responses, including one clade that prefers growth at low light, but survives this temporary light shock, consistent with its persistence during deep vertical mixing events in the ocean. Bioinformatic approaches were employed to explore the complex evolutionary history of a gene family that might be part of genomic adaptation to different light conditions in *Prochlorococcus*, the high-light-inducible (*hli*) genes, small photosystem-associated proteins involved in the cyanobacterial response to high light and other forms of stress. Finally, the distribution and cell properties of *Prochlorococcus* populations across an Eastern South Pacific transect were analyzed in the context of the light environment, showing dramatic differences from the rich coastal waters to the extremely clear waters of the South Pacific Gyre. The studies presented here provide several new perspectives on the role of light in *Prochlorococcus* physiology, the nature of genetic variation across *Prochlorococcus* and its functional and ecological consequences, making progress towards understanding the processes driving evolution in this important organism.

Thesis Supervisor: Sallie W. Chisholm
Title: Professor of Civil and Environmental Engineering

Acknowledgements

First and foremost, I'd like to thank Penny Chisholm for sharing with me her brilliant vision of our planet, its oceans and its creatures, and at the same time for teaching me an attention to scientific detail and rigor that will serve me throughout my career. I would like to thank my committee members. Janelle Thompson welcomed me to MIT in my first rotation and since has served as a role model and a source of support and sensible advice. I thank Ed Delong for his thoughtful advice and kind interest in my opinions, as a colleague and teaching assistant. I thank Jing-Ke Weng whose diverse experiences and vision of evolution have assisted in the final phases of creating this thesis. I thank Colleen Cavanaugh, first for taking me on 10 years ago, when I literally showed up at her lab's door asking for a summer research experience, which proved to be the start to my path in marine microbiology, and now for serving as my external advisory thesis committee member, bringing this PhD process to a close. We are grateful for the support that enabled this work, from the National Science Foundation's Center for Microbial Oceanography Research and Education, The Gordon and Betty Moore Foundation's Marine Microbiology initiative, and individual grants from the National Science Foundation to Penny Chisholm.

I thank Alan Grossman and David Schauer, who is dearly missed, for founding the Microbiology program, filling an important gap in educational opportunities at MIT and for choosing me to be a part of its first class. I would also like to thank all of the professors who taught me classes during my time at MIT, especially Aviv Regev, Mike Laub, Roman Stocker, Eric Alm, Alan Grossman, David Schauer, Jacquin Niles, Ed Delong and Scott Edwards. My coursework was an inspiring welcome to this exciting place. The classes were so good, and in many instances practical, that they were an important part of launching my work in microbiology and what I learned in them influences my thinking every day. I've had so many wonderful teachers over the years, and I would like to thank them all, including Mrs. Cavicchio of the Lilja school for her important role in teaching me to read (the most important skill I ever learned). I thank Dr. Mitchell, for daring me to take a biology class when I couldn't find a college major, and I thank Ted Cox for teaching that class, an inspiring introduction to modern biology. I thank Francois Morel, Pat McGinn and Yan Xu for mentoring me through my first sustained research efforts in college and introducing me to the world of phytoplankton.

I thank my labmates, all the people who've shared the Chisholm Lab over my time here, the ones from the very beginning who convinced me I wanted to join the lab, the ones from the early years who taught me what I know about *Prochlorococcus* and how to do research, and the new ones who keep bringing new ideas to the sphere of *Prochlorococcus*. I thank all the people who work in Parsons, eat in Parsons, and talk in Parsons. I am particularly grateful to the best cubicle buddies ever, Teresa and Patricia and Dave. I thank also MIT for being such a good place – full of neat talks, pretty libraries, movies with popcorn and the Muddy Charles. I thank my friends Alison Takemura, who has been with me this whole PhD path, and Jess Lander and Amy Glynn, old friends who stick with me through long unexplained microbiological absences, feed me, listen to me and help me see the bigger world around me.

I thank my husband's family, Madeline Kaczmarczyk, Jerry Berta, Darlene Kaczmarczyk, all the Bertas, and Amy, Brian and Claire Bengtson, for fun trips, for their understanding and support, by putting up with my working over holidays and sending wonderful treats in the mail, and for the joy of getting to know the wonderful baby Claire toward the end of this PhD.

I thank my parents, for everything that led me here, particularly their loving attention and passionate advocacy in the details of my education over the years and for the new relationship we've forged as adults, since I've moved back to Boston for graduate school. I thank them for their constant love and support. I also thank them for their heroic copy-editing assistance with this thesis. I thank my brother Neil, for all our wonderful years together. I thank all my whole family, all the Thompsons and Weidemiers, for support, love and fun.

I thank my husband Zach for technical assistance in coding, plotting, installing software, and navigating statistics, for talking to me for hours and hours about my work and my feelings about it, for draft reading and copy-editing this thesis. I thank Zach for for feeding me and helping me to sleep, for help in caring for the plants and making me laugh. I thank Zach for his presence in my life and his love.

Biographical Note

Jessie Thompson grew up outside of Boston, playing in the backyard and growing things in the garden with her family. This lead to an early and lasting love for all things photosynthetic, starting with peas and daffodils, thanks to her parents and grandparents. Some wonderful teachers along the way helped develop a pleasure in learning and an interest in science. In college, studying biology and art, this developed into a passion for phytoplankton, thanks to the mentorship of the whole lab of Francois Morel, the beauty and complexity of marine diatoms, and the sheer, unexpected fun of doing bench research. She was impressed at the intricacies of life that modern biology research lets us see. That led to continued learning about small, important things with the MIT Microbiology program, as part of the very first class, and then the opportunity to join the lab of Penny Chisholm, which led to a whole graduate education in and by *Prochlorococcus*. She is married to an astronomer, which is a powerful thing for remembering the bigger-than-global implications for life, photosynthesis and evolution. She finds it useful in the practice of research to have another scientist to talk to at any time of the day or night, but more importantly she is grateful to have such a kind and loving person to share her life. Throughout graduate school they've lived in North Cambridge, and grown green things on the porches of a triple decker, although thesis writing this spring has had negative effect on this season's seed starts. Moving forward, she is looking forward to a fulfilling career as a scientist, and many other adventures.

Table of Contents

Chapter I. Introduction

Prochlorococcus and its role in the global oceans

Phytoplankton, microscopic aquatic photosynthetic organisms, exert a commanding influence on the nature of the air and oceans through their roles in the biogeochemical cycling of elements and the flow of the Sun's energy throughout our planet, today and over much of the history of the Earth (Hays et al., 2005, Falkowski and Isozaki, 2008, Falkowski and Raven, 2007, Falkowski et al. 2004, Bang and Chisholm, 2012, Blank et al., 2010). They collectively fix as much carbon and produce as much oxygen annually as all plants on land, supporting the marine food web (Field et al., 1998, Falkowski and Raven, 2007). *Prochlorococcus*, a unicellular marine cyanobacterium, contributes significantly to these processes due to its remarkable abundance in the vast oligotrophic mid-latitude oceans (Figure 1.1, Figure 1.2, Figure 1.6; Partensky et al., 1999; Follows et al., 2007, Zwirglmaier et al., 2008, Flombaum et al., 2013). *Prochlorococcus* is the smallest of the phytoplankton (0.5-0.7 μm diameter), with the smallest genome (1.7-2.7 Mb), features thought to be among its many adaptations to its nutrient-poor environment (Batut et al., 2014, Rocap et al., 2003, Dufresne et al., 2003, Coleman and Chisholm, 2007, Van Mooy et al., 2009, Partensky et al., 1999, Partensky and Garczarek, 2010).

Figure 1.1. The *Prochlorococcus* habitat
The midlatitude open oceans are characterized by very clear blue water (Morel et al., 2007) and very low nutrient concentrations (e.g. Karl et al., 2002). Sargasso Sea, July 2009, from my first visit to the oligotrophic open oceans, with the Microbial Oceanography course at the Bermuda Institute of Ocean Sciences.

Prochlorococcus and its closest relative, the marine *Synechococcus*

The closest relative of *Prochlorococcus* is marine *Synechococcus*, the second most abundant cyanobacterium in the oceans (Figure 1.2, Flombaum et al., 2013). *Prochlorococcus* has diverged from *Synechococcus* in many ways, including through dramatic changes in the photosynthetic light gathering antennae (Ting et al., 2002).

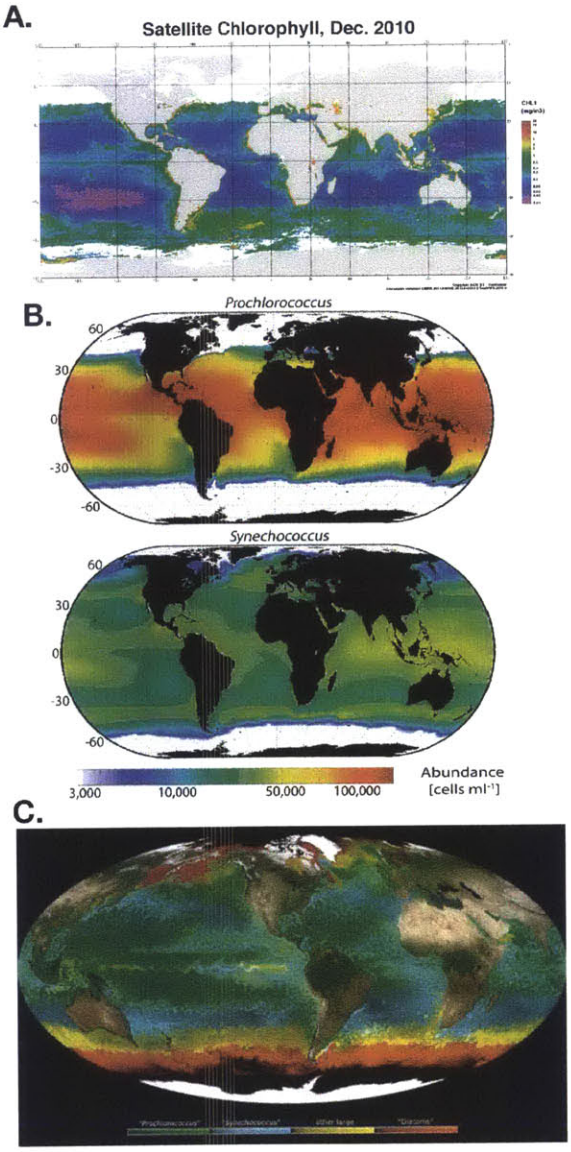


Figure 1.2. The global distribution of *Prochlorococcus* in relation to other phytoplankton
Prochlorococcus lives in the midlatitude open oceans, in low productivity gyres, and moderately productive equatorial regions. In some places it is the dominant phytoplankter, in other cases it coexists with many other species. (A) Chlorophyll viewed from space as a proxy for productivity in the oceans, MERIS/MODIS/SeaWiFS merged chlorophyll measurements for the month of December 2010, (from http://hermes.acri.fr/index.php?class=archive, chl1 AVW algorithm, weighted average) (B) Distributions of *Prochlorococcus* and *Synechococcus* from Flombaum et al. 2013, based on a global model built from empirical counts integrated with measurements of environmental variables, scaled up to the global ocean. (C) The product of a global ocean phytoplankton model, built from agents analogous to *Prochlorococcus*, *Synechococcus*, diatoms and other eukaryotic phytoplankton, allowed to populate a sophisticated chemical and physical ocean model, reproducing natural patterns (Follows et al., 2007).

In *Prochlorococcus* the large, flexible protein-pigment complex known as the phycobilisome used for light harvesting in most cyanobacteria has been replaced by a new family of proteins, homologous to stress-induced alternate light gathering proteins, which use unusual chlorophyll derivatives, divinyl chlorophyll A and B, as primary light gathering pigments (Ting et al., 2002, Scanlan et al., 2009, Partensky et al., 1999). This is a less costly strategy, in terms of nitrogen use, that limits *Prochlorococcus* compared with the spectral range achievable with phycobilisomes but is well adapted to the low nutrient water with primarily blue light that characterizes most *Prochlorococcus* habitat (Ting et al., 2002, Morel et al., 2007). Numerous additional differences have been observed across their genomes, including in metabolism, nutrient acquisition strategies, and stress response, with an overall trend of smaller genomes in *Prochlorococcus*, achieved through a complex history of gene gain and loss (Scanlan et al., 2009, Kettler et al., 2007). *Prochlorococcus* and *Synechococcus* have overlapping but distinct ecologies; *Prochlorococcus* reaches higher abundances in the tropical and subtropical open ocean, but *Synechococcus* is capable of living in a wider range of habitats, including coastal and high latitude regions (Figure 1.2; Flombaum et al., 2013, Follows et al., 2007).

The paradox of the plankton and niche adaptation

Prochlorococcus is a fine example of the paradox of plankton writ small (Hutchinson et al., 1961). The paradox of the plankton poses the question of how can the ocean, a relatively homogenous, seemingly simple, liquid environment support the staggering diversity of phytoplankton that we observe, in light of the principles of competitive exclusion and niche adaptation (Barton et al., 2010, Hutchinson 1961, Macarthur, 1958). This body of theory holds that two organisms cannot coexist if they share the same niche, defined by Hutchinson as a property of the organism, not the environment, an n-dimensional hyperspace, each axis of which represents an environmental variable, the ranges of which set the organism's potential and limits (Barton et al., 2010, Hutchinson 1961, Macarthur, 1958, Hutchinson, 1957, Colwell et al., 2009). The solution to the paradox of the plankton is not a simple one - there are many answers for why there are so many spectacular phytoplankton - but, for a start, their coexistence is enabled by complexity in the marine environment in chemistry, physics, community structure and in changes over time, enabling organisms to carve out complex, unique niches (Hutchinson et al., 1961). This collection of ideas is part of our fundamental framework as we continue to explore the complexity of phytoplankton ecology and evolution in the *Prochlorococcus* system.

***Prochlorococcus* diversity: ecotype and habitat adaptations**

Prochlorococcus can be divided into ecotypes, phylogenetic clusters that display different physiological attributes and distributions in the environment (Figure 1.3; Moore et al., 1998, Moore and Chisholm 1999, Zinser et al., 2006, Johnson et al., 2006). Broadly these ecotypes can be divided into the high-light-adapted (HL), proliferating near the surface, and low-light-adapted (LL), found deeper in the water column, where there is less light but higher nutrient concentrations (Moore et al., 1998, West and Scanlan, 1999, Coleman and Chisholm, 2007). Within the HL group, one ecotype has a lower temperature range, consistent with its higher-latitude distribution (Zinser et al., 2007). Within the LL group, one clade distinguishes itself in tolerance of high light shock, consistent with its ability to persist in the water column following deep mixing events which expose it to surface light (Zinser et al., 2007, Malmstrom et al., 2010). Adaptations to various dimensions of the *Prochlorococcus* niche space correspond to phylogeny to differing degrees (Martiny et al., 2009). In sequenced genomes and environmental genome fragments, variation in nutrient assimilation pathways is observed in hypervariable genomic islands, with signatures of phage-mediated horizontal gene transfer (Coleman et al., 2006). For example, a comparison of *Prochlorococcus* sequences from metagenomic samples in the Atlantic and Pacific found that most gene content was similar at the two sites, but phosphate

uptake-related genes were significantly more abundant in the low-phosphorus Atlantic (Coleman et al., 2010). Local selection pressures result in differential fixation of some genes at the two sites (Coleman et al., 2010). Selection is thought to act very efficiently on the large population sizes and relatively high growth rates of *Prochlorococcus* in the wild, tuning these genomic complements to their environments at a very fine level (Kashtan et al., 2014). Diversity within *Prochlorococcus* enables its widespread distribution, through adaptation of distinct lineages to different habitats, over depth and geography, on multiple evolutionary timescales (Biller et al., 2015, Martiny et al., 2009).

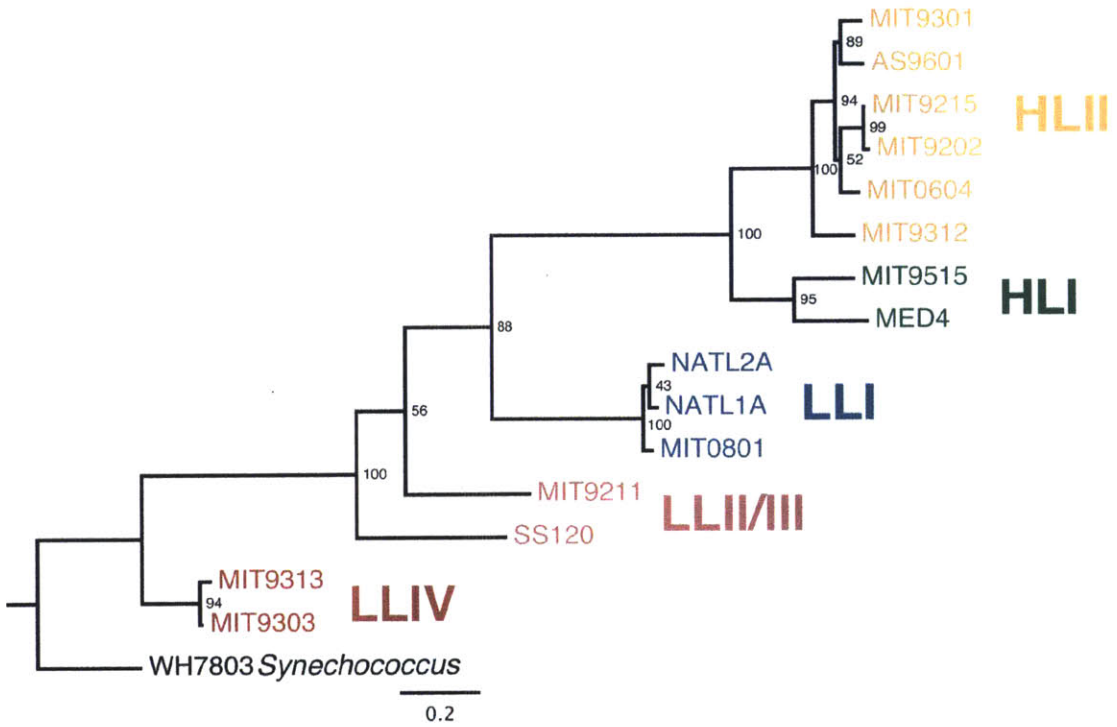


Figure 1.3. Phylogenetic relationships between *Prochlorococcus* ecotypes
Approximate *Prochlorococcus* lineage phylogeny, showing ecotype designations at right (e.g. HLI) and ecotype colors, which will be used throughout this thesis. This is a GyrB DNA gyrase DNA-based maximum likelihood phylogeny (phyML), which has been shown to be a useful marker for *Prochlorococcus*-wide phylogeny (Mühling et al., 2012). The particular genomes represented here are the set of currently fully closed *Prochlorococcus* genomes. This is similar (but slightly expanded) to the set of genomes available at the start of this thesis work; now we have more than 40 genomes (and counting), mostly of draft quality. The marine *Synechococcus* WH7803 serves here as an outgroup.

The power of the *Prochlorococcus* system to answer fundamental questions in microbial evolution

Diverse approaches available in the *Prochlorococcus* tool kit, in the field, in genomes and in culture-based laboratory studies, have made it a model in microbial ecology and evolution, contributing to our basic understanding of how microbial genomes evolve (Coleman and Chisholm, 2007, Scanlan et al., 2009, Biller et al., 2015). The unique flow cytometry signature of *Prochlorococcus* enables rapid identification, enumeration and sorting of *Prochlorococcus* populations (Figure 1.4; Chisholm et al., 1988, Moore et al., 1998). Unlike many marine microbes, *Prochlorococcus* is cultivable (Figure 1.5), which has lead to a rich body of work exploring differences in physiology within *Prochlorococcus* diversity and between *Prochlorococcus* and *Synechococcus*, their interactions with other bacteria and the cultivation and study of the viruses that infect them (Rappé and Giovannoni, 2003, Chisholm et al., 1992, Rippka et al., 2000, Moore et al., 2007,

Partensky and Garczarek, 2010, Biller et al., 2015). There are currently 41 published sequenced genomes of *Prochlorococcus* cultured strains, all within 3% 16S rRNA sequence divergence, and more single-cell derived partial genomes from wild samples, which together have revealed dramatic patterns of ecologically significant differentiation (Rocap et al., 2003; Dufresne et al., 2003; Kettler et al., 2007, Thompson et al., 2011, Coleman et al., 2006, Biller et al., 2014, Morris et al., 2008, Malmstrom et al., 2013).

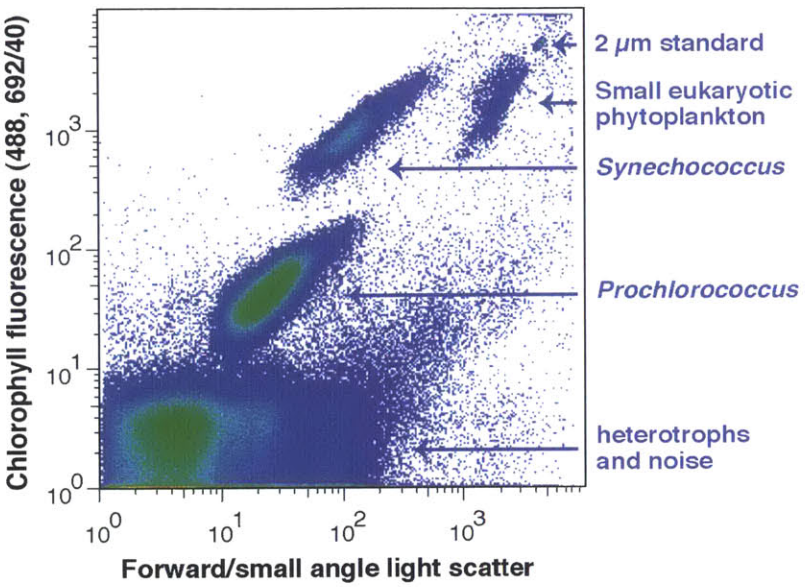


Figure 1.4. Seawater viewed through a flow cytometer: small chlorophyll containing particles Seawater phytoplankton, viewed through flow cytometry based on chlorophyll fluorescence and forward angle scatter (a rough size proxy), for a sample from a mesotrophic site, off the coast of Chile, 24m, 21°S, 76°W. *Prochlorococcus* is the smallest particle in seawater with chlorophyll fluorescence. The flow cytometer led to the discovery of *Prochlorococcus* as the unique and abundant organism it is, and it has since remained a critical tool for the study of *Prochlorococcus* (Chisholm et al., 1988).

The abundance of *Prochlorococcus* in the environment makes it a major component of many marine whole-community metagenomic and metatranscriptomic sequencing efforts, which, combined with the availability of many high-quality reference genomes spanning *Prochlorococcus* diversity, enables detailed study of the distribution and selection of genomic traits in the wild (e.g. Frias-Lopez et al., 2008, Hewson et al., 2009, Coleman et al., 2010, Rusch et al., 2010, Ottesen et al., 2014). *Prochlorococcus* is not the most tractable system, compared with some other cyanobacteria; attempts at genetic manipulation have been met with limited success and culturing still presents some challenges (Biller et al., 2015, Tolonen et al., 2006, Moore et al., 2007). Its study is ultimately motivated by its contributions to the open ocean ecosystems and its unique properties among phototrophs (Partensky et al., 1999). We strive to understand *Prochlorococcus* both because it has the ability to bring us fundamental insight about how microbes live and evolve in the ocean and to better understand this specific organism, as a critical part of the processes that influence the world's oceans and air.

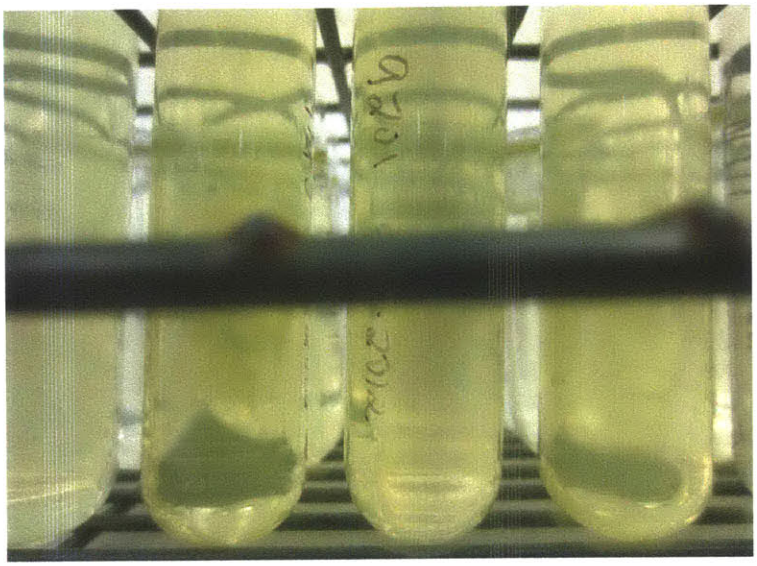


Figure 1.5. ***Prochlorococcus*** **cultures**
High density *Prochlorococcus* batch cultures (e.g. MIT9201 at center right) growing under typical conditions - in seawater amended with inorganic nutrients, with their characteristic lime green pigmentation.

Overview of questions and work presented in this thesis

Here, we set out to continue exploration of *Prochlorococcus* ecology and evolution, through genomic, field and lab techniques, toward understanding the role of light in *Prochlorococcus* biology and how *Prochlorococcus* has evolved to fill its diverse ocean habitats. First, through enrichment efforts targeting low-light adapted *Prochlorococcus* and the application of a recently developed purification technique (Berube et al., 2014), we isolated new strains of *Prochlorococcus* from the North Pacific and sequenced their genomes. These strains come primarily from the LLIV clade of *Prochlorococcus*, and they significantly expand our coverage and understanding of this clade. Second, we explored a family of ideas surrounding the ability of the LLI ecotype of *Prochlorococcus* to survive transient, severe increases in light, which other LL groups cannot, and which may explain the persistence of the LLI group during deep winter mixing events (Malmstrom et al., 2010). Using the growing culture collection, we ask how diverse *Prochlorococcus* respond to transient exposure to high light. The high-light-inducible genes, a family of chlorophyll-binding cyanobacterial stress response proteins, are good candidates to be part of the genomic adaptation behind this trait (among others), and the number of genes from this family varies within and between ecotypes (Coleman and Chisholm, 2007, Kettler et al., 2007). Through analysis of the recently expanded genome collection, we investigate how the high-light-inducible gene family has evolved to its present complexity across *Prochlorococcus*. Finally, we describe a collection of samples from a transect across the Eastern South Pacific, analyzed using flow cytometry, and present cell properties and the distribution of *Prochlorococcus*, both vertically and geographically, in relation to the light environment, which changes dramatically across this transect. As a whole, we hope that this work contributes to a better understanding of *Prochlorococcus* diversity and adaptation.

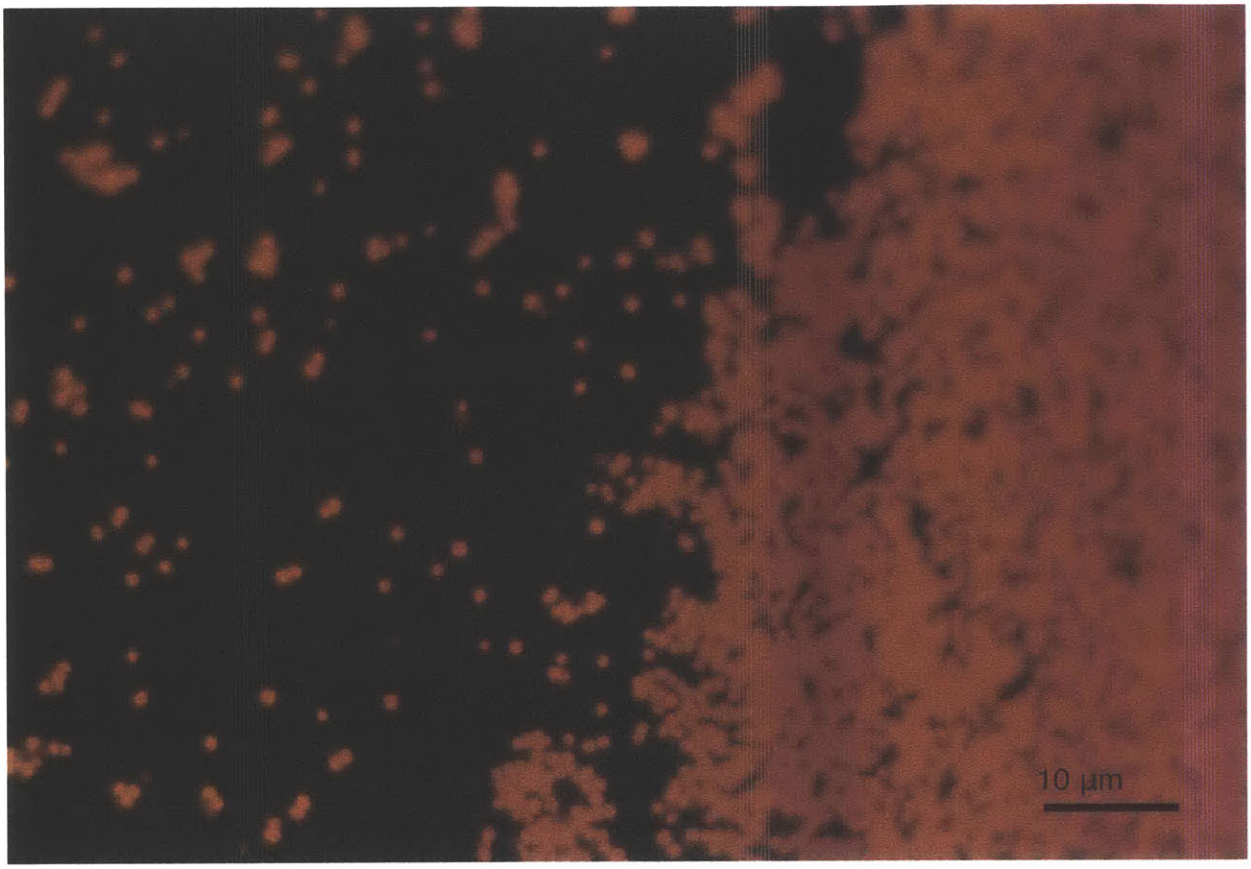


Figure 1.6. Small, round *Prochlorococcus* cells
Chlorophyll autofluorescence microscopy, edge of a pellet from a culture of strain NATL2A.

References

Bang, M., and Chisholm, P. (2012). Ocean Sunlight: How Tiny Plants Feed the Seas (New York, NY: The Blue Sky Press, an imprint of Scholastic).

Barton, A.D., Dutkiewicz, S., Flierl, G., Bragg, J., and Follows, M.J. (2010). Patterns of diversity in marine phytoplankton. Science *327*, 1509-511.

Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol

Berube, P.M., Biller, S.J., Kent, A.G., Berta-Thompson, J.W., Roggensack, S.E., Roache-Johnson, K.H., Ackerman, M., Moore, L.R., Meisel, J.D., et al. (2014). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. ISME J

Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., Awad, L., Roache-Johnson, K.H., Ding, H., Giovannoni, S.J., et al. (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *1*, 140034.

Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015). *Prochlorococcus*: the structure and function of collective diversity. Nat Rev Microbiol *13*, 13-27.

Blank, C.E., and Sánchez-Baracaldo, P. (2010). Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. Geobiology *8*, 1-23.

Chisholm, S., Frankel, S., Goericke, R., Olson, R., Palenik, B., Waterbury, J., West-Johnsrud, L., and Zettler, E. (1992). *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b. Archives of Microbiology *157*, 297-300-.

Chisholm, S.W., Olson, R.J., Zettler, E.R., Waterbury, J.B., Goericke, R., and Welschmeyer, N. (1988). A novel free-living prochlorophyte occurs at high cell concentrations in the oceanic euphotic zone. Nature *334*, 340-43.

Coleman, M.L., and Chisholm, S.W. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. Trends Microbiol *15*, 398-407.

Coleman, M.L., and Chisholm, S.W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. Science *311*, 1768-770.

Colwell, R.K., and Rangel, T.F. (2009). Hutchinson's duality: the once and future niche. Proc Natl Acad Sci U S A *106 Suppl 2*, 19651-58.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V., Duprat, S., Galperin, M.Y., Koonin, E.V., et al. (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A *100*, 10020-25.

Falkowski, P.G., and Isozaki, Y. (2008). Geology. The story of O2. Science *322*, 540-42.

Falkowski, P.G., and Raven, J.A. (2007). Aquatic Photosynthesis (Princeton University Press).

Falkowski, P.G., Katz, M.E., Knoll, A.H., Quigg, A., Raven, J.A., Schofield, O., and Taylor, F.J. (2004). The evolution of modern eukaryotic phytoplankton. Science *305*, 354-360.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. Science *281*, 237-240.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proceedings of the National Academy of Sciences *110*, 9824-29.

Follows, M.J., Dutkiewicz, S., Grant, S., and Chisholm, S.W. (2007). Emergent biogeography of microbial communities in a model ocean. Science *315*, 1843-46.

Frias-Lopez, J., Shi, Y., Tyson, G.W., Coleman, M.L., Schuster, S.C., Chisholm, S.W., and Delong, E.F. (2008). Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A *105*, 3805-810.

Hays, G.C., Richardson, A.J., and Robinson, C. (2005). Climate change and marine plankton. Trends Ecol Evol *20*, 337-344.

Hewson, I., Paerl, R.W., Tripp, H.J., Zehr, J.P., and Karl, D.M. (2009). Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. Limnology and Oceanography *54*, 1981-994.

Hutchinson, G.E. (1957). Concluding Remarks. Cold Spring Harbor Symposia on Quantitative Biology *22*, 415-427.

Hutchinson, G.E. (1961). The Paradox of the Plankton. The American Naturalist *95*, 137-145.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M., and Chisholm, S.W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science *311*, 1737-740.

Karl, D.M. (2002). Nutrient dynamics in the deep blue sea. Trends Microbiol *10*, 410-18.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science *344*, 416-420.

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., et al. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet *3*, e231.

MacArthur, R.H. (1958). Population Ecology of Some Warblers of Northeastern Coniferous Forests. Ecology *39*, 599-619.

Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. ISME J *4*, 1252-264.

Malmstrom, R.R., Rodrigue, S., Huang, K.H., Kelly, L., Kern, S.E., Thompson, A., Roggensack, S., Berube, P.M., Henn, M.R., and Chisholm, S.W. (2013). Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. ISME J *7*, 184-198.

Martiny, A.C., Tai, A.P., Veneziano, D., Primeau, F., and Chisholm, S.W. (2009). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. Environ Microbiol *11*, 823-832.

Moore, L.R., and Chisholm, S.W. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus* : ecotypic differences among cultured isolates. Limnol Oceanogr *44*, 628-638.

Moore, L.R., Coe, A., Zinser, E.R., Saito, M.A., Sullivan, M.B., Lindell, D., Frois-Moniz, K., Waterbury, J., and Chisholm, S.W. (2007). Culturing the marine cyanobacterium *Prochlorococcus*. Limnology and Oceanography: Methods *5*, 353-362.

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature *393*, 464-67.

Morel, A., Claustre, H., Antoine, D., and Gentili, B. (2007). Natural variability of bio-optical properties in Case 1 waters: attenuation and reflectance within the visible and near-UV spectral domains, as observed in South Pacific and Mediterranean waters. Biogeosciences *4*, 913-925.

Morris, J.J., Kirkegaard, R., Szul, M.J., Johnson, Z.I., and Zinser, E.R. (2008). Facilitation of robust growth of *Prochlorococcus* colonies and dilute liquid cultures by "helper" heterotrophic bacteria. Appl Environ Microbiol *74*, 4530-34.

Mühling, M. (2012). On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. Environ Microbiol *14*, 567-579.

Ottesen, E.A., Young, C.R., Gifford, S.M., Eppley, J.M., Marin, R., Schuster, S.C., Scholin, C.A., and DeLong, E.F. (2014). Ocean microbes. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. Science *345*, 207-212.

Partensky, F., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. Ann Rev Mar Sci *2*, 305-331.

Partensky, F., Blanchot, J., and Vaulot, D. (1999). Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. Bulletin de l'Institut Oceanographique, Monaco *19*

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev *63*, 106-127.

Rappé, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu Rev Microbiol *57*, 369-394.

Rippka, R., Coursin, T., Hess, W., Lichtle, C., Scanlan, D.J., Palinska, K.A., Iteman, I., Partensky, F., Houmard, J., and Herdman, M. (2000). *Prochlorococcus marinus* Chisholm et al. 1992 subsp. pastoris subsp. nov. strain PCC 9511, the first axenic chlorophyll a2/b2-containing cyanobacterium (Oxyphotobacteria). Int J Syst Evol Microbiol *50 Pt 5*, 1833-847.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature *424*, 1042-47.

Rusch, D.B., Martiny, A.C., Dupont, C.L., Halpern, A.L., and Venter, J.C. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. Proceedings of the National Academy of Sciences *107*, 16184 -116189.

Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., Post, A.F., Hagemann, M., Paulsen, I., and Partensky, F. (2009). Ecological Genomics of Marine Picocyanobacteria. Microbiology and Molecular Biology Reviews : MMBR *73*, 249-299.

Thompson, A.W., Huang, K., Saito, M.A., and Chisholm, S.W. (2011). Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. ISME J *5*, 1580-594.

Ting, C.S., Rocap, G., King, J., and Chisholm, S.W. (2002). Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. Trends Microbiol *10*, 134-142.

Tolonen, A.C., Liszt, G.B., and Hess, W.R. (2006). Genetic manipulation of *Prochlorococcus* strain MIT9313: green fluorescent protein expression from an RSF1010 plasmid and Tn5 transposition. Appl Environ Microbiol *72*, 7607-613.

Van Mooy, B.A., Fredricks, H.F., Pedler, B.E., Dyhrman, S.T., Karl, D.M., Koblizek, M., Lomas, M.W., Mincer, T.J., Moore, L.R., et al. (2009). Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature *458*, 69-72.

West, N.J., and Scanlan, D.J. (1999). Niche-Partitioning of *Prochlorococcus* Populations in a Stratified Water Column in the Eastern North Atlantic Ocean. Appl Environ Microbiol *65*, 2585-591.

Zinser, E., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D., and Chisholm, S.W. (2007). Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. Limnol Oceanogr *52*, 2205-220.

Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J., and Chisholm, S.W. (2006). *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol *72*, 723-732.

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F., Massana, R., Ulloa, O., and Scanlan, D.J. (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. Environ Microbiol *10*, 147-161.

Chapter II. Targeted isolation and genomic sequencing of new low-light adapted *Prochlorococcus* strains

Jessie W. Berta-Thompson[1,2], Andrés Cubillos Ruíz[1,2], Jamie W. Becker[1], Kristin N. LeGault[1], Sallie W. Chisholm[1,3]

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
[2]Microbiology Graduate Program, Massachusetts Institute of Technology
[3]Department of Biology, Massachusetts Institute of Technology

Abstract

The globally abundant marine cyanobacterium *Prochlorococcus* is amenable to cultivation, forming dense green cultures in inorganic-enriched seawater under controlled light and temperature conditions in the laboratory. In vitro cultures enable the convenient study of countless properties of a microbe, including its physiology under different conditions, its genome, its interactions with other organisms and direct comparisons with other microbes. However, the process of isolation in *Prochlorococcus*, coaxing cells from the ocean to grow in pure culture in the laboratory, is not yet routine. There are approximately 60 *Prochlorococcus* strains reported in culture, which have proven tremendously valuable to our understanding of the organism, yet this number represents only a tiny fraction of wild *Prochlorococcus* diversity. To expand the *Prochlorococcus* culture collection, we performed isolation efforts targeting low-light adapted *Prochlorococcus*, a group of interest to several areas of *Prochlorococcus* research, including comparative light physiology, genome evolution and secondary metabolite production, for which we have only a limited set of strains. From seawater collected from 150m at the well-characterized Station ALOHA, site of the Hawaii Ocean Time Series in the North Pacific gyre, we isolated 12 new low-light adapted *Prochlorococcus* strains and one high-light adapted strain, many of them already in the form of axenic, clonal cultures. We employed a novel combination of established techniques, (i) *Prochlorococcus* enrichment through inorganic amendment of seawater and (ii) dilution to extinction in organic-rich media, to purify multiple axenic strains efficiently from the same initial water sample. We have sequenced their genomes of many of these new strains, to draft quality. The majority of these strains are from the LLIV (or e9313) ecotype, the deepest branching ecotype in the *Prochlorococcus* phylogeny, with the largest genomes and the most strain-specific flexible genome content, often found deeper in the water column than other *Prochlorococcus* ecotypes. This new set of strains includes both wide ranging diversity within the LLIV clade, sampling from multiple subclades observed in the wild, as well as fine scale variants with identical marker gene sequences but genome-wide variation at the single nucleotide polymorphism level. Genomic information derived from these new cultures has already begun to expand our understanding of the LLIV ecotype, and we are confident these strains will be a valuable addition to the *Prochlorococcus* toolkit moving forward.

2.1 Introduction

The value of culturing *Prochlorococcus*

Prochlorococcus is the most abundant photoautotroph in the oceans and a major contributor to biogeochemical cycles globally (Partensky et al., 1999). It is unique among the cyanobacteria for its small cell and small genome size, in some ways representing a minimal photosynthetic system (Partensky and Garczarek, 2010). The study of *Prochlorococcus* has blossomed in recent years through the application of a large array of modern biological technologies. We can study *Prochlorococcus* in the field, taking advantage of its unique flow cytometric signature, which enables its rapid identification, counting and sorting (Chisholm et al., 1988, Moore et al., 1998, Rodrigue et al., 2009), and the fact that, due to its abundance, it makes up a large component of open ocean whole community metagenomic and metatranscriptomic sequencing efforts (Hewson et al., 2009, Rusch et al., 2010, Ottesen et al., 2014, Coleman and Chisholm, 2010, Coleman and Chisholm, 2007). We can also study *Prochlorococcus* in the laboratory, through culturing and the vast array of methodologies that come with it (Biller et al., 2015, Moore et al., 2007, Chisholm et al., 1988, Coleman and Chisholm, 2007). Many other important marine microbes are as of yet uncultured, or more difficult to culture than *Prochlorococcus* (Stewart, 2012). For example, SAR11 is the most abundant bacterium in the sea, the heterotrophic counterpart of *Prochlorococcus* in the vast low nutrient open oceans, but it grows slowly and has only been studied by a few research groups (Carini et al., 2013, Giovannoni and Stingl, 2007, Rappé and Giovannoni, 2003, Dupont et al., 2012). Cultivation-independent molecular methods have made great strides towards understanding the ecological patterns and processes of microbes in the ocean independent of culturing, but many questions can be more effectively addressed with laboratory cultures, like sequencing complete genomes, measuring nutrient usage profiles, demonstrating traits, and separating the distinctive contributions of individuals to processes in complex communities. *Prochlorococcus* cultures have given us a nuanced understanding of light and nutrient physiology across the remarkable phenotypic variation within the group (e.g. Moore et al., 1998, Moore and Chisholm, 1999, Moore et al., 2002, Berube et al., 2014). With the advent of genomics, whole genome sequences from cultures have provided a vivid understanding of *Prochlorococcus* diversity and evolutionary processes (Rocap et al., 2003, Dufresne et al., 2003, Kettler et al., 2007, Coleman et al., 2006, Scanlan et al., 2009, Biller et al., 2014). *Prochlorococcus* cultures also enable the propagation of their phage and the study of the infection process, a critical part of carbon flow and mortality in the ecosystem and of *Prochlorococcus* evolutionary dynamics (Sullivan et al., 2003, Mann, 2003, Mann, 2005, Avrani et al., 2011). Each strain in culture represents just one cell from the 10,000-100,000 *Prochlorococcus* that reside in each milliliter of the tropical surface ocean, but serves as an infinite resource of biomass and information about one clonal lineage.

How is *Prochlorococcus* isolated?

Prochlorococcus isolation has mostly been performed by filtering *Prochlorococcus*-containing seawater to remove larger phototrophs (>0.8 μm), taking advantage of its small size to perform this selective step, then adding nitrogen, phosphorus and trace metal amendments to the seawater, incubating at low light, which generally favors *Prochlorococcus* over other phytoplankton, and then monitoring these enrichments for growth. In one case, two genetically distinct strains of *Prochlorococcus* were isolated from the same water sample by flow sorting two distinct populations from seawater, based on their different light scattering properties and chlorophyll content (Moore et al., 1998). Dilution-to-extinction methods, developed for the isolation of oligotrophic marine bacteria, combine dilution with high throughput culturing techniques to separate individual bacteria, enabling isolation of slow growing bacteria without competition from other members of the community (Giovannoni and Stingl, 2007). These techniques were recently applied to isolate *Prochlorococcus*, incubating dilution samples at low light, which resulted in successful isolation of

several *Prochlorococcus* strains (Biller et al., 2014). Targeted isolation of nitrate-utilizing *Prochlorococcus* cultures were obtained for the first time, by replacing the urea and ammonia used previously in enrichment protocols with nitrate as the sole nitrogen source, clarifying a long standing complication in our understanding of *Prochlorococcus* nitrogen use (Martiny et al., 2009b, Berube et al., 2014). The yields for these *Prochlorococcus* isolation attempts have been low, usually a few strains from each effort, and many failed efforts go unreported, but the principles are simple, and no different than for any other bacterial isolation: give the organism what is needs to grow, and reduce competition and predation. Over time, we are moving from haphazard isolation successes grateful for any *Prochlorococcus* that grows, to an increasingly targeted and robust *Prochlorococcus* isolation practice.

Culturing and purification

Culturing *Prochlorococcus* is increasingly routine and widespread, but still presents some challenges compared to other more easily manipulable bacteria and algae (Moore et al., 2007). *Prochlorococcus* is primarily grown in liquid batch culture in seawater amended with ammonia, phosphate and trace metals (Moore et al., 2007, Rippka, 1988). *Prochlorococcus* can be grown on solid media, primarily through pour plating in soft agar, but with low recovery rates and variable success in isolating colonies (Moore et al., 2007). In most bacteria, the formation of colonies on solid media from a single cell is the usual route to cultivation of clonal, axenic lineages. Without access to this tool for *Prochlorococcus*, purification has been challenging. *Prochlorococcus* (and other phytoplankton) strains often begin as contaminated unialgal cultures, consisting of one photoautotroph strain as well as heterotrophic bacteria. We can isolate single *Prochlorococcus* strains through serial dilution to statistical clones, or in some cases stable unialgal isolates have emerged from enrichments through selection or drift over many passages in batch culture (Moore et al., 2007). The next step is to remove heterotrophic bacterioplankton and create an axenic culture. Axenic strains have been obtained through streak plating unialgal cultures, serial dilution and flow sorting (Saito, 2001, Moore et al., 2005, Moore et al., 2007, Rippka et al., 2000). Research on the interactions between *Prochlorococcus* and heterotrophs has provided considerable insight into why it is often difficult to obtain and maintain axenic *Prochlorococcus* in the past – the contaminants can provide benefits relieving redox stress, and perhaps more (Morris et al., 2008, Morris et al., 2011, Sher et al., 2011). In one study, axenic strains were purified by plating streptomycin resistant *Prochlorococcus* mutants with streptomycin-sensitive 'helper' heterotrophic bacteria, picking colonies, then taking advantage of differential antibiotic susceptibility to remove the heterotrophs (Morris et al., 2008). Another highly effective method of obtaining axenic *Prochlorococcus* to come out of co-culture work is a modification of the dilution-to-extinction method, in which *Prochlorococcus* is diluted into a natural seawater-based medium amended with inorganic nutrients, vitamins, pyruvate, glycerol, acetate and lactose (ProMM media, Berube et al., 2014). ProMM media has two properties that have made the process of obtaining axenic *Prochlorococcus* easier than traditional dilution -to-extinction methods. The presence of pyruvate in the medium enables *Prochlorococcus* to grow from lower cell densities than in its typical medium, which is thought to be due to the fact that pyruvate acts as a quencher of hydrogen peroxide (Berube et al., 2014, Keen et al., 2012). *Prochlorococcus* lacks the enzyme, catalase, which is needed to safely destroy hydrogen peroxide, and hydrogen peroxide can occur at toxic levels in typical culture conditions (Tichy and Vermaas, 1999). This is thought to be one reason why *Prochlorococcus* is more difficult to grow in the absence of other bacteria – the co-occurring heterotrophic contaminants can assist with stress induced by reactive oxygen species (Morris et al., 2011). The addition of peroxide quenching chemicals, like pyruvate, improves the recovery of cultures from dilutions. Sodium sulfide, another peroxide quenching agent, is routinely added to solid cultures of nonaxenic *Prochlorococcus* for the propagation of phage, and another quencher, sodium thiosulfate, has long been an additive in media for growing other microalgae (Lindell, 2014, Wang et al., 2002, Vermaas et al., 1987). Additionally, the fact that

this medium contains high concentrations of organics means that many of the typical marine heterotrophic bacteria present in non-axenic *Prochlorococcus* cultures will grow to high densities if present. Thus, if a well contains these heterotrophic bacteria, it rapidly becomes turbid, and can be removed from further monitoring efforts. Occasionally heterotrophic strains that do not grow to high density in ProMM avoid detection, but can be identified in downstream purity test in additional media types.

Why is *Prochlorococcus* difficult to culture?

We know that *Prochlorococcus* is sensitive to changes in light (Malmstrom et al., 2010) and to redox stress including readily produced hydrogen peroxide (Morris et al., 2011). We know that it is sensitive to chemical contaminants (Mann et al., 2002), which results in tracemetal cleaning techniques being applied to cultureware (Moore et al., 2007). In some shipboard incubation experiments, researchers report that *Prochlorococcus* moved from the ocean to a bottle for hours or days of incubation often quickly disappear or die in controls, which is also true of some other phytoplankton (Fernández et al., 2003, Calvo-Díaz et al., 2011, Gieskes et al., 1979). This 'bottle effect' (Gieskes, et al., 1979) might relate to the chemical changes that occur in confinement – loss of mixing and dilution and loss of grazers, trace contaminants on the bottles themselves, or simply to the challenge of moving *Prochlorococcus* out of the ocean without the stress of changing light, temperature or chemical conditions. As we continue to learn more about *Prochlorococcus* and its interactions with its chemical, physical and biological surrounds, our culturing methods will likely continue to improve.

The *Prochlorococcus* culture collection; the relationship between the culture collection and the oceans

At the time of the most recent major genome sequencing effort (Biller et al., 2014), all *Prochlorococcus* cultures in the MIT collection had their genomes sequenced, bringing the total number of isolates with sequenced genomes up to 39. There are additional cultures without sequenced genomes (e.g. Rocap et al., 2002, Ahlgren et al., 2006a, Roache-Johnson, 2013, Chisholm et al., 1992), and we estimate that the culture collection consists of approximately 60 reported strains. However, there may be many more than are currently published, in labs around the world. There are many potential additional strains under development just in the Chisholm group, but they are not stable, classified or purified yet, so await future work for publication (personal communications, Jamie Becker, Kristen LeGault, Steve Biller, Paul Berube). Many but not all of the established strains are available for purchase through major culture collections (e.g. the Roscoff Culture Collection, the National Center for Marine Algae and Microbiota), so the work of individual laboratories is important for maintaining strain diversity for the research community at large.

Prochlorococcus can be divided into genetically and ecologically distinct groups called ecotypes. These ecotypes fall into two broader groups, the low-light adapted (LL, clades LLI-VII) and the high-light adapted (HL, clades HLI-VI) *Prochlorococcus* (Figure 2.1), based on their distribution over the water column and range of light supporting growth for cultured representatives (Rocap et al., 2002, Moore et al., 1998, Biller et al., 2015). Existing *Prochlorococcus* cultures are distributed over six of these clades, (LLI-IV and HLI-II) and from this collection we have learned a tremendous amount about the properties distinguishing ecotypes and individual strains and the evolution of *Prochlorococcus*. The scale of *Prochlorococcus* diversity in the oceans is vast. All *Prochlorococcus* are united as a distinct group by their unique pigment and photosystem properties, their open ocean habitat range and phylogenetic affiliation. Yet, there is substantial variation in traits, from the level of deeply branching ecotypes, to variation in nutrient uptake traits within ecotypes linked to selection under different environmental conditions, to allelic and gene content variation even among co-occurring strains with identical marker gene sequences.

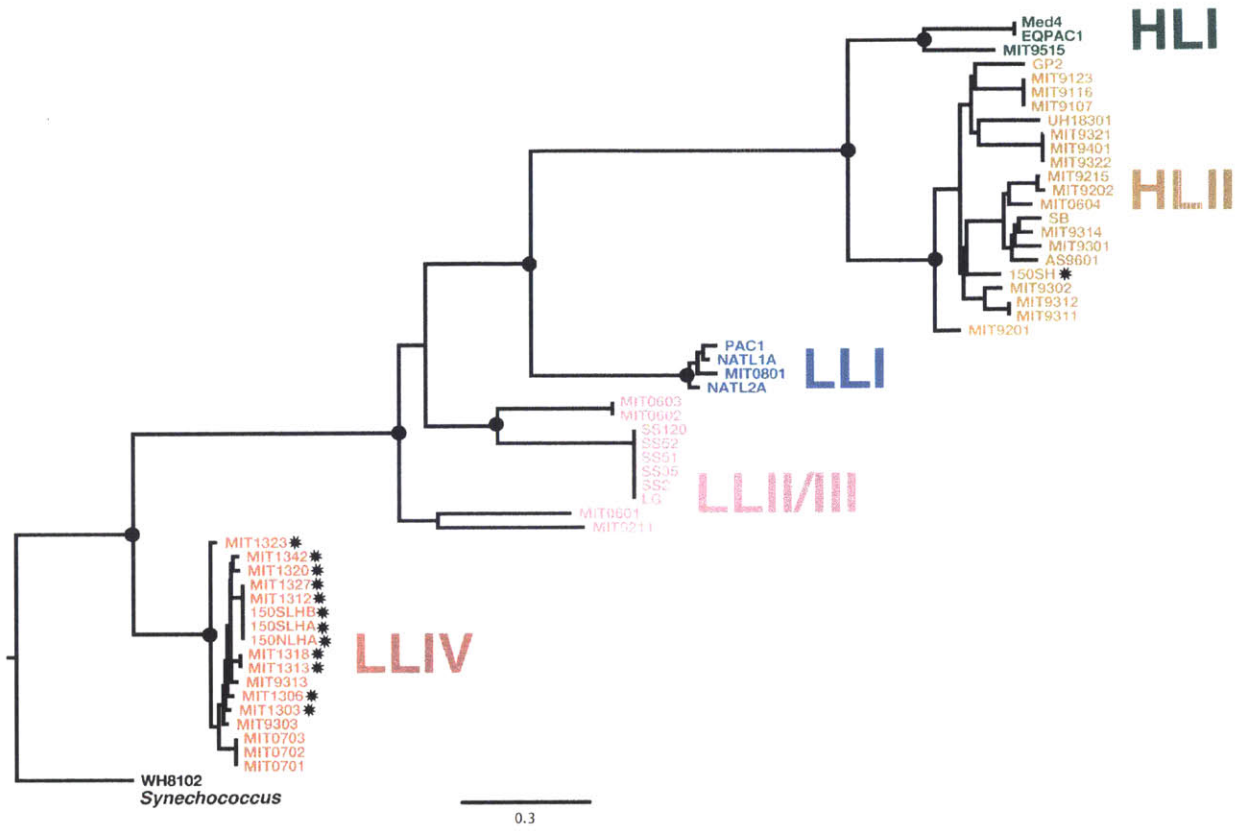


Figure 2.1. Phylogeny of *Prochlorococcus* cultures
This phylogeny contains all the *Prochlorococcus* strains with sequenced genomes, which, remarkably, at this point in time, include most of the *Prochlorococcus* strains in culture i.e. only a dozen or so published *Prochlorococcus* strains do not have sequenced genomes, although there are no doubt unpublished strains awaiting analysis (Biller et al., 2014, e.g. Ahlgren and Rocap, 2006, Roache-Johnson 2013). New strains isolated in this study are marked with stars. HLI, HLII, LLI, LLII, LLIII and LLIV are the cultured *Prochlorococcus* ecotype names, labeling their corresponding clade, strain names at tree are leaves are colored by ecotype to match. This phylogeny was built with GyrB DNA, which is a good phylogenetic marker for *Prochlorococcus*, approximating the lineage phylogeny (Mühling et al., 2012). Maximum likelihood phylogeny (phyml TN93+pinv+gamma4 with 100 bootstrap replicates). Major nodes with greater than 90% bootstrap support (ecotypes) are labeled with a black dot, nodes within ecotypes are sometimes well supported and sometimes not, this marker gene is not ideal for resolving differences at this scale (see instead ITS phylogeny, Figure 2.11).

Each *Prochlorococcus* genome contains roughly 1,800-3,000 genes, which include a set of about 1,200 genes shared by all *Prochlorococcus*, referred to as the core genome, and hundreds of genes with variable distributions across *Prochlorococcus*, or the flexible genome. The *Prochlorococcus* pangenome (all the genes that are in all the *Prochlorococcus*), is currently at 13,000 genes (based on roughly 150 genomes, including single cell genomes), and each new genome reveals additional genes (Biller et al., 2015). A theoretical analysis estimated that the *Prochlorococcus* pangenome in the global oceans contains 80,000 genes (Baumdicker et al., 2012). We have only begun to sample from the sheer vastness of *Prochlorococcus* diversity present within a drop of water. Metagenomics and sequencing of PCR clone libraries of the internal transcribed spacer (ITS) between 16S and 23S bacterial rRNA (a high resolution phylogenetic marker commonly used for within-genus or within-species comparisons or barcoding) has revealed many uncultured clades of *Prochlorococcus*, on the same deep branching phylogenetic scale as the cultured ecotypes. The NC1 (also known as LLVII) clade is found deeper in the water column at many sites, (Martiny et al., 2009b, Shibl et al., 2014, Jiao et al.,

2014). The LLV and LVI ecotypes are found in oxygen minimum zone waters, where an anoxic region overlaps the euphotic zone (Lavin et al., 2010, Astorga-Eló et al., 2015). The HLIII, HLIV, HLV (also called HNLC) clades occur in high-nutrient low-chlorophyll iron-limited regions of the ocean (West et al., 2011, Rusch et al., 2010, Malmstrom et al., 2013). The HLVI clade, sister to the well-studied most abundant HLII clade, is found deeper in the water column, intermediate in depth (but not phylogeny) between other HL clades and LL clades (Huang et al., 2011). Bringing any of these clades into culture would enable us to test hypotheses about their traits that have been suggested by their ecological distributions and genomic information gleaned for these clades from metagenomes and single-cell genomes.

To track the distribution of *Prochlorococcus* ecotypes through time and space in the oceans, ecotype-specific ITS qPCR and probe methods have been widely applied, targeting regions of the ITS that are shared within ecotypes but variable between them (Bouman et al., 2006, Ahlgren et al., 2006, Zwirglmaier et al., 2007). For surface waters, the numbers of *Prochlorococcus* measured through qPCR summed over all ecotype primer sets match cell counts obtained via flow cytometry well for many locations, indicating that existing primers capture the full population. Deeper in the water column, however, at the based of the euphotic zone where low light *Prochlorococcus* dominate, flow cytometry counts consistently indicate that there are more cells than our primers can recognize – more diversity we have not sampled (Ahlgren et al., 2006, Zinser et al., 2006, Zinser et al., 2007, Johnson et al., 2006). Part of this population could be explained through the NCI/LLVII clade, but clades not yet observed or variation at priming sites within known clades could also contribute. The base of the euphotic zone is a good place to try to isolate novel *Prochlorococcus* strains. These uncultured clades only describe uncultured diversity in terms of the phylogenetic divisions of *Prochlorococcus*. The nature of *Prochlorococcus* genome evolution, with a large pool of variable gene content (the flexible genome) results in traits that are not all vertically inherited; particularly for nutrient acquisition strategies, the environment plays a larger role than phylogenetic affiliation in determining distribution of genes (Martiny et al., 2009abc, Coleman and Chisholm, 2007, Scanlan et al., 2009). Our picture of what is represented in culture or not can also be viewed in terms of these traits and metabolic potential. Sampling from new locations with different chemical conditions, and using different selective nutrient sources in isolation protocols (as for the nitrate utilization example described earlier) would allow us to expand our culture collection from a trait-centric viewpoint.

Targeting Low-light adapted *Prochlorococcus*

We wanted to address the substantial gaps in our *Prochlorococcus* culture collection, by expanding the number of strains in culture through targeted isolation of low-light adapted *Prochlorococcus*. Apart from the gaps represented by uncultured clades, we have only a few representatives in culture for each of the low-light adapted *Prochlorococcus* clades – which are rich in diversity in terms of genetic distances, gene content and functional traits. Low-light *Prochlorococcus* can live deeper in the water column than most phytoplankton, and have larger genomes and cell sizes. All LL ecotypes share the ability to grow at lower light levels, but other aspects of their light physiology, light-related gene content and depth distribution are observed to vary significantly. In qPCR surveys of the water column at many sites, the LLIV clade peaks deepest, at the bottom of the euphotic zone, the LLII and III clades are found with similar distributions, either precisely in step with the LLIV or peaking slightly above them, always below mixed layer, and the LLI ecotype peaks below HL, but above the other LL, and is found in deeply mixed waters (Ahlgren et al., 2006, Zinser et al., 2006, Johnson et al., 2006, Zinser et al., 2007, Malmstrom et al., 2010). LL Prochlorococcus have larger genomes than HL, containing more flexible genome variation in each genome we have sampled so far. Some of this variation is shared across ecotypes, and some is specific to individual strains. Although there are smaller numbers of LL cells compared to HL in the global oceans, each genome contributes more to the pangenome, the pool of genes moving among *Prochlorococcus* populations. Recent publications have

cited the need for the isolation and sequencing of additional LL genomes, because more examples would help us sort out the ecotype-defining traits from environment-specific evolutionary pressures, and they likely contain vast reserves of flexible gene content, as each new LL genome has widely expanded what we know about the metabolic potential of *Prochlorococcus*. We set out to perform targeted isolations of low-light adapted *Prochlorococcus*, on a research cruise in the North Pacific, following established enrichment methods with several modifications. After developing complex enrichments of exponentially growing *Prochlorococcus*, we applied dilution-to-extinction in ProMM medium. These efforts led to the efficient isolation and purification of a collection of different strains from the same water sample, substantially expanding our sample set for the LLIV clade of *Prochlorococcus*. Here, we describe this process, the logic used to target LL strains, early genomic findings and recommendations for future isolation efforts.

2.2 Materials and Methods

Cruise

Samples were obtained during the HOE-PhoR cruise, which took place over May 22-June 5 (e.g. del Valle and Karl, 2014). Samples from 150m resulting in successful isolations were taken 6-2-2013 at Station Aloha. Full information on enrichment conditions prepared at sea are in Supplementary Table S2.2.

Dilution-to-extinction

For successful enrichment sample 150mS, original isolation conditions were 1um filtration of raw seawater, and addition of Pro2 nutrients + 1uM thiosulfate. For subsequently transfers in the lab, the subcultures used for subsequent dilution experiments were grown in media with Hawaii seawater, Pro2 nutrients, 1uM thiosulfate under continuous light of 1-3 μmol photons $m^{-2}s^{-1}$. 150mN sample conditions started out in the original enrichments as nitrite as the only nitrogen source, but by the time of the dilution experiment, the most successful subcultures were growing in Woods Hole seawater-based ESL Pro99 (Moore et al., 2007)

Light selection simplification

Variants of the two enrichments that yielded cultures in dilution experiment (150mS and 150mN), were maintained in batch culture in Pro99 Sargasso Sea water at low light (approximately 1 μmol photons m^{-2} s^{-1}), and even more than a year after sampling from the sea, these remained complex, with multiple *Prochlorococcus* flow cytometry signatures and too complex a population to obtain a clean ITS sequence through PCR products (indications of multiple *Prochlorococcus* genotypes). In March 2014, in order to select for different subsets of *Prochlorococcus* diversity, these enrichments were split in to two conditions, either moved to higher light (approximately 12 μmol photons $m^{-2}s^{-1}$), still low in the range of *Prochlorococcus* growth but an order of magnitude higher than acclimated light, or kept at low light (approximately 1 μmol photons m^{-2} s^{-1}), under continuous illumination, 24°C. Some *Prochlorococcus* in the 150mS sample survived the transition to higher light, and this sample now contains what appears to be a unialgal, nonaxenic HLII strain. This exercise was repeated a few months later (June 2014), because the samples at low light continued to grow as complex enrichments. This time, two duplicate aliquots each of the 150mN and 150nS enrichments growing at low light were moved to higher light (10uE). Some *Prochlorococcus* in all four samples survived; three of these have stable LLIV ITS sequences (150NLHA, 150SLHA, 150SLHB) and genomic sequences consistent with a highly simplified population (perhaps unialgal), but the fourth remained mixed.

ITS-rRNA PCR conditions

For sequencing the ITS-rRNA marker gene, for initial characterization of strains and subsequent checks for stability, we used conserved primers targeting *Prochlorococcus* and *Synechococcus* conserved regions at the end of the 16S and 23S rRNA genes, enabling robust amplification of the variable ITS region, which is a good high resolution barcode for *Prochlorococcus* strains. For routine PCR from *Prochlorococcus* cultures, it is not necessary to perform a full DNA extraction – we get good amplification from the following rough extraction method, akin to direct colony PCR, but with a step to remove seawater salts, which can interfere with PCR, and concentrate biomass. First, we spin down 1.0 ml of a culture (10^6-10^8 cells ml^{-1}, usually anything with visible color forms a pellet, the denser the culture the better), in a microcentrifuge tube for 15 minutes at 16,000g, RT, or until a pellet forms, sometimes as long as 30min, depending on the strain and density of the culture. All spent media was removed with a pipette, and the pellet was spun down again, for 1 minute at 13,000-16,000g, and again, residual seawater was removed with a pipette. The pellet was resuspended in 25-100ul (depending on size of pellet) of Tris-HCl pH 8.0, Tris-EDTA or PCR quality 18MΩ

water, by pipetting up and down and vortexing. The cells were lysed by boiling for 10 minutes. The resulting lysed cell mixture was centrifuged for 5.0 min at 4C, to pellet cell debris, and the supernatant was removed, stored at -20C and used as PCR template. PCR primers (ITS-F: 5'-CCGAAGTCGTTACTYYAACCC-3' and ITS-R 5'-TCATCGCCTCTGTGTGCC-3') and conditions are as in Rodrigue et al, 2009. PCR products were purified and sequenced by Eton Biosciences, Cambridge, MA.

Flow cytometry conditions

Flow cytometry was performed on a BD/Cytopeia Influx Cell Sorter, using 488nm excitation argon laser, and emission filters primarily for chlorophyll red fluorescence (680nm/40 bandpass) and phycoerythrin orange fluorescence (580nm/30 bandpass). Occasionally in the course of monitoring axenicity, cultures were run stains with SYBR-green, which fluoresces when bound to double-stranded DNA, monitored with a 530nm/40 emission filter near the molecule's peak emission, to compare the number of likely DNA-containing particles with the chlorophyll-based *Prochlorococcus* measurements. All flow cytometry data was analyzed using the FlowJo software package (www.flowjo.com).

Axenicity tests

We assessed the axenicity of our cultures using a panel of organic media, one minimal seawater-based and two rich broths – ProMM, ProAC and Marine Purity Test Broth (Berube et al., 2014, Morris et al., 2008, Saito, 2002), and through flow cytometric analysis of populations of phototrophs and heterotrophs, using chlorophyll channels (488ex, 690/40em, 635ex, 690/40em) and the DNA stain SYBR-green (488ex, 530/40em). In some cases we also used microscopic inspection to assess the presences of heterotrophs.

Culturing conditions

After initial enrichment in specialized media, cultures and enrichments described here were maintained in batch culture for purposes of DNA extraction and culture maintenance in standard *Prochlorococcus* culture conditions, using the Pro99 media (Moore et al., 2007) based on either Vineyard Sound coastal seawater obtained from the Environmental Systems Lab of the Woods Hole Oceanographic Institute, or Sargasso Seawater obtained from cruises.

Genome library prep and sequencing and assembly

Genomic DNA sequencing libraries were prepared as in Rodrigue et al., 2010, and sequenced on an Illumina MiSeq at the MIT BioMicroCenter, in two separate batches, with library insert sizes around 350 basepairs. Genomes were assembled with Spades (version 3.1.1). Contigs shorter than 500 bp were removed. For non-axenic cultures BLAST against the NCBI nr (non-redundant) database was used to identify and remove contaminating heterotrophic sequences.

GyrB phylogeny

Alignment of GyrB DNA sequences for all available *Prochlorococcus* genomes was performed with muscle (default settings; Edgar, 2004), which performed well. Maximum likelihood tree was built using phyml (Guindon et al., 2010)., with TN93 + pinv +gamma4 model and 100 bootstraps, which produced the same major branching topology as the GTR model, with slight variations within ecotypes - not really enough resolution in this gene for that scale anyway). Figtree visualization.

ITS phylogeny

The fine scale ITS phylogeny, just the LLIV cultures (figure) was produced from a muscle (v 3.8.31, default settings) alignment (Edgar, 2004). The LLV OMZ-associated outgroup sequence was identified from

Lavin et al., 2010. Phylogeny was produced in PhyML 3.0 (Guindon et al., 2010), with the Tamura-Nei 1993 model of nucleotide evolution, 4 gamma-distributed rate categories and a modeled proportion of invariant residues, with 100 bootstrap replicates. Figtree was used to render all trees as images (http://tree.bio.ed.ac.uk/software/figtree/). Previously cultured strains include two LLIV strains from the North Atlantic, MIT9313 and MIT9303 (Moore et al., 1998), both with published genomes, one axenic (MIT9313), and a group of seven closely related isolates (mostly identical ITS sequences) from the South Atlantic (MIT0701, MIT0702, MIT0703, SA-B7, SA-B5, SA-C4, SA-C8; Biller et al., 2014), three of which have sequenced genomes, none of which are axenic.

Ecotype primer analysis

To assess the possible sensitivity of ecotype enumeration primers to our cultures, we aligned the ITS sequences from LLIV cultures to the qPCR primers for the e9313 ecotype, known as 93031f, CAACGAGCCAATGGTGAGAA and 93133r, GGCTTCAATCTCAAACTTCTCC, originally from Ahlgren et al., 2006, used subsequently for characterization of ecotype distributions in Johnson et al., 2006, Zinser et al., 2006, Zinser et al., 2007 and Malmstrom et al., 2010. Alignment was performed in Geneious 6.0.5 (Kearse et al., 2012), and the alignment image in Figure 2.13 was produced in Jalview (Waterhouse et al., 2009)

Comparison of LLIV cultures with uncultured ITS sequences from wild clone libraries

To obtain LLIV ITS sequences from published clone libraries, for the purpose of placing our cultures in the larger context of wild diversity, we performed a blast search (blastn algorithm) using the MIT9313 full length ITS as query agains the nr (nonredundant) database, which includes several published *Prochlorococcus* ITS clone libraries (search implemented 3/28/2015, through web server http://blast.ncbi.nlm.nih.gov/Blast.cgi). We took the top 500 hits, which included LLV and LLVI OMZ-associated *Prochlorococcus* clades, indicating sufficiently deep sampling to include all LLIV representatives in the database and then some more distant sequences. To refine this set to only LLIV and closely related sequences, for more accurate fine scale comparisons (it is difficult to align the ITS across large distances and many sequences), we performed a rough pairwise clustering (UPGMA, based on pairwise global alignments, implemented in mafft, Katoh et al., 2013), and used this to select only sequences in a large cluster containing LLIV cultures, LLV and LLVI uncultured sequences, excluding more distantly related sequences, which also had lower GC content typical of other Prochlorococcus clades. Using this set of 386 uncultured ITS sequences, which represent several publications and many ocean samples, along with our 17 cultures, we built a multiple alignment in muscle (v3.8.31, defaults produced a reasonable alignment, Edgar, 2004) and, based on this alignment, trimmed all sequences to the full length ITS start and stop positions. Some wild sequences were not the full length, but remain in the analysis. The trimmed sequences were then realigned with muscle using default settings (Edgar, 2004). This alignment was used to construct the phylogeny in Figure 2.14, with the fasttree approximate likelihood method (Price et al., 2010), using the gtr evolutionary model and gamma option for likelihood calculations and rates. Tree was rooted on the LLVI clade (based on the branching relationships between LLIV, LLV, and LLVI in Lavin et al., 2010), and visualized in Figtree (http://tree.bio.ed.ac.uk/software/figtree/). We used a BLAST search to gather LLIV sequences from the NCBI database containing published *Prochlorococcus* ITS clone libraries (nr), and sampled deep enough to also capture the LLV and LLVI clades, the closest related clades, ensuring that we spanned the full range of LLIV variation. These clone libraries certainly do not represent the full diversity in the oceans; there may be depths of undiscovered diversity within the clade, and the geographic range is biased towards the North Atlantic (where MIT9303 and MIT9313 came from) and North Pacific (where the strains reported here came from).

Genome and marker gene nucleotide identity comparisons

Average nucleotide identity (ANI) was calculated for each pair of LLIV genomes according to the blast-based method of Goris et al., 2007, using the ANI webserver (http://enve-omics.ce.gatech.edu/ani/). Default options were employed: minimum length 700 bp, minimum identity 70%, minimum alignment 50 bp, window 1000 bp, step size 200 bp. Values shown represent 2-way ANI. Additional distances are based on muscle (v. 3.8.31 default, Edgar 2004) alignments of the ITS and 16S regions for all LLIV cultures, with distances calculated in Geneious (Kearse et al., 2012), rendered as images using matplotlib (Hunter, 2007).

Mauve alignments for whole genome comparison

For a rough assessment of shared and unique genomic content, genomes were aligned using the ProgressiveMauve algorithm. Custom python tools were used for calculating the fraction of each genome in the alignment from the mauve backbone file containing genomic coordinates for regions contained in the alignment. For assessing clonality from whole genome assemblies, mauve alignments were analyzed for SNPs, using the SNP calling function and visualization of locations and density of SNPs across contigs in the Mauve alignment viewer and Geneious. Strains were called clones if there were only 10s or 100s of SNPs located primarily at the ends of contigs (low quality/ambiguous parts of assembly), and were called different strains if there were hundred or thousands of SNPs located at the center of contigs; these strains usually had indels too. In the future we will analyze SNPs in the raw Illumina data to confirm clonality of the culture's population.

2.3 Results and Discussion

2.3.1 A program for targeted isolation of low-light adapted *Prochlorococcus*

Targeted enrichment program for low light adapted *Prochlorococcus*

Taking advantage of relatively frequent and accessible sampling opportunities in the oligotrophic North Pacific Subtropical Gyre, we set out to isolate new low-light *Prochlorococcus* strains on the HOE-PhoR cruise to the Hawaii Ocean Time series sampling site, Station ALOHA, May 22- June 5. This site has been extremely well characterized (Karl and Church, 2014), including a times series following *Prochlorococcus* ecotype distributions over the water column over several years, using qPCR (Malmstrom et al., 2010). By looking at this timeseries data (Figure 2.2), for the time of year of our sampling (between the historical May and June sampling dates), we could predict that to sample populations containing significant numbers of LL *Prochlorococcus*, we should sample below 100m, and to sample LL-dominated populations, before the total numbers of *Prochlorococcus* decline too much (Figure 2.3), we should sample around 150m. At this site at this time of year, the other globally abundant marine picocyanobacterium *Synechococcus* is approximately 2 orders of magnitude lower in abundance than *Prochlorococcus* (Figure 2.3, Malmstrom et al., 2010).

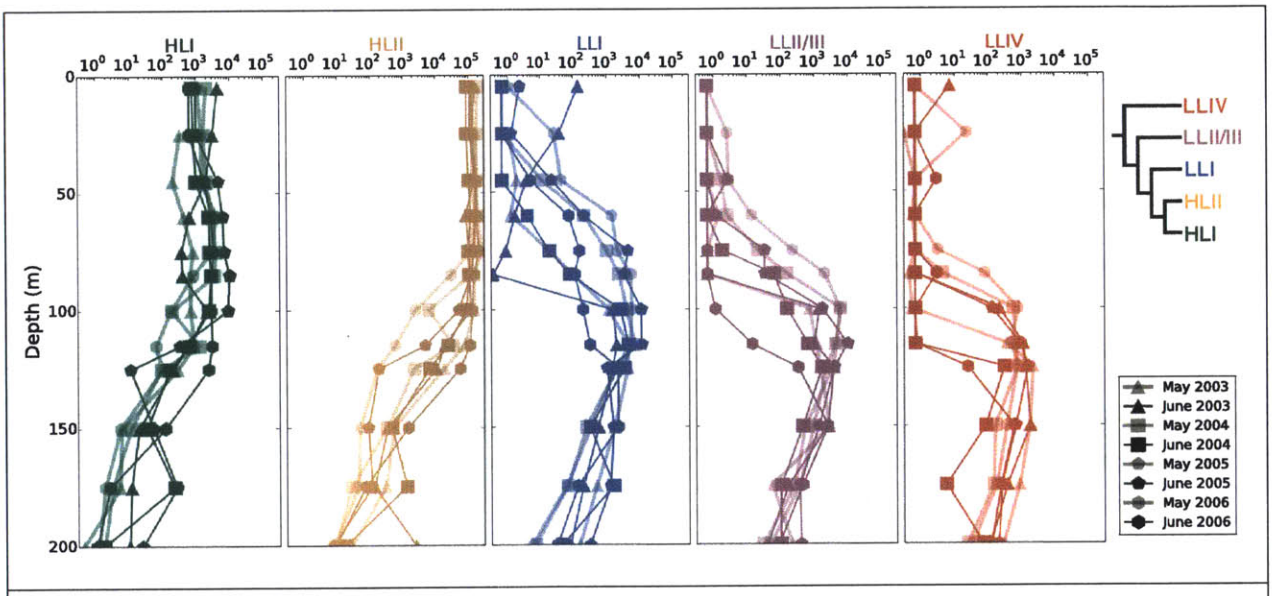


Figure 2.2 Four years of May and June Station ALOHA ecotype time series data from Malmstrom et al., 2010
Raw data from Malmstrom et al., 2010 provided for our analysis by Allison Coe, Chisholm Lab.

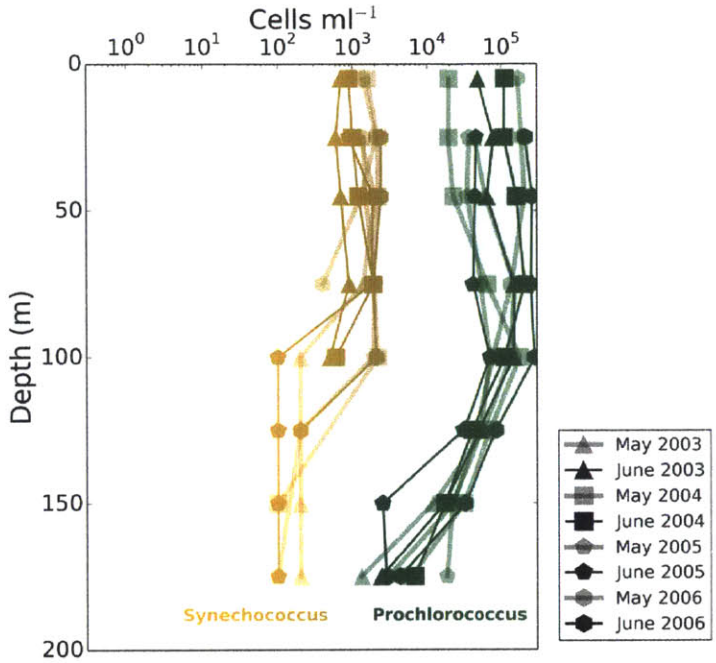


Figure 2.3 Four years of May and June Station ALOHA *Synechococcus* and *Prochlorococcus* counts
Raw data from Malmstrom et al., 2010 provided our analysis by Allison Coe, Chisholm Lab.

We obtained the small samples required for isolations and started the enrichment process at sea (see Figure 2.4). We followed established protocols for isolation (Moore et al., 2007) with some modifications. To accommodate the characteristics of LL *Prochlorococcus* we paid special attention to sample depth, cell size selection and light conditions. We collected water from deep in the euphotic zone (150m) and filtered it with a larger pore size (1.0 μm) than often used in the past for *Prochlorococcus* isolation (0.6-0.8 μm), to ensure passage any of the the slightly larger LL-adapted *Prochlorococcus*, while still removing large phytoplankton and grazers. *Synechococcus* tends to be slightly larger than *Prochlorococcus*, overlapping this size cutoff. Whether because *Synechococcus* concentrations tend to be low at this site and time (Figure 2.3), or because this size selection effectively removed *Synechococcus* from our samples, we did not observed *Synechococcus* in enrichments upon their return to lab (Figure 2.5). We took these samples toward the end of the cruise, to minimize the time between sampling at sea and controlled incubation conditions in the laboratory. Because LL *Prochlorococcus* are sensitive to high light, care was taken to control light conditions, from ocean to lab. We sampled into opaque bottles, performed all manipulations in low indoor light, incubated samples at sea in containers covered with blue gels to allow less light in, and maintained controlled, low light levels through transport until the strains were in controlled incubators in the lab. A light meter was used throughout this process to ensure that cells mostly experienced irradiation at 1 μmol photons m^{-2} s^{-1}, equivalent to 0.1% surface irradiance of typical daytime sunlight, similar to the cells' original habitat and selective for low-light strains based on measurements of light-growth rate relationships across *Prochlorococcus* strains.

In some of the original isolations we tried using hydrogen peroxide quenching agents (thiosulfate, pyruvate) as additives in initial enrichments, as well as amino acids and the antibiotic nalidixic acid, to which *Prochlorococcus* is resistant. We tried some enrichments with nitrite as the only added nitrogen source because most LL strains tested to date are able to use nitrite as a sole N source, while this trait has a patchier distribution across HL cells (Moore et al., 2002, Berube et al., 2014, Supplemental Table S2.2). We

attempted several samples with amino acid additions in this study, which encouraged heterotrophic growth as a carbon source, but *Prochlorococcus* also grew, in some samples for some time, although never into a robust culture. Getting live seawater from sea back to the lab without exposing cells to major changes of temperature or other stressors is a challenge. We split the samples between hand-carried luggage and an overnight box container with the rest of the materials. Samples from both contained plenty of *Prochlorococcus* when they arrived in the lab, and samples from both methods showed some growth in early transfers. For the shipping containers, cells were given light from an inexpensive LED booklight, shaded with blue gels to a light level 1 μmol photons m^{-2} s^{-1}. For the hand-carried luggage, samples were wrapped in blue gel and occasionally exposed to ambient light.

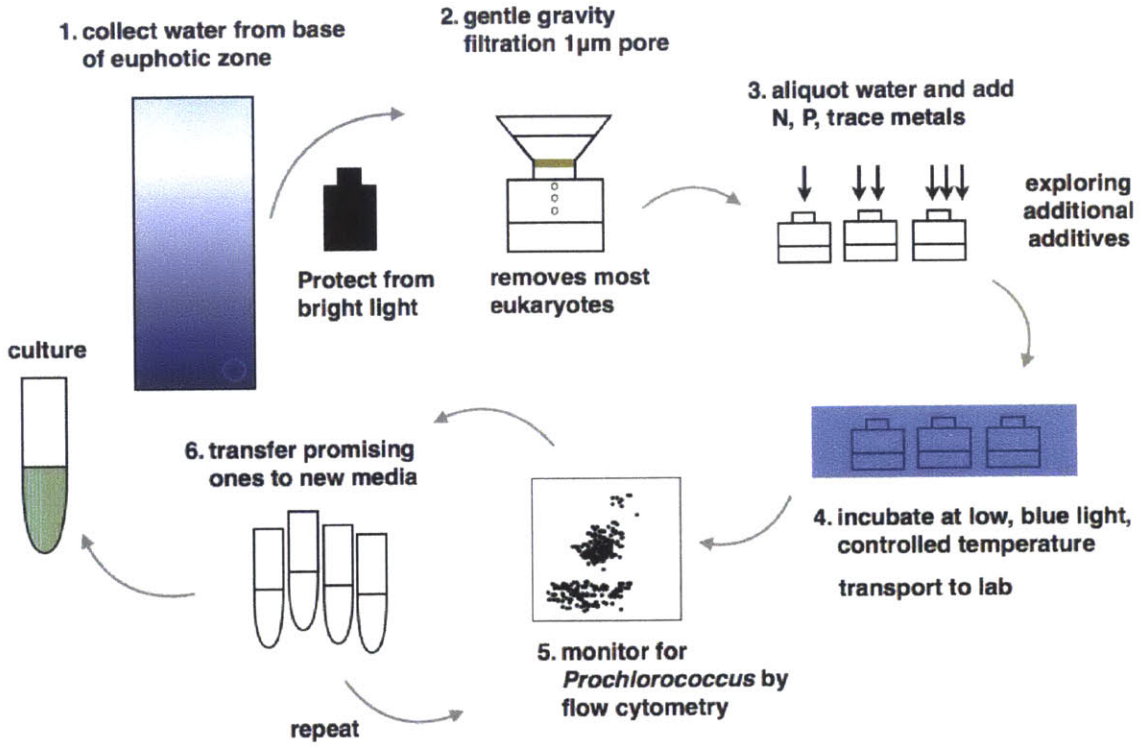


Figure 2.4. A process for targeted enrichment of LL *Prochlorococcus*: from dimly lit seawater to the robust growth of culture in the lab
Here we outline the basic enrichment procedure followed in this work, described in detail in the text. This procedure is largely similar to the isolation methods reviewed in Moore et al., 2007, but with several choices made and steps added to specifically target LL *Prochlorococcus* ecotypes (sampling depth, filtration size, light selection, flow cytometry evaluation of ecotype).

Initial enrichment conditions result in a low yield of successfully growing enrichments

When back in the lab, enrichment samples were placed in incubators at low light, with controlled temperature, and monitored by flow cytometry for signs of surviving or growing *Prochlorococcus*. Each time we detected *Prochlorococcus* in a sample, we transferred a portion into fresh media. Sometimes after a few weeks the *Prochlorococcus* would disappear from those transfers, other times they would survive. At this stage the cultures were colorless, and the chlorophyll was below the level of detection with our bulk fluorometer

but easily detectable by flow cytometry (Figure 2.5). Eventually some turned green. During these transfers, we were continually splitting enrichments into new conditions, slightly different light, different seawater bases for the media (ESL coastal or SSW oligotrophic), different incubators with different temperatures or diel/continuous light patterns, partly to hedge the system toward our goal of finding successful conditions, and partly with the idea of selecting for different traits from the initial enrichments. We did not formally test each of these, but tried a plurality to hedge and explore space.

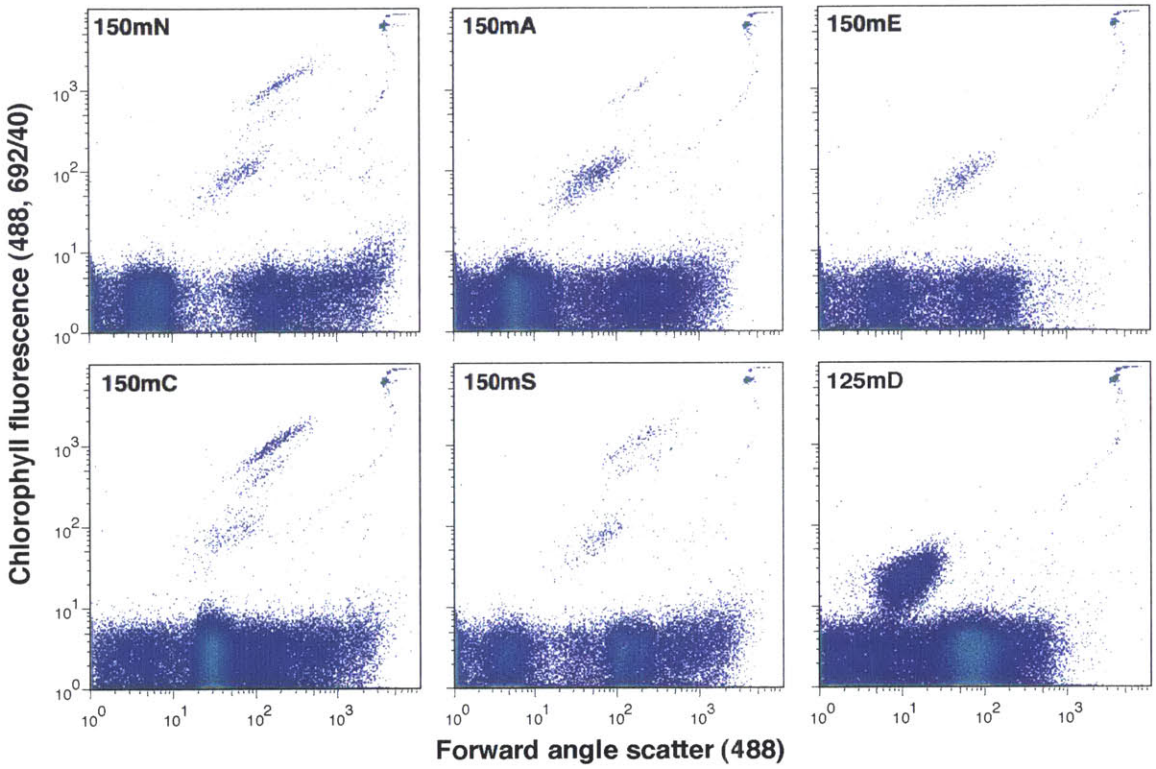


Figure 2.5. What do enrichment samples, fresh from the sea, new to the lab, look like?
Flow cytometry signatures (forward angle scatter vs chlorophyll fluorescence) show that these early enrichments, viewed 2 weeks after original sampling from the sea, contained *Prochlorococcus*, an important indication of success at this initial stage. These signatures are somewhat similar to raw seawater samples (e.g. Chapter IV). However, they show reduced *Prochlorococcus* populations, and both the *Prochlorococcus* and heterotroph populations are highly variable from sample to sample, indicating that these communities have been influenced by their diverse chemical and physical handling in the time from sea. By way of illustration, these plots represent only a small subset of the cultures which returned to the lab, most of which still contained some *Prochlorococcus*, with similar signatures. Of particular note, samples 150mN and 150mS below eventually grew to be the only successful high-density enrichments from these efforts, despite humble beginnings as average samples in early stages.

Two of our enrichments eventually turned green, both the original tubes and transfers from those tubes (Supplemental Figure S2.1). Visible color roughly corresponds to a density $>10^6$ cells ml^{-1} (above natural ocean densities but convenient densities for biomass applications). Approximately ten others grew successfully through multiple transfers, but never reached densities detectable with bulk fluorometer or by eye, and so were labor-intensive to determine status and transfer timing through flow cytometry. We let these cultures go, and focused efforts on higher density cultures in this round of efforts, as a matter of convenience. High density is not an absolute requirement – there can be valuable cultures that do not reach

high densities in the lab, like the SAR11 system, that is monitored with high throughput flow cytometry. Such a method could be applied to *Prochlorococcus* in the future. Further work is need to rigorously test conditions of additives for routine culturing and controlled selective conditions, but one thing we learned from this work is the value of a plurality of conditions. Nearly every sample still had *Prochlorococcus* when it arrived at the laboratory, many showed growth or persistence in different enrichment conditions over time, but only two samples achieved high-density cultures (Figure 2.6, Supplemental Figure S2.1). All subsequently purified strains came from these two enrichments.

2.3.2 Dilution-to-extinction experiments result in purification of multiple strains from two enrichments

Dilution-to-extinction to maximize diversity: choosing complex enrichments

Our dense enrichments contained complex flow cytometric signatures, consistent with multiple HL/LL ecotype-level genetic diversity, and the fact that they had only been out of the ocean for a few months, gave us hope that there might be still other scales of diversity present in them. So, we applied the dilution to extinction in ProMM protocol, previously used to get axenic *Prochlorococcus* from unialgal cultures, to attempt to purify multiple *Prochlorococcus* strains from our complex enrichments. The pair of samples we chose for dilution experiments were typical of all the enrichments and subenrichments that achieved high *Prochlorococcus* densities after five months in the lab (all ultimately derived from the two initial samples 150mN and 150mS), multiple *Prochlorococcus* populations, multiple chlorophyll-free heterotrophic bacterial populations (Figure 2.6). These cultures look very different from established pure cultures, and very different from seawater, which is more like the early enrichment samples (Figure 2.5). These particular cultures were chosen out of about a dozen subcultures of the 150N and 150S enrichments, all in different incubators (diel/continuous) and media (original enrichment chemistry or standard *Prochlorococcus* media, in different seawater backgrounds), because they were growing well and had particularly complex flow cytometry signatures, likely to contain diverse *Prochlorococcus* populations.

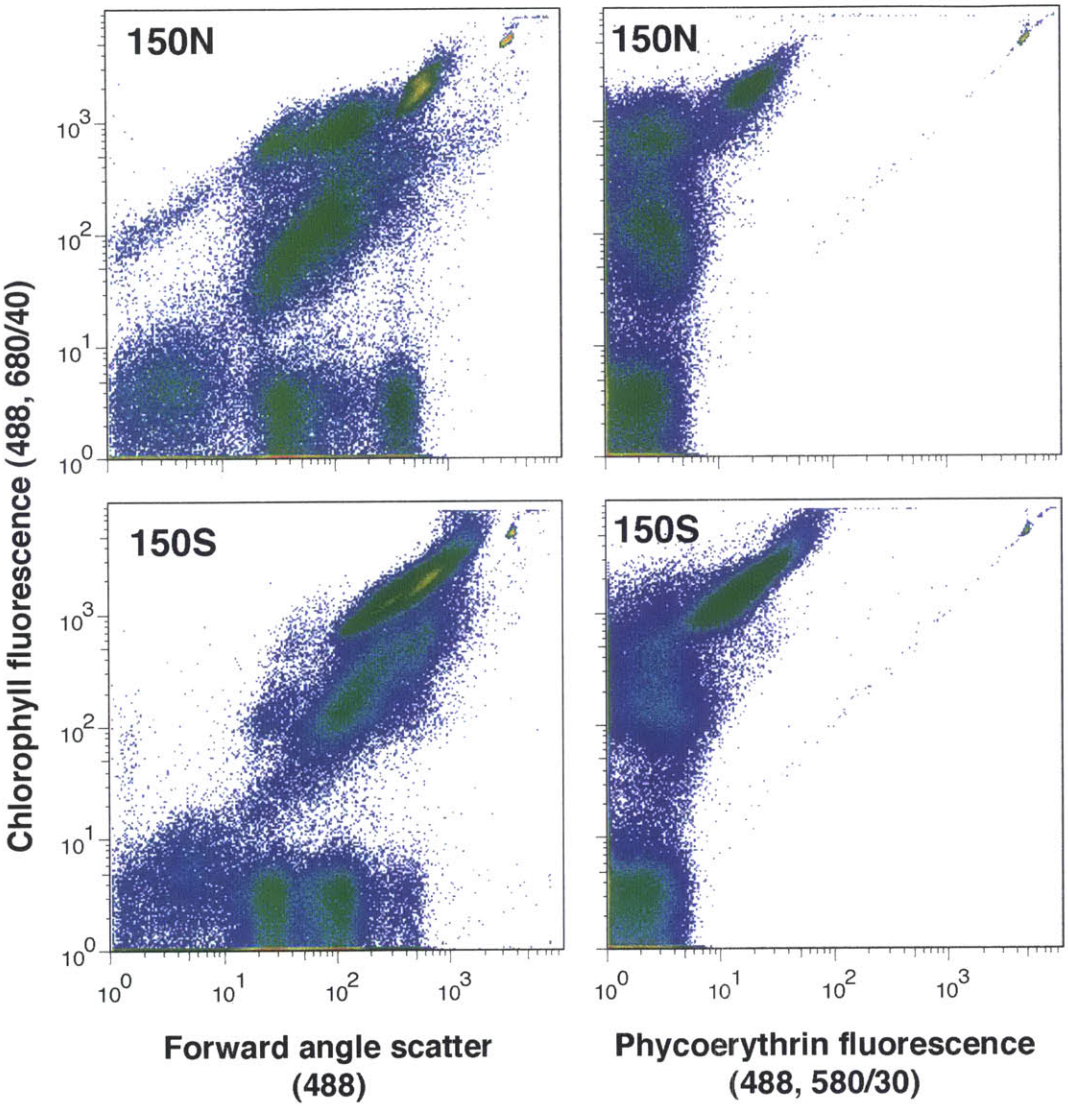


Figure 2.6. Successful, complex *Prochlorococcus* enrichments
Mature enrichments used in the dilution-to-extinction experiments, viewed as sampled on the day of the dilution experiment. These flow cytometry signatures were taken from two successful *Prochlorococcus* enrichments (150mS and 150mN), after four months in the lab and many transfers. When in late stationary phase, these enrichments formed dense green cultures with *Prochlorococcus* concentrations in excess of 10^8 cells ml^{-1}, just like established *Prochlorococcus* cultures. These enrichments have complex flow cytometry signatures with multiple *Prochlorococcus* populations, and there are more *Prochlorococcus* cells than heterotrophs. The y axis represents chlorophyll fluorescence per cell for all four plots (as presented elsewhere throughout this thesis). For the pair of plots stacked on the left, the x-axis represents forward angle scatter, a rough proxy of size. For the pair of plots on the right, the x-axis represents orange phycoerythrin fluorescence, a remnant from ancestral phycobilisomes, present in different amounts across *Prochlorococcus* strains and conditions, with generally more in LL strains, (Hess et al., 1996, Hess et al., 1999, Roache-Johnson 2013, Hess et al., 1996, Scanlan et al., 2009, Steglich et al., 2005, Steglich et al., 2003, Wiethaus et al., 2010). The phycoerythrin data was useful for distinguishing diversity within *Prochlorococcus* enrichments- some populations in these enrichments contain phycoerythrin, and other populations do not, so we diluted these particular ones because we knew, from flow cytometry alone, that they contained substantial *Prochlorococcus* diversity. Standards, YG fluorescent 2um beads, appear as a tight population in the upper left corner of each plot.

Dilution-to-extinction: procedure and results

We took those two promising enrichments and diluted them to about one cell per well for many plates, mimicking the chemical conditions of the original sample amendment (Figure 2.7). In the first few weeks we monitored the plates, marking cloudy wells (contaminated). After six weeks we started to see green wells, and transferred them, first in 96 well plates, to maintain their conditions as much as possible, and then transitioning them into standard *Prochlorococcus* culture conditions (Moore et al., 2007). We continued to see new green wells, even as some wells started to evaporate, up to three months after the start of the dilution experiment. For these new cultures, we tested their axenricity (all but one were clean), and sequenced their ITSrRNA and viewed them on the flow cytometry, all of which data showed we had obtained LLIV *Prochlorococcus* isolates (Figure 2.8, Figure 2.9). We fulfilled our goal of targeted isolation of new low-light adapted *Prochlorococcus* strains.

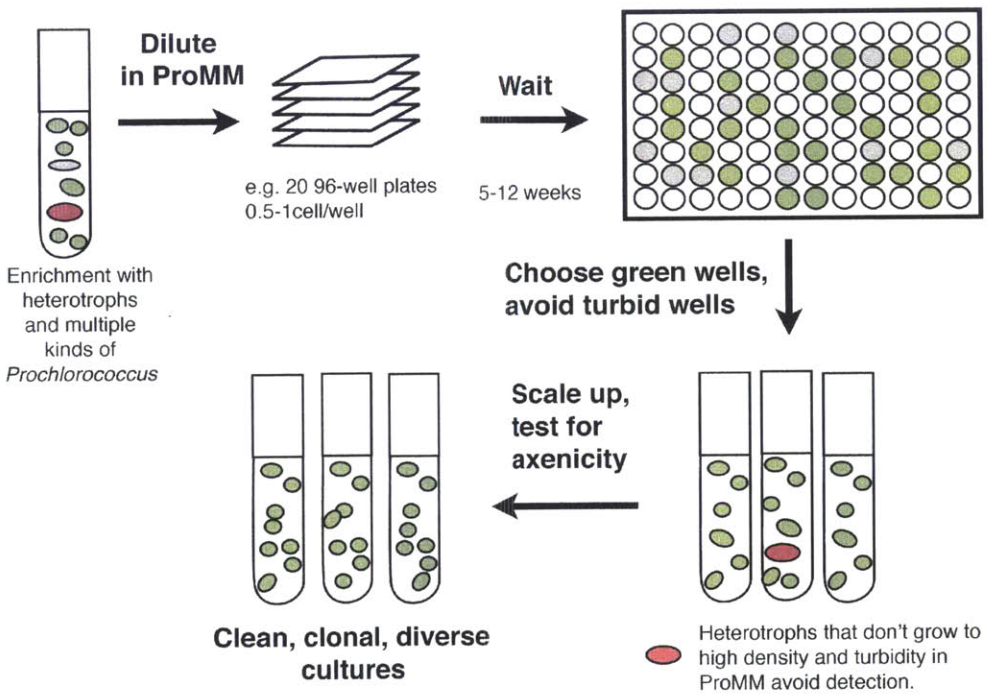


Figure 2.7. Dilution in pyruvate-containing media to obtain axenic, clonal cultures
Outline for the basic dilution-to-extinction procedure followed in this work, described in detail in the text. This procedure is largely similar to the purification method described in Berube et al., 2014, but instead of diluting an established unialgal culture to obtain an axenic version of that culture, here we apply the method to a complex enrichment containing rich *Prochlorococcus* diversity, only a few months out of the sea, to ultimately purify multiple different axenic strains in a single experiment.

Purified cultures: officially named, cryopreserved, stably maintained throughout the lab

Nine unique new LLIV strains came out of dilution purifications from HOT HOE-PhoR enrichments (Table 2.1, Figure 2.8). Eight of these are axenic. They have been stably maintained throughout the laboratory, and are cryopreserved for long term safety of the genetic stock. From this experience, we conclude that early monitoring and transferring of enrichments is of critical importance, and quickly following that enrichment process, while cultures still contain diverse mixtures, with dilution-to-extinction in ProMM is a successful approach for simultaneously isolating multiple diverse *Prochlorococcus* strains.

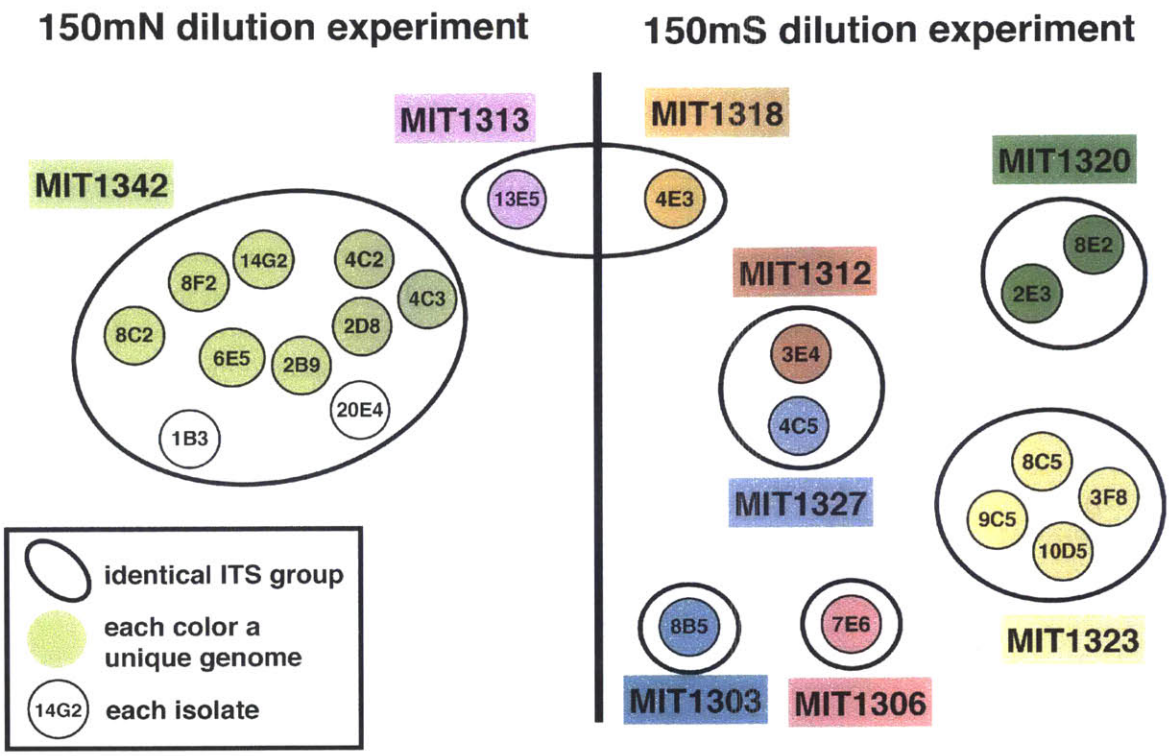


Figure 2.8. Yields of *Prochlorococcus* enrichment dilution-to-extinction experiments
The two high density *Prochlorococcus* enrichment cultures for which we performed ProMM organic dilution-to-extinction yielded 11 isolates each (circles above). 150mN and 150mS refer to the original enrichment samples prepared at sea, grown up in the lab. Each dot in the figure represents a LLIV strain that came out of these dilution to extinction experiments (one successful green well in a plate). Some of these strains were clones of each other (based on genome sequencing) indicated as identical colored dots. Strains with the same ITS sequence, but not necessarily full genome identity, are contained within larger black ovals. 4 additional green wells from these experiments did not take in subsequent transfers, and were lost (not shown). Within circles are identifiers for each green well that came up. The two strains in white, 1B3 and 20E4 have not yet had their genomes sequenced, so they may or may not be clones with their identical-ITS group (although chances are...).

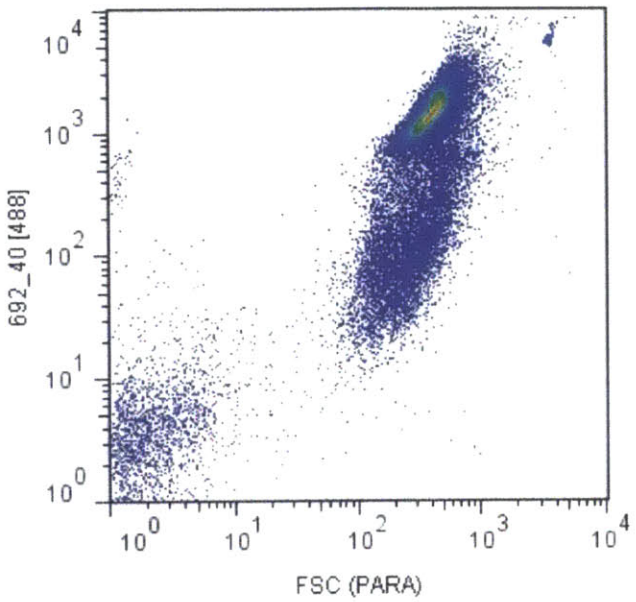


Figure 2.9. Flow cytometric signature of axenic culture MIT1320
After dilution plating, we obtained cultures with clean flow cytometry signatures typical of axenic LLIV *Prochlorococcus*. Note the absence of signal near the x-axis, where in previous enrichments, numerous particles without chlorophyll fluorescence, indicative of heterotrophic bacteria were visible. 2 um standards appear in the upper right corner.

2.3.3 Light selection for further simplifying complex enrichments

Light is a strong selective agent among *Prochlorococcus* lineages

In parallel with the dilution experiment described above, we continued to transfer the complex enrichments from which they came (150mN and 150mS), and they remained complex enrichments (based on complex FCM profiles and our inability to obtain a clean sequence from ITS-PCR products). We find it remarkable that these have remained so complex, under the limited conditions of the lab, and given that historically unialgal cultures were obtained through serial transfer alone. This persistent complexity may be attributable to the continued transfer at very low light (<1 μmol photons m^{-2} s^{-1}), not typical of *Prochlorococcus* culture conditions (10-60 μmol photons m^{-2} s^{-1}, Moore et al., 2007). Growth at low light was originally chosen to be selective for LL strains, but perhaps this slow growth also slows the rate at which competition simplifies *Prochlorococcus* enrichments.

In past *Prochlorococcus* isolation, strains were obtained just by waiting, subculturing and transferring for months and years until competition and bottlenecks at transfer resulted in unialgal, nonaxenic cultures. None of our cultures had reached that stage after one year. Curious if we could culture more diverse strains out of these enrichments before they simplified on their own, we tried applying a stressor – a transition to higher light – on aliquots of the enrichments, to select for a subset of the population that could withstand light transition. The cultures acclimated to 1 μmol photons m^{-2} s^{-1}, and then moved to higher light (8-15 μmol photons m^{-2} s^{-1}). Transitions to higher light are known to be stress for phototrophs, and for LLIV *Prochlorococcus* especially. We hypothesized this stressor would be one way to simplify the population, selection favoring light stress tolerant strains, a phenotype that varies across *Prochlorococcus* (Biller et al., 2015).

Additional strains purified through differential survival of light transitions

We performed this stress selection process twice. The first time, 150mN and 150mS enrichments grown (since sea) at low light (1μmol photons m^{-2} s^{-1}) were transferred each in duplicate to high light (10μmol photons m^{-2} s^{-1}); only 1 of the 4 tubes grew up (visual check only, took ~2 months to reach green). This strain had a single pure ITS, and a simple flow cytometry signature, both stable over time, supportive of the idea that the culture was simplified to a unialgal strain, but not conclusively so, and it was not axenic. This was a HLII *Prochlorococcus* strain, provisionally named 150SH (Figure 2.1, Table 2.1). This is exciting because it means we have now isolated representatives from two ecotypes, HL and LL, from exactly the same water sample, a sympatric pair (as in Moore et al., 1998), which could be valuable for studying the matrix of traits that are differentially specific to environment and phylogeny. In repeating this experiment, moving aliquots of the same low light acclimated enrichments to the same high light conditions again several months later, this time all four tubes transitioned (150N and 150S replicates) grew, with some subset of the complex enrichment surviving the transition. Over 6 months following their initial growth, three of these four are stable likely unialgal cultures, by repeated flow cytometry and ITS sequencing, but this time all three are LLIVs, identical in ITS to MIT1327 and MIT1312. For these three strains, provisionally named 150NLHA, 150SLHA and 150SLHB, we proceeded to genome sequencing. The assembled genomes revealed that although these strains have the same ITS as isolates from the dilution experiment, there are hundreds or thousands of SNPs and indels distinguishing their genomes. These sets may correspond to the nearly-identical-ITS backbone subpopulations, that characterize fine-scale population structure in *Prochlorococcus* in the wild (Kashtan et al., 2014). We have not yet assigned these cultures proper strain nomenclature (MIT13__) because they have not been through dilution-to-extinction purification to improve the likelihood of clonality, but the fact that they assembled neatly into complete genomes of the expected size for a single *Prochlorococcus*, suggests they are highly simplified populations; a little more work is required to make them axenic and statistically clonal.

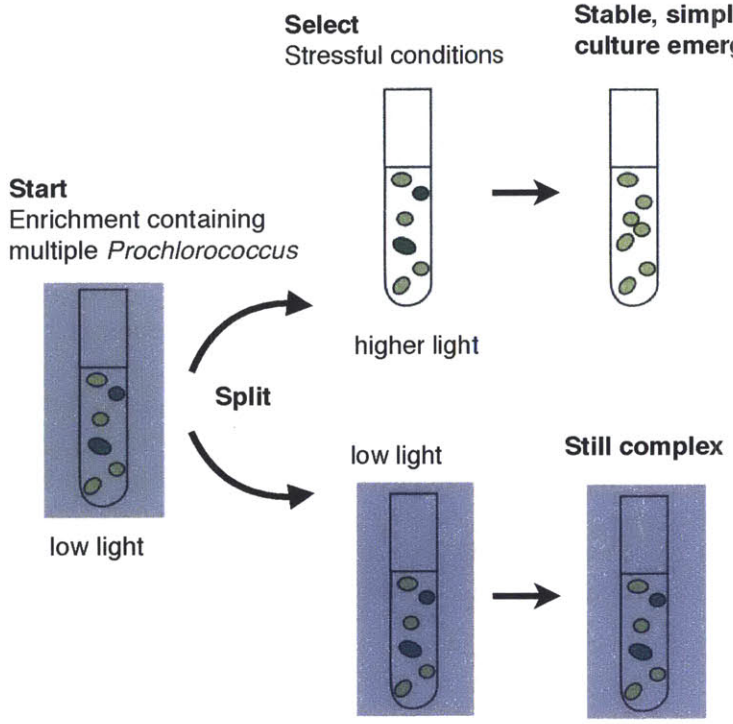


Figure 2.10. Selection through light stress to simplify enrichments Schematic of our approach to further mine our enrichments for novel strains. Enrichments growing with a complex mixture of different *Prochlorococcus* and heterotrophs, were split into aliquots of the and some were transitioned to higher light, which represent stressful, selective conditions for some *Prochlorococcus*, resulting in simplified, and in several cases, unialgal cultures.

Although we know little about within-ecotype light adaptation, it is possible that HL ecotype cells from this deep in the water column, below the mixing layer, like our new 150SH, could be different, in light and nutrient adaptive ways from surface ones (Kashtan et al., 2014). For the LLVI 150SLHA group, light selection may have selected a variant of the LLVI which is more tolerant than others isolated under strictly low light conditions, again opening possibilities for testing the question of whether there is within-ecotype light adaptation variation.

Ongoing mining of enrichments and purification of cultures through additional methods

Nearly two years after their removal from the oceans, some of the enrichments from this study are still growing at very low light (1 μmol photons m^{-2} s^{-1} or less) or moderately low light (10 μmol photons m^{-2} s^{-1}), and still contain complex mixtures of *Prochlorococcus*, based on our inability to obtain a clean ITS sequence using picocyanobacterial-specific primers and, in some cases, still multiple flow cytometry populations indicative of different ecotypes. We are still working to mine these enrichments for yet more strains, through flow sorting and dilution, in the hopes they may contain additional strains we have yet to isolate. Further dilution-to-extinction in ProMM efforts are currently underway to purify the non-axenic light-selected strains and the one nonaxenic (MIT1313) from the dilution experiment away from their heterotrophs.

2.3.4 What have these isolation efforts contributed to the diversity of our culture collection and our knowledge of the LLIV clade?

What did we isolate?

We have so far isolated thirteen unique strains (Table 2.1), all from the same water sample taken from 150 meters in the North Pacific. Twelve of these are members of the LLIV clade (Figure 2.1), the clade found deeper in the water column than other *Prochlorococcus*. One is from the HLII clade, the most abundant group of *Prochlorococcus* at stratified sites, and probably across the world (Bouman et al., 2006, Johnson et al., 2006). Prior to these efforts the LLIV clade had only a small number of reported strains in culture, two from the North Atlantic and three from the South Atlantic. Now, our LLIV collection represents three oceans (Figure 2.11). Eight of our new strains are axenic, tripling the number of axenic LLIV cultures (Figure 2.11).

Table 2.1. New *Prochlorococcus* isolates described in this study

Strain	Ecotype	Axenic?	Initial enrichment	Followup process	ITS group
1312	LLIV	yes	150mS	Dilution-to-extinction in ProMM	1327/1312
1327	LLIV	yes	150mS	Dilution-to-extinction in ProMM	1327/1312
1313	LLIV	no	150mN	Dilution-to-extinction in ProMM	1313/1318
1318	LLIV	yes	150mS	Dilution-to-extinction in ProMM	1313/1318
1303	LLIV	yes	150mS	Dilution-to-extinction in ProMM	unique
1306	LLIV	yes	150mS	Dilution-to-extinction in ProMM	unique
1320	LLIV	yes	150mS	Dilution-to-extinction in ProMM	unique
1323	LLIV	yes	150mS	Dilution-to-extinction in ProMM	unique
1342	LLIV	yes	150mN	Dilution-to-extinction in ProMM	unique
150SHL	HLII	no	150mN	LL->HL transition	unique
150S LH A	LLIV	no	150mS	LL->HL transition	1327/1312
150S LH B	LLIV	no	150mS	LL->HL transition	1327/1312
150N LH A	LLIV	no	150mN	LL->HL transition	1327/1312

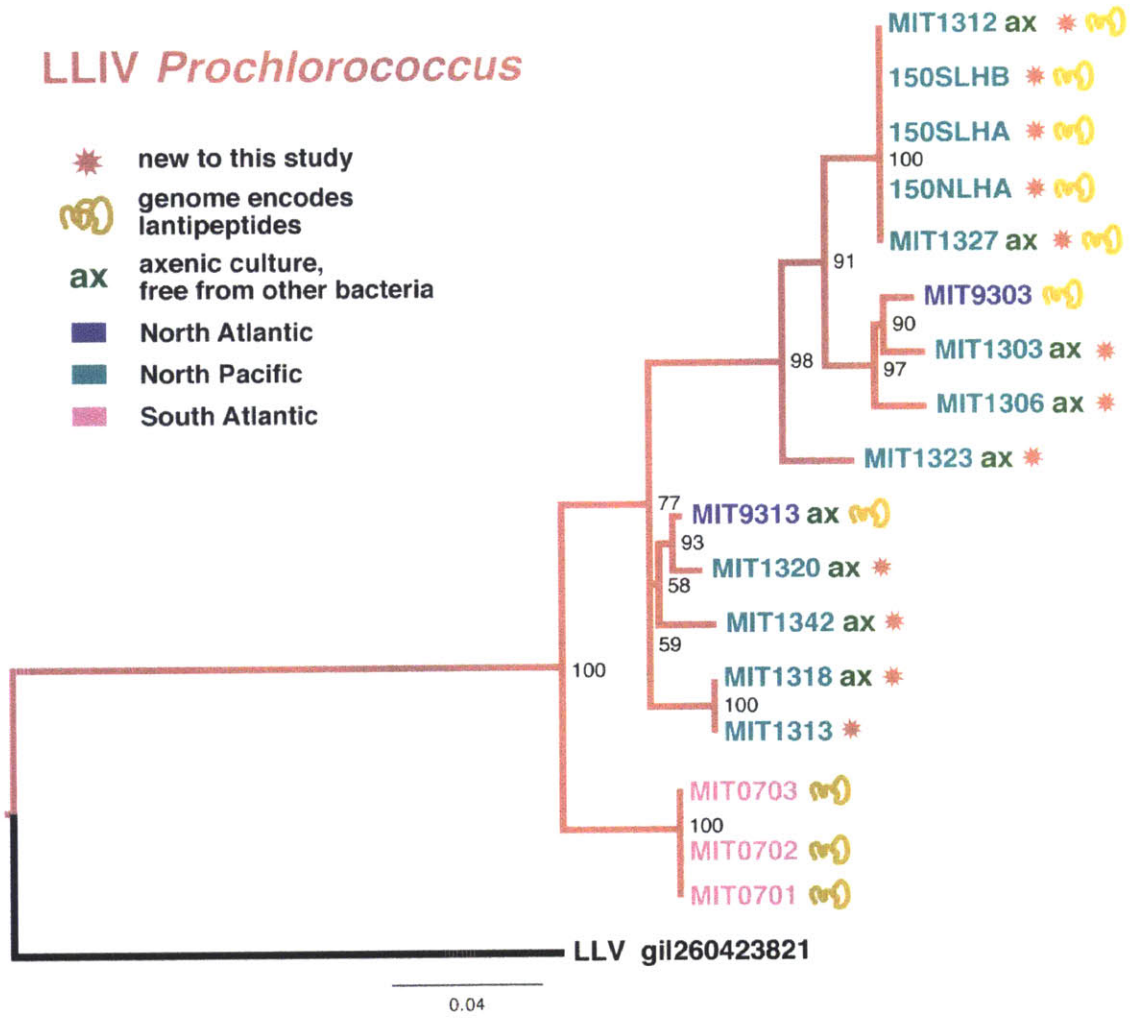


Figure 2.11. Fine-scale phylogeny of LLIV clade cultures and distribution of lantipeptide genes
ITSrRNA approximation of the lineage phylogeny for our LLIV cultures. New genomes isolated and genome sequenced for this study, genomes containing lantipeptides, and axenic strains are marked. Strains are colored by ocean of origin (which correspond to study of origin, Moore et al., 1998, Biller et al., 2014, this study). ITS maximum likelihood phylogeny (phyml, TN93, pinv+ gamma 4), bootstrap values appear at nodes. Rooted with a sequence from the uncultured LLV clade, sister to the LLIV (marked with genbank accession number). We chose this root because other *Prochlorococcus* or *Synechococcus* strains from more distantly related cultured groups gave different rooting patterns (e.g. Figure 2.1).

What distinguishes the LLIV clade of *Prochlorococcus*?

Although we did not capture any uncultivated clades in these isolation efforts, we were very excited by the recovery of LLIV clade *Prochlorococcus*. The LLIV clade has the largest genomes among *Prochlorococcus*, including large numbers of genome-specific genes (unique to each genome). The more divergent genomes we have so far spanning this clade differ from each other by hundreds of genes (Biller et al., 2014). This is substantially more gene content variation than is observed in any other ecotype, despite the small distances between their 16S and ITS markers; by traditional measures, these are very closely related strains, but genome-wide, they show significant variation and evolutionary distance. For this reason,

two recent genomic analysis papers (Kettler et al., 2007, Biller et al., 2015) called for more LLIV genomes. The LLIV clade is the most deeply branching clade of the cultured *Prochlorococcus* lineages, sharing an ancestor with the rest of *Prochlorococcus* only shortly after the divergence of *Prochlorococcus* from marine *Synechococcus* (Figure 2.1). The genomes of the LLIV clade have a GC content of 50%, like *Synechococcus*, unlike the rest of *Prochlorococcus* genomes, which have 30-40% GC (reviewed in Biller et al., 2015). The LLIV genomes share more orthologs with *Synechococcus* than other *Prochlorococcus*, and sometimes phylogenies of individual genes cluster with *Synechococcus*. However, they are also missing many genes shared by all *Synechococcus*, and are distinctly *Prochlorococcus*, with their small cells, pigment content, cellular properties and genome-wide phylogenetic affiliation with other *Prochlorococcus*.

In the field, LLIV *Prochlorococcus* are found exclusively at the base of the euphotic zone below the mixed layer, where nutrient concentrations are higher than surface waters, but little light penetrates. Among the ecotypes for which we have molecular assays, these are found the deepest, in some sites tracking with the depth distribution of the LLII/III and in some sites occurring deeper, their peak abundance offset by several meters (Zinser et al., 2007, Malmstrom et al., 2010). They often peak in abundance between the 1%-0.1% surface irradiance levels, which is deep not only among *Prochlorococcus*, but for all phytoplankton – they live at an extreme edge of photosynthetic life. The integrated abundance of the LLIV clade over the water column is usually 1-2 orders of magnitude lower than the dominant HLII *Prochlorococcus*, but they are nearly always present, globally.

One LLIV strain, MIT9313, has been extensively characterized, usually in comparative studies with a HLI strain Med4. MIT9313 can grow with an order of magnitude less iron than HL Med4, and tolerates an order of magnitude higher copper concentration toxicity (Thompson et al., 2011, Mann et al., 2002). MIT9313 does not handle light shock or growth at high irradiance (Kettler, 2011, Moore et al., 1998). In pairwise co-cultures with a panel of heterotrophic marine bacteria, while Med4 growth was largely unaffected by the presence of co-cultured bacteria, MIT9313 displayed varied responses to co-culture, from enhancement of growth to total inhibition (Sher et al., 2011). We do not know if these properties are clade-wide although genomics gives some clues to the genes behind them. Among the expanded genomic complement of the LLIV clade, they have more transcription factors, more transporters, more nutrient utilization pathways compared to other *Prochlorococcus* (Kettler et al., 2007, Scanlan et al., 2009, Biller et al., 2015). Members of the LLVI clade have the remarkable ability to produce lanthipeptides – natural products of unknown function, consisting of short peptides, ribosomally constructed, which form complex structures when a dedicated modifying enzyme forms lanthionine bridges between cysteine and serine or threonine residues (Li et al., 2010). This ability is shared only with certain strains of marine *Synechococcus;* no *Prochlorococcus* other than LLIV members encode these peptides (Li et al., 2010). The genes encoding lanthipeptides are so diverse that no two are alike in existing genomes. These molecules are exported from the cell, a costly activity in the dilute environment of the oligotrophic open ocean. The possible functional significance and evolution of this trait is a subject of substantial interest and ongoing research – more genomes from the LLIV clade are in high demand. Given the wide array of functions possessed by LLIV *Prochlorococcus*, and their wide genome-to-genome variation, we imagine these new strains will be rich in new functions and they will be able to help us answer some questions about the clade: What functions are characteristic of the LLIV clade, shared by all? What traits vary between LLIV from different oceans? What range of nutrient acquisition strategies do they possess? How do lantipeptides evolve, and what do they do? Our new strains will likely contribute many new genes to our sample of the *Prochlorococcus* pangenome, and these will enrich our understanding of the capabilities of *Prochlorococcus* populations, and help to answer basic questions about the nature of this interesting clade.

Genome sequencing and assembly

To begin characterizing and assess clonality of our new isolates, we sequenced their genomes to draft quality. Genome statistics for all LLVI *Prochlorococcus* are listed in Table 2.2, including new strains from this study. Based on their size and number of genes, the new draft genomes likely capture all or most of each genome, but they lack the full physical mapping that allows comprehensive study of genomic arrangements and the certainty of gene presence and absence that come with fully closed genomes. It would be useful in the future to close these genomes, perhaps through the application of new long-read technologies, like the Pacific Biosciences sequencing that recently assisted in closing two *Prochlorococcus* genomes (Biller et al., 2014).

Table 2.2. LLIV *Prochlorococcus* genome assembly statistics and basic properties

Strain	Number of contigs	Total bp	Longest contig length	Average contig length	N50	Percent GC	Number of proteins	Percent coding	Reference
MIT9313	1	2,410,873	2,410,873	2,410,873	2,410,873	50.7	2,551	83.9	Rocap et al., 2003
MIT9303	1	2,682,675	2,682,675	2,682,675	2,682,675	50	2,732	84	Ketter et al., 2007
MIT0701	53	2,592,571	414,082	48,916	84,463	50.6	2,666	82.3	Biller at al., 2014
MIT0702	61	2,583,057	345,502	42,345	76,101	50.6	2,659	82.2	Biller at al., 2014
MIT0703	61	2,575,057	295,777	42,214	81,186	50.6	2,643	81.9	Biller at al., 2014
MIT1303	47	2,560,150	725,082	54,471	135,805	51	2,610	83	this study
MIT1306	12	2,498,944	772,618	208,245	486,153	51	2,514	84	this study
MIT1313	28	2,590,341	687,899	92,512	296,926	50.0	2,625	82.9	this study
MIT1318	27	2,584,744	816,149	95,731	232,924	50.0	2,627	82.7	this study
MIT1320	26	2,500,454	839,702	96,171	494,475	50	2,604	84	this study
MIT1323	26	2,440,679	502,848	93,872	326,490	51	2,503	83	this study
MIT1342	27	2,548,000	800,664	94,370	391,365	50	2,610	83	this study
MIT1312	53	2,561,499	408,456	48,330	263,603	50.5	2,656	83.2	this study
MIT1327	34	2,591,587	715,496	76,223	328,388	50.3	2,627	83.6	this study
150NLHA	45	2,512,699	264,269	55,837	109,417	50.6	2,607	83.2	this study
150SLHA	66	2,558,254	229,474	38,761	137,856	50.6	2,670	82.4	this study
150SLHB	106	2,472,965	164,075	23,329	70,051	50.8	2,607	82.7	this study

N50 is the size of the contig for which 50% of the genome is contained in contigs of that size or larger. Percent coding was calculated here to include proteins, rRNAs, tRNAs and tmRNAS. Protein count and percent coding are based on prokka annotations performed for this study to support direct comparisons of old and new data; these differ from RAST annotations, but are similar to previously published. Sets of genomes with identical ITS sequences but variation across the genome in SNPs and indels are listed consecutively and tinted gray, separated by strains with unique ITS sequences in white. Differences in assembly statistics reflect variable coverage, which is influence by the sequencing depth itself and by the axenicity of the culture, which can dilute the sequencing across organisms.

Assessing relationships between highly similar cultures

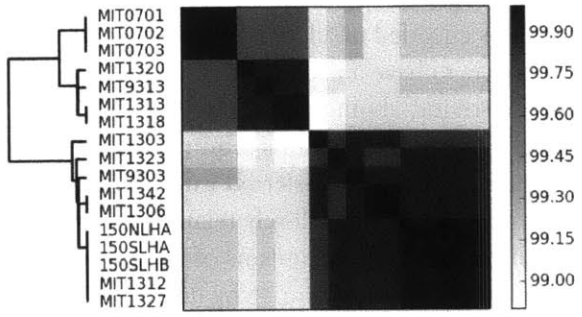
We knew from ITS-rRNA sequencing that some of our strains had sequences identical to each other at this marker, but that alone does not mean a set of strains are clones; wild *Prochlorococcus* populations include variants within identical-ITS sequences and mutations genome-wide (Kashtan et al., 2014). For nearly all of our successful isolations (a few were purified too late for this analysis), we performed

whole genome sequencing, which among many other things, allows us to assess relationships between closely related strains, and determined which sets of identical ITS sequences represented identical clones, and which included genome-wide variation, at a fine scale. In a few cases, our dilution to extinction experiment resulted in isolation of sets clones, which is consistent with the exponential growth of enrichments prior to dilution, but we also obtained considerable diversity, including two groups of strains with identical ITS sequences but variable genomes (Figure 2.12). Within these groups, we observed hundreds to thousands of SNPs and a few indels (data not shown, approach described in Materials and Methods). To more rigorously assess the clonality of these cultures, it would be useful in the future to look at the raw sequencing reads, prior to assembly, to assess complexity of populations in these cultures.

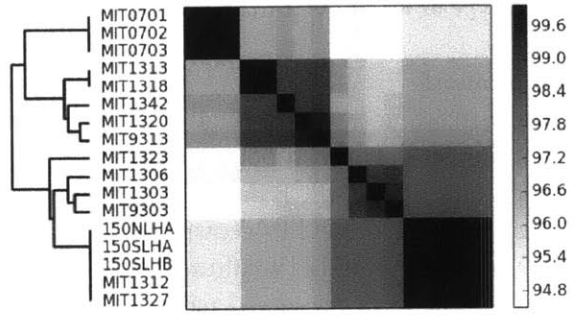
How do these genomes vary? Preliminary genome comparisons

All of our LLIV cultures have 16SrRNA pairwise identities above 98.9% (Figure 2.12A); these are highly similar microbes by traditional 16SrRNA standards (Stackebrandt et al., 1998). Using the ITS marker all samples are above 94% identical to each other, and we can begin to see substructure within this group with several distinct clades with approximately 98% within-clade identity (Figure 2.12B). However, when we calculated the genome-wide average nucleotide identity (ANI; Goris et al., 2007) we found relatively low identity values between strains (95%; Figure 2.12C), and in some cases sets of these strains would be classified as different species by traditional ANI measures (Goris et al., 2007). This clade does not follow the usual relationship between 16SrRNA and ANI established from comparisons of many bacteria, but this remarkable genome-wide variation is consistent with our prior knowledge that LLIV genomes contain great pairwise variation (Kettler et al., 2007, Scanlan et al., 2009, Biller et al., 2015). For the three identical-ITS sets – (MIT0701, MIT0702, MIT0703), (MIT1318, MIT1313) and (MIT1327, 150NLHA, 150SLHA, 150SLHB, MIT1312) – there is actually a range of within group similarity visible in ANI values. This is masked by the scale in Figure 2.12C, but visible in the rescaled version Figure 2.12D. The MIT0701 group is are approximately 100% identical genome-wide, although there are some SNPs and indels (Biller et al., 2014). The MIT1318 and MIT1327 groups range from 99.85% to 100% ANI similarity, spanning a range of fine-scale diversity observed in the wild (Kashtan et al., 2014) but not yet studied in culture.

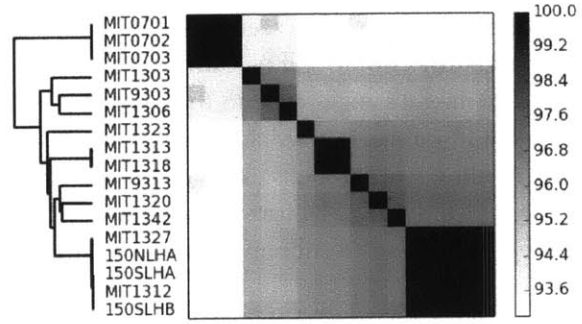
A. 16S-rRNA percent identity



B. ITS percent identity



C. Whole genome average nucleotide identity



D. Whole genome ANI scaled for identical ITS sets

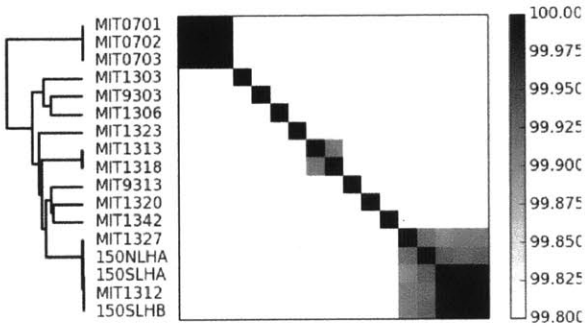


Figure 2.12. Similarity between LLIV isolates:: 16S, ITS, ANI
Each image (A-C) is a symmetric matrix of similarity measures, for all pairwise comparisons of 17 cultures. The order across the top is the same left to right as the labels from top to bottom (self-self identity comparisons along the diagonal). The strains are ordered base on hierarchical clustering (average linkage), shown on the left (not robust phylogenies). Note that the scale bars at right are different for each metric, to represent full range of its variation. For the 16S-rRNA (A), all members of the LLIV clade are at least 98.9% identical to each other. For the ITSrRNA (B), we have greater resolution; all strains are more than 94.5% identical at this marker, but more complex relationships between strains emerge. The genome-wide average nucleotide identity (C), calculated based on the method of Goris et al., 2007 shows that some of our samples are very similar, and other remarkably divergent, given usual relationships between ANI and 16SrRNA (Goris et al., 2007). For the strains that have identical ITS sequences, the ANI values vary between pairs, some nearly clonal, some with thousands of SNPS, viewed on the finer scale version at left, D.

What have these new strains taught us so far?

Although only newly isolated, these cultures have begun to contribute to our picture of *Prochlorococcus* adaptation. These new LLIV strains were rapidly integrated into ongoing projects. For example, we have learned that lantipeptide production, which was encoded in all five previously sequenced LLIV genomes (and no other *Prochlorococcus* genomes), now appears to have a patchy distribution across the clade. Within the newly expanded collection of co-isolated LLIV strains, about half encode the ability to make lantipeptides (Figure 2.11). Mapping this trait onto the phylogeny of the strains requires several gain and loss events. This is not an ecotype-wide trait, but part of the horizontally transferred variable gene content of the clade. Combined with the older cultures, there are some pairs of closely related strains with and without these genes, giving us a more detailed picture of the molecular mechanisms of these loss processes. Although we do not know the functions of these lantipeptides yet, their patchy distribution suggest they are valuable under certain conditions, not universally part of the life of a LLIV cell.

2.3.5 How do our new LLIV cultures compare to the LLIV ecotype as we know it in the oceans?

Would these strains be detected by published qPCR ecotype primers?

Given that our initial interest in culturing LL strains was partially motivated by the the qPCR-flow cytometry mismatch indicating that our current primers do not capture the deepest populations of *Prochlorococcus* well, we wondered, would we have detected these strains in our qPCR data? Or are these part of the unknown? These primers were designed in 2006 specifically to detect the LLIV ecotype as known at the time (Ahlgren et al., 2006). Aligning these primers to the current full set of LLIV cultures' ITS sequences (Figure 2.13) shows that for the forward primer, the site is perfectly conserved across all our LLIV cultures. For the reverse primer, however, there is some variation. Three of our new cultures have one SNP each, and the recently isolated MIT0701, MIT0702, MIT0703 strains have indels and a SNP. Although phylogenetically and genomically falling well within the LLIV clade, the SA strains would not be amplified and counted by these primers, so they certainly qualify in the category of deep cells seen by FCM but not qPCR. A single SNP is tolerable in PCR, so it is likely that these cultures would still be counted, but possibly with differential amplification efficiencies. This kind of variability might explain the improved results reported by increasing the primer concentration in one qPCR publication (Malmstrom et al., 2010) – changing PCR conditions could overcome the slightly lower binding affinity. Cells like our new culture were mostly likely included in published counts of the LLIV ecotype at HOT (Malmstrom et al., 2010). So, these new cultures are not part of the mysterious unknown deep *Prochlorococcus* subpopulations, and there is still a great deal of LL diversity not sampled in culture.

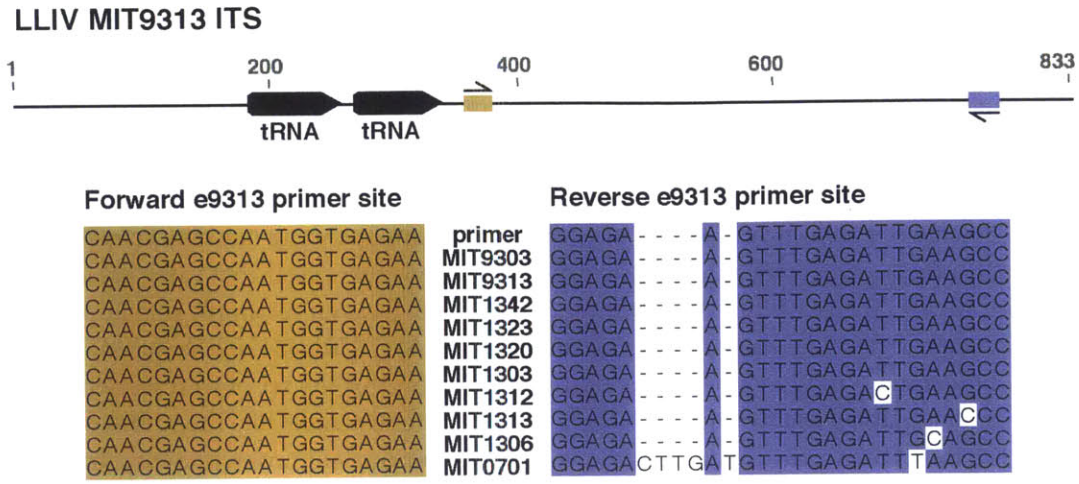


Figure 2.13. Would our ecotype qPCR primers have captured these new cultures? Probably.
For detection and enumeration of *Prochlorococcus* ecotypes in the wild, there exists a collection of primers specific to each ecotype, but broad enough to capture as many members of the ecotype as possible (Ahlgren et al., 2006). The LLIV or e9313 ecotype primers are here aligned with the homologous ITS region for each unique LLIV ITS sequence now in our culture collection. The schematic at the top shows the position of priming sites at between-ecotype variable regions within the ITS within the overall structure of the ITS. The alignment at left (orange) shows that the forward priming site is fully conserved. The reverse priming site (blue) is not conserved across all members of the ecotype: MIT0701 and related sequences would not be amplified by this primer pair, due to an indel. The new cultures include several SNPs in the reverse priming site, but only one per sequence, indicating they are probably similar enough the to the primers to allow successful PCR, given robust conditions, interesting in light of the fact one recent study reported a increase in the performance of these primers through increasing primer concentrations (Malmstrom et al., 2010)

How do the LLIV cultures fit into the LLIV ecotype population structure in the wild?

The range of ITS diversity in the new LLIV cultures falls in a tight cluster with previous strains – spanning the variation between them though not sampling beyond (Figure 2.11). This could be because this phylogenetic structure reflects what the clade really looks like, and we are sampling it well, or it could be that through coincidence or bias, we are only sampling a small part of wild diversity in the clade, despite the fact that these cultures were sampled 20 years and half a planet apart. We were curious how these cultures relate to what we know about LLIV diversity in the oceans – how well are we sampling the wild distribution of LLIV diversity in our cultures? Are these cultures all very closely related, on the scale of the wild LLIV clade, or are we sampling across the full range of diversity? What are we still missing? To address this, there are many ITS clone libraries from the ocean that allow us to say more about the phylogenetic structure of this clade. Through a simple BLAST fishing trip into the NCBI nt (nucleotide) database, we gathered all available uncultured seawater clone library derived ITS sequences clustering with the LLIV ecotype.

We compared our cultures to ITS rRNA sequences from published clone libraries representing hundreds of uncultured *Prochlorococcus* sequences from several oceans, not all the cells in the sea, but the best of our current knowledge about how the LLIV clade is structured. By building a rough phylogeny of these uncultured sequences together with our cultures (Figure 2.14), we learned that we have sampled remarkably broadly across the diversity of this clade. This is exciting from a comparative genomics perspective, given the remarkable variation in gene content in the LLIV clade (Biller et al., 2015, Kettler et al., 2007), this improved sampling across the LLIV phylogeny should help us resolve the time scale of gene gain and loss events within the clade, and reveal many new genes of the *Prochlorococcus* flexible genome.

Such an analysis will help us understand the extent to which variation in the large LLIV flexible genome is influenced by phylogeny and common descent of past events or by more recent environment-specific selection pressure and recent horizontal gene transfer.

On the scale of this wild diversity (Figure 2.14), our collection contains sets of nearly identical strains with close wild relatives. Our collection contains sets of strains with highly similar but distinct ITS sequences and sets with moderately related sequences in the same subclade, a pattern that largely characterizes the population structure of wild sequences, in this dataset and in the HL strains studied in Kashtan et al., 2014. Finally, our collection contains distantly related pairs spanning diversity across deep subclade divisions within the ecotype. Not surprisingly, this analysis also makes clear that there are several distinct subclades with the LLIV observed in the environment which we have yet to sample in culture, so further culturing efforts targeting LLIV strains would be valuable. The MIT0701, MIT0702 and MIT0703 sequences are deeply branching – not just among our cultures, but across this full set of wild sequences. They come from the under-sampled South Atlantic – there may be more sequences like them that we have yet to see. Compared to the ITS-diversity observed in clone libraries, our new cultures represent a broad, but still incomplete sampling of the LLIV clade.

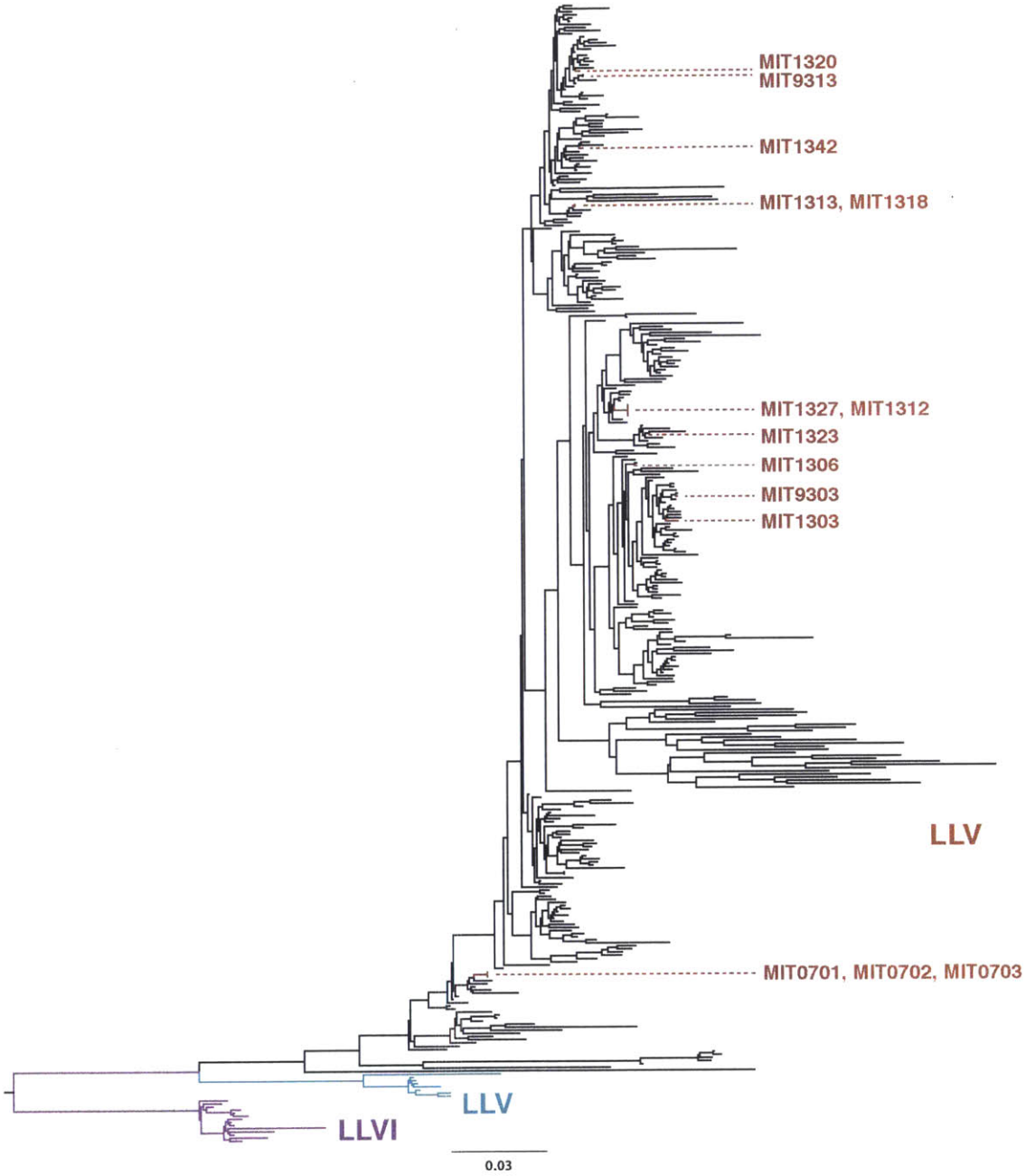


Figure 2.14. LLIV cultures in the phylogenetic context of wild sequences
The LLIV strains we have in culture span the diversity observed for this clade in the wild, as we know it from ITS clone library sequences, including samples from several subclades, but we are still missing other subclades in culture. ITS rRNA approximate likelihood phylogeny based on multiple alignment of sequences in the LLIV clade accessed through NCBI Blast searches, (alignment implemented in muscle, phylogeny implemented in fasttree, GTR + gamma). LLV and LLVI sequences from the uncultured oxygen-minimum zone associated clades, the most closely related ecotypes to the LLIV, were used to root the LLIV phylogeny.

2.4 Conclusions and Future Directions

What will these new isolates teach us?

We isolated thirteen unique strains, all from the same water sample taken from 150 meters from the North Pacific, including twelve new LLIV strains and one new HLII strain (Figure 2.1). How many genes will they add to the expanding the *Prochlorococcus* pangenome? The process of integrating these strains into the previously developed framework for clustering *Prochlorococcus* groups of orthologous genes is underway, so we will soon be able to compare the protein complements of these strains. It will be interesting to see how these strains contribute to the picture of LLIV gene content diversity – whether they have large complements of unique genes like the previously sequenced ones, and what those unique traits might be. Based on our genome alignments and previous work on this clade, we expect extensive inter-strain variability and many new genes. What nitrogen and phosphorus sources will these strains be capable of using? Genomic comparisons across these strains may generate functional hypotheses about new traits and functional differences between strains, that can then be tested with the cultures in the lab. During these analyses, it will be interesting to take into account that these are sympatric strains – we know they all shared a small volume of water, their habitat and a set of environmental conditions, for at least the moment of sampling. This collection of co-isolates may help us to untangle effect of phylogeny and inherited traits shared across the ecotype or within subclades from environmentally selected traits in the flexible genome. What genes are shared by all these strains but absent from the Atlantic LLIV? It will be particularly interesting to compare the co-isolated HL and LL (as in Moore et al, 2003), to see if there are environment-specific traits that span ecotypes. Pairs of closely related strains with and without lantipeptides may help us to begin to untangle the functional consequences of of encoding them, and the evolutionary processes behind their gain and loss. From some of the early enrichment samples of this project, oligotrophic heterotrophs capable of growth on the dissolved organic matter in seawater alone were isolated in tandem, which will enable sympatric co-culture studies. It will also be interesting to explore the differences between the sets of very close relatives (identical ITS groups), both genomically and through functional work in the lab, because this kind of variation is characteristic of how wild *Prochlorococcus* populations are structured.

Future of *Prochlorococcus* culturing: what will we isolate next?

There are an estimated 10^{27} *Prochlorococcus* in the oceans, encompassing vast reserves of genotypic diversity (Flombaum et al., 2013, Biller et al., 2015). This diversity is the product of 150 – 500 million years of evolution since divergence of the *Prochlorococcus* lineage from the rest of cyanobacteria (Dufresne et al., 2003, Blank et al, 2010). We have only just begun to sample from this diversity into our culture collection; the *Prochlorococcus* of the oceans have much more to teach us. One challenge moving forward will be to isolate cultures from the several major uncultured clades that we know about from molecular field data, but have yet to bring into the lab. We would like to study the unique traits they have been hypothesized to carry, obtain full genome sequences for them, and look for ecologically significant functional differences among ecotypes, as we have done for the five cultured ecotypes. For some of these uncultured clades, like the HLIII, HLIV and HLV clades associated with iron-limited high-nutrient, low-chlorophyll ocean regions or the LLV and LLVI clades associated with oxygen minimum zones, obtaining cultures will primarily be a matter of sampling from the right waters – those with distinct chemical patterns in known geographic regions. The difficulty is a function of limited geographic access to the waters they inhabit, but our expanding molecular description of the oceans will help to target these efforts. For others, like the NC1/LLVII clade, there is no geographic pattern reported, but they are found in many places deep in the euphotic zone (Martiny et al., 2009, Jiao et al., 2014), including at the more easily accessible time series stations in the North Pacific gyre (HOT) and the North Atlantic gyre (BATS) – depth, not geography should guide sampling to target this

clade. Efforts like the one described here, targeted at the base of the euphotic zone, may succeed in capturing this clade. While we had a remarkable amount of prior information in this study to guide sampling before even leaving port, with the ecotype qPCR time series (Malmstrom et al., 2010), without this information, it is still be possible to perform targeted isolations of HL and LL *Prochlorococcus* with the aid of a shipboard flow cytometer, since the flow cytometry signatures of these groups are distinctive, usually with a transition from HL to LL-dominated populations somewhere below the chlorophyll maximum (as in Moore et al., 1998).

The longstanding mismatch between physical flow cytometry counts of *Prochlorococcus* and qPCR primer-dependent counts of ecotypes (dependent on prior knowledge of molecular makers) at the base of the euphotic zone – but not the surface – tells us we know there are more LL *Prochlorococcus* out there that we know little about. These could be explained by the NCI/LLVII clade, for which qPCR primers have not yet been designed, or by LLIV cultures which are not perfect matches to the existing LLIV primers, or by similar diversity within the other LL clades. Still another possibility, there could still be more uncultured deeply branching clades we have yet to sample. In any case, the base of the euphotic zone will be a good place to look for deep pools of unsampled diversity.

Another approach to isolation is to frame the search not in terms of phylogeny and lineage, but in terms of traits, especially the nutrient acquisition strategies that map onto marine environment, which are among the critical niche-specifying traits of cells. Typically in *Prochlorococcus* genomes, diverse nutrient acquisition strategies are part of the flexible genome, more a function of environment than phylogeny (Martiny et al., 2006, Martiny et al., 2009abc, Coleman and Chisholm, 2010). Through targeted isolations we can attempt to select for isolates with specific traits – how we isolate controls what we get. We do not know what in our isolation efforts so far has selected for the limited subset of the wild diversity in culture, so the open-minded application of diverse light and temperature conditions and diverse forms of nutrients will be important for continuing to expand the culture collection. Expanding basic biological understanding of *Prochlorococcus*, like the discovery that heterotrophs detoxify hydrogen peroxide for *Prochlorococcus* (Morris et al., 2011), will improve our techniques moving forward, unlocking new abilities in isolation and purification. There is room to improve in the realm of solid-state culturing, in understanding the factors that control culture density limits and the onset of stationary phase, and in the range of nutrient forms that support *Prochlorococcus* growth.

Finally, in future work, based on our experiences mining complex enrichments for new strains over the course of many months, and finding different strains at different times, it would be fascinating to study the enrichments themselves over the course of their life in the lab, from sampling, growth, serial passaging and subculturing. Following this process more closely, through deep sequencing and flow cytometry, would help address the fundamental unknown of culture bias, how our laboratory choices in influence trajectories toward isolation of different types of *Prochlorococcus*. Studying enrichments could also provide insights into *Prochlorococcus* competition and interactions with each other and co-isolated heterotrophs, in limited microcosms, an intermediate between studying real communities in the field, and the simplified system of two-strain co-cultures.

Due to its global significance and tractability, *Prochlorococcus* has become a major model for the study of microbial ecology and evolution, contributing to our understanding of basic principles of how life adapts to the many environmental conditions of the oceans (Biller et al., 2015). The power of the *Prochlorococcus* system for addressing fundamental questions in biology arises from the combination of its extraordinary abundance, diversity and contributions to primary production in the oceans and our ability to study it both in the lab and the field. In field we can measure wild diversity, count populations and follow their distributions over geography and depth, and in the lab, we can study traits in individual strains, picking apart differences and relating physiology to genomes – culturing is a critical part of our ability to

study this system. Barbara McKlintock advocated for the scientific value of her intimacy with her research organism, maize, arguing that some of her insights were only possible through having a 'feeling for the organism,' an idea which has come to represent a vital facet of the modern biologist's practice. For *Prochlorococcus*, perhaps the best way to get this feeling would be onboard a ship surrounded by the sparkling tropical sun and deep, deep blue of an oligotrophic ocean, with a flow cytometer humming through samples spanning depth and distance. That might not be possible as often as we would like, so the next best thing for getting a feeling for *Prochlorococcus* is through culturing, which brings these beautiful marine organisms into the sphere of human experience, growing on the scale of our days and weeks, changing clear seawater to a rich bright green visible to the photon gathering powers of our human eyes. We do not have complete control over *Prochlorococcus* in culture, and the isolation of strains is as still substantial and risky undertaking, but we are moving towards a more intentional and routine *Prochlorococcus* isolation and culturing techniques.

Acknowledgements

This work was supported by a grant to Sallie W. Chisholm from the Center for Microbial Oceanography Research and Education (CMORE). Many thanks to the crew and CMORE scientific organizing team for the HOE-PhoR cruise. Thanks to Zachory Berta-Thompson for assistance with plot rendering and copy editing in this manuscript. Particular thanks to coauthors Kristin Legault and Jamie Becker for taking responsibility for the maintenance of a copy of these cultures during the writing of this thesis.

References

Ahlgren, N.A., and Rocap, G. (2006). Culture isolation and culture-independent clone libraries reveal new marine *Synechococcus* ecotypes with distinctive light and N physiologies. Appl Environ Microbiol *72*, 7193-7204.

Ahlgren, N.A., Rocap, G., and Chisholm, S.W. (2006). Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. Environ Microbiol *8*, 441-454.

Astorga-Eló, M., Ramírez-Flandes, S., DeLong, E.F., and Ulloa, O. (2015). Genomic potential for nitrogen assimilation in uncultivated members of *Prochlorococcus* from an anoxic marine zone. ISME J

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates *Prochlorococcus*-virus coexistence. Nature *474*, 604-08.

Baumdicker, F., Hess, W.R., and Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol *4*, 443-456.

Berube, P.M., Biller, S.J., Kent, A.G., Berta-Thompson, J.W., Roggensack, S.E., Roache-Johnson, K.H., Ackerman, M., Moore, L.R., Meisel, J.D., et al. (2014). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. ISME J

Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., Awad, L., Roache-Johnson, K.H., Ding, H., Giovannoni, S.J., et al. (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *1*, 140034.

Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015). *Prochlorococcus*: the structure and function of collective diversity. Nat Rev Microbiol *13*, 13-27.

Blank, C.E., and Sánchez-Baracaldo, P. (2010). Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. Geobiology *8*, 1-23.

Bouman, H.A., Ulloa, O., Scanlan, D.J., Zwirglmaier, K., Li, W.K., Platt, T., Stuart, V., Barlow, R., Leth, O., et al. (2006). Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. Science *312*, 918-921.

Calvo-Díaz, A., Díaz-Pérez, L., Suárez, L.Á., Morán, X.A., Teira, E., and Marañón, E. (2011). Decrease in the autotrophic-to-heterotrophic biomass ratio of picoplankton in oligotrophic marine waters due to bottle enclosure. Appl Environ Microbiol *77*, 5739-746.

Carini, P., Steindler, L., Beszteri, S., and Giovannoni, S.J. (2013). Nutrient requirements for growth of the extreme oligotroph '*Candidatus* Pelagibacter ubique' HTCC1062 on a defined medium. ISME J *7*, 592-602.

Chisholm, S., Frankel, S., Goericke, R., Olson, R., Palenik, B., Waterbury, J., West-Johnsrud, L., and Zettler, E. (1992). *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll a and b. Archives of Microbiology *157*, 297-300-.

Chisholm, S.W., Olson, R.J., Zettler, E.R., Waterbury, J.B., Goericke, R., and Welschmeyer, N. (1988). A novel free-living prochlorophyte occurs at high cell concentrations in the oceanic euphotic zone. Nature *334*, 340-43.

Coleman, M.L., and Chisholm, S.W. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. Trends Microbiol *15*, 398-407.

Coleman, M.L., and Chisholm, S.W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. Science *311*, 1768-770.

del Valle, D.V., and Karl, D.M. (2014). Aerobic production of methane from dissolved water-column methylphosphonate and sinking particles in the North Pacific Subtropical Gyre. AQUATIC MICROBIAL ECOLOGY *73*, 93-105.

Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I.M., Barbe, V., Duprat, S., Galperin, M.Y., Koonin, E.V., et al. (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A *100*, 10020-25.

Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.J., Richter, R.A., Valas, R., Novotny, M., Yee-Greenbaum, J., Selengut, J.D., et al. (2012). Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. ISME J *6*, 1186-199.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792 -11797.

Fernández, E., Marañón, E., Morán, X.A.G., and Serret, P. (2003). Potential causes for the unequal contribution of picophytoplankton to total biomass and productivity in oligotrophic waters. Marine Ecology Progress Series *254*, 101-09.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proceedings of the National Academy of Sciences *110*, 9824-29.

Gieskes, W.W.C., Kraay, G.W., and Baars, M.A. (1979). Current 14C methods for measuring primary production: Gross underestimates in oceanic waters. Netherlands Journal of Sea Research *13*, 58-78.

Giovannoni, S., and Stingl, U. (2007). The importance of culturing bacterioplankton in the 'omics' age. Nat Rev Microbiol *5*, 820-26.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol *57*, 81-91.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol *59*, 307 -3321.

Hess, W.R., Partensky, F., van der Staay, G.W., Garcia-Fernandez, J.M., Börner, T., and Vaulot, D. (1996). Coexistence of phycoerythrin and a chlorophyll a/b antenna in a marine prokaryote. Proc Natl Acad Sci U S A *93*, 11126-130.

Hess, W.R., Steglich, C., Lichtle, C., and Partensky, F. (1999). Phycoerythrins of the oxyphotobacterium *Prochlorococcus* marinus are associated to the thylakoid membrane and are encoded by a single large gene cluster. Plant Mol Biol *40*, 507-521.

Hewson, I., Paerl, R.W., Tripp, H.J., Zehr, J.P., and Karl, D.M. (2009). Metagenomic potential of microbial assemblages in the surface waters of the central Pacific Ocean tracks variability in oceanic habitat. Limnology and Oceanography *54*, 1981-994.

Huang, S., Wilhelm, S.W., Harvey, H.R., Taylor, K., Jiao, N., and Chen, F. (2011). Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. ISME J

Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science and Engineering *9*, 90-95.

Jiao, N., Luo, T., Zhang, R., Yan, W., Lin, Y., Johnson, Z.I., Tian, J., Yuan, D., Yang, Q., et al. (2014). Presence of *Prochlorococcus* in the aphotic waters of the western Pacific Ocean. BG *11*, 2391-2400.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M., and Chisholm, S.W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science *311*, 1737-740.

Karl, D.M., and Church, M.J. (2014). Microbial oceanography and the Hawaii Ocean Time-series programme. Nat Rev Microbiol *12*, 699-713.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science *344*, 416-420.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol *30*, 772-780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647-49.

Keen, O., Dotson, A., and Linden, K. (2012). Evaluation of Hydrogen Peroxide Chemical Quenching Agents following an Advanced Oxidation Process. J Environ Eng *139*, 137-140.

Kettler, G.C. (2011). Genetic diversity and its consequences for light adaptation in *Prochlorococcus*. MIT PhD Thesis

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., et al. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet 3, e231.

Lavin, P., González, B., Santibáñez, J.F., Scanlan, D.J., and Ulloa, O. (2010). Novel lineages of *Prochlorococcus* thrive within the oxygen minimum zone of the eastern tropical South Pacific. Environmental Microbiology Reports *2*, 728-738.

Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P.J., Joewono, I., Rusch, D., Chisholm, S.W., and van der Donk, W.A. (2010). Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. Proc Natl Acad Sci U S A *107*, 10430-35.

Lindell, D. (2014). The Genus *Prochlorococcus*, Phylum Cyanobacteria. In The Prokaryotes, E. Rosenberg, E. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Springer Berlin Heidelberg).

Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. ISME J *4*, 1252-264.

Malmstrom, R.R., Rodrigue, S., Huang, K.H., Kelly, L., Kern, S.E., Thompson, A., Roggensack, S., Berube, P.M., Henn, M.R., and Chisholm, S.W. (2013). Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. ISME J *7*, 184-198.

Mann, E.L., Ahlgren, N., Moffett, J.W., and Chisholm, S.W. (2002). Copper toxicity and cyanobacteria ecology in the Sargasso Sea. Limnol Oceanogr *47*, 976-988.

Mann, N.H. (2003). Phages of the marine cyanobacterial picophytoplankton. FEMS Microbiol Rev *27*, 17-34.

Mann, N.H. (2005). The third age of phage. PLoS Biol 3, e182.

Martinez, A., Osburne, M.S., Sharma, A.K., Delong, E.F., and Chisholm, S.W. (2011). Phosphite utilization by the marine picocyanobacterium *Prochlorococcus* MIT9301. Environ Microbiol

Martinez, A., Tyson, G.W., and Delong, E.F. (2010). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. Environ Microbiol *12*, 222-238.

Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. Proc Natl Acad Sci U S A *103*, 12552-57.

Martiny, A.C., Huang, Y., and Li, W. (2009). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. Environ Microbiol *11*, 1340-47.

Martiny, A.C., Kathuria, S., and Berube, P.M. (2009). Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. Proc Natl Acad Sci U S A *106*, 10787-792.

Martiny, A.C., Tai, A.P., Veneziano, D., Primeau, F., and Chisholm, S.W. (2009). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. Environ Microbiol *11*, 823-832.

Moore, L.R., and Chisholm, S.W. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus* : ecotypic differences among cultured isolates. Limnol Oceanogr *44*, 628-638.

Moore, L.R., Coe, A., Zinser, E.R., Saito, M.A., Sullivan, M.B., Lindell, D., Frois-Moniz, K., Waterbury, J., and Chisholm, S.W. (2007). Culturing the marine cyanobacterium *Prochlorococcus*. Limnology and Oceanography: Methods *5*, 353-362.

Moore, L.R., Ostrowski, M., Scanlan, D.J., Feren, K., and Sweetsir, T. (2005). Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. AQUATIC MICROBIAL ECOLOGY *39*, 257-269.

Moore, L.R., Post, A.F., Rocap, G., and Chisholm, S.W. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. Limnology and Oceanography *47*, 989-996 .

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature *393*, 464-67.

Morris, J.J., Johnson, Z.I., Szul, M.J., Keller, M., and Zinser, E.R. (2011). Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. PLoS ONE *6*, e16805.

Morris, J.J., Kirkegaard, R., Szul, M.J., Johnson, Z.I., and Zinser, E.R. (2008). Facilitation of robust growth of *Prochlorococcus* colonies and dilute liquid cultures by "helper" heterotrophic bacteria. Appl Environ Microbiol *74*, 4530-34.

Mühling, M. (2012). On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. Environ Microbiol *14*, 567-579.

Ottesen, E.A., Young, C.R., Gifford, S.M., Eppley, J.M., Marin, R., Schuster, S.C., Scholin, C.A., and DeLong, E.F. (2014). Ocean microbes. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. Science *345*, 207-212.

Partensky, F., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. Ann Rev Mar Sci *2*, 305-331.

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev *63*, 106-127.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Rappé, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu Rev Microbiol *57*, 369-394.

Rippka, R. (1988). Isolation and purification of cyanobacteria. Methods Enzymol *167*, 3-27.

Rippka, R., Coursin, T., Hess, W., Lichtle, C., Scanlan, D.J., Palinska, K.A., Iteman, I., Partensky, F., Houmard, J., and Herdman, M. (2000). *Prochlorococcus marinus* Chisholm et al. 1992 subsp. pastoris subsp. nov. strain PCC 9511, the first axenic chlorophyll a2/b2-containing cyanobacterium (Oxyphotobacteria). Int J Syst Evol Microbiol *50 Pt 5*, 1833-847.

Roache-Johnson, K.H. (2013). Characterization of Phycoerythrin Physiology in Low-Light Adapted *Prochlorococcus* Ecotypes. University of Maine, MS Thesis

Rocap, G., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol *68*, 1180-191.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature *424*, 1042-47.

Rodrigue, S., Malmstrom, R.R., Berlin, A.M., Birren, B.W., Henn, M.R., and Chisholm, S.W. (2009). Whole genome amplification and de novo assembly of single bacterial cells. PLoS One *4*, e6864.

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J., and Chisholm, S.W. (2010). Unlocking Short Read Sequencing for Metagenomics. PLoS ONE *5*, e11840.

Rusch, D.B., Martiny, A.C., Dupont, C.L., Halpern, A.L., and Venter, J.C. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. Proceedings of the National Academy of Sciences *107*, 16184 -116189.

Saito, M.A. (2001). The biogeochemistry in the Sargasso Sea. MIT/WHOI Joint Program PhD Thesis

Saito, M.A., Moffett, J.W., Chisholm, S.W., and Waterbury, J.B. (2002). Cobalt limitation and uptake in *Prochlorococcus*. Limnol Oceanogr *47*, 1629-636.

Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., Post, A.F., Hagemann, M., Paulsen, I., and Partensky, F. (2009). Ecological Genomics of Marine Picocyanobacteria. Microbiology and Molecular Biology Reviews : MMBR *73*, 249-299.

Sher, D., Thompson, J.W., Kashtan, N., Croal, L., and Chisholm, S.W. (2011). Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. ISME J *5*, 1125-132.

Shibl, A.A., Thompson, L.R., Ngugi, D.K., and Stingl, U. (2014). Distribution and diversity of *Prochlorococcus* ecotypes in the Red Sea. FEMS Microbiology Letters *356*, 118-126.

Stackebrandt, E., and Goebel, B.M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16s rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic Bacteriology *44*, 846-49.

Steglich, C., Frankenberg-Dinkel, N., Penno, S., and Hess, W.R. (2005). A green light-absorbing phycoerythrin is present in the high-light-adapted marine cyanobacterium *Prochlorococcus* sp. MED4. Environ Microbiol *7*, 1611-18.

Steglich, C., Mullineaux, C.W., Teuchner, K., Hess, W.R., and Lokstein, H. (2003). Photophysical properties of *Prochlorococcus marinus* SS120 divinyl chlorophylls and phycoerythrin in vitro and in vivo. FEBS Letters *553*, 79 - 84.

Stewart, E.J. (2012). Growing Unculturable Bacteria. J Bacteriol *194*, 4151-160.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003). Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature *424*, 1047-051.

Thompson, A.W., Huang, K., Saito, M.A., and Chisholm, S.W. (2011). Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. ISME J *5*, 1580-594.

Tichy, M., and Vermaas, W. (1999). In vivo role of catalase-peroxidase in *Synechocystis* sp. strain PCC 6803. J Bacteriol *181*, 1875-882.

Vermaas, W.F., Williams, J.G., and Arntzen, C.J. (1987). Sequencing and modification of psbB, the gene encoding the CP-47 protein of Photosystem II, in the cyanobacterium *Synechocystis* 6803. Plant Mol Biol *8*, 317-326.

Wang, Z., Xu, Y., Yang, Z., Hou, H., Jiang, G., and Kuang, T. (2002). Effect of Sodium Thiosulfate on the Depletion of Photosynthetic Apparatus in Cyanobacterium *Synechocystis* sp. PCC 6803 Cells Grown in the Presence of Glucose. Photosynthetica *40*, 383-387-.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-191.

West, N.J., Lebaron, P., Strutton, P.G., and Suzuki, M.T. (2011). A novel clade of *Prochlorococcus* found in high nutrient low chlorophyll waters in the South and Equatorial Pacific Ocean. ISME J *5*, 933-944.

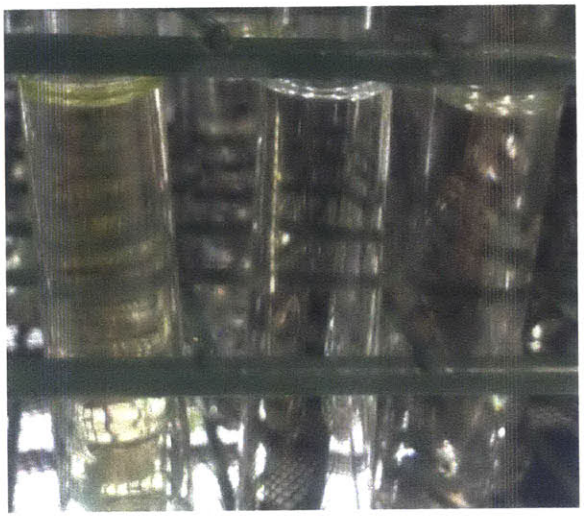
Wiethaus, J., Busch, A.W., Dammeyer, T., and Frankenberg-Dinkel, N. (2010). Phycobiliproteins in *Prochlorococcus marinus*: biosynthesis of pigments and their assembly into proteins. Eur J Cell Biol *89*, 1005-010.

Zinser, E., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D., and Chisholm, S.W. (2007). Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. Limnol Oceanogr *52*, 2205-220.

Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J., and Chisholm, S.W. (2006). *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol *72*, 723-732.

Zwirglmaier, K., Heywood, J.L., Chamberlain, K., Woodward, E.M., Zubkov, M.V., and Scanlan, D.J. (2007). Basin-scale distribution patterns of picocyanobacterial lineages in the Atlantic Ocean. Environ Microbiol *9*, 1278-290.

Supplemental Figures and Tables



Supplemental Figure S2.1. First visible high density cultures
Prochlorococcus cultures are largely identical in appearance, with some variation in a yellow-to-green color range, turbidity shifts in dense or old cultures and color intensity, varying with light conditions and density. Visual inspection is efficient for monitoring growth of large numbers of slow growing cultures that reach high density. Bulk chlorophyll fluorescence detection is slightly more sensitive than the eye (by about half an order of magnitude), and flow cytometry is much more sensitive (about four orders of magnitude). A culture that reaches this density is much easier to work with in practice, because routine transfers can be performed at the right time without instrument-based monitoring of growth.

Supplementary Table S2.1. HOE-PhoR cruise isolation manifest: a *Prochlorococcus* enrichment kit

Materials	Quantity	Purpose
Acid washed, autoclaved squat Nalgene polycarbonate bottles, 25ml	14	Hold enrichments
Acid washed, autoclaved round bottom Nalgene polycarbonate tubes, 30ml	18	Hold enrichments
Acid washed, autoclaved round bottom Nalgene polycarbonate tubes, 10ml	21	Hold enrichments
Acid washed reusable plastic gravity filtration units, 47mm diameter, Millipore	2	Size fractionation
Acid washed, autoclaved teflon bottles, 125ml	3	Sample water
6 square feet of window screening	1	Control light
1 square foot pieces of blue plastic gels	3	Control light
Filters: GF/C 1.2um, 47mm diameter, glass microfiber, Whatman	1	Size fractionation
Filters: GF/D 2.7um, 47mm diameter, glass microfiber, Whatman	1	Size fractionation
Filters: GF/F 0.7um, 47mm diameter, glass microfiber, Whatman	1	Size fractionation
Filters: Nucleopore track etch polycarbonate membrane 1.0um, 47mm, Whatman	1	Size fractionation
Filters: Nucleopore track etch polycarbonate membrane 0.8um, 47mm, Whatman	1	Size fractionation
Filters: Polycarbonate membrane filters, 0.6um, 47mm, Poretics	1	Size fractionation
4by4 30ml test tube rack	1	Hold enrichments
4by10 10ml test tube rack	1	Hold enrichments
LED book light, Mighty Bright	1	Ship enrichments
Teal lab tape, VWR	1	Organize
1 quart ziploc bag	5	Ship enrichments
1 gallon ziploc bag	5	Ship enrichments
Isolation lab notebook with references	1	Plans and record
Sodium pyruvate 100mM, 50 ml, 0.2 μm filter sterilized	1	Peroxide quencher
Sodium pyruvate 100uM, 40 ml, 0.2 μm filter sterilized	1	Peroxide quencher

Materials	Quantity	Purpose
Sodium thiosulfate 100mM, 50ml, 0.2 μm filter sterilized	1	Peroxide quencher
Sodium thiosulfate 100uM, 40 ml, 0.2 μm filter sterilized	1	Peroxide quencher
Urea, 10mM, 60ml, 0.2 μm filter sterilized	1	N source
Ammonium chloride, 5mM, 60ml, 0.2 μm filter sterilized	1	N source
Sodium phosphate monobasic, 1mM, 60ml, 0.2 μm filter sterilized	1	P source
Sodium nitrite, 15mM, 100ml, 0.2 μm filter sterilized	1	N source
Urea, 100mM, 100ml, 0.2 μm filter sterilized	1	N source
Ammonium chloride, 50mM, 100ml, 0.2 μm filter sterilized	1	N source
Sodium phosphate monobasic, 10mM, 60ml, 0.2 μm filter sterilized	1	P source
100X Pro99/Pro2 Trace Metal Mix, 0.2 μm filter sterilized	1	Metal source

Supplemental Figure S2.2 Initial enrichment conditions for all seawater samples

Date	Depth, Name	Volume	Filter	Pro2	1/10 Pro2	15uM NO_2^- 1uM P	Pyruvate 1mM	Pyruvate 1uM	Thiosulfate 1uM	Thiosulfate 1mM	Amino acid mix total	Nalidixic acid	Trace Metals
5/25	100m A	15 ml	1.0 μm	V			V						V
5/25	100m B	15 ml	1.0 μm		V			V					V
5/25	100m C	15 ml	1.0 μm			V		V					V
5/25	100m D	20 ml	1.0 μm	V					V				V
5/25	100m E	20 ml	1.0 μm		V					V			V
5/25	100m F	20 ml	1.0 μm			V				V			V
5/25	100m G	7 ml	1.0 μm			V		V					V
5/25	100m H	7 ml	1.0 μm			V			V				V
5/25	100m I	7 ml	1.0 μm			V	V						V
5/25	100m J	7 ml	1.0 μm			V				V			V
5/28	125m A	15 ml	0.8 μm			V	V						V
5/28	125m B	15 ml	0.8 μm		V		V						V
5/28	125m C	15 ml	0.8 μm	V			V						V
5/28	125m D	15 ml	0.8 μm			V				V			V
5/28	125m E	20 ml	0.8 μm			V	V						V
5/28	125m F	20 ml	0.8 μm		V		V						V
5/28	125m G	20 ml	0.8 μm	V			V						V
5/28	125m H	20 ml	0.8 μm			V				V			V
5/28	125m I	7 ml	0.8 μm			V	V						V
5/28	125m J	7 ml	0.8 μm		V		V						V
5/28	125m K	7 ml	0.8 μm	V			V						V
5/28	125m L	7 ml	0.8 μm			V				V			V
5/29	125m Aaa	15 ml	0.8 μm			V					pre-inc		V

Date	Depth, Name	Volume	Filter	Pro2	1/10 Pro2	15uM NO_2^- 1uM P	Pyruvate 1mM	Pyruvate 1uM	Thiosulfate 1uM	Thiosulfate 1mM	Amino acid mix total	Nalidixic acid	Trace Metals
5/29	125m Baa	15 ml	0.8 µm			V					pre-inc	V	V
5/29	125m Caa	7 ml	0.8 µm		V		V				pre-inc		V
5/29	125m Daa	7 ml	0.8 µm		V		V				pre-inc	V	V
5/29	125m Eaa	7 ml	0.8 µm	V					V		pre-inc		V
5/29	125m Faa	7 ml	0.8 µm	V							pre-inc		V
5/29	125m Gaa	7 ml	0.8 µm	V							pre-inc	V	V
6/2	150m A	15 ml	0.8 µm	V									V
6/2	150m B	15 ml	0.8 µm		V								V
6/2	150m C	15 ml	0.8 µm			V							V
6/2	150m D	15 ml	0.8 µm	V							100nM	V	V
6/2	150m E	15 ml	0.8 µm		V						100nM		V
6/2	150m F	20 ml	1.0 µm	V			V						V
6/2	150m G	20 ml	1.0 µm		V		V						V
6/2	150m H	20 ml	1.0 µm			V	V						V
6/2	150m I	20 ml	1.0 µm			V		V					V
6/2	150m J	20 ml	1.0 µm			V			V				V
6/2	150m K1	20 ml	1.0 µm			V					100nM		V
6/2	150m K2	20 ml	1.0 µm			V				V			V
6/2	150m L	20 ml	1.0 µm		V			V			100nM		V
6/2	150m M	20 ml	1.0 µm		V				V		100nM	V	V
6/2	150m N	7 ml	1.0 µm			V							V
6/2	150m O	7 ml	1.0 µm			V		V			100nM		V
6/2	150m P	7 ml	1.0 µm	V			V						V
6/2	150m Q	7 ml	1.0 µm		V		V						V
6/2	150m R	7 ml	1.0 µm			V	V						V
6/2	150m S	7 ml	1.0 µm	V					V				V
6/2	150m T	20 ml	1.0 µm			V				V	100nM		V
6/2	150 m U	7 ml	none			V	V					V	V

Filter: pore size of polycarbonate filter used to remove larger phytoplankton, bacteria and detritus from seawater used for *Prochlorococcus* enrichments. Polycarbonate filters have defined pore sizes useful for size fractionation; all filtration was performed with GF/C glass fibre filter backing filters (nominal cutoff size of fiber matrix - 1.2μm).
All samples TM: Amino acid mix? Pro2: phosphate, ammonia, urea 1/10 Pro2
15ml samples in 30ml widemouth nalgene polycarbonate bottles, 20ml samples in 30 ml tube oakridge polycarbonate, 7ml samples in 10 ml tube oakridge polycarbonate.
pre-inc amino acids: different kind of experiment, inspired by recent work on amino acid use by *Prochlorococcus*: unfiltered seawater was incubated 24 hours with 500 uM amino acid mix, exposed to 1 μmol photons $m^{-2}s^{-1}$ irradiance, room temperature, then filtered through 0.8μm polycarbonate filter (with GF/C glass fibre backing filter); these samples had many heterotrophs, but also some healthy Pro made it back to shore. None grew to high density *Prochlorococcus* enrichments.

Supplementary Table S2.3. LLIV Strains isolated and purified from HOE-PhoR cruise samples

Strains[1]	Original sample[2]	Axenic	Genome sequence	Clone group[3]	Official name[4]
8E2	150m S	Yes	Yes	clone 1320	
8C5	150m S	Yes	Yes	clone 1323	
3E4	150m S	Yes	Yes	unique	MIT1312ax
3F8	150m S	Yes	Yes	type clone 1323	MIT1323ax
4E3	150m S	Yes	Yes	unique	MIT1318ax
2E3	150m S	Yes	Yes	type clone 1320	MIT1320ax
4C5	150m S	Yes	Yes	unique	MIT1327ax
7E6	150m S	Yes	Yes	unique	MIT1306ax
10D5	150m S	Yes	Yes	clone 1323	
14G2	150m N	Yes	Yes	type clone 1342	MIT1342ax
4C2	150m N	Yes	Yes	clone 1342	
4C3	150m N	Yes	Yes	clone 1342	
9C5	150m S	Yes	Yes	clone 1323	
8F2	150m N	Yes	Yes	clone 1342	
2D8	150m N	Yes	Yes	clone 1342	
2B9	150m N	Yes	Yes	clone 1342	
6E5	150m N	Yes	Yes	clone 1342	
13E5	150m N	No	Yes	unique	MIT1313
8B5	150m S	Yes	No	unique	MIT1303ax
8C2	150m N	Yes	Yes	clone 1342	
20E4	150m N	Yes	No	ITS identical to clone group 1342	unknown
1B3	150m N	Yes	No	ITS identical to clone group 1342	unknown

[1] Strain designations based on name of plate and well, each representing one green well from dilution experiment successfully propagated as a batch culture, listed in order of appearance of green wells (over several months).
[2] Refers to original enrichment - two dilution experiments were performed. 150m S was prior to dilution grown in and diluted in Pro2 media based on Hawaii seawater + 1μM thiosulfate; dilution experiment conducted in a continuous light incubator. 150mN was grown in and diluted in Pro99 based on Woods Hole Vineyard Sound seawater; diel cycling incubator used for dilution experiment.
[3] Based on genome sequencing, strains assigned to clone groups, if highly similar.
[4] Each unique lineage (one type strain for each clone group) received an official MIT strain name; ax for axenic.

Chapter III. The high-light inducible genes of the marine cyanobacterium *Prochlorococcus*: a diverse and dynamic gene family

Jessie W. Berta-Thompson[1,2], Greg C. Kettler[1,3], Steven J. Biller[1], Simon J. Labrie[1], Julie M. Miller[1], Nadav Kashtan[1], Sara E. Roggensack[1], Sallie W. Chisholm[1,3]

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
[2]Microbiology Graduate Program, Massachusetts Institute of Technology
[3]Department of Biology, Massachusetts Institute of Technology

Abstract

High-light-inducible (*hli*) genes encode a family of small photosystem-associated, chlorophyll-binding cyanobacterial stress response proteins. Genomes from different clades of the globally abundant marine cyanobacterium *Prochlorococcus* vary widely in the number of *hli* family member genes they carry, from 8 to 43 per genome. These genes likely play a role in niche adaptation within *Prochlorococcus*, as well as differentiating *Prochlorococcus* from other cyanobacteria. Many *Prochlorococcus*-infecting phages also carry *hli* genes, related to a distinct subset within host gene diversity. Here we trace the evolutionary dynamics of *hli* genes in *Prochlorococcus*, the closely related marine cyanobacterium *Synechococcus*, and the phage that infect them, by examining the distribution and diversity of *hli* genes in cultured strains and DNA from the wild. The number of *hli* genes and the assortment of sequence variants differ between the genetically and ecologically distinct clades (ecotypes) that make up the *Prochlorococcus* radiation and also between closely related strains within ecotypes. The genomic context, distributions across taxa, sequence variants and phylogenies of *hli* genes suggest that duplications, rearrangements and horizontal gene transfers have played a role in generating their complex distribution, and that high numbers of *hli* genes in different ecotypes arose through a combination of shared history and independent events. The arrangement of *Prochlorococcus hli* genes into head-to-tail tandem arrays appears to facilitate shuffling of different combinations of genes and the gain and loss of *hlis* in multi-gene sets. We also expanded analyses of the light shock response of cultures developing the idea that certain ecotypes are more tolerant of intense, transient light shock than others. This phenotype corresponds roughly with the number of *hli* genes in the genomes, along with other high-light related genes, consistent with the hypothesis that *hli* genes might play a role in adaptations to fluctuating light conditions, especially in the low-light preferring but high-light tolerant LLI ecotype. The *hli* genes of *Prochlorococcus* have a complex history of expansion, varying in number over several time scales in *Prochlorococcus* evolution, an exception to the paradigm of genome streamlining and loss of paralogs in oligotrophic marine bacteria with small genomes.

3.1 Introduction

***Prochlorococcus* diversity enables its broad distribution over space and depth**

The cyanobacterium *Prochlorococcus* plays a central role in the cycling of nutrients and flow of energy through oligotrophic marine ecosystems (Partensky et al. 1999a). At different sites across its habitat range in the tropical and subtropical open oceans, *Prochlorococcus* contributes 5-80% of phytoplankton primary productivity, often reaching densities of 10^5 cells/ml or 10% of the bacterial population, for an estimated 10^{27} cells globally (Goericke and Welschmeyer, 1993, Liu et al. 1997, Li, 1994, Partensky et al. 1999a, Flombaum et al. 2013). Its abundance and broad distribution is enabled in part by traits that make *Prochlorococcus* particularly adapted to the low nutrient concentrations of the open oceans, such as its small size, and by its diversity, supporting survival of different *Prochlorococcus* under many conditions changing over time, depth and geography (Moore et al. 1998, Scanlan et al. 2009, Biller et al. 2014, Partensky and Garczarek, 2010, Kashtan et al., 2014). At the broadest scale of this diversity, *Prochlorococcus* can be divided into genetically and ecologically distinct ecotypes (Figure 3.1), the low-light adapted (LL, clades LLI-VII) and high-light adapted (HL, clades HLI-VI), based on depth distributions, pigment characteristics, phylogenetic relationships, and, for clades with cultured representatives, the range and optima for growth as a function of light intensity (Urbach and Chisholm, 1998, Moore et al., 1998, Moore et al., 1999, Rocap et al., 2002).

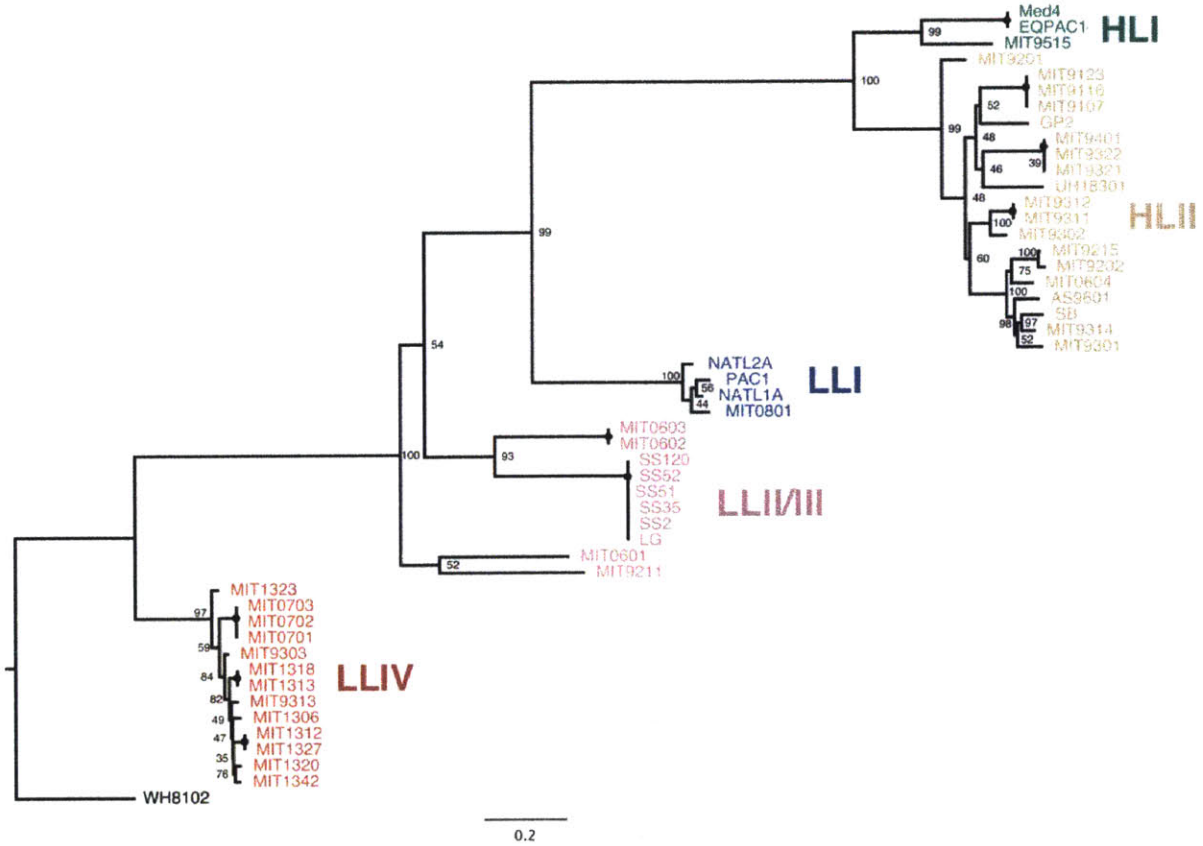


Figure 3.1. ***Prochlorococcus* ecotypes, phylogeny and genomes**
Phylogenetic structure of *Prochlorococcus* ecotypes using the DNA gyrase subunit B gene, a good marker for *Prochlorococcus* phylogenetic resolution (Mühling, 2012), including all available sequenced genomes. The low-light (LL) adapted ecotypes are more deeply branching; the high-light (HL) adapted ecotypes form a derived clade. Maximum likelihood phylogeny, 100 bootstrap replicates, at nodes. Outgroup *Synechococcus* WH8102.

The high-light adapted strains are capable of growth at higher light intensities, grow faster at higher light intensities, and are unable to grow at very low light intensities, compared to low light adapted strains (Moore et al., 1999). Functional variation in temperature adaptation differentiates some ecotypes, and at finer scales of phylogenetic diversity, within ecotypes, differences in nutrient uptake traits enable *Prochlorococcus* populations to adapt across their many chemically distinctive ocean habitats (Zinser et al., 2007, Martiny et al., 2009, Coleman and Chisholm 2007, Scanlan et al., 2009).

Adaptation to changing light conditions distinguishes one *Prochlorococcus* clade

The surface of the open ocean is characterized by a mixed layer, tens of meters deep, created by turbulent mixing processes, identified by its uniform temperature, salinity and density, which is vertically mixed on the timescale of a few days (de Boyer Montégut et al., 2004, Brainerd and Gregg, 1995, Denman and Gargett, 1983). Seasonal temperature changes result in variations in the depth of this mixed layer, negligible in some regions and over 100m in others, with the deepest mixed layers occurring in the winter at seasonally variable sites (Malmstrom et al., 2010 , Giovannoni and Vergin, 2012, Bathen, 1972). Field observations of *Prochlorococcus* ecotype distributions show that one ecotype, the LLI clade, often persists during deep winter mixing events, when other LL ecotypes all but disappear (Johnson et al., 2006, Zinser et al., 2007, Malmstrom et al., 2010, Giovannoni and Vergin, 2012). Deep mixing brings many changes in the physical, chemical and biological aspects of a cell's surroundings, including exposure to light. The light environment of *Prochlorococcus* spans four orders of magnitude in photon flux, from the surface to the base of the euphotic zone (Moore et al., 1999, Zinser et al., 2007). Superimposed on diel cycling, a cell in the mixed layer can experience daily variation in light due to vertical mixing of one order of magnitude in a stratified water column, or 2-4 orders of magnitude during deep mixing events (Figure 3.2, Denman and Gargett, 1983). Below the mixed layer, where LL populations typically reach their maximum population sizes under stratified conditions, the light field is relatively stable, although other mechanisms can vertically perturb these water parcels, including eddies, upwelling and internal waves (Denman and Gargett, 1983).

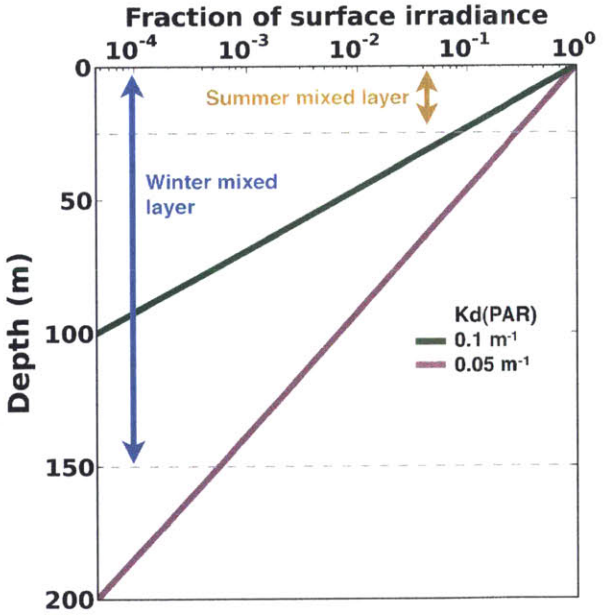


Figure 3.2. The physical light environment of *Prochlorococcus*

A simplified scheme for the *Prochlorococcus* light environment: the idealized extinction of photosynthetically active radiation (PAR) over the water column for 2 typical values of the PAR extinction coefficient measured in the oceans in regions with *Prochlorococcus* (Morel et al., 2007, Flombaum et al., 2013). A K_d(PAR) value of 0.1m^{-1} corresponds to moderately productive ocean conditions (e.g. upwelling near continents or the equator); 0.05m^{-1} is typical of oligotrophic habitats. Examples of mixed layer depths are shown for stratified and well-mixed water columns representing a seasonal range, although values differ significantly with location (Giovannoni and Vergin, 2012, de Boyer Montégut et al., 2004). Vertical advection in the mixed layer occurs over tens of meters over hours or days (Denman and Gargett, 1983).

While absorbing light is an essential part of life for a photoautotroph, excess light can damage the photosystem, slowing or stopping photosynthesis, and generate reactive oxygen species, causing damage throughout the cell (Adir et al., 2003, Long et al., 1994). Shifts in light not only affect the energy balance of the cell, but can also cause stress, photoinhibition and sometimes cell death, so life in fluctuating light presents a challenge (Long et al., 1994, Tikkanen et al., 2012). HL *Prochlorococcus* strains survive intense, transient light shocks, while most LL strains do not, with the exception of the LLI ecotype (Malmstrom et al., 2010, Six et al., 2004, Kettler, 2011). The LLI ecotype shares some features of its light physiology with other LL clades, including optimal growth at low-to-moderate light intensity and the ability to grow at very low light intensities that do not support HL ecotypes (Zinser et al., 2007). However, the LLI ecotype often has an intermediate depth distribution in stratified water columns, peaking in abundance deeper than the HL ecotypes but above other LL *Prochlorococcus* and sometimes appearing in the mixed layer with the HL, so it is sometimes referred to as an intermediate light ecotype (Johnson et al., 2006, Zinser et al., 2006, Zinser et al., 2007, Malmstrom et al., 2010, Partensky and Garczarek, 2010). The genomic GC content and phylogenetic position of the LLI clade is also intermediate between the HL and other LL (Kettler et al., 2007).

With the advent of genomics, one striking feature distinguishing the LLI clade emerged - high numbers of high-light inducible (*hli*) family genes. These genes are multicopy in all cyanobacteria, but in *Prochlorococcus* the number of *hli* genes varies widely by ecotype, 8-15 in most LL genomes, 15-25 in HL genomes, and 25-43 in the LLI clade (Kettler et al., 2007, Bhaya et al., 2002, Coleman and Chisholm 2007, this work). *Hli* genes are small proteins involved in the cyanobacterial response to many stresses, which are in the chlorophyll A/B binding (CAB) superfamily that includes the light harvesting antennae of plants, (Dolganov et al., 1995, Muramatsu and Hihara, 2012, Engelken et al., 2010). The first high-light inducible gene was named for its rapid increase in expression after a shift to high light (*Synechococcus elongatus* PCC7942 *hliA*), and functional work supports a critical role for these genes in light shock and acclimation to high light (Dolganov et al. 1995, He et al., 2001). However, subsequent work, primarily in the major cyanobacterial model strain *Synechocystis* PCC6803, revealed that *hli* genes are also induced in response to nutrient starvation (nitrogen and sulfur) and low temperature, and that they can be expressed under low light conditions, resulting in an alternate nomenclature of small CAB-like proteins, *scp* genes (Dolganov et al., 1995, He et al., 2001, Funk and Vermaas, 1999).

What is the function of *hli* genes?

The precise function of the proteins in the *hli* family are still a subject of active research, but there has been extensive genetic and biochemical exploration of these genes in freshwater model cyanobacteria, developing several ideas about their roles in the cell. *Synechococystis* PCC 6803 has 4 *hli* genes (hliA, B, C, D or scp C, D, B, E, Funk and Vermaas, 1999, He et al., 2001). A mutant with all four of these genes knocked out cannot survive low- to high-light transitions that the wild type can, but the genes are not essential for growth under low-light conditions (He et al., 2001). In these *Synechococystis* experiments, low-light generally refers to something around 40 μmol photons $m^{-2}s^{-1}$, a value well below the saturating irradiance for growth, and high-light refers to something that would be high in the environment, near, at or slightly above the optimal light for acclimated *Synechococystis* growth, e.g 300 μmol photons $m^{-2}s^{-1}$ (Muramatsu and Hihara, 2012, Kopecná et al., 2012). Full midday midlatitude sunlight is around 2000 μmol photons $m^{-2}s^{-1}$, rapidly attenuating under water (e.g. Figure 3.2). Because cells take time to acclimate to different light conditions, during transitions to higher light, a cell can experience photodamage and photoinhibition even under light conditions compatible with growth, but *hli* genes seem to help with this process (Muramatsu and Hihara, 2012).

Hli proteins have been shown to be physically associated with both photosystem I (PSI) (Wang et al 2008) and photosystem II (PSII) (Yao et al., 2007, Kufryk et al., 2008, Promnares et al., 2006), photosystem assembly intermediates (Promnares et al., 2006, Yao et al., 2007, Knoppová et al., 2014), and each other (Storm et al, 2008, Yao et al., 2007). The chlorophyll-containing D1 protein in the oxygen evolving complex of photosystem II sustains particularly heavy damage during high light stress, and a key part of the cell's response to excess light is the rapid replacement of damaged D1 protein (reviewed in Nixon et al 2010, Adir et al., 2003, Muramatsu and Hihara, 2012). Based on the reduction of chlorophyll half-life in *hli* knockout mutants, and other findings, one functional hypothesis is that *hlis* may bind chlorophyll to prevent its degradation during D1 protein replacement, enabling recycling of chlorophyll and repair of damaged chlorophyll (Yao et al., 2012, Vavilin et al., 2007, Nixon et al., 2010, Storm et al., 2008, Promnares et al., 2006, Yao et al., 2007, Chidgey et al., 2014). During light stress, the 4x*hli* knockout strain produces more singlet O_2, one of the toxic products of free chlorophyll photochemistry, which led to the idea that *hlis* may be protecting the cell from the generation of reactive oxygen species by free chlorophyll, simultaneously protecting and disarming this powerful molecule (Sinha et al., 2012, Latifi et al. 2009, Apel and Hirt, 2004). Carotenoids, photoprotective pigments, have also been found associated with *hli* proteins (Daddy et al., 2015, Storm et al., 2008). These pigments act as energy scavengers; in this context, they could be dissipating energy absorbed by the chlorophyll bound to the *hli* protein, or they could be protecting the photosystem by absorbing light directly (Chidgey et al., 2014, Promnares et al., 2006, Yao et al., 2007). The *hli* knockout accumulates the chlorophyll precursor chlorophyllide (which is also a chlorophyll damage product) and has reduced chlorophyll concentrations, consistent with impaired chlorophyll recycling in the absence of *hli* genes or possibly a role for *hlis* in the regulation of chlorophyll synthesis (Yao et al., 2012, Havaux et al., 2003, Xu et al., 2002a and 2002b).

Are there specific roles for different *hli* proteins?

The four different *hlis* of *Synechococystis* have distinct, but overlapping roles. In expression measurements over a range of conditions, hliA, hliB and hliC show similar profiles, but hliD is different (He et al., 2001). In protein sequence, *Synechococystis* hliA and hliB are similar proteins, the product of a relatively recent duplication, and the other two are divergent (Funk and Vermaas, 1999). Among the many *Prochlorococcus hli* genes it is possible to identify five in each genome that are more closely related to these characterized *Synechococystis* proteins (Figure 3.3B, Lindell et al. 2004, Bhaya et al., 2002, Kettler, 2011). In *Synechococystis*, *hli* mutants have growth defects in high light, but the quadruple knockout has a more severe phenotype than single, double or triple mutants, indicating some functional redundancy among these proteins (Havaux et al, 2003, He et al. 2001, Wang et al. 2008). Elegant recent work has led to a more refined understanding of the functional roles of specifically *hliC* and *hliD* (Chidgey et al., 2014, Knoppová et al., 2014). Most photosystem proteins are synthesized on membrane-bound ribosomes and cotranslationally inserted into the thylakoid membrane, and chlorophyll is cotranslationally inserted into apoproteins (reviewed in Sobotka, 2013). *Hli* proteins assist in this process through the delivery and insertion of chlorophyll into newly synthesized apoproteins (Chidgey et al., 2014). Immunoprecipitation of chlorophyll synthase (chlG), an integral membrane protein that performs the last step in chlorophyll synthesis, showed that it is associated with hliD, as well as ribosomes, the SecY/YidC machinery that inserts the proteins into the membrane, and Ycf39, a PSII assembly factor (Chidgey et al., 2014). Ycf39 forms a complex with hliC, hliD, chlorophyll and beta carotenoid, which transiently associates with PSII assembly intermediates (Knoppová et al., 2014). This Ycf39/HliC/HliD complex is thought to safely deliver chlorophyll, both new and recycled, to newly synthesized D1 protein, and also to protect the new D1 and nascent photosystem from light damage, a process of increased importance during the rapid D1 turnover and PSII repair of light shock conditions (Knoppová et al., 2014, Chidgey et al., 2014).

Basic properties of *hli* genes

The *hli* gene family encompasses a broad level of protein diversity, united by a small number of conserved features. These proteins are short, mostly between 30-90 amino acids (Figure 3.3C). They share a conserved transmembrane helix hydrophobic domain and a chlorophyll binding motif, common to all CAB superfamily proteins, and their N-terminal region is highly variable, in both length and amino acid composition (Figure 3.3A; Hess et al., 2001, Dolganov et al., 1995, He et al., 2001, Bhaya et al., 2002). This results in very low protein identity among some members of the family (Figure 3.3B).

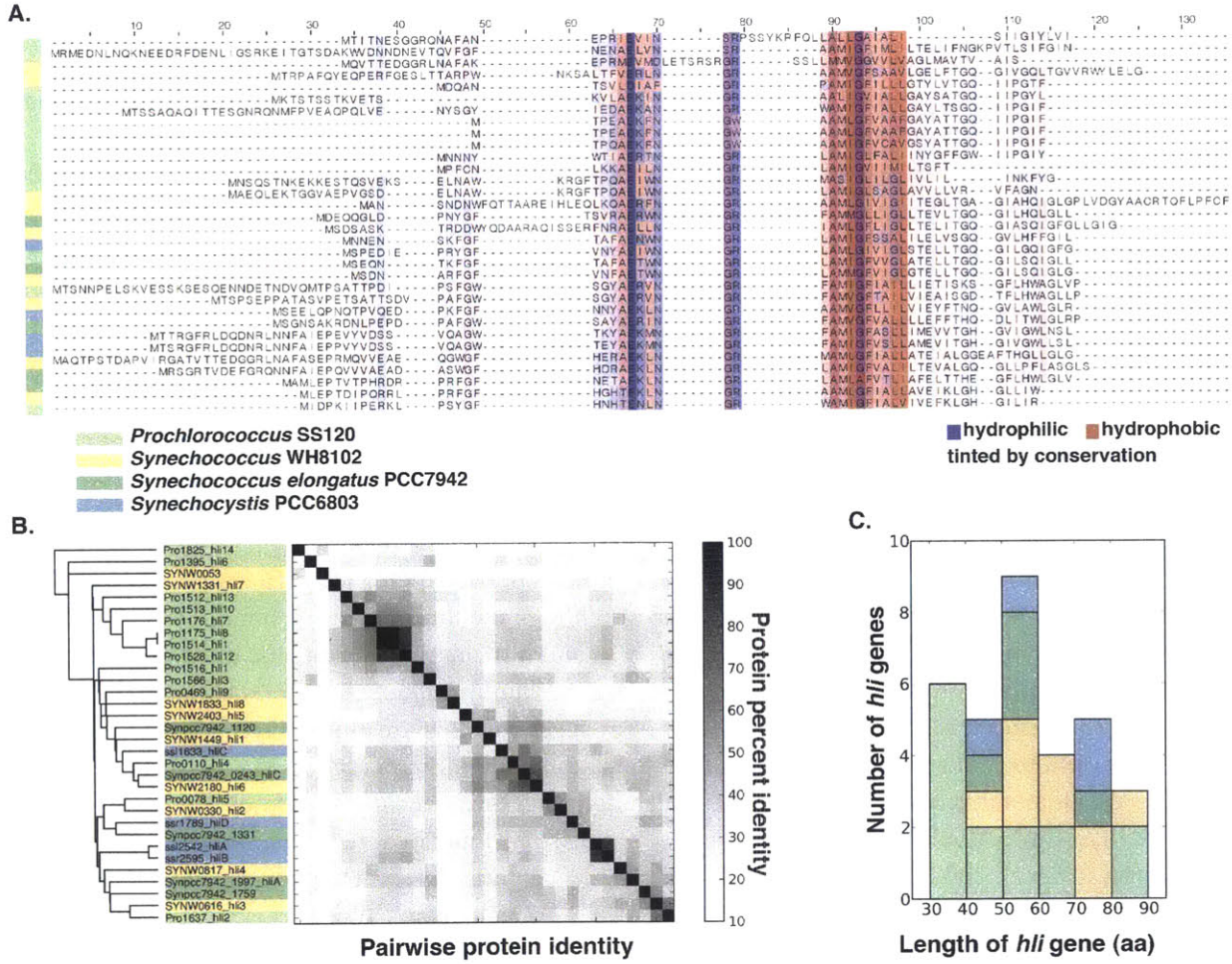


Figure 3.3. Basic Properties of diverse *hli* proteins: all the *hli* genes from four cyanobacterial genomes
(A) Multiple alignment of all the *hlis* from four genomes: one marine *Synechococcus*, WH8102, one *Prochlorococcus*, the LLII SS120, and the two freshwater model cyanobacteria in which most *hli* biochemistry has been done, *Synechococcus elongatus* PCC7942 and *Synechocystis* PCC6803 (colored by genome). Alignment colored by hydrophobicity - shows transmembrane domain, and tinted by conservation, residues with conservation about 15% are tinted, more intense colors are more conserved. Sequences are ordered based on a refined UPGMA clustering (muscle), shown at left of distance matrix in (B). This is not a robust phylogeny, which would be difficult to infer from proteins this diverse. *hlis* are annotated in published databases, either Cyanobase or NCBI. Protein percent identity calculated as identical residues/(aligned residues + internal gaps). (C) Length distributions for these genes.

hli evolution in the *Prochlorococcus* flexible genome and in phages

Phages infecting marine cyanobacteria often carry host-derived genes (referred to as auxiliary metabolic genes or host/phage shared genes), including photosynthesis genes, which are collectively thought to enhance cellular metabolism during infection toward the production of more phage (Sullivan et al., 2003, Mann et al., 2003, Clokie and Mann, 2006,Thompson et al., 2011b). *Hli* genes are in this group, found in the genomes of many phage infecting *Prochlorococcus* and *Synechococcus*. During infection, phage-encoded *hli* genes are expressed and host-encoded *hli* genes are induced, likely both contributing to an expanded pool of *hli* transcripts and proteins (Lindell et al., 2005, Lindell et al., 2007). *Prochlorococcus hli* genes can be broadly divided into two groups, those that cluster with orthologs in freshwater cyanobacteria and occur as single copy core genes in *Prochlorococcus*, and those that cluster with phage genes, without close homologs in freshwater species, and occur in multiple copies per genome (Lindell et al., 2004, Bhaya et al, 2002, Kettler, 2011). Some of the multicopy *Prochlorococcus hli* proteins have a highly conserved C-terminal motif, not found in freshwater *hlis*, which suggests a unique evolutionary trajectory of the *hli* proteins of *Prochlorococcus*, and the possibility of altered function or binding partners specific to this group of proteins (TGQIIPGF/IF, Bhaya et al, 2002). Based on their distribution across *Prochlorococcus* ecotypes and proposed functions, it has been hypothesized that the copy number and sequence variation in *Prochlorococcus hli* genes are part of the genomic adaptation behind light physiology differences between *Prochlorococcus* ecotypes: the light-shock tolerance of LLI strains and general high-light adaptation of HL strains (Bhaya et al. 2002, Coleman and Chisholm, 2007, Kettler et al, 2007).

Each *Prochlorococcus* genome has between 1,900 and 3,000 total genes (Kettler et al., 2007, Biller et al., 2014). About 1,200 of these genes that are shared by all *Prochlorococcus* make up the *Prochlorococcus* core genome, required by all *Prochlorococcus*, containing essential gene and genes needed for common stress conditions, defining the genus (Kettler et al., 2007, Biller et al., 2015). The rest of the genes in each genome are part of what is called the flexible genome, varying from strain to strain, including genes related to specific environmental conditions, and these genes occur in hyper-variable genomic islands (Coleman et al., 2006). Together, the core genome and all the flexible genes of all *Prochlorococcus* make up a set of genes that is called the pan-genome, which currently stands at 13,000 genes for *Prochlorococcus*, based on the genomes sampled so far, and probably contains thousands more genes in the wild (Biller et al., 2015, Baumdicker et al., 2012, Tettelin et al., 2005). Multicopy *hli* genes are often found in *Prochlorococcus* genomic islands, where gene gain and loss in the *Prochlorococcus* flexible genome occurs (Coleman et al, 2006, Kettler 2011). In these islands, *hli* genes are often arranged in tandem, with multiple *hli* genes in a row, head to tail, structures which may enable coordinated expression and duplication or transfer of multiple hlis at once (Bhaya et al., 2002). In a recent study of fine-scale evolution, *hlis* were among the small number of genes which distinguished co-occurring subpopulations of closely related HLII *Prochlorococcus*, very recent change compared to ecotype differentiation, but old enough to represent fixed differences among subpopulations, likely a product of selection (Kashtan et al. 2014).

Hli gene expression in *Prochlorococcus*

Hli genes are an important part of the *Prochlorococcus* expression response to stress, with some members of the gene family differentially expressed in almost every perturbation experiment to date, as in *Synechococystis*, including nitrogen starvation, phage infection, iron starvation, and of course, changing light intensity and color (see Supplemental Figure S3.2 for a summary; Berg et al., 2011, Steglich et al. 2006; Tolonen et al. 2006; Lindell et al. 2007; Thompson et al. 2011a). The only two exceptions were carbon-oxygen ratio manipulation and phosphorus starvation experiments, where *hli* genes did not play a significant part in the response (Bagby et al. 2015, Martiny et al. 2009). In two instances, specific regulatory sequences have been detected upstream of *hli* genes, the NtcA nitrogen regulator in *Prochlorococcus* genomes

(Tolonen et al. 2006), and the pho box phosophorus-uptake master regulator in phage (Kelly et al. 2013), indicating that *hli* gene expression is likely a coordinated part of these major stress-response regulons. Across these experiments, overlapping subsets of the *hli* gene family responded under different conditions, but in all cases it was the *Prochlorococcus*-specific, multicopy *hli* genes that respond to stress in *Prochlorococcus*, not the single-copy *hli* genes more closely related to freshwater *hlis* (Supplemental Figure S3.2; Kettler, 2011). It appears that the conserved, single copy, freshwater-shared genes are no longer responding to stress in *Prochlorococcus* the way their orthologs in *Synechocystis* do, although they are maintained in the genome. Perhaps the newer *Prochlorococcus* specific multicopy genes have taken the stress responsive part of *hli* functionality, and the single copy freshwater-like genes provide housekeeping functions in chlorophyll trafficking and photosystem assembly. Diel expression measurements, over the course of a day-night cycle in a culture, indicate that the expression of *hli* genes is part of the daily cycle of gene expression in the cell growing under optimal conditions, not just during stress, with different *hli* genes cycling in expression throughout the light-dark cycle with different phases and amplitudes (Zinser et al., 2009). In a metatranscriptomic study sampling over the course of the day-night cycle at 23m, near Hawaii where *Prochlorococcus* is a major component of the community, 16 distinct *Prochlorococcus hli* gene clusters were detected, again with peaks at different times over a day-night cycle, demonstrating that the expression of these genes is a basic part of life for *Prochlorococcus* in surface waters (Ottesen et al. 2014).

An evolutionary framework for *hli* genes across *Prochlorococcus*, *Synechococcus* and their phage

After observing the recurring importance of *hli* genes in many areas of *Prochlorococcus* biology, particularly their copy number variation and connection to the unique properties of the LLI clade, we were curious about the evolutionary dynamics of *hli* genes among *Prochlorococcus*. Our work is guided by the following questions: How are *hli* gene variants distributed across genomes and ecotypes? How do the *hli* genes in *Prochlorococcus* and marine *Synechococcus* compare, and what can this tell us about the ancestral state and ecological roles of these genes? When, relative to the *Prochlorococcus* phylogeny, did duplications of different *hli* genes occur? By what molecular mechanisms have *hli* complements changed through time? In particular, how did the high copy number in the LLI arise? How often, and from which host clusters, have phage acquired *hli* genes? To address these questions we first re-annotated *hli* genes across a large dataset of culture-based and wild DNA, to build a consistent and comprehensive picture of the distribution of *hlis* in *Prochlorococcus*, *Synechococcus* and their phage. To bring order to the many *hli* genes of *Prochlorococcus*, we performed a clustering analysis, sorting them into sets of deeply branching orthologous proteins, so that we could begin to treat them as the diverse and distinct proteins that they are, instead of a single group. This exercise uncovered the nature and timing of expansions and major protein innovations over the course of *Prochlorococcus hli* evolution. We then compared chromosomal arrangements and gene phylogenies within these clusters to build hypotheses about the events giving rise to the observed *hli* genes, which include a variety of evolutionary mechanisms. We also continued to explore ecotype light physiology, an important part of the context for interpreting these genomic differences, and by screening additional strains for their response to light shock, we gathered substantial additional support for the hypothesis that LLI strains are more light shock tolerant than other LL adapted *Prochlorococcus*, as well as uncovering new features of the response to light shock.

3.2 Materials and Methods

Single-cell amplification, DNA sequencing and assembly

Cells from the Hawaii Ocean Time Series and Bermuda-Atlantic Time Series sites were collected as described in (Kashtan et al., 2014), and sorted and amplified at the Bigelow Center for Single Cell Genomics. Cells from the Eastern Tropical South Pacific oxygen minimum zone were collected in November, 2010, during the CMORE BiG RAPA cruise, from 55m depth in the secondary chlorophyll maximum of Station 1, in the OMZ. Each 1 ml seawater was mixed with 10% glycerol and flash frozen, then stored at -80C. A single sample was thawed for this single-cell sort. Sequencing libraries were constructed as described in Rodrigue et al., 2010, with epicentre phi29 DNA polymerase, and sequenced on the illumina GAII and HiSeq platforms, in several batches. Raw data was trimmed based on base call quality scores (CLC quality trim), adapters were removed using the cutadapt program, and genomes were assembled using SPAdes 3.0.0.

Phage genomes selected for analysis

The phage genomes used in our analysis include all available genomes from phage isolated on *Prochlorococcus*, and subgroup 5.1, 5.2 and 5.3 *Synechococcus*, which are mostly marine, but also include a few freshwater and brackish strains that are phylogenetically affiliated with marine groups. Likewise the phage are mostly from marine (euryhaline) samples, but in a few cases came from brackish or freshwater samples, but are phylogenetically affiliated with marine cyanophage or isolated off marine hosts.

Annotation of *hli* genes: Whole genome re-annotation and ORF calling

To ensure consistent annotations of *hlis* and their surrounding genomic context, all sequences (published and new) used in this study were first re-annotated using the Prokka annotation pipeline (Seemann et al., 2014), which uses the prodigal gene caller (Hyatt et al., 2010) combined with provisional functional annotation based on several protein databases, and runs a suite of external tools for identification of additional features. Prokka was run using aragorn for tRNA/tmRNA identification (Laslett and Canback, 2004), Barrnap for rRNA identification (vicbioinformatics.com/software.Barrnap.shtml), specifying kingdom bacteria/virus depending on the sample, and otherwise with default settings, without ncRNA or signal peptides searches and without the use of a genus-specific database.

Hli genes are occasionally missed by automated gene callers, due to their small size or because their presence in genomic islands and horizontal transfer histories result in properties inconsistent with the majority of the genome. Prodigal (Hyatt et al., 2010) builds training sets from the genome itself for start codon preference, RBS motifs and distances, and incorporates information on GC content and dinucleotide frequency, all of which could be out of character for a recently transferred gene. To ensure sensitive detection of *hli* ORFs, we performed a second ORF calling method, a naive search for start/stop codon pairs, allowing for alternate start codons TTG, GTG and CTG, retaining all ORFs longer than 30 amino acids and preferring the longest possible ORF faced with a choice of start codons. Many of the products of this method are spurious ORFs or excessively long ORFs, and in general the prodigal method is much better for accurately calling genes, but we used it only to expand our search set of possible *hli* proteins, and it enabled the recovery of several additional *hli* genes per genome missed by prodigal, including a few previously annotated ones. Different gene callers make different choices among possible start codons, resulting in orthologous genes or identical with substantively different lengths. In all cases, where the naive gene calls and prodigal gene calls shared a stop codon but differed in the choice of start codon, the more informed prodigal gene calls were used.

Annotation of *hli* genes: Building sets of *hli* genes as input for hidden Markov models

To identify *hlis* from our set of candidate ORFs, we used a set of hidden markov models (HMM) constructed from a set of previously annotated *hli* genes, following the approach developed in Greg Kettler's dissertation, with some modifications (Kettler 2011). Previous *hli*-specific studies have used motif searches (Lindell et al. 2004) or a single HMM for all *hli* genes (Bhaya et al. 2002). These methods rely on the small number of highly conserved residues to identify the genes, which is a good method. However, the motif breaks down in some places, and we do not known enough biochemically to evaluate the meaning of such events to place hard limits on the definition of an *hli* gene. So, we chose to use a gene-specific homology based approach targeted for *Prochlorococcus*, *Synechococcus* and phage *hlis*, building custom HMMs based on different clusters of genes from these taxa, so that the full length of the genes inform the homology search, not just the small region of the genes that are conserved across all *hlis*.

We gathered the HMM input search data set from published annotations of *hlis* built using diverse methods from past studies. We used 13 *Prochlorococcus* genomes, 11 *Synechococcus* genomes and 36 myophage and podophage isolated on *Synechococcus* or *Prochlorococcus* with annotated hlis (listed in Supplementary Table 3.5). We took any gene in published versions of these annotations (downloaded from ncbi databases) annotated as 'high-light inducible' or a variant thereof (HLIP, *hli*, CAB-like). To this set we added any gene in a Version 3 ProPortal *Prochlorococcus/Synechococcus* ortholog cluster (Kelly et al., 2012) for which one member of the clusters was annotated in the ProPortal functional notes as a high-light inducible gene (or variant). Some of these past annotations were motif-based, some homology based, using different databases available over time.

Next, we wanted to sort these proteins into different sequence clusters representing different orthologs; published ortholog clusters with homology and length settings chosen for most genes did not perform well on this gene family, overclustering and underclustering different parts of *hli* diversity. The full set of these protein sequences were aligned with the gap-friendly global alignment MAFFT parameter set 'einsi' (Katoh et al;, 2013), and roughly phlyogenetically clustered using the FastTree maximum likelihood approximation (Price et al., 2010). Sub-clusters of this tree were identified first using taxa distributions, based on the idea that ancient orthologs are distributed across more taxa. This led to approximate 30% identity cutoffs between sequence groups. We inspected these potential clusters as multiple sequence alignments (MAFFT-einsi) to assess shared conserved residues in the variable N terminal region, indicating gene-wide homology beyond the *hli* motif region. These groups of related genes that could be aligned over their full length were used as search inputs for HMM, 16 different groups of *hli* genes. Any genes in clusters with fewer than three examples were left out of these models.

One additional set of genes was necessary to avoid misclassifying the ferrochelatase gene as an *hli*. Ferrochelatase catalyzes the insertion of iron into protoporphyrin to produce heme, a critical branch point between heme and chlorophyll synthesis in protoporphyrin metabolism (Sobotka et al., 2008). In cyanobacteria, the primary catalytic domain of ferrochelatase is homologous to other bacterial ferrochelatase genes, but the enzyme contains an additional domain highly similar to *hli* genes (likely derived from a gene fusion), hypothesized function as a regulatory domain influencing ferrochelatase activity, perhaps based on chlorophyll binding (Funk and Vermaas, 1999, Sobotka et al., 2008, Storm et al., 2013). This domain is still similar to *hli* genes, causing some confusion in annotations based on homology, especially in single cells which contain many partial genes. We built a HMM using 12 ferrochelatase genes spanning *Prochlorococcus* diversity. The gene is sufficiently conserved that the *Prochlorococcus* genes identify the ferrochelatase in every *Synechococcus* genome as well. We used this information to exclude ferrochelatase genes from our *hli* analyses.

Annotation of *hli* genes: implementation of hidden Markov models for *hli* annotation

To build HMMs, we aligned each group of *hli* genes with with MUSCLE 3.6 with default parameters, converted to Stockholm alignment format with BioPython (Cock et al., 2009), and ran hmmbuild (part of HMMer 3.0) with default parameters (Finn et al., 2010, Edgar et al., 2004). We ran hmmsearch (version 3.0) with e-value cutoff (-domE) 0.001 to identify significant matches to each *hli* gene cluster (Finn et al., 2010).

***hli* motif searching for comparison and additional classification**

We performed several motif searches to place our annotation and *hli* clustering results in the context of past work and assess conservation of key residues. In the following motifs, "x" represents any one amino acid and "/" means 'or.' These searches include (i) the motif most recently used for systematic annotation of *hli* genes: a match of at least six of 10 amino acids in the motif AExxNGRxAMIGF (Lindell et al. 2004 and Kettler et al. 2007), (ii) a related motif referenced in certain *hli* ncbi genbank annotations, citing Bhaya et al., 2002, visible in that paper's alignments but not explicitly discussed, ExxNGxxAMxG, (iii) the motif identified as common in the C-terminal of certain *Prochlorococcus hli* proteins, TGQIIPGF/IF and (iv) the biochemically supported CAB chlorophyll binding motif ExxN/HxR, which is sufficient for chlorophyll binding, even in a tiny peptide (Eggink and Hoober 2000). All motif searches were implemented in BioPython, run on new and old *hli* annotations described above.

Clustering *hli* proteins to identify ortholog groups

After identifying a large set of possible *hli* genes through the above annotation processes, to identify meaningfully similar groups of orthologs, we clustered raw gene finds based on a UPGMA clustering based on pairwise global comparisons. This method does not rely on a multiple alignment, which we decided was preferable after viewing many low quality multiple alignments of the dataset. We identified clusters based on the topology of the UPGMA clustering, informed by the span of taxa in ancient orthologs, an approximate 30% protein identity cutoff between ortholog clusters, and inspection of multiple sequence alignments within clusters to assess the extent of conservation across full protein length. These came out similar to the rough clustering used to build the HMM search clusters, but were not in all cases identical, since many new genes were added to the set. These groupings are similar to the input search clusters described above in construction of gene-by-gene HMMs, but not identical. In all the simpler cases, the same clusters were recreated, just expanded with more data. We found it easier to re-cluster the data than to fit all the new data into the old models, because some genes not in that previously annotated set were not accounted for in the previous clustering, and some clusters were unstable with more data.

***Prochlorococcus* GyrB phylogeny**

For the phylogeny in Figure 3.1 illustrating relationships among *Prochlorococcus* genomes used in this study, we gathered gyrB DNA gyrase subunit B DNA sequences (from Prokka annotations described above). This gene was chosen to build a rough idea of relationships among genomes because an analysis of marker and core genes found that it does a good job of resolving phylogeny, is easy to align and generally agrees in phylogenetic branching with other core genes and markers (Mühling 2012). Sequences were aligned in muscle v 3.8.31, using default settings, which performed well for this gene (Edgar, 2004a and Edgar 2004b). Phylogeny was inferred using phyml, with the HKY model (allowing for different frequencies of nucleotides and different transition and transversion rates), with an estimated proportion of fixed sites and four gamma distributed rate categories, and 100 bootstrap replicates (Guindon et al., 2010).

Note on ecotype assignments

It is useful to note that the ecotype nomenclature applied to *Prochlorococcus* has changed over time and between publications. In some publications the ecotypes were referred to using an "e" before the name of a type strain belonging to a particular clade (Ahlgren et al., 2006, Zinser et al., 2007, Malmstrom et al., 2010). Here we use the HL and LL followed by a number notation (West and Scanlan, 1999, Biller et al., 2015), which has come to be accepted as a better naming scheme as more and more *Prochlorococcus* sequences are gathered from the environment. Through clone libraries and metagenomic sequencing, several clades have emerged that do not have cultured representatives, so the HL/LL terminology seems more useful moving forward. The equivalencies between the old and new terminologies for our purposes are: eMIT9313 = LLIV, eMIT9211 = LLIII, eSS120 = LLII, eNATL or eNATL2A = LLI, eMed4 = HLI, eMIT9312 = HLII. Here we refer to LLII/III as a single group in some places; they are under-sampled groups that we know little about (Biller et al., 2014), and they appear to be similar and form a larger clade by some phylogenetic methods.

Light shock experiments: Basic culture conditions

For light shock experiments, *Prochlorococcus* and *Synechococcus* cultures were grown in batch culture in Pro99 media (Moore et al. 2007) made from Sargasso Sea water. A single batch of seawater collected in April 2013 was used for all experiments. Maintenance culture volumes ranged from 15-35ml, in borosilicate glass tubes. Prior to shock experiments, cultures were acclimated for at least two months to growth at 21 ± 0.5 °C and 27 ± 3 µmol quanta m^{-2} s^{-1} continuous illumination under white fluorescent lamps. Under these conditions, all strains used grow consistently, though at different rates, and the light intensity is well below photoinhibition levels (Moore et al., 1999, Zinser et al., 2007). Cultures were transferred to fresh media with approximately 1 ml of culture diluted into 20 ml fresh media every 5-14 days, at late log phase, to avoid stationary phase and keep cultures as near to continuously growing as possible. Growth was monitored intermittently to determine transfer timing and acclimation status, and daily during experiments, using a chlorophyll fluorometer (10-AU and TD-700 models, Turner Designs, Sunnyvale, CA).

Light shock experiments: Strains

All the strains used have sequenced genomes. We used axenic strains when available (WH8102ax, WH7803ax, SBax, MIT9301ax, MIT0801ax, NATL1Aax, NATL2Aax, Med4ax, MIT9313ax), but to cover the full range of *Prochlorococcus* ecotype diversity, we also included several nonaxenic strains (MIT9303, SS120, MIT9211). Axenic cultures are difficult to obtain in *Prochlorococcus*, the result of diverse purification methods with patchy results, but hopefully over time more axenic representatives from these clades will become available (Moore et al. 2007, Berube et al. 2014). While not ideal, this work with nonaxenic strains turned out to be somewhat fortuitous, revealing an interesting interaction between heterotrophs and light shock, without entirely compromising our ability to study light shock in these strains (see Results). Axenicity was monitored before and during each experiment using three different purity test broths, Marine Purity Broth (Saito et al., 2002), ProMM (Berube et al., 2014) and ProAC (Morris et al., 2008), supplemented with occasional checks by flow cytometry (BD/Cytopeia Influx). Strain identity was checked prior to experiments via PCR and sequencing of the ITS marker region as in Rodrigue et al., 2009. Sequence was performed by Eton Biosciences, Cambridge, MA.

Light shock experimental design

Additional lamps were used to create a high-light space in the same incubator used for acclimated growth. The addition of numerous fans to normal incubator air flow was necessary to maintain even temperature across these dramatic light gradients. For each light shock experiment, six identical cultures

were inoculated from a single parent culture to a low but detectable initial target fluorescence, and allowed to grow at 27 ± 2 μmol quanta m^{-2} s^{-1} for three days for *Prochlorococcus* or 1.5 days for the faster growing *Synechococcus* strains, to establish exponential growth (to early- or mid-log phase) and separate the effects of transfer and light shock. Three of the cultures were then moved to 300 ± 15 μmol quanta m^{-2} s^{-1} for four hours of light shock, then returned to 27 ± 2 μmol quanta m^{-2} s^{-1}, and monitored for one week (or more). Daily measurements or samples were taken for bulk culture fluorescence, flow cytometry and fluorescence induction and relaxation analysis for the first eight days of the experiment, or for fast-growing *Synechococcus*, at twice the cadence for half the time (same number of samples), in all cases with additional samplings on the day of the shock, one immediately before the shock and one immediately after. Bulk fluorescence measurements were continued for at least 11 days after transfer, by which point most control cultures reached stationary phase.

Light intensity was monitored regularly, several times throughout the course of each experiment, using a photosynthetically active radiation sensor (LI-COR, Lincoln, NE), by vacating each rack position and measuring where the center of the culture would sit, so our values describe ambient light, not accounting for tube glass or water. Each strain was tested in turn in the same space, not at the same time, due to space and volume constraints, but in the same incubator space, same conditions and similar timing. Experimental culture volumes started at 35 ml, and were reduced to 25 ml after sampling; sampling was limited because volumes less than this interfere with fluorescence readings in our fluorometers. For each datapoint raw chlorophyll fluorescence measurements were modified by subtraction of average of triplicate background readings of sterile media, and the contribution of this measurement's error was added in quadrature with the sample standard deviation for plotted data error bars.

Co-culture experiment

For the co-culture experiment, we used the strain *Alteromonas macleodii* AIA, which was isolated from *Prochlorococcus* NATL2A, has a sequenced genome and has been used extensively in co-culture experiments (Allison Coe and Steven Biller, in preparation). This culture was grown up from frozen stock in ProMM media, then inoculated into an axenic MIT9313ax culture. The co-culture was grown to log phase, allowing the *Prochlorococcus* and *Alteromonas* to adjust to each other's presence, and used to inoculate six cultures at the same initial fluorescence value (requiring slightly different dilutions). These were grown for three days into log phase, alongside an axenic control. Then three of each treatment were exposed to light shock as described above. Samples for flow cytometry and fast induction and relaxation fluorometry measurements were taken for eight days after inoculation (not yet analyzed).

3.3 Results and Discussion

3.3.1 *Prochlorococcus* response to light shock

Is light shock tolerance an ecotype specific trait?

The relationship between *Prochlorococcus* ecotypes, light physiology and relative abundance during deep mixing events led to the hypothesis that the LLI clade tolerates changing light conditions better than other low-light adapted strains (Zinser et al., 2007, Malmstrom et al., 2010, Kettler, 2011). It has been suggested that the relatively high number of *hli* genes in LLI genomes may be a genomic adaptation contributing to this phenotype (Malmstrom et al., 2010). The culture-based evidence establishing that LLI strains tolerate severe transient light shocks better than other LL strains was based on a small set of type strains representing broad ecotype classes (Malmstrom et al. 2010, Kettler et al. 2011). Here we explore the light shock physiology of a set of strains representing all cultured ecotypes and multiple strains within ecotypes (Table 3.1). We ask whether there is within-ecotype variation in light shock tolerance, or whether it is truly an ecotype-wide trait. To put this in the larger context of picocyanobacterial evolution, we included two marine *Synechococcus* strains, which tolerate higher light than *Prochlorococcus* (Mella-Flores et al. 2012, Moore et al. 1995).

Table 3.1. Strains used in light shock experiments.

Strain[1]	Genus	Clade	Isolation region	Isolation depth	Number of *hli* genes[2]	References[3]
WH8102ax	*Synechococcus*	5.1A III	Sargasso Sea	n/a	9	Palenik et al. 2003, Waterbury et al. 1986
WH7803ax	*Synechococcus*	5.1B V	North Atlantic	25m	10	Dufresne et al. 2008, Waterbury et al. 1986
MIT9313ax	*Prochlorococcus*	LLIV	Gulf Stream	135m	10	Rocap et al. 2003, Moore et al. 1998
MIT9303	*Prochlorococcus*	LLIV	Sargasso Sea	100m	10	Kettler et al. 2007, Moore et al. 1998
MIT9211	*Prochlorococcus*	LLIII	Equatorial Pacific	83m	13	Kettler et al. 2007, Moore et al. 1999
SS120	*Prochlorococcus*	LLII	Sargasso Sea	120m	14	Dufresne et al. 2003, Chisholm et al. 1992
NATL1Aax	*Prochlorococcus*	LLI	North Atlantic	30m	43	Kettler et al. 2007, Partensky et al. 1993
NATL2Aax	*Prochlorococcus*	LLI	North Atlantic	10m	43	Kettler et al. 2007, Scanlan et al. 1996
MIT0801ax	*Prochlorococcus*	LLI	Sargasso Sea	40m	40	Biller et al. 2014
Med4ax	*Prochlorococcus*	HLI	Mediterranean Sea	5m	24	Rocap et al. 2003, Moore et al. 1995
MIT9301ax	*Prochlorococcus*	HLII	Sargasso Sea	90m	17	Kettler et al. 2007, Rocap et al. 2002
SBax	*Prochlorococcus*	HLII	Western Pacific	40m	19	Biller et al. 2014, Shimada et al. 1995

[1]ax following strain name denotes axenic; if missing, strain is non-axenic.
[2]Numbers are based on homology searches described later in this chapter. While precise numbers differ depending on method of annotation, the trends of varying *hli* copy number across *Prochlorococcus* ecotypes are robust across different methods.
[3]Citations refer to genome publication and strain reference, in that order, which for one case is the same.
Colors correspond to ecotype designations, matched to Figure 3.1.

Light shock experiments

We screened this diverse set of isolates for the ability to survive a severe transient increase in light intensity, using timing and light conditions similar to those previously shown to be diagnostic for differences between ecotypes (Malmstrom et al. 2010). For each light shock, cells acclimated to moderate light (27 μmol photons $m^{-2}s^{-1}$), were moved for four hours to an order of magnitude brighter conditions (300 μmol photons $m^{-2}s^{-1}$), then returned to the moderate light conditions of acclimation. This roughly mimics the light shock corresponding to a trip up and down in the water column by tens of meters. We monitored growth for a few days before the shock, to ensure cells were growing exponentially, and a few days after, to assess the effect of light shock. All of these strains grow well at 27 μmol photons $m^{-2}s^{-1}$, although not at the same rates, and this level is well below photoinhibition light intensity for all strains (Zinser et al. 2007, Moore et al. 1995, Moore et al. 1999, Moore et al. 1998). HL *Prochlorococcus* and *Synechococcus* can grow well at 300 μmol photons $m^{-2}s^{-1}$, although in some cases it is above their optimal light for growth, but LL strains, including LLI, cannot grow under these conditions (Zinser et al. 2007, Moore and Chisholm 1999).

Strains from the same ecotype have the same light shock response and LLI strains are shock-tolerant

Cultured representatives from the LLI and HL clades of *Prochlorococcus* and marine *Synechococcus* fully tolerate this light shock, while the LLII/III and LLIV representatives show signs of severe inhibition (Figure 3.4). This finding is consistent with past work and supports the hypothesis that the LLI ecotype is adapted to changing light conditions, compare to other LL strains (Malmstrom et al. 2010). Our data support the idea that the light shock phenotype is shared across members of the same ecotype. The light shock tolerant phenotype corresponds with high *hli* copy number across *Prochlorococcus*, consistent with the predicted role of *hli* genes in surviving this form of stress, although many other genomic features likely also contribute to this phenotype. The marine *Synechococcus* tested have only a modest number of *hli* genes (around 10), the same as LLIV *Prochlorococcus*; their ability to withstand high light likely occurs through different mechanisms, since they have very different photosystems (Scanlan et al. 2009). Because we do not have genetic tools for the creation of targeted knockouts in *Prochlorococcus*, we cannot directly test relationships between genotype and phenotype. Nonetheless, this kind of exploration of natural variation in phenotype, and its association with genotype, helps shape hypotheses about the significance of genomic adaptations, and the functional implications of molecular diversity in the wild.

An unexpected result in this series of experiments was the difference in behavior between strains from the LLIV clade and the LLII/III clades (Figure 3.4). The former nearly bleach after the light shock and remain at fluorescence values near background for a week after the shock, while the latter lose fluorescence for one day following the shock, then slowly recover, an intermediate phenotype. In the wild, LLII/III populations sometimes reach their maximum abundance slightly above LLIV populations in stratified water columns, which hints at a slightly different light physiology (Ahlgren et al., 2006, Maelstrom et al., 2010, Zinser et al., 2006, Zinser et al., 2007). Both occur at vanishingly low concentrations during deep mixing events, consistent with their distinct light-shock sensitive phenotype relative to other groups tested here. These LLII/III genomes have several more *hli* genes than the LLIV (Figure 3.8), consistent with the hypothesis that these genes confer increased light shock tolerance. In the *Prochlorococcus* phylogeny, the LLIV clade is the most deeply branching (sharing a common ancestor with the rest of *Prochlorococcus* the longest time ago), followed by the LLII/III, then the LLI, and finally the more recently emerging HL clades. In this context, it appears that more derived groups gradually acquired photoinhibition tolerance traits.

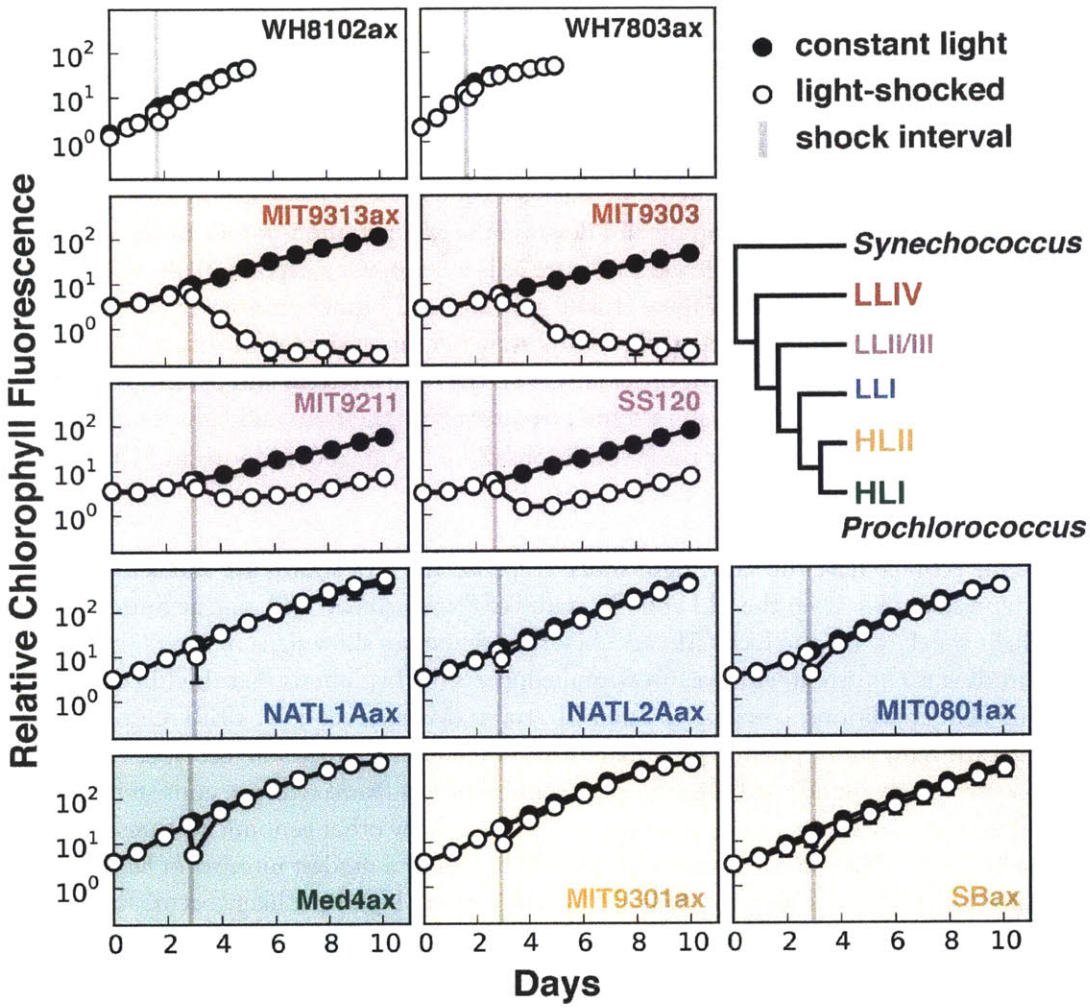


Figure 3.4. Light shock response phenotypes across *Prochlorococcus* diversity
Each box represents one strain, grown at 27 μmol photons $m^{-2}s^{-1}$ for 3 days to exponential growth phase, exposed to 300 μmol photons $m^{-2}s^{-1}$ for 4 hours, and returned to previous conditions. Schematic phylogeny at right shows relationships among ecotypes, in colors matching representative strains (see also Table 3.1 for these assignments). Measurements represent bulk culture chlorophyll fluorescence of shocked cultures and control cultures growing under constant light over time. Each point is an average of triplicate cultures, and error bars (some smaller than points) represent standard deviation with an added noise component from the subtraction of background measurements. The *Synechococcus* strains used here grow much faster than *Prochlorococcus* under these conditions. To capture these cultures over exponential phase, samples were taken at twice the cadence, and the duration of the experiment was half the time of the *Prochlorococcus* experiments, but the shock duration was still four hours. These experiments were performed in a background of growth at continuous light, to avoid the complications of the cells' shifting light physiology over a daily cycle.

Nature of the measurements

These data represent bulk chlorophyll fluorescence measurements of the shocked and control cultures. This quantity is a product of photosystem quantity and functional state, shifting with the acclimation state of cells and changes in photon exposure, but it increases proportionally with biomass when cells are in balanced exponential growth. For the cultures that are relatively unaffected by light shock, bulk fluorescence

dips over the course of the 4 hour shock, but recovers nearly to the level of the unshocked control after 24 hours, after which point exponential increases in fluorescence matches that of the control, in balanced growth, indicating a return to growth after the shock. By contrast, for the LLII/III and LLIV strains bulk fluorescence exhibits a continued decline at 24 hours, indicating a departure from balanced growth, and cultures appear partially bleached, with very pale color compared to unshocked controls. We took additional samples (not yet analyzed) that will help to disambiguate the effects of cell growth from physiological responses to light shock: samples for flow cytometry, which enables cell counts and per cell chlorophyll fluorescence measurements, and measurements with fast repetition rate fluorometry, which probes photosynthetic efficiency and photosystem absorptive cross section. Nonetheless, we can infer that the dramatic differences between ecotypes observed in bulk fluorescence patterns are indicative of different light-shock response phenotypes, in some cases robustly tolerant, in other cases severely affected by this perturbation.

Long term behavior after light shock of axenic and non-axenic low light cultures

One potentially complicating factor in these experiments is that three of the strains used from the LLII/III (SS120, MIT9211) and LLIV (MIT9303) clades were not axenic (see Methods). The use of nonaxenic strains was necessary to achieve our goal of exploring light shock across a broad phylogenetic range of *Prochlorococcus*, because only one LLIV and no LLII/III strains were available as axenic cultures at the time. While we might not expect heterotrophs to affect the immediate photophysiology of light shock, the presence of certain heterotrophic bacteria is known to alleviate redox stress in *Prochlorococcus* cultures by removing hydrogen peroxide, which *Prochlorococcus* does not encode the necessary enzymes to detoxify, and perhaps through other unknown mechanisms as well (Morris et al. 2011, Morris et al. 2008). Hydrogen peroxide might be among the many reactive oxygen species generated during photoinhibition (reviewed in Nishiyama et al. 2006, Muramatsu and Hihara, 2012), so it is reasonable to imagine that heterotrophs could interact with the effect of light shock on cells. Although we focused our sampling efforts on the days immediately surrounding the light shock for our primary object of observing the cultures' response to light shock, for most experiments we continued to take basic culture bulk fluorescence measurements for longer, until cultures reached stationary phase. For all three nonaxenic LLII/III and LLIV strains tested, cultures made a full recovery, after initial inhibition by light shock, while the one axenic LLIV strain, MIT9313ax did not (Figure 3.5). This strain dropped to background and limit of detection on the fluorometer at which point we stopped monitoring it; at the same time point in MIT9303, recovery had already begun. This suggested to us the possibility that the presence of heterotrophs may be alleviating some of the stress of light shock, enabling survival of at least some cells and detoxification of conditions to a sufficient degree to support later growth. From this data alone, we cannot say whether this difference is definitely due to heterotrophs, or genetic differences between strains, or if perhaps the axenic 9313 might have eventually recovered if sampled longer. So we did a controlled experiment with an axenic strain and a heterotroph to explore this phenomenon further.

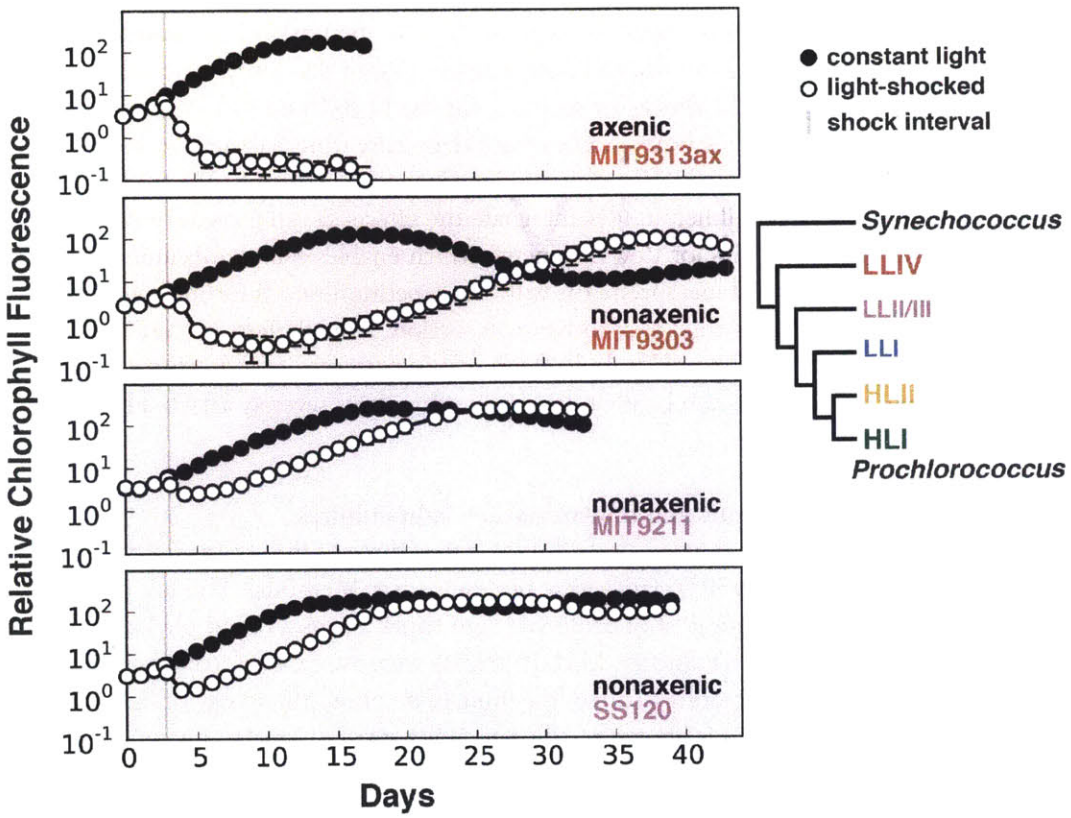


Figure 3.5. Long term behavior of cultures over light shock experiments with LLII/III and LLIV clade strains. For the same experiments described in Figure 3.16, some cultures were followed past the primary targeted time range of the experiment, which reveals that some inhibited cultures eventually recover and grow, to the same stationary phase culture fluorescence as the controls 1-3 weeks later. The first panel shows an axenic culture, which we stopped monitoring when the shocked sample reached background fluorescence (culture cleared). The next three show long term results for the three nonaxenic cultures used in these experiments, which we measured longer than initially planned when they began to show recovery.

The effect of beneficial heterotrophs on light shock

We ran an experiment to address the hypothesis that the difference in recovery after light shock between our two LLIV strains could be due to the presence of heterotrophs and to determine the extent to which heterotrophic bacteria influence light shock response behavior. We intentionally contaminated the LLIV axenic strain MIT9313ax with a 'helper' contaminant isolated off another nonaxenic *Prochlorococcus* culture (*Alteromonas* AIA from NATL1A), which has been previously used in co-culture experiments (Steve Biller and Allison Coe, personal communication). The axenic strain was bleached following light shock, and did not recover. In contrast, the presence of the helper heterotroph resulted in eventual full recovery of the co-culture following light shock (Figure 3.6). This supports the idea that contaminating heterotrophs could be responsible for the long term recovery from light shock observed in nonaxenic LLII/III and LLIV *Prochlorococcus* (Figure 3.5) relieving some aspect of the stress of light shock over the course of the culture. However, in the immediate aftermath of light shock in our co-culture experiment, the decrease in fluorescence in axenic and nonaxenic cultures is highly similar, suggesting that the short term behavior is independent of the presence of heterotrophs, instead governed by the effect of photoinhibition damage on *Prochlorococcus*. This piece of information supports our interpretation of the intermediate phenotype LLII/

III results above (Figure 3.4). If the immediate days after light shock are more likely to represent *Prochlorococcus* physiology, the differences in this phase between LLII/III strains and LLIV strains may genuinely represent intermediate phenotype governed by ecotypic *Prochlorococcus* genetic differences, despite the cultures' contaminating heterotrophs.

This co-culture's high cell densities and very simple two-species composition are far from realistic ocean conditions. Still, this relationship fits into the growing paradigm of co-dependence between the streamlined *Prochlorococcus* genome missing basic functions and other members of the community, which *Prochlorococcus* photosynthesis helps to feed (Morris et al., 2011). We can add light shock to the growing collection of stressors that helper heterotrophs alleviate, along with growth at low cell densities, growth in environmentally relevant concentrations of hydrogen peroxide and growth on solid media (Morris et al. 2011, Morris et al. 2008). Even for a phototrophic-specific stress like light shock, *Prochlorococcus* may be reliant on other organisms for detoxification.

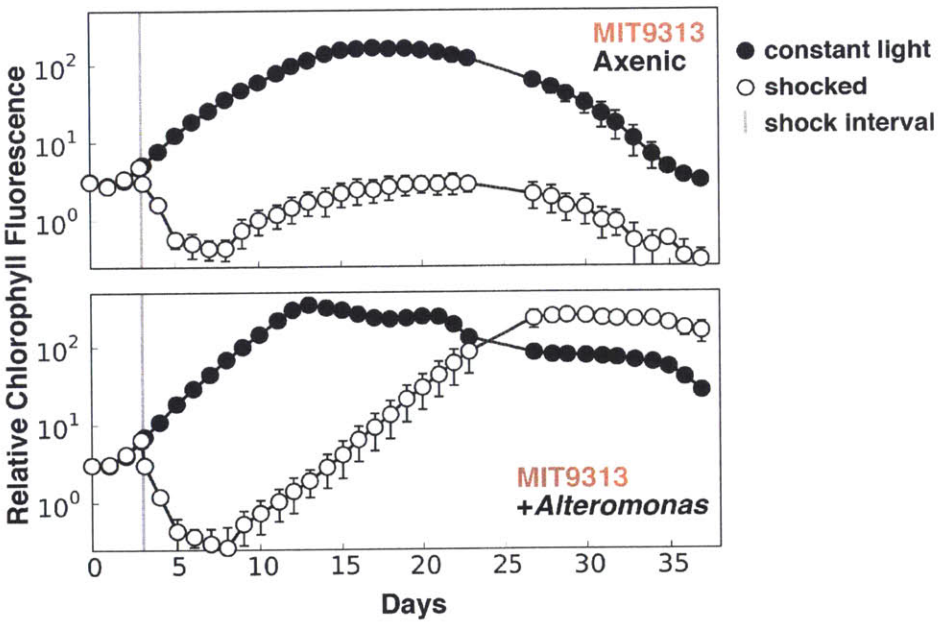


Figure 3.6. MIT9313ax + *Alteromonas* AIA co-culture Light shock response
The axenic control in this co-culture experiment represents an independent repetition of the MIT9313ax experiment described above, but followed longer in time. While later in this course there was a slight increase in fluorescence readings in the shocked axenic culture, it never reached even the culture's initial starting fluorescence value or regained green color.

3.3.2 Annotation and copy number variation of *hli* genes in *Prochlorococcus*, *Synechococcus* and cyanophage

Challenges and goals for annotation of *hli* genes

To approach our goal of studying the patterns and mechanisms of *hli* evolution in *Prochlorococcus*, we first need a set of good quality *hli* gene annotations. Some of the properties that make *hli* genes interesting also complicate their annotation; they are not annotated accurately by automated homology-based pipelines. To generate a consistent dataset for further analysis, we re-annotated *hli* genes across the recently expanded *Prochlorococcus*, marine *Synechococcus* and cyanophage genomic datasets. *hli* genes are short thus open reading frame (ORF) calling algorithms, that operate with size cutoffs to avoid spurious short ORFs, can miss *hli* genes (left as intergenic space). The *hli* gene family is diverse, sharing only a small conserved region while the rest of the protein is variable in amino acid composition and length, so protein identity cutoffs for homology can fail, depending on the reference set and details of parameter choices (Hess et al., 2001). Historically, this challenge has been addressed through focused searches for the conserved regions of the protein family. In the case of Bhaya et al, (2002), a single HMM based on all known *hli* genes, and thus sharing only conserved residues, was used along with the motif ExxNGxxAMxG (where x is any amino acid) summarizing those conserved residues. A later *Prochlorococcus* and phage-specific study used a slightly different motif to call *hli* genes, requiring at least six matches to the motif AExxNGRxAMIGF (Lindell et al. 2004). From research on plant CAB proteins, there is biochemical support for a small subset of these residues ExxH/NxR, as a minimal peptide pattern sufficient for chlorophyll-binding (Eggink and Hoober, 2000), but there has been no site-directed mutagenesis or other biochemical work directly investigating the importance of these motif *hlis* in any marine system.

While the motif-search approach has proved an effective strategy, it is possible that true homologs in this gene family might evolve away from the motifs, and borderline cases could be excluded; it's difficult to assess exactly what these motifs should be without information on the biochemical constraints of *hli* function. So, to create high quality, consistent and comprehensive annotations of *hli* genes for this study, we decided to use a set of multiple HMM based on previously annotated *hli* genes from *Prochlorococcus*, *Synechococcus* and their phage, treating different members of this protein family as individual proteins, while at the same time keeping an eye on motif patterns from previous methods. Starting with unannotated DNA sequences, we first applied two methods of gene calling to increase sensitivity (after finding that different gene calling methods find different sets of *hli* genes): [1] a best-practices sophisticated gene-caller and [2] a naive open reading frame search (start/stop codon pairs with ORFs longer than 30 amino acids), used cautiously to fill in a few genes. We applied a HMM search method, based on 16 different multiple sequence alignments each representing a different *hli* sequence cluster, built from previously annotated *Prochlorococcus*, marine *Synechococcus* and cyanophage *hli* genes, annotated by a variety of different motif and homology-based methods by different researchers over time. The different sequence clusters were chosen based on a rough protein phylogeny, with the aim of finding sets of proteins that can be meaningfully aligned over their full length, sharing some homology within each cluster beyond the chlorophyll binding motif. The motifs used previously (Bhaya et al., 2002, Lindell et al., 2004, Kettler et al., 2007) are still a strong part of these models because they are the most highly conserved parts of each multiple sequence alignment, but this sensitive and specific approach also uses information from the full length of the protein to search for homologous sequences.

We applied these *hli* annotation methods to the available genome sequence data from *Prochlorococcus*, as well as the closely related group *Synechococcus*, and the phage that infect each of these groups, as their *hli* gene histories are intertwined with those of their hosts (data summarized in Table 3.2, detailed in Supplemental Tables 3.1, 3.2 and 3.3 at the end of this chapter, see also Figure 3.1; Lindell et al.,

2004, Scanlan et al., 2009). This set of genomes enables us to resolve evolutionary events at several different scales, within and among *Prochlorococcus* ecotypes, covering a wide range of *Prochlorococcus* diversity. As this set of genomes is bound by what strains we have isolated into culture, there is patchy representation across *Prochlorococcus* groups compared to what we know exists in the wild. Some clades are better sampled than others, some clades have no cultured representatives; in some cases we have sets of nearly identical isolates, in other cases only a few divergent members represent an ecotype (Figure 3.1, Moore et al., 2007, Biller et al., 2015).

Table 3.2 Summary of genomic sequences used in this study

Data type	Number	Notes
Prochlorococcus culture genomes	49	Spanning the 5 cultured HL and LL ecotypes
Prochlorococcus single cell genomes new to this study	30	Primarily expanding coverage of LLI genomes, with some HLII as well
Marine *Synechococcus* genomes	24	Sub-clusters 5.1, 5.2, 5.3, including coastal, open ocean and brackish *Cyanobium*-group strains
Genomes of cyanophages infecting *Prochlorococcus* and *Synechococcus*	80	Myo-, podo-, siphoviridae families

Environmental sequence data lets us escape the limitations of culturing, without the benefits of functional follow-up. The LLI clade, the one with the highest number of *hli* genes (Kettler et al. 2007, Coleman and Chisholm 2007) and unusual intermediate light physiology (Scanlan et al., 2009, Malmstrom et al., 2010), was particularly under-sampled among cultured isolates and genomes (Kettler et al., 2007, Biller et al., 2014). As such, for this clade we decided to supplement our culture genomes with single-cell genomes, DNA amplified from individual cells from seawater samples, to maximize our sampling for phylogenetic resolution in hopes of resolving transfer and duplication events leading to the high numbers of *hlis*. Single-cell genomics enables only partial recovery of a genome, so it is not useful particularly in the study of genome-wide copy number variation, but even partial genomes can add to our understanding of the sequencing evolution of groups of proteins and the molecular events that contribute to change in this ecotype and gene family. These single cells were chosen for sequencing from existing libraries from three oceans to expand coverage of the LLI clade, particularly sampling deeper branching members (Figure 3.7 and Supplemental Table 3.1).

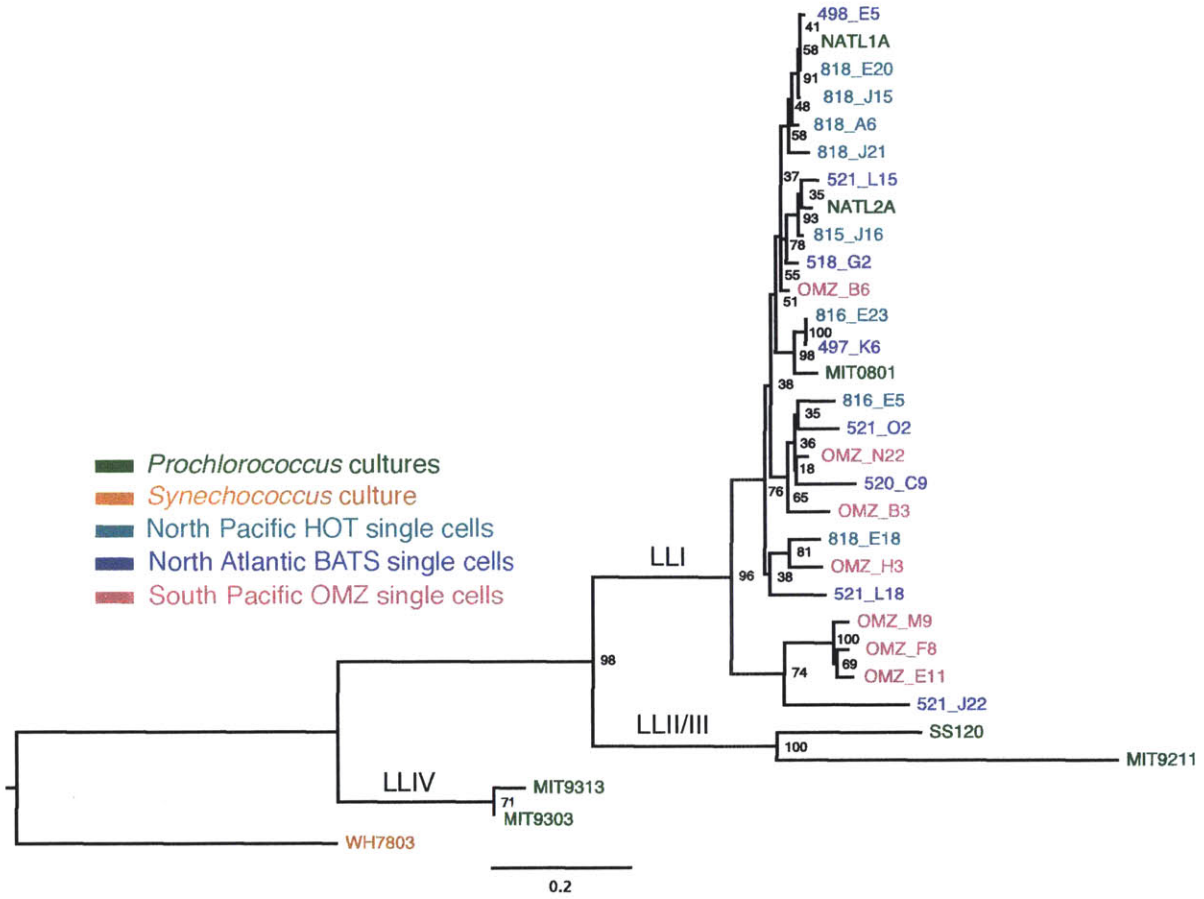


Figure 3.7. LLI single cell phylogeny: using wild cells to sample more deeply branching members of the LLI clade Note in the LLI clade (uppermost) the three cultures in green, relative to the rest of the LLI sequences, which represent single cell genomes from three separate oceans. These single cells allow us to access major subclasses not currently in culture, akin to sampling deeper back in time in our study of the *hli* evolutionary trajectory. ITSrRNA phylogeny (maximum likelihood, phyML, mafft, einsi alignment, trimmed to length of shortest sequence). Outgroups LLII/III, LLIV, *Synechococcus* WH7803, to give a sense of scale for the depth of diversity within the LLI clade observed in these wild single cell genomes.

The distribution of *hli* gene family members across *Prochlorococcus* and *Synechococcus* genomes

With more data and a sensitive, consistent annotation approach, we can re-assess how many *hlis* are in each *Prochlorococcus* genome and how *hli* copy number varies with ecotype. Our updated annotations confirm the major trends in *hli* copy number from smaller datasets that motivated this project (Bhaya et al., 2002, Coleman and Chisholm, 2007, Kettler et al., 2007, Kettler 2011), that the LLI strains have the most *hlis*, followed by the HL strains, and then the other LL strains (Figure 3.8). With this expanded and refined dataset a few interesting patterns emerge. The HL strains have consistently high numbers of *hlis*, compared to LLII/III and LLIV or other cyanobacteria, but they also exhibit a very large range within the ecotype (17 - 26 *hli* family members). One pattern made clearer with more genomes is that the LLII/II ecotypes (the second most deeply branching clade) consistently have more *hlis* than the LLIV ecotype (the most deeply branching clade). This suggests the possibility that the process of expansion of *hli* started earlier in *Prochlorococcus* evolution than the major expansions of the LLI/HL clade - the growth of the *hli* family in *Prochlorococcus* may have been a gradual process, ongoing in some lineages throughout the evolutionary

history of the genus. Recently added genomes (Chapter II of this thesis) include examples of the LLIV clade with 10 or 11 *hlis*, while older genomes contain 8 or 10; even in this clade with the minimum *Prochlorococcus hli* complement, *hlis* are variable. Marine *Synechococcus* for which genomes are currently available mostly have moderate numbers of *hlis*, closer to the low end observed in *Prochlorococcus*, but they also show considerable variation in *hli* counts from genome to genome, suggesting that in *hlis* are a dynamic part of the flexible genome and perhaps involved in niche adaptation in *Synechococcus* as well.

Within the LLI *Prochlorococcus* ecotype, which on average contains the most *hli* genes, this new data set revealed that not all LLI strains contain the record high numbers in the 40s. One of the new strains contains only 25 *hli* genes - similar to the numbers found in the HL strains. Thus 40+ *hlis* is not a LLI defining trait, but 25 is still on the high end of what we observe in HL strains, broadly consistent with the hypothesis that these genes may be involved in an ecotype wide trait related to light shock tolerance. One caveat about interpreting these copy numbers is that the 3 LLI genomes with 40, 43 and 43 (MIT0801, NATL1A, NATL2A) *hlis* are all fully closed genomes, while the one with 25 (PAC1) is in draft form, split over 20 contigs (Biller et al., 2015). While it is possible once closed this genome will contain more *hlis*. analysis of the core genes for the draft quality genome suggested that this assembly represents a full protein complement for a *Prochlorococcus* cell (and the same is true for all the draft genomes presented here) thus we suspect we are catching all the *hlis* (Biller et al., 2015).

The relationship between *hli* count and light physiology of each ecotype is consistent with the previously postulated hypothesis that these genes could contribute to high light adaptation, either with respect to light preferences for growth, or for the light shock survival phenotypes described above (Rocap et al., 2003, Bhaya et al., 2002, Coleman and Chisholm, 2007, Kettler et al., 2007), but the variation within ecotypes suggest they might also be doing more. On a finer scale, there is variation in *hli* copy number within all ecotypes and few pairs genomes have the same set of these genes, showing change in this gene family on multiple time scales (see Results 2.3.3). Although for any individual genome we might be observing some evolutionary noise, on the whole this pattern of variable *hli* family flexible genome content is consistent with the idea that some of these *hli* genes play a role in recent environment-specific adaptation (Kashtan et al., 2014, Kettler et al., 2007, Biller et al., 2015, Coleman et al., 2010, Coleman et al., 2006).

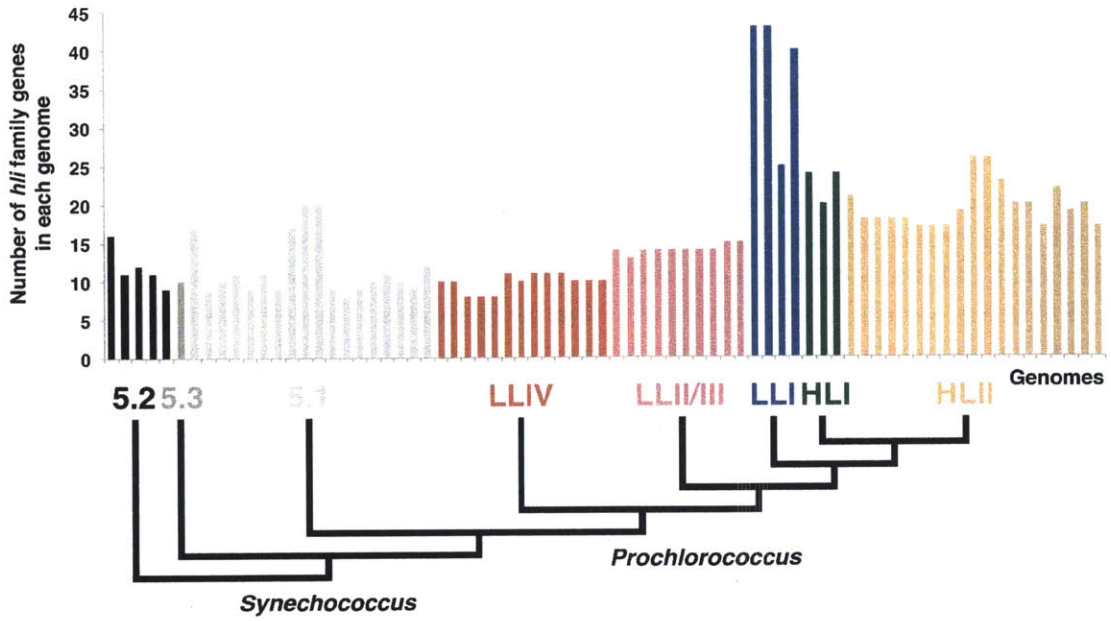


Figure 3.8. ***Hli*** **family copy number in genomes of cultured strains of marine** *Synechococcus* **and** *Prochlorococcus*
Each bar represents the number of *hli* genes in one genome, for 49 *Prochlorococcus* genomes available for this study and 29 marine *Synechococcus*. A schematic phylogeny is shown below the histogram for reference, and the *Prochlorococcus* clades are color coded as in Figure 3.1.

The distribution of *hli* genes across phages infecting *Prochlorococcus* and *Synechococcus*

Hli genes are frequently part of the suite of host-derived auxiliary metabolic genes carried by cyanophage, along with other genes involved in photosynthesis, core carbon metabolism and nutrient uptake genes (Lindell et al., 2004, Clokie and Mann, 2006, Kelly et al., 2013). We applied our new *hli* annotation approach to the genomes of a collection of phage that infect *Prochlorococcus* and marine *Synechococcus* (Fig. 3.9) to see what patterns might emerge. These phage examined span three families, the myoviridae, podoviridae and siphoviridae, and all are lytic. The cyanomyophage (T4-like, contractile tails) have relatively large genomes (~200kbp), and these generally carry the largest collections of host genes (Clokie et al., 2010, Kelly et al., 2013). Our hunt for *hlis* in these genomes were overall consistent with previous annotations, but in a few cases we were able to find a few more *hlis* with our more sensitive search approach. Every myovirus isolated on *Prochlorococcus* or *Synechococcus* has at least one and up to six *hli* genes. *hli* genes have become a vital part of their genome, fixed across all cyanomyoviruses sampled to date. The podoviridae have smaller genomes, ~50kbp, and generally carry only a few host genes per genome, but most carry one or two *hli* genes (Labrie et al., 2013). For the siphoviridae, our limited sample includes a few phage isolated on *Synechococcus* which carry host genes, including *hlis* in two cases, and phage isolated off *Prochlorococcus* that mostly have very small genomes and no host-like genes, representing a different infection strategy (Frois-Moniz, 2014).

In the myoviruses and podoviruses, we observed a striking relationship between the host of isolation at the genus level (*Prochlorococcus* or *Synechococcus*) and the number of *hli* genes in each phage genome. Myophage isolated on *Synechococcus* have one or two *hli* genes, with one exception (S-SSM4 has four *hlis*), while myophage isolated on *Prochlorococcus* have three to six *hli* genes. Podophage isolated on *Synechococcus* have zero or one *hli* genes; those isolated on *Prochlorococcus* have one or two (Figure 3.9). This is somewhat

unexpected in light of the fact that in the laboratory, many of these phage have complex patterns of cross infection between *Prochlorococcus* and *Synechococcus* (Sullivan et al., 2003); they are by no means strictly adapted to one host genus. For example, there are some phage that infect only a subset of *Prochlorococcus* ecotypes, and some that infect both *Prochlorococcus* and *Synechococcus* but not every member of each genus (Sullivan et al., 2003, Labrie et al., 2013). Further complicating this picture, host range is a trait that can change rapidly, under strong selection as host and phage co-evolve in a frantic arms race (Stoddard et al., 2007, Avrani et al., 2011, Avrani and Lindell, 2015). The pattern we observe suggests there could be some relationship between host genus and and phage *hli* complement, with *Prochlorococcus*-isolated phage using more *hli* genes, like *Prochlorococcus* does, which may be worth exploring in the future.

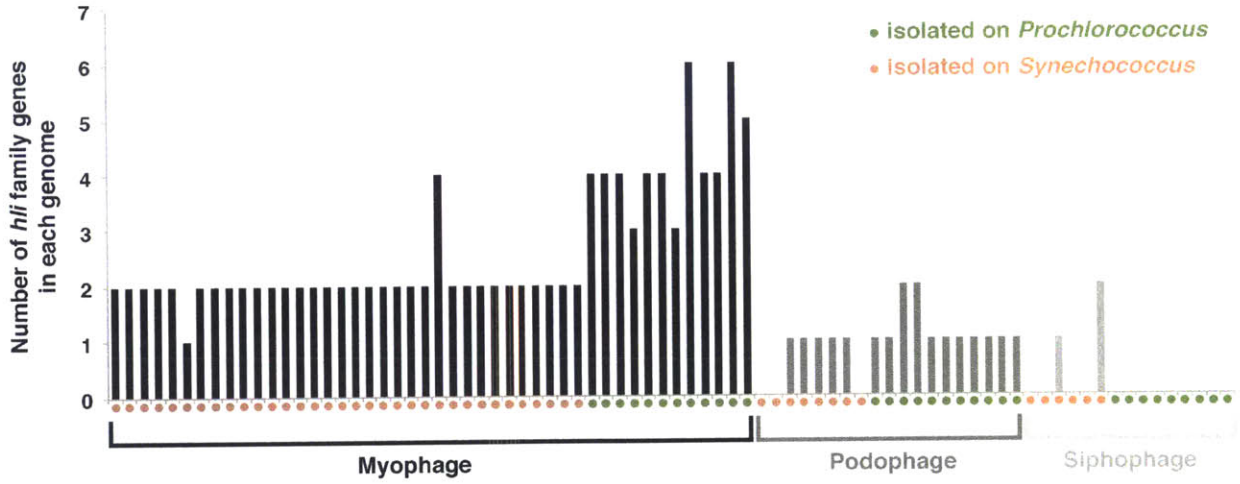


Figure 3.9. *Hli* family copy number by genome across phage isolated off marine *Synechococcus* and *Prochlorococcus*
The distribution of *hlis* across cyanophage, based on the targeted annotation methods in this work. This collection of genomes includes all the phage isolated off marine *Synechococcus* and *Prochlorococcus* with available genomic sequences. The sequences (genomes across the x-axis) are sorted first based on phage family, then by the genus of the host on which the phage was isolated, and then order is arbitrary.

Based on their wide distribution across phage, as part of a small subset of host genes that are fixed in phage genomes, it seems that *hlis* are an important part of the viral toolkit for productive infections in marine picocyanobacteria. As in hosts, phage *hli* genes include both conserved, core genes present in every genome within certain groups, and flexible genes, whose copy number varies depending on lineage, which may be a product of selection by the particular environmental conditions and hosts encountered for each phage lineage. What are the functional benefits for phage carrying *hli* genes? In the larger framework for the role of host-derived genes in phage, it has been suggested that the purpose of these host genes is to influence the cell's metabolism towards gathering the building blocks needed to produce phage particles, to maximize phage production rates, through the enhanced performance of the light reactions of photosynthesis, producing ATP and reducing power (Thompson et al. 2011b, Clokie and Mann 2006, Lindell et al., 2005), and through nutrient uptake – particularly phosphorus (Kelly et al., 2013, Zeng et al., 2012). While their precise functions are still open, the general role of *hli* genes in supporting photosynthesis, building and repairing photosystems, especially in times of stress, fits into this picture of phage encouraging cellular productive capacity. In one set of *Prochlorococcus* infections, NADPH/NADP ratios were found to be higher in infected cells than uninfected controls, suggesting that the cell is in a relatively reduced state during infection (Thompson et al., 2011b). Under high light conditions, electron

transport chains are fully reduced, and excess energy finds alternate pathways to flow, some to safety valves, and some to damage, which *hli* proteins help to alleviate; the phage, in mimicking the reduced state of a full electron transport chain in high light for the purposes of maximizing production may benefit from an expanded pool of *hli* proteins in a similar way.

3.3.3 The structure of the *hli* gene family in *Prochlorococcus*

Clustering *hli* genes: bringing order to a complex gene family

Treating these genes as simple counts belies the vast diversity within this protein family. What we really would like to know is how many types of proteins there are, and which types are in which genomes. How many deeply divergent proteins are there? How many recent paralogs and exact duplicates? How are *hli* gene variants, in sequence and structure, distributed across *Prochlorococcus*, *Synechococcus* and phage? What can observing this evolutionary process indicate about potential functional and adaptive properties of these genes? To resolve this picture we clustered the hli annotations to identify distinct orthologs within the family and and examined their distributions among host and phage.

Clustering is not straightforward in paralog-rich gene families. Early work (Bhaya et al 2002) analyzing the first two *Prochlorococcus* genomes explored their *hlis* only in relation to other cyanobacteria. More recently, Lindell et al (2004) developed the concept of dividing *Prochlorococcus hlis* into two groups, single copy, core, fresh-water-shared, and multi-copy phage-shared genes. After trying a few approaches, we settled on a UPGMA clustering based on pairwise alignments (see Methods), because aligning over the whole gene family proved unreliable (many gaps, differences in clustering sensitive to alignment parameter choice). We used a combination of host taxa distributions, similarity between proteins (which worked out to ~30% or higher within a cluster) and inspection of multiple alignments for N-terminal conservation to pick out groups of orthologs. This resulted in 19 clusters of genes (Figure 3.10). Five are shared between *Prochlorococcus* and *Synechococcus*, and the rest are specific to a genus. The 5 *Prochlorococcus-Synechococcus* shared clusters mostly correspond to the single copy-core/freshwater shared *hlis* previously reported by Lindell et al (2004). For one of the shared clusters, PS3, all proteins in the cluster have clear homology across the group, but only a few members of the cluster have the conserved *hli* motifs used for identification of the gene family in previous studies. They are included in these analyses as possible *hli* genes with this caveat, and further analysis is required to fully explore the relationship between motifs, biochemical data supporting roles for individual residues and protein level conservation in this gene family. A few unclassified genes were too divergent to fit into a cluster (long branches in Fig. 3.10), or formed a cluster of 5 or fewer genes; in some cases these are due to major mutations, frameshifts, or may represent false positives, but some are genuine *hlis* that are highly divergent, perhaps the products of HGT or recombination events (see Results 3.3.4).

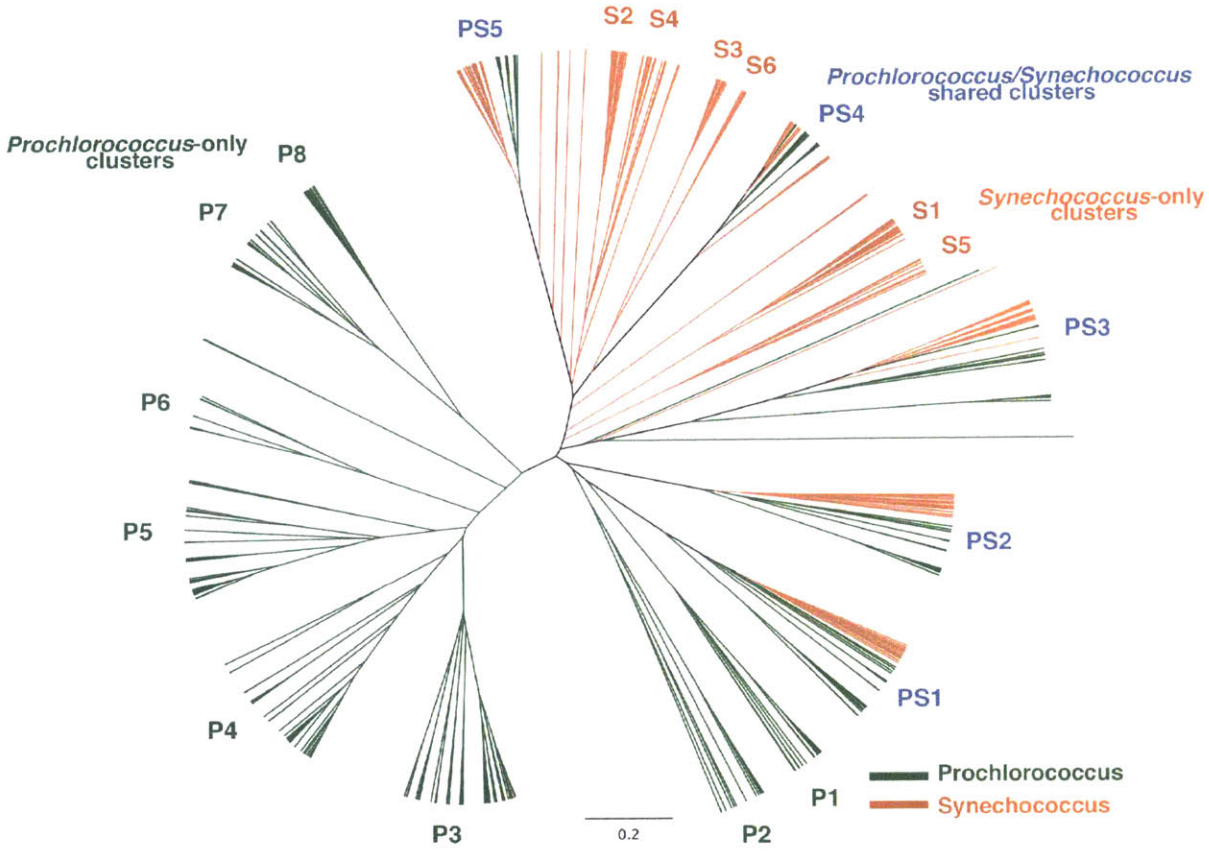


Figure 3.10. Clustering of *hli* protein sequences from *Prochlorococcus* and *Synechococcus* genomes
For (49) *Prochlorococcus* and (29) *Synechococcus* genomes from culture isolates (see Supplemental Tables 3.1, 3.2 for list), this is a UPGMA clustering based on global pairwise alignments between sequences. The nodes are colored based on taxa in each clade, which shows a distinct distribution of genes unique to *Prochlorococcus*, genes unique to *Synechococcus* and shared genes, which were assigned to discrete clusters (labels around edge of dendrogram), each representing one deeply divergent ortholog group, based on this data.

Distribution of *hli* clusters across *Prochlorococcus* genomes and ecotypes: implications for their evolution

Genes within *Prochlorococcus* genomes can be grouped into three broad categories: (1) genes shared with *Synechococcus* – all of which are single copy core genes; (2) core genes present in all *Prochlorococcus*, which can be present at variable copy number, and (3) flexible genes present in some but not all *Prochlorococcus*. Where do the *hli* clusters fit in this schema? Some *hli* genes are core, ancient, single copy and stable, and some are dynamic, changing in copy number (and presence/absence) between and within ecotypes (Figure 3.11). The cluster distribution shows us that the expansion of *hli* genes in the LLI ecotype – the strains that typically have 40+ *hli* genes – occurred through massive duplication of about five existing genes (P3, P5, P4, P7, P6 in Figure 3.11). Duplication events within shared clusters and one highly divergent new gene give rise to the set of genes in the HL ecotypes. LLI and HL have reached high copy number mostly through expansion in the same few clades. The LLII/III clade has most of deeply branching protein diversity present in all *Prochlorococcus*, just lower copy numbers, so the common ancestor with LLII/III may have been the source of major innovations (Figure 3.11). But even before that, the LLIV clade contains *hli* proteins absent from *Synechococcus*. Within clades, few genomes have exactly the same complement. Some genes are single copy genes in many *Prochlorococcus*, but are not in *Synechococcus* (P1, P2,

P8, Figure 3.11). Some *hlis* are present in all *Prochlorococcus*, but massive duplication occurred in certain lineages, some *hlis* were lost in some lineages, duplicated in others. HL and LLI share some history, but not all: shared and parallel paths to high copy number.

The conserved single copy core genes are more likely to be housekeeping genes, or genes needed in conditions encountered by all *Prochlorococcus*, to select for their continued presence (Zhang et al., 2003, Tettelin et al., 2005, Kettler et al., 2007, Biller et al., 2015). Genes that change and move over time, in the flexible genome, are more likely to have niche-specific roles, valuable under certain circumstances, and their presence is selected to different extents in different genomes and ecotypes (Tettelin et al., 2005). It is interesting that these *Prochlorococcus* flexible genome *hlis* show variation at both deeply branching ecotype scales - there are clear differences between the sets of *hlis* carried by each ecotype - and at the shallowest scales - it is rare in this data set to find two genomes (except for sets of nearly identical strains) that have quite the same set of genes.

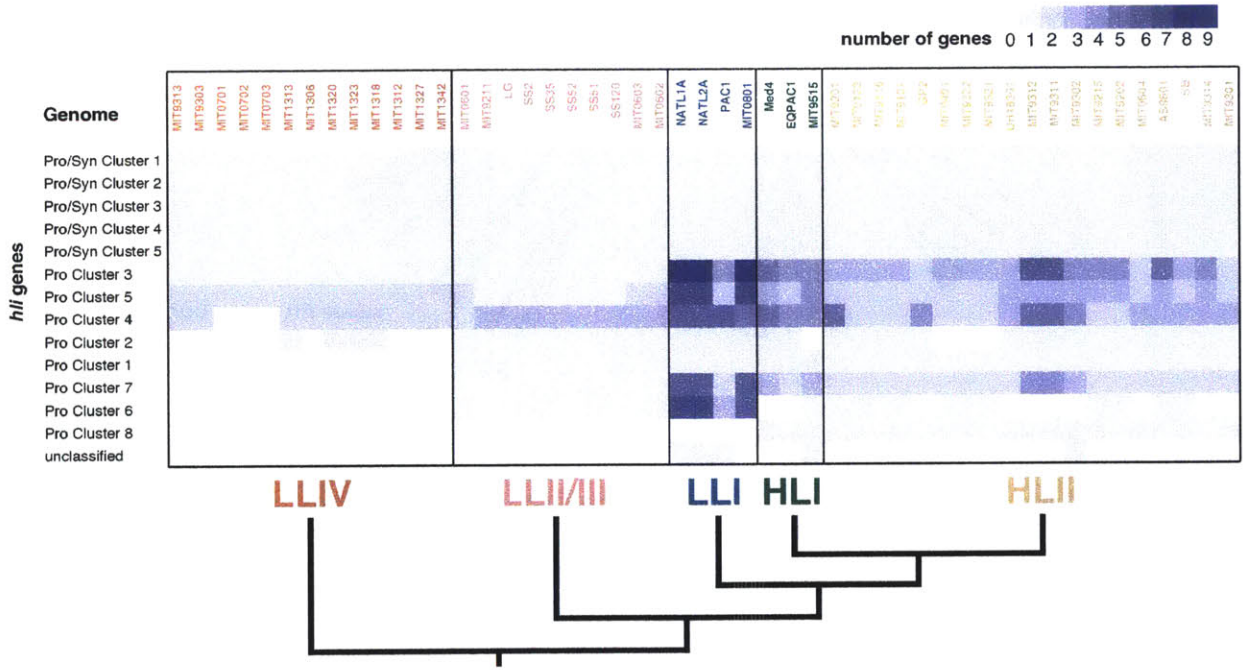


Figure 3.11. Distribution of *hli* gene clusters across *Prochlorococcus* ecotypes
Each column represents all the *hlis* genes in one genome, and each row represents one gene cluster (as defined in Figure 3.10). Color of box indicates copy number - darker colors represent higher numbers of genes in that genome assigned to that cluster. Genomes are sorted by ecotype, Clusters are sorted first by whether they are *Prochlorococcus*/*Synechococcus* shared clusters (Pro/Syn cluster) or specific to *Prochlorococcus* (Pro cluster) then by breadth of representation across genomes.

Distribution of *hli* gene clusters across *Synechococcus*

How do the *hli* complements in marine *Synechococcus* compare to *Prochlorococcus*? This is of interest both for understanding the ancestral shared state and subsequent trajectories, and also in the context of *Synechococcus* ecology, which spans a huge range of phylogenetic diversity and environmental conditions, in which *hli* variability could also possibly play an adaptive role. Previously we had noticed that LLIV *Prochlorococcus* and *Synechococcus* have similar numbers of *hlis*, but a closer examination of these sequences shows that they are not the same genes (Figure 3.12). How are *hli* variants distributed across *Synechococcus*

diversity? Four of the five *Prochlorococcus-Synechococcus* shared gene clusters are also single copy core genes in *Synechococcus*. The fifth one (cluster PS5) is variable in copy number in some *Synechococcus*; all genomes have at least one and there are up to five copies in some genomes. There are an additional three clusters which are core genes in *Synechococcus*, but absent from all *Prochlorococcus*, which are single copy except for the deeply divergent WH5701. Finally, there are five flexible *hli* genes with patchy presence/absence across genomes, and to a lesser extent than we saw in *Prochlorococcus*, copy number variation.

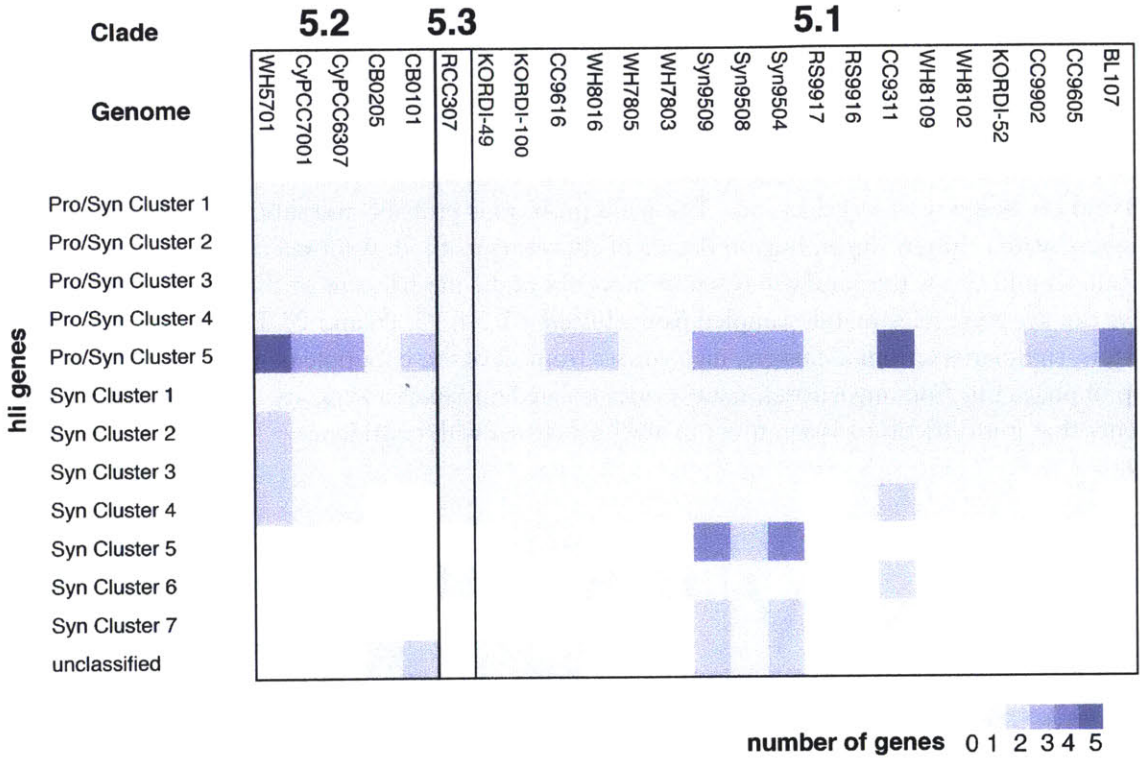


Figure 3.12. Distribution of hli gene clusters across *Synechococcus* genomes
Each column represents all the *hlis* genes in one genome, and each row represents one gene cluster (as defined in Figure 3.10). Color of box indicates copy number - darker colors represent higher numbers of genes in that genome assigned to that cluster. Genomes are sorted by ecotype, Clusters are sorted first by whether they are *Prochlorococcus/ Synechococcus* shared clusters (Pro/Syn cluster) or specific to *Synechococcus* (Syn cluster) then by breadth of representation across genomes.

The remarkable feature of these flexible *hli* genes in *Synechococcus* is that they are not the same genes as those in the *Prochlorococcus* flexible genome; they are an entirely different set. So, it looks as though in *Synechococcus* the same type of *hli* evolutionary dynamics are happening, with conserved cyanobacterial and genus-specific genes, and variable strain-specific genes, but with a different pool of protein variants. Thus the overall variation in this gene family is driven less by copy number variation of closely related sequences, and more by patchy distributions of deeply divergent proteins. The *Synechococcus* strains with unusually high numbers of *hli* genes are MITS9504, MITS9508, MITS9509, strains isolated from the equatorial Pacific in a single isolation effort, from the CRD1 clade associated with tropical and subtropical upwelling (Ahlgren and Rocap, 2012), and CC9311, a coastal strain characterized by large genomic repertoire of genes for sensing and responding to its environment, of which these genes are a part (Palenik et al., 2006). *Synechococcus* is a large, complex genus, inhabiting more different kinds of environments than *Prochlorococcus*

(including coastal and polar regions), without the relatively straightforward genome-ecology ecotype structuring of *Prochlorococcus* (Dufresne et al., 2008, Ahlgren and Rocap, 2012). Rather it has a complex biogeography made up of many, many clades with different distributions, and for now, limited genomic representation (Scanlan et al., 2009, Ahlgren and Rocap, 2012). We can imagine that the *hli* evolutionary dynamics observed here contribute to *Synechococcus* functional variations and adaptation to diverse light, mixing and other conditions across their diverse habitats.

What is the origin of phage *hli* genes?

Phage get their *hlis* from a subset of the *Prochlorococcus*–specific clusters (Figure 3.13, Supplementary Figure 3.1). This is true except for one deeper branching clade, where we can not tell exactly which host cluster where it came from - it has similar distances from a few. Work on phage clusters is still underway, because they did not fit neatly into the clusters established above; the addition of large amounts of environmental sequence data from single cells and fosmids made 2 of the clusters unstable and we're still working to find the best way to sort these out. The main problem is probably recombination events that make sequences switch clusters depending on details of clustering method. A formal consideration of recombination should clarify this, and will teach us more about the mechanisms of change in *hli* genes. However, we can say that phage mainly sampled from clusters P3, P4, P5, P6 and P7. Podo viruses sample from 2 clusters, siphoviruses from 2 clusters, myo viruses from all of these (Supplementary Figure 3.1). One small group of phage hlis from myoviruses, mainly ones isolated on *Synechococcus*, are different enough from any host genes that it is difficult to assign them to any host cluster with confidence.

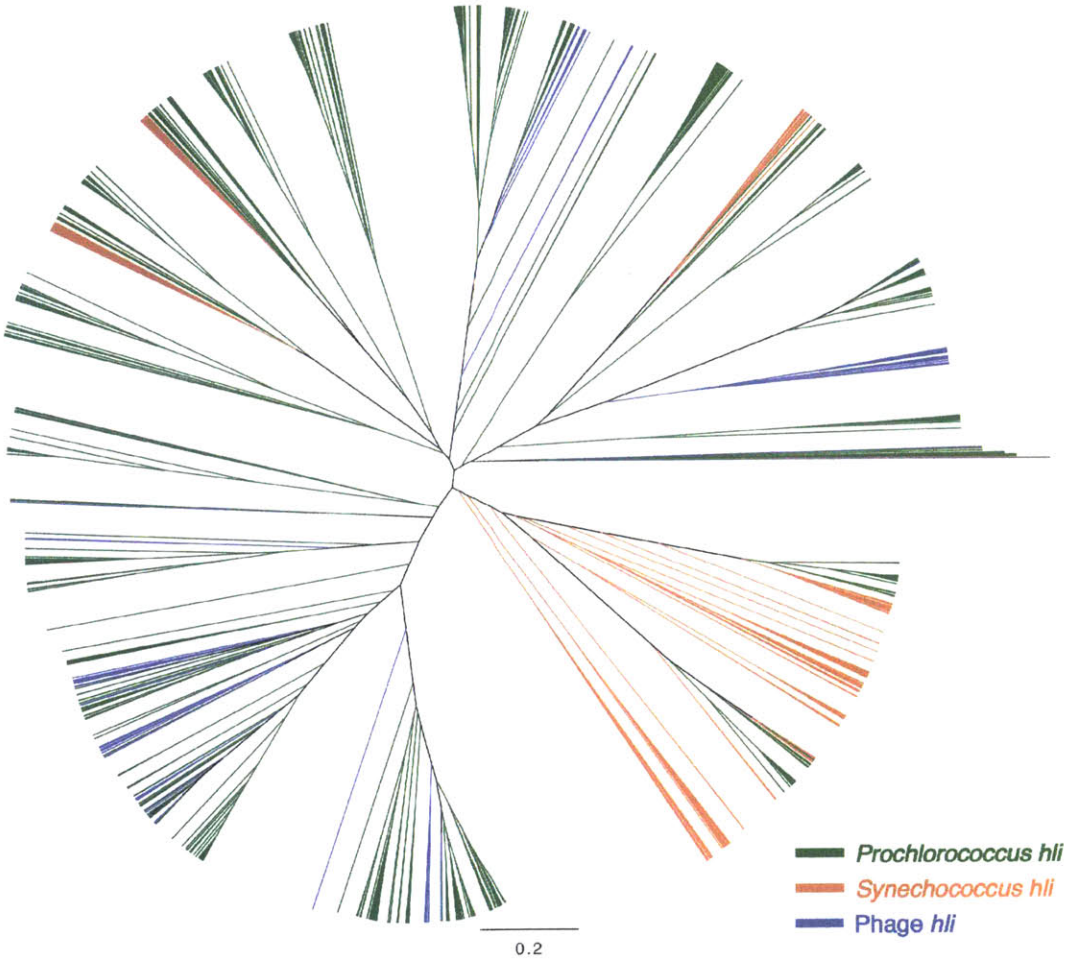


Figure 3.13. Phage *hli* clustering, with *Prochlorococcus* and *Synechococcus*
To answer where phage are getting their *hlis*, we clustered all host and phage genes using the rough pairwise alignment-based UPGMA method (as above, Materials and Methods). Whether the phage was isolated on *Prochlorococcus* or *Synechococcus*, phage *hli* genes (in blue) cluster with *Prochlorococcus* genes (green), never with *Synechococcus* genes (red) or *Prochlorococcus/Synechococcus* shared clusters. Of interest, there is one cluster of phage *hli* genes which is highly divergent from the nearest *Prochlorococcus* gene (right side of clustergram), which does not fit well into any host gene cluster (see Supplemental Figure 3.1). There is a small caveat in that this analysis is incomplete (we have not assigned discrete clusters), because recombination events in some phage *hlis* make them difficult to force into discrete ortholog clusters with host genes. We have not yet completed the formal recombination analyses which will be incorporated into this cluster framework to clarify this issue, and give recombinants their own clusters.

What does the distribution of *hli* clusters among strains mean for roles of genes, functionally and in defining groups?

The distribution of *hli* genes across *Prochlorococcus* indicates that their evolution is intricately connected to the evolution of *Prochlorococcus*, on several time scales. As outlined in Figure 3.14, some of these genes are ancient, conserved genes that are shared between all marine *Synechococcus* and *Prochlorococcus*, and some of these are probably shared by all *cyanobacteria*. Their gene products are billions of years old, likely functioning as basic tools – core to the assembly and maintenance of the photosynthetic machinery, the safe manipulation of chlorophyll or used in times of frequently encountered light stress.

Then, later in the evolution history, *Prochlorococcus*-specific, part of the set of genes that differentiate *Prochlorococcus* from *Synechococcus* - each have their own set of *hli* genes which are core in the genus, ultimately derived from older genes, but the precise history is not clear, but they diverged in protein sequences; these are part of the toolkit that differentiates these two groups, which have very different photosystems - perhaps these genes interact with the very different antennae, the most dramatic difference between thew two genus, or they could related to the overlapping but different environmental distributions - *Prochlorococcus* is more abundant deeper in the water column and in more oligotrophic regions, *Synechococcus* has a much larger geographic range, thriving in coastal and high latitude environments where Pro is absent. These *hli* genes are part of what makes *Prochlorococcus* distinct from other cyanobacteria. Then, a larger part of this genus differentiating gene pool is the flexible genes, not conserved. There are still more genes, which are not core in *Prochlorococcus*, but different pools or flexible genes are circulating in the variable genomes of *Prochlorococcus* and *Synechococcus*. Most of the deep branching diversity is in fact present in all but the deepest branching LLIV clade - here these proteins being to separate ecotype. But most of the ecotype type differentiating power of *hli* genes isn't at the level of deeply different protein evolution, but in copy number variation - LLII/III, LLI and HL have different numbers of the same basic protein types, so in these cases the number of *hli* genes is an ecotype-defining gene set. Finally, *hli* genes are involved in within ecotype variation, at the most recent scale of Prochlorococcus evolution. Particularly in the HLII clade, which is the most abundant in the oceans and the one we have the most genomes from, *hli* copy number varies widely, from 17 to 26; all relatively high compare to other *Prochlorococcus*, but leaving large opportunity for functional variation within the ecotype.

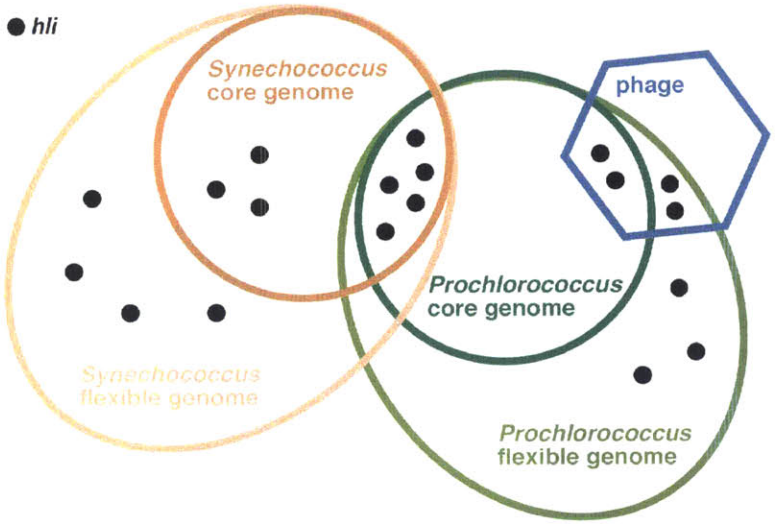


Figure 3.14: Schematic distribution of *hlis* in picocyanobacterial genome space
A qualitative Venn diagram, to illustrate the many categories that *hlis* fit into over marine picocyanobacterial evolution: part of *Prochlorococcus* core genome, *Synechococcus* core genome, and the shared *Prochlorococcus/Synechococcus* core - part of defining set of genes for all cyanobacteria, and each genus here, and part of the flexible genome differentiating lineages within each group. Each dot represents a cluster, which isn't perfect, because in the *Prochlorococcus* core genes - shared among all *Prochlorococcus* - are also part of the flexible genome, in multiple copies. The phage *hlis* are mostly derived from the *Prochlorococcus* flexible set.

3.3.4 Arrangement and rearrangement of *hlis* across the *Prochlorococcus* genome

Where are the many *hli* genes of *Prochlorococcus* located in the genome?

The genomic context for a gene can provide insight into its history (is it stable or changing?), mechanisms of movement (are mobile elements involved?), and operon structure. For some genomic islands, genes with related functions can travel in cassettes, horizontally transferred not as single genes, but small collections of genes that confer a selective advantage to a cell under the same environmental conditions. Thus we investigated the locations of *hlis* in the genome, in hopes of gaining further insight into their histories of duplication and horizontal transfer. The conserved *Prochlorococcus-Synechococcus* shared *hli* genes are found singly, scattered throughout the genome, in largely stable genomic contexts, like most other core genes (Figure 3.15, and in Kettler, 2011). A small number of the *Prochlorococcus*-specific *hli* genes, are also found singly (Figure 3.15, and in Kettler, 2011). The majority of the *Prochlorococcus*-specific multicopy *hli* genes and are located in genomic islands, like other flexible genome content in *Prochlorococcus*, and their surrounding gene context changes from genome to genome, between and within ecotypes (Coleman et al., 2006, Kettler, 2011). These genes are not found alone, but in tandem arrays of several head-to-tail *hli* genes, which are likely operon structures, as first reported in Bhaya et al. 2002. This arrangement has remarkable implications both for the function of these *hlis*, suggesting they could act in units of sets of genes working together rather than individual proteins, and for their evolution, because this repeat structure becomes susceptible to complex rearrangement through homologous recombination.

Among cyanobacteria, this tandem array structure of *hlis* has been found so far only in *Prochlorococcus* (Bhaya et al. 2002). This provokes the question: when in the evolution of *Prochlorococcus* did this feature of tandem arrays of *hli* genes appear? We first examined the arrangement of *hlis* in marine *Synechococcus*, and we found occasional instances of pairs of *hlis* near each other, separated by several thousand base pairs, but not in tandem arrays, even for the strains with high *hli* copy numbers (Figure 3.16). A tandem array appears in LLIV *Prochlorococcus*, which has a single tandem array of four *hli* genes (Figure 3.16). In the LLII/III *Prochlorococcus* genomes, the next most deeply branching clade of the *Prochlorococcus* phylogeny, there are two or three tandem array structures scattered throughout the genome, each with 2, 3 or 4 *hlis* (Figure 3.16). The parsimonious explanation for this distribution is that the first *hli* array appeared in the ancestor of all *Prochlorococcus* after divergence from the *Synechococcus* group, and following that physical gathering of genes, arrays began to move around and undergo remodeling. In the LLI *Prochlorococcus*, there are many of these arrays, seven to nine per genome, with 2, 4 or 5 members each (Figure 3.16). This reveals another part of the mechanism by which LLI came to have so many *hli* genes; the copy number of *hlis* in these genomes expanded to 40 or more not through duplications or horizontal transfers of individual *hli* genes, but in units of whole arrays, four or five at a time, requiring only a few molecular events. Among the three closed LLI genomes that we have, the MIT0801 genome differs from the NATL1A and NATL2A genomes in their *hli* complements by missing an entire array (Figure 3.16); that is, sets of *hlis* come and go through individual molecular gain or loss events. The HL *Prochlorococcus* share this pattern, but also look different: they have several arrays of four genes, but many more of two or three genes, and more *Prochlorococcus*-specific *hlis* that are not in arrays but appear singly in the genome (Figure 3.16). These diverse arrangements of *hlis* on the genome further support the idea of independent trajectories of the *hli* gene family expansion in the LLI and HL lineages; their *hlis* are not arranged in the same structures.

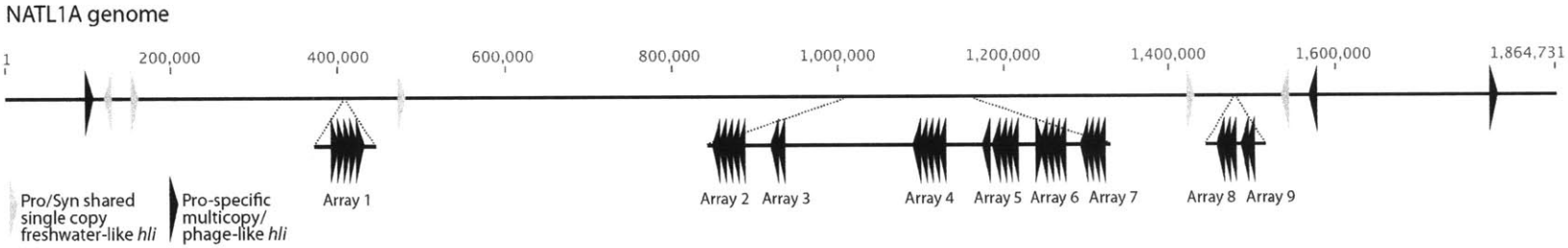


Figure 3.15. Arrangement of *hli* genes in the NATL1A LLI *Prochlorococcus* genome
The many *hlis* of NATL1A (the *Prochlorococcus* with the most *hlis*), are scattered throughout the genome. Each triangle represents one *hli* gene, at its position along the full NATL1A genome (base-pair positions above). Inserts are shown where there are too many *hlis* too close together to be represented on the main scale. The Pro/Syn shared single copy freshwater-like *hli* genes all appear singly in the genome. The Pro-specific multicopy/phage-shared *hli* genes occasionally appear singly, but usually appear in tandem arrays of head-to-tail *hli-hli-hli*. These arrays are numbered from start to finish on the genome, and referenced in Figure 3.18. Most of these arrays (2 through 7) occur in one region, a large genomic island conserved across the LLI genome architecture (the big island), which is rich in flexible genome content and highly variable between genomes (Kettler et al., 2007, Kettler, 2011).

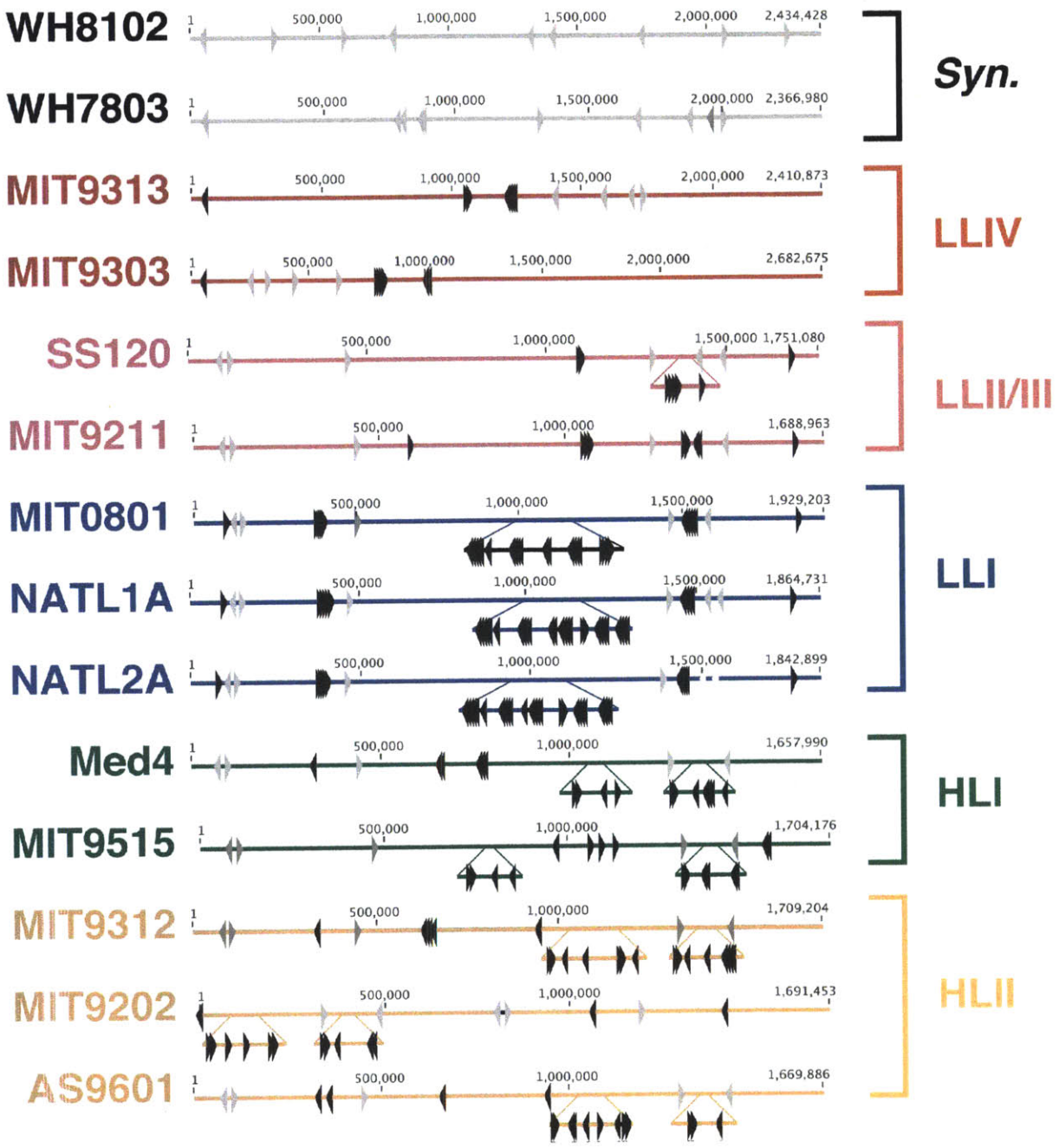


Figure 3.16. Arrangement of hli genes for *Prochlorococcus* genomes
As in Figure 3.15, all the *hlis* in the locations in the genome, showing arrays, for most of the available closed genomes. Each triangle represents an hli gene. For *Prochlorococcus* genomes, the *Prochlorococcus*-specific *hlis* are in black, the *Prochlorococcus/Synechococcus* shared hlis are in grey. For the two *Synechococcus* genomes at the top, all hlis are grey for simplicity, although they include both *Synechococcus*-specific genes and *Prochlorococcus/Synechococcus* shared genes. Where hlis are crowded, zoomed insets below main genome plot show arrangement.

How are sequence variants arranged in arrays?

Our next question, looking more deeply into these arrangements, was how are *hli* sequence variants (i.e. the orthlog clusters, the sets of identical repeated genes) distributed across array structures? It has been shown that some *Prochlorococcus hli* arrays can be composed of diverse *hli* variants (Bhaya et al. 2002 and Kettler 2011). Figure 3.17 shows all the *hlis* of NATL1A, a member of the LLI ecotype and currently the record holder with the most *hli* genes, at 43. The *hlis* are clustered by protein similarity, a visualization that shows at a glance the history of expansion. This collection of genes contains about 15 basic protein types, and six genes which have recently multiplied to many copies each, often with identical protein sequence. Figure 3.18 shows these same NATL1A genes in their array contexts. Most of the genes in arrays are the multicopy, but a few more distant variants are also present. The first striking feature of this collection of arrays from a single genome is that some arrays are nearly identical. They contain copies of the same proteins in the same order, although not necessarily with perfect DNA repeats. This supports the idea that the *hli* gene family has expanded through duplication or horizontal transfer of whole arrays. The arrays that are not the same contain different sets of genes in a combinational fashion. The combinations are not random; they have constraints. For example, arrays start with either genes from clusters P4 or P5, every array longer than two contains a gene from the cluster P6, and every array contains either a gene from P3A or P3B or both (Figure 3.18). We hypothesize that each array is some kind of functional unit, with several *hli* proteins acting together, but that the units are somewhat interchangeable and redundant, so the precise protein composition is not under strong conservative selection.

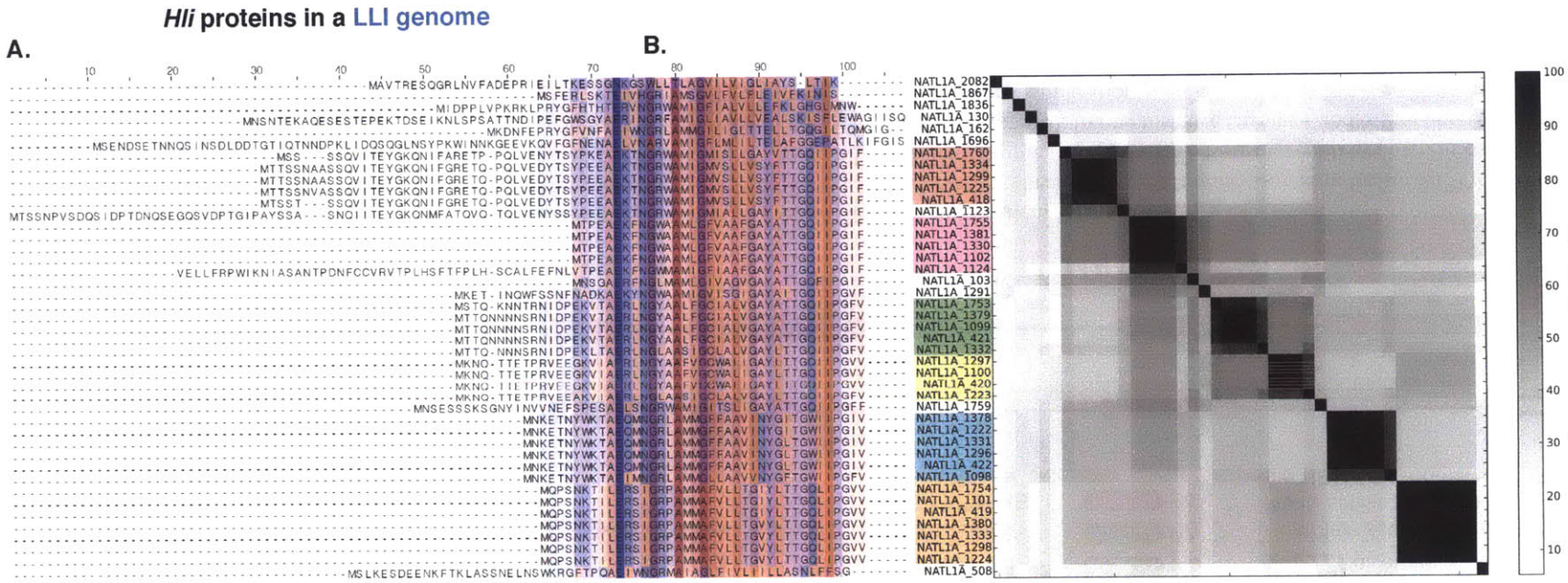


Figure 3.17. The 43 *hlis* of NATL1A, a LLI strain

This two part figure highlights the similarity, diversity and structure of the *hlis* in the NATLA1 genome.

(**A**) A protein multiple sequence alignment, with sequences ordered such that, in accordance with the clustering scheme in (B). Residues are colored by hydrophobicity (red = hydrophobic, blue = hydrophilic), and tinted by conservation across the alignment (saturated color = high conservation, no color = conservation <15%). Several of the highly conserved residues correspond to chlorophyll binding residues, as established in plant proteins. The transmembrane domain is visible as a hydrophobic patch, and the *Prochlorococcus*-specific C-terminal motif TGQIIPGF/IF is visible for most sequences. Note many of these multicopy genes are identical at the protein level, while others show some variation. (**B**) Protein identity matrix based on multiple alignment in (A), in which each row corresponds to values for adjacent sequence in alignment at left, clustered according to a modified UPGMA method (mafft). This is a symmetrical matrix, such that the order of labeled sequences from top to bottom is reiterated from left to right across the top. The clustering defines groups of nearly identical proteins, which are colored at the taxa labels to the left. These correspond to the *Prochlorococcus*-specific gene clusters in Figure 3.10, from top to bottom: red = cluster P5, magenta = cluster P4, green = P3A, a subset of cluster P3, yellow = P3B, a subset of cluster P3, teal = cluster P7, orange = cluster P6. Cluster P3 was defined on a useful scale for our *Prochlorococcus*-wide analysis (Figure 3.10), but for this NATL1A-specific analysis, we chose to split cluster P3, because, based on their N-termini, P3A and P3B represent two distinct proteins, which may be of significance in their function and evolution.

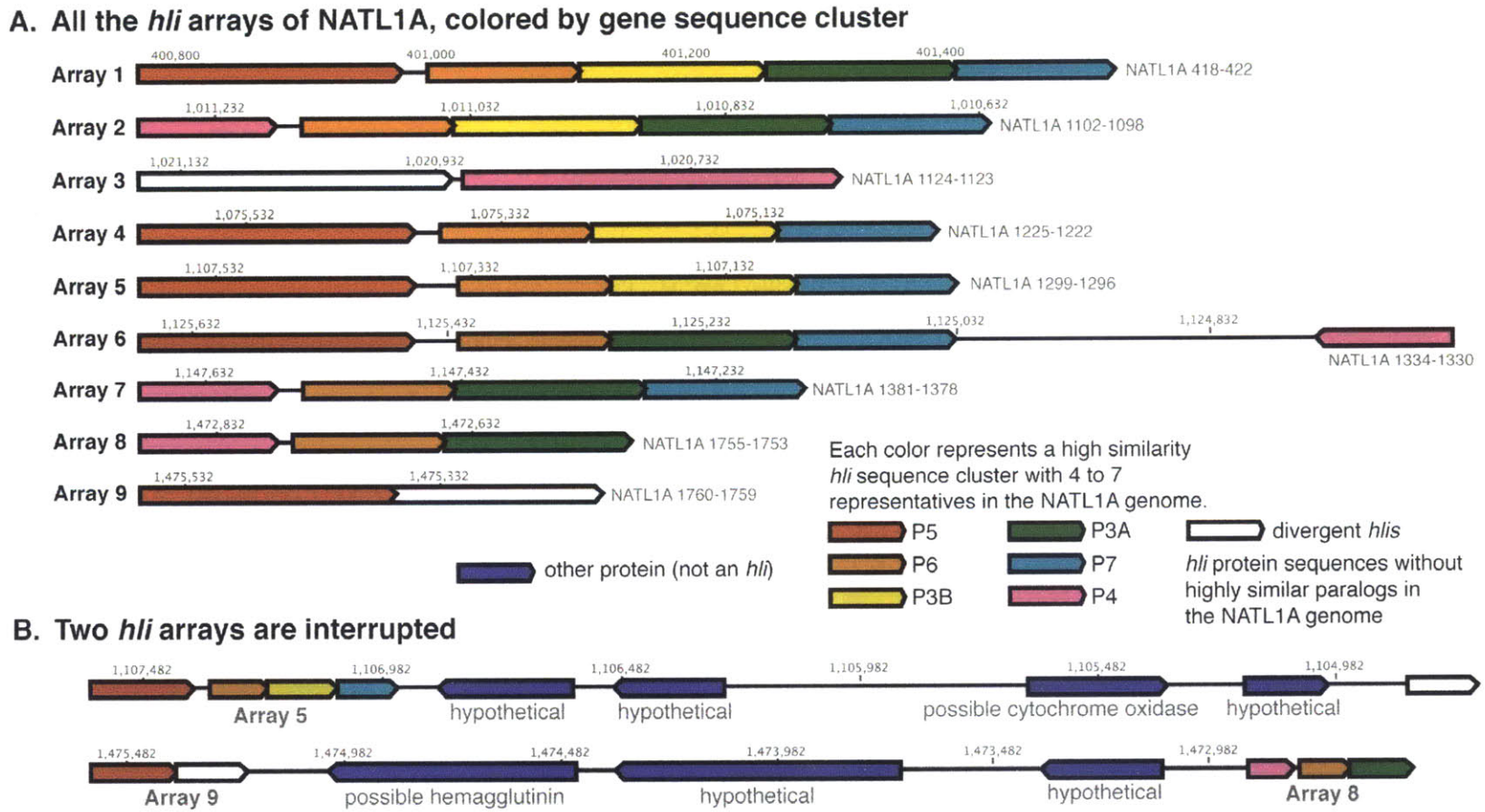


Figure 3.18. The *hli* arrays of the NATL1A genome and the structure of protein diversity within them

(A) These genome segments marked with genes represent the *hli* arrays of the NATL1A genome (numbered in their genomic context in Figure 3.15). Each *hli* gene in each array is colored according to its gene ortholog cluster as defined in Figure 3.17 (also see legend in this figure). These arrays contain different combinations of *hli* genes, but have some repeating patterns (e.g. arrays often start with P5/red or P4/magenta proteins). In one instance (Arrays 4 and 5), we observe the same array structure twice, indicating that the *hli* complement of the genome can grow through whole array duplication events. In another instance (Arrays 7 and 8), we observe very similar arrays that differ by the loss or gain of a gene, suggesting that arrays can also change through expansion and contraction. On the whole, arrays explore a combinatorial but non-random space in protein variants and gene order. **(B)** In two instances, *hli* arrays in NATL1A are interrupted by other genes (*hli-hli*-other gene-*hli*). By one parsimonious interpretation, these could represent events in which several genes have been inserted into *hli* arrays, perhaps through homologous recombination between *hli* genes from different sources. If this is the case, we can start to think about *hlis* as short repeated DNA sequences found in many places in *Prochlorococcus* and cyanophage genomes, which is a relatively rare molecular phenomenon in this system largely devoid of repeated DNA (except tRNAs; Coleman et al., 2006, Kelly et al., 2013). So, we hypothesize that *hlis* themselves may function as HGT facilitators. It is of interest that in both cases, *hli* genes flanking the putative insertions do not fall into the usual array *hli* sequence clusters, which may be a result of recombination within those genes.

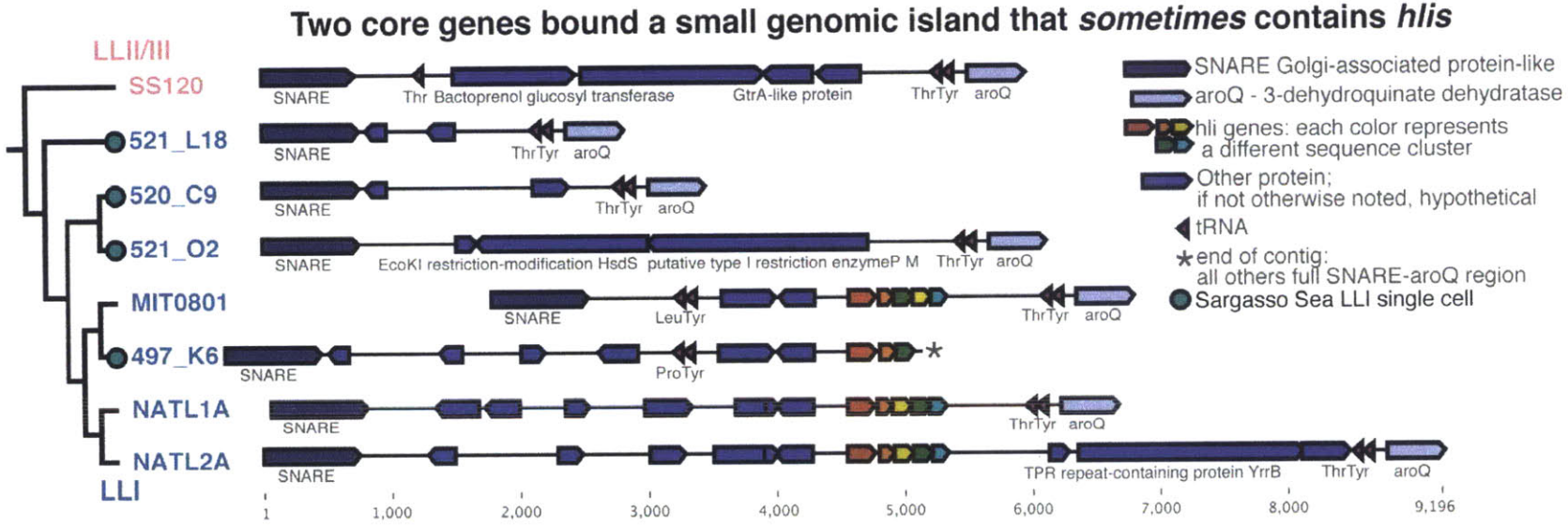


Figure 3.19. Sampling a small, dynamic genomic island across single cells captures the arrival of an *hli* array

These are genome segments bounded by a pair of core genes, on the left the SNARE protein and on the right aroQ, which is adjacent to a pair of tRNAs, notorious hotspots for horizontal gene transfer (Juhas et al., 2008). Between these core genes, is a small genomic island, visible by the sheer variety of protein number and arrangement in this region. In some LLI genomes this island contains an *hli* array, in NATL1A this is Array 1 from Figures 3.15 and 3.18. Most of the genes in this island are marked above in blue; most are annotated as hypothetical proteins, and the few with putative annotations are noted. The hli arrays are visible as multicolor cassettes; each *hli* gene is colored according to its gene cluster from 3.17. Four of these sequences come from LLI single cell genomes of the Sargasso Sea (521_L18, 520_C9, 521_O2, 497_K6), chosen to sample more deeply into the LLI clade than our limited, closely-related cultured examples (MIT0801, NATL1A, NATL2A). The phylogenetic relationships between these genomes are shown in the schematic tree at left, based on the ITSrRNA marker (Figure 3.7). By sampling deeper branching genomes within the LLI clade, and analyzing what is shared and different among genomes, we can make inferences about ancestral states and the timing of evolutionary events. In this case, we observe one derived clade that contains an *hli* array (MIT0801, 497_K6, NATL1A, NATL2A, the bottom four sequences), and several more deeply branching examples of this island without *hlis*, indicating an insertion event of this array (by parsimony analysis). This shows that Array 1 is not an ecotype-wide gene cassette. In single cell 521_O2 and the outgroup SS120, we observe interesting other examples of *Prochlorococcus* flexible genome genes, a restriction modification system involved in the defense against foreign DNA, and a cell surface glucosylation gene cassette, respectively. Both of these play a role in the rapid phage evolutionary arms race (e.g., Avrani et al, 2011, Kashtan et al., 2014). In this context, we can think of *hlis* as cutting edge evolutionary tools within the LLI clade, on par with the dynamism of host-phage coevolution. Single cell sequencing allows only partial recovery of genomes, so we are showing the subset of our samples for which we recover this particular island region.

This also has implications for the molecular mechanisms of *hli* evolution. Some tandem arrays expand and contract through internal recombination, making identical head to tail pairs. This does not seem to be happening in *hli* arrays. Rather, recombination within or between arrays is shuffling diverse proteins into new collections. Although the *hli* genes are divergent, there are still some homologous conserved DNA regions, in the C terminal region of the genes (see Figure 3.17), that could enable recombination between different *hli* gene clusters. Or, these recombination events could occur between nearly identical proteins in different arrays. More work is required to perform targeted analysis of recombination, which may explain the history of each array in more detail, and may explain some of the divergent genes that do not fit well into any clusters. In any case, each recombination event could allow sampling from a large pool of genes, into different functional units.

In two instances, the *hli* arrays of NATL1A have been interrupted by other genes: we see a few *hli* genes, followed by a few other genes, then more *hli* genes (Figure 3.18). This suggests that tandem arrays of *hlis*, as some of the few repeats in the lean *Prochlorococcus* genome, could function as sites for recombination or horizontal gene transfer of other genes. With each recombination between *hli* genes from two different pieces of DNA, other genes might tag along. In this case the genes are mostly hypothetical (typical of flexible genome in nonmodel organisms), but one is a cytochrome oxidase (in Array 5 in Figure 3.18); some cytochrome oxidases function to reduce O_2 as an energy valve during excess or unbalanced electron flow, like high-light conditions. This is a weak annotation, but it may be a hint that *hlis* could travel in units with other light shock or redox stress genes. The second instance contains a possible hemagglutinin annotation (in Array 9 in Figure 3.18), which is a phage protein relatively commonly found in *Prochlorococcus* genomes. This may indicate a potential historical connection or recombination between phage and host *hli* genes. The *hli* genes on the edges of these array interruptions might be so divergent (not fitting in existing clusters), because the recombination events generated altered proteins. These arrays are changeable units, shifting in *hli* content and receiving other genes, perhaps through homologous recombination between the many *hli* genes scattered throughout this system.

Consequences of tandem array structures for expression and translation

In bacterial genomes, genes adjacent on the chromosome can form operons: sets of genes in sequence transcribed as a single polycistronic mRNA, with shared transcriptional regulatory control. This is common for separate subunits of the same functional protein machine, which are useless without each other, or for proteins that function independently, but are required under the same circumstances. There is some evidence that *hli* genes are expressed as operons (Steve Biller personal communication). For *hlis*, the expressional coupling that comes with entering the same operon, could be explained by a benefit to the cell of having five different *hli* genes acting at the same time with slightly different functions all useful under the same conditions. Alternatively, it could be useful if the five genes act physically in concert, forming heteromultimers to carry out their functions. With the caveat that these are not direct orthologs of the multicopy, *Prochlorococcus*-specific *hli* genes in these array structures, the biochemical work on *Synechocystis hli* genes (which are scattered throughout the genome) contains evidence that *hlis* can be found in complexes together with each other as well as with other proteins (He et al., 2001, Storm et al., 2008, Chidgey et al., 2014). The presence of many hli arrays could enable the same basic genes to be tunable expression in different conditions, with independent regulatory regions. This comes with a small caveat on annotation methods, because we observed that different gene callers can choose different start sites for the same *hli* gene. In general, the slightly overlapping start is probably the most likely in the larger context of bacterial operon evolution, but to really know the operon structure you need measurements of transcription start and stops (possible now genome-wide through careful construction of RNA seq libraries) and direct measurements of protein to confirm exactly what is translated. *Hli* genes are so short, and more highly

expressed in stress, so they did not form a significant part of published proteins of *Prochlorococcus* cultures in log phase growth at moderate light. Upcoming and ongoing RNA seq work will allow us to address operon structures more precisely soon, on a large scale.

The *hli* tandem arrays also include many instances of slightly overlapping genes, either by one or four bp [e.g A(TGA) stop, (ATG)A start]. With the second gene poised in the right position for translation initiation just as the first gene is finishing, the coupled translation between adjacent *hlis* likely creates similarly quantities of the proteins. Such coupling is important for proteins serving as subunits in part of larger functional machine. This spacing structure is thought to evolve over time, from adjacent genes, to overlapping genes via deletion events - this conserves spaces, but also has functional consequences. Operon arrangement and spacing directly influences gene expression. Spacing between genes is a selected feature of operons, allowing for fine tuned translational control through either intervening space or coupling.

How are *hli* genes arranged in phage genomes?

Myophage and podophage both organize their genomes so that the host derived genes, the auxiliary metabolic are found together, in islands, in one or several places throughout the genome, so *hlis* in phage are generally found near and transcribed in concert with the other photosynthesis and carbon metabolism genes during infection (Labrie et al., 2013, Sullivan et al., 2005). Most podophage only have one *hli* gene, but for the ones with two (a pair of relatively distantly related Red Sea isolates) one has 2 *hlis* separated on opposite ends of the genome, and the other has its pair head-to-tail (Labrie et al., 2013, this work). In the siphovirus with two hlis they appear near each other but not in tandem, separated by three hypothetical proteins. Just as in *Prochlorococcus* genomes, for the myovirus genomes with many *hli* genes (2-6), the *hlis* are organized into tandem arrays (Lindell et al., 2004 Kelly et al, 2014, this work, data not yet shown). And, as for the host arrays, these are composed of diverse sequence variants and gene clusters, which shift from genome to genome, and the composition in terms of clusters is similar to hosts, although some *hli* genes in phage have diverged significantly at the protein level from their ancestral host homologs (Lindell et al., 2004, Kelly et al, 2014, this work, data not yet shown).

How do *hli* arrays change over evolutionary time?

Comparing some cultured and single cell genomes at one *hli* array locus (Figure 3.19), sampling deeply into the LLI lineage, we managed to observe a gain event of an *hli* locus. This small genomic island, bounded by core genes, contains tRNAs, hallmarks of horizontal gene transfer, for phage and host recombination events. For one small clade of our sample, four genomes have the *hli* array; for the other clades, and more deeply branching members of the group, and for a LLII outgroup, other genes, or no genes, fill the space between these core genes (Figure 3.19). So this *hli* array is not an ecotype-defining gene, rather, it was gained at some point after the LLI clade diverged from the rest of *Prochlorococcus*. We found it interesting that the other genes in this location in other lineages include a glycosyl transferase and a restriction-modification system, both fast-changing flexible genome tools that might have a role in the evolutionary arms race with phage. We find it remarkable that it appears that *hlis* are changing on the same time scale and genomic structural fashion as these genes (Kashtan et al., 2014). A similar analysis using large fragments of DNA for a different *hli* mini-genomic island showed that arrays also change through expansion and contraction, one gene at a time (Kettler, 2011).

3.4 Conclusions and Future Directions

Findings of this study: *hli* genes across *Prochlorococcus*, *Synechococcus* and their phage.

Our analyses confirm that the number of *hlis* and the assortment of sequence variants differ between deeply branching clades, and that most *Prochlorococcus*-infecting phage carry multiple *hli* genes. *Hli* genes are an important part of the *Prochlorococcus* flexible genome adaptive toolkit, frequently remodeled, changing dramatically in sequence and copy number across *Prochlorococcus* genomes on several timescales. *Prochlorococcus* evolution has given rise to a major radiation in the *hli* gene family. The *hlis* of *Prochlorococcus* have long been divided into two categories, the core genome, single copy, freshwater cyanobacterial orthologs at stable genome locations, and the multicopy, phage-shared flexible genome *hlis* in genomic islands, which respond to stress in *Prochlorococcus* expression perturbation experiments. Here we built curated annotations of *hli* genes for *Prochlorococcus*, *Synechococcus* and phage genomes. We organized these genes into finer-scale ortholog clusters, helping to describe the evolutionary events that create the overall distribution of these genes. By a comparison with *Synechococcus*, sister group to *Prochlorococcus*, we showed that the multicopy group of *hlis* are a specific addition to the *Prochlorococcus* lineage. While *Synechococcus* also have *hlis* in their flexible genomes, they are sampling from a different pool of proteins than the *Prochlorococcus* flexible genome multicopy *hlis*. Even the deeply branching LLIV *Prochlorococcus* ecotype, which generally shares more gene content with *Synechococcus* than other *Prochlorococcus* (Scanlan et al., 2009, Kettler et al., 2007) and has the same number of *hlis* as most *Synechococcus*, contains the *Prochlorococcus*-specific *hli* variants absent from *Synechococcus*.

We found that the massive expansion of *hli* genes to 40 copies in some LLI genomes occurred through repeated duplication of a few sequence clusters, by duplication or acquisition through horizontal transfer in a few tandem arrays. Why do the LLI *Prochlorococcus* carry so many *hli* genes? The remarkably high numbers in the LLI clade could be compensating for the lack of other mechanisms for dealing with high light present in HL strains, a brute force copy number approach, compared to the larger, concerted collection of adaptations in the HL strains (Coleman and Chisholm, 2007, Scanlan et al., 2009, Hess et al., 2001). Given their expression profiles, responding to a range of stressors, these genes could could be of adaptive value in a variety of different conditions, not just high light and changing light, a flexible genomic tool preferentially used by this clade.

The tandem array structure allows this gene family to change rapidly in units of whole arrays, not just one gene at a time, and allows arrays to change over time by shuffling array contents through homologous recombination, within and between genomes. The LLI and HL high copy numbers of *hli* complements evolved in part through shared history in these repeats, and in part through independent duplication or horizontal transfer events, reaching high copy numbers in parallel. We add to this picture the idea that the LLII/III clade may also contain the earliest seeds of *Prochlorococcus* adaptation to increasing light; although they have relatively low numbers of *hli* genes, they contain members of the same basic ortholog groups that later expand in the LLI and HL lineages. Also considering our physiology data showing the moderately robust response to LLII/II strains to light shock, these results suggest a more gradual evolutionary trajectory from LL to HL across the *Prochlorococcus* radiation, than previously thought, informing our interpretation of observations of these ecotype distributions in the wild. Many different factors play a role in growth and change in the *hli* family of *Prochlorococcus*, including phage, recombination, deletion/insertions, horizontal transfer, duplication, mutation and their tandem array structure. We hope that this study of the structure of this gene family and its evolution will ultimately contribute to a better functional understanding of these genes and their role in niche adaptation in *Prochlorococcus*.

Ecotype evolution and light shock in the environment

Here we tested, for one set of conditions, the response of a number of *Prochlorococcus* strains to light shock, and showed clear differences between LL ecotypes. What we tested, tolerance of brief intense periods of light, is a challenge for the cells that is distinct from sustained growth at these high light intensities. Although many molecular mechanisms for the two processes are shared, the timescales are different. Even strains capable of growth at high light do not always survive a transition directly from low to high light, so cultures are routinely stepped up gradually through intermediate light intensities to high light intensities (Moore et al. 1999, Moore et al., 2007).

There is a wide space to explore of possible combinations of acclimation light intensity and shock light intensity and duration; a huge range of variations across this parameter space occurs in the environment, we've tested one set of conditions here. There are many open questions surrounding how different lineages handle mixing. Conditions may exist that would separate the degree of light shock tolerance of LLI from HL, but from these results, the LLI clade shows robust tolerance of light shock. This difference between the LLII/III and LLIV strains suggests that there may be differences in light adaptations between these ecotypes as well, and that the ability to withstand transient changes in light may have begun to evolve earlier in *Prochlorococcus* history than we had previously thought. Based on these results and others, we are increasingly seeing that light physiology is not a strictly a HL/LL binary state in *Prochlorococcus*. Although there are strong features separating those two groups, we can lay a more gradual range of phenotypic adaptation on top of that division. For cells in the ocean this manifests as a cascade of overlapping habitat ranges for these ecotypes in the stratified water column. Most of the time, LL cells are below the mixed layer and experience relatively stable low light conditions, while HL ecotypes can also experience not only higher light but also constantly changing light in the mixed layer. However, the water column changes over time, with seasonal mixing events or smaller scale effects of changing temperature, wind and waves, so even LL cells can experience dynamic light conditions, providing another dimension in the niche space of *Prochlorococcus* in which ecotypes distinguish themselves.

Expansion in a contracting genome: adaptive implications of gene duplications

The *hli* genes of *Prochlorococcus*, with their a complex history of expansion, represent a marked exception to the paradigm of genome streamlining and paralog loss in oligotrophic bacteria. Genomes of some free-living bacteria adapted to oligotrophic environments are small, the smallest of all genomes except for obligate symbionts that can give up genes without a fitness cost because hosts provide functions (Batut et al., 2014). There is a theoretical framework for the process of genome reduction in oligotrophic free-living organisms, known as genome streamlining, suggesting that smaller genomes may be an adaptation to the low nutrient environment (Strehl et al., 1999, Dufresne et al., 2003, Rocap et al., 2003, Giovannoni et al., 2005). Random deletions occur in typical genome replication (the deletional bias), and if those genes are not essential, they can be safely lost (Giovannoni et al., 2005). A gene's essentiality is often a function of the environment. In the relatively homogenous and stable environment of the tropical open ocean, more genes may be dispensable than in the spatial and temporal complexity of a soil environment or the gut, where nutrients are plentiful and genes for many different conditions might be required frequently, and genomes tend to be bigger (Morris et al., 2012, Giovannoni et al., 2005). The shrinking genome comes with a possibly adaptive advantage when N and P are scarce, in a lower demand for nucleotides and fewer proteins to make (Batut et al., 2014). This is not true of all ocean microbes - many other adaptations occur across the whole community - but it does seem to apply to *Prochlorococcus*, to different extents from the larger genomes of the LL, which have access to more nutrients at depth, and smaller genomes of HL, who live in the lowest nutrient conditions (Batut et al., 2014, Yooseph et al., 2010). An interesting new level of complexity to this theory is that some of the lost functions can be provided by other members of the community (Morris et al., 2012). Inspired by the *Prochlorococcus* beneficial heterotroph interactions described above (one providing carbon, one providing peroxide removal services), this is a more holistic, multifactorial version of cheater

theory, in which members of a community share the burden of providing resources, each contributing (and not contributing) essential factors, bringing a picture of ecological connectivity to the level of microbes exchanging chemicals, enabling each to minimize its burden (Morris et al., 2012). Another part of this genome streamlining process is that paralogs have been preferentially lost from *Prochlorococcus* genomes: gene families, in general, have become progressively smaller in *Prochlorococcus* (Luo et al., 2011) where any functional redundancy exists.

Gene duplications provide a form of genetic variation, the raw material for selection, that can be adaptive in several ways (Henikoff et al., 1997, Hastings et al., 2009). Going from one copy of a gene to two, if both are expressed, increases the protein dosage, which could be beneficial, detrimental or neutral (Kondrashov et al., 2002, Papp et al., 2003). In the case of *hlis*, in some cases several of the copies of a gene in the same genome are nearly identical, and they occur at such high numbers in the genome, some of which are expressed under the same conditions, that there is likely a dosage increase at work. Another fate of a duplicated gene is that the same gene can be rapidly placed under different transcriptional control, allowing a gene, like the *hli* that might be useful under more than one circumstance, to join a new regulon without leaving its old one (Hastings et al., 2009). The *hlis* of *Prochlorococcus* have a range of different expression patterns. In some cases the same genes are expressed in many conditions, but in other cases similar proteins encoded by different genes are expressed in specific conditions, perhaps cases of duplication resulting in diversified expression. More expression data across diverse strains under different light conditions, both transient shocks and acclimated stable conditions, would inform our understanding of *hli* function and the link between these genes and light. Finally, perhaps the most powerful fate of a duplicated gene is that with the creation of a second copy, an important gene can remain under strict negative selection and keep up its function, while the second copy is released from negative selection by the presence of its paralog and is free to explore mutational space. Most of this exploration is deleterious, but some can lead to the birth of new protein functions (Zhang, 2003). This is how many new protein functions arise (Zhang, 2003). In the *hlis* of *Prochlorococcus*, there is so much protein variation, particularly between ortholog groups, that it is easy to imagine that some of that variation could relate to new functions, like the Prochlorococcus-specific *hli* C terminal motif, new to this group of proteins, which now appears to be conserved in its own right.

Why is there this proliferation of *hli* protein diversity in *Prochlorococcus*? What could these proteins' functions be?

Prochlorococcus evolution has given rise to a major radiation of deeply divergent hli proteins (roughly eight new genes). Based on expression data (Kettler, 2011), and even without knowing their specific functions, it appears that *hlis* represent small protein resources allocated to the photosystem in times of stress, supporting the cell's core machinery under many difficult conditions. What might be the functional consequence of all this diversity? A single amino acid change can affect the function of a protein, at a key residue, or a hundred amino acid changes can have no effect on a protein's function, so without biochemical work, this is speculative. For the most part, the key hydrophobic region and chlorophyll binding motifs are conserved across all of these proteins, which indicates some likelihood of shared chlorophyll-binding functionality. In some of our protein clusters the *Prochlorococcus*-specific C terminal motif is conserved, which might indicate some other, unknown, but conserved function specific to those proteins. Usually *hlis* from different clusters show little homology in the N-terminal region, which is highly variable in both length and protein sequence, but within ortholog clusters sometimes the full length of the protein, N-terminal included, can be conserved. One possible functional hypothesis, based on the body of biochemical work on *hli* function in other systems, is that these different *Prochlorococcus hlis*, positioned in the photosynthetic membranes by their hydrophobic region, could be binding and ferrying chlorophyll to

and from different members of the large set of chlorophyll binding proteins of the photosynthesis machinery, while at the same time protecting the cell from free chlorophyll. Perhaps the variable N-terminal sequence determines the hli protein's binding partners, both apoproteins awaiting chlorophyll and other proteins that act as assembly factors.

Why does *Prochlorococcus* have so many *hli* proteins that are so different from those in other cyanobacteria?

The most significant difference between *Prochlorococcus* and all other cyanobacteria is in the light harvesting apparatus of the photosystem; the evolution of their *hlis* occurred in the context of this sea change. Most cyanobacteria use a phycobilisome antennae to gather light, a large complex of proteins bound to phycobilin pigments in a range of colors that absorb light and transduce that energy to the photosystem (Ting et al., 2002, Six et al., 2007). In *Prochlorococcus*, there are no large phycobilisome structures, just a few remnant phycobiliproteins. Instead, the primary light gathering antennae is composed of unique prochlorophyte chlorophyll binding proteins (*pcbs*) which are derived from the CP43/CP47/IsiA family of photosystem components (Zhang et al., 2007), and which use chlorophyll to gather light (Hess et al., 2001, Ting et al., 2002). The chlorophyll used in these antennae is primarily divinyl chlorophyll a and b, pigments not found in most other cyanobacteria. Perhaps this new group of chlorophyll binding antennae proteins required different *hlis* to orchestrate chlorophyll delivery and insertion, and whatever other functions *hlis* might provide. The conserved C-terminal motif specific to the multicopy *Prochlorococcus* proteins (TGQIIPGF/IF) could be involved interactions with *pcbs*. Also, depending on exactly how the *hli* proteins bind chlorophyll molecules (which is not currently known), the new kinds of divinyl chlorophyll molecule employed in the *Prochlorococcus* system could require variations in *hli* proteins to compensate for their sidegroup modifications. This is strictly a speculative hypothesis based on the evolutionary patterns of *hlis*, chlorophyll binding proteins which include divergent sequences unique to *Prochlorococcus*, and *pcbs*, chlorophyll binding proteins which are unique to *Prochlorococcus*. But, if these *hlis* can function as chlorophyll trafficking proteins, with specific interactions with target apoproteins (as in the Ycf39/hliC/hliD/PSII example from *Synechocystis*) it stands to reason that modified proteins would be required for this vast array of new chlorophyll binding proteins. Furthermore, in the switch from a phycobilin-dependent to a chlorophyll-dependent antennae, perhaps the cell's reliance on and demand for *hli* proteins increased - they are chlorophyll-related whatever their function. To test this hypothesis would require a great deal of careful biochemistry, perhaps starting with a study of the interactions between *hli* proteins, *pcb* proteins, pigments and the rest of the photosystems and thylakoid-associated proteins. This collection of proteins represents a rich model for future biochemical work, an evolutionary testing ground for functional potential of *hli* genes being put to use in the oceans.

***hli* evolution, *Prochlorococcus* adaptation and ocean selection**

Based on the distribution of *hli* sequence diversity in the phylogenetic context of *Prochlorococcus* lineages, we can imagine a possible history of these genes, intimately connected to the evolutionary trajectory of *Prochlorococcus*. First, in the ancestor of all *Prochlorococcus* or shortly thereafter, some new deeply divergent *hlis* evolved, probably through mutation of a duplicated or horizontally transferred gene from existing *hli* groups. These new genes together formed a tandem array. During the course of the *Prochlorococcus* radiation thereafter, many events occurred, before, during and after the emergence of each ecotype, moving those genes and arrays within and between genomes, duplicating them, and further mutating them into different protein sequences.

Hli genes assist in the response to diverse stressors. Based on their expression profiles, what we know about their functions, and their distribution across ecotypes, the *Prochlorococcus hli* genes may contribute to the process of acclimation to higher light, to the maintenance of growth at high light, to survival of

transient light shock, and to other dimensions of a phytoplankter's niche space, including iron and nitrogen starvation and other forms of oxidative stress (e.g. Tolonen et al., 2006, Steglich et al., 2006, Chidgey et al., 2014, Lindell et al., 2007, Kettler, 2011 and Supplementary Figure 3.2). These are all conditions which take a toll on the photosynthetic machinery, which *hlis* probably alleviate. We believe that over time, diverse environmental pressures across the many *Prochlorococcus* habitats have resulted in selection of different *hli* gene complements (in sequence variants and in the number of genes), across different *Prochlorococcus* lineages, contributing to the definition of niche space for each lineage, on both ancient and recent time scales of *Prochlorococcus* evolution. *Hli* genes are a flexible genomic tool used throughout history of *Prochlorococcus*, contributing to the resiliency of *Prochlorococcus* individuals and populations in the dynamic marine environment.

Acknowledgements

This work was funded by grants to S.W.C from the Gordon and Betty Moore Foundation Marine Microbiology Initiative (GBMF495), National Science Foundation Evolutionary Ecology and Biological Oceanography Programs (Award #1145734) and the NSF Center for Microbial Oceanography Research and Education (C-MORE). G.K. was supported in part by the NIH Pre-Doctoral Training Grant T32GM007287. We thank Duygu Kasdogan of York University for her assistance with the light shock experiments and helpful conversations. We thank Allison Coe, Rogier Braakman and Andres Cubillos-Ruiz for more valuable discussions. We thank the Bermuda Atlantic Time-series and the Hawaii Ocean Time-series personnel for sample collection, the Bigelow Laboratory Single Cell Genomics Center for single-cell sorting and whole-genome amplification and the MIT BioMicro Center for sequencing.

References

Adir, N., Zer, H., Shochat, S., and Ohad, I. (2003). Photoinhibition - a historical perspective. Photosynth Res *76*, 343-370.

Aguirre von Wobeser, E., Ibelings, B.W., Bok, J., Krasikov, V., Huisman, J., and Matthijs, H.C. (2011). Concerted changes in gene expression and cell physiology of the cyanobacterium *Synechocystis* sp. strain PCC 6803 during transitions between nitrogen and light-limited growth. Plant Physiol *155*, 1445-457.

Ahlgren, N.A., and Rocap, G. (2012). Diversity and Distribution of Marine *Synechococcus*: Multiple Gene Phylogenies for Consensus Classification and Development of qPCR Assays for Sensitive Measurement of Clades in the Ocean. Front Microbiol *3*, 213.

Ahlgren, N.A., Rocap, G., and Chisholm, S.W. (2006). Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. Environ Microbiol *8*, 441-454.

Allakhverdiev, S.I., Tsvetkova, N., Mohanty, P., Szalontai, B., Moon, B.Y., Debreczeny, M., and Murata, N. (2005). Irreversible photoinhibition of photosystem II is caused by exposure of *Synechocystis* cells to strong light for a prolonged period. Biochim Biophys Acta *1708*, 342-351.

Andersson, U., Heddad, M., and Adamska, I. (2003). Light stress-induced one-helix protein of the chlorophyll a/b-binding family associated with photosystem I. Plant Physiol *132*, 811-820.

Apel, K., and Hirt, H. (2004). Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annu Rev Plant Biol *55*, 373-399.

Avrani, S., and Lindell, D. (2015). Convergent evolution toward an improved growth rate and a reduced resistance range in *Prochlorococcus* strains resistant to phage. Proceedings of the National Academy of Sciences

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates *Prochlorococcus*-virus coexistence. Nature *474*, 604-08.

Bagby, S.C., and Chisholm, S.W. (2015). Response of *Prochlorococcus* to varying CO2:O2 ratios. ISME J

Bailey, S., Clokie, M.R.J., Millard, A., and Mann, N.H. (2004). Cyanophage infection and photoinhibition in marine cyanobacteria. Res Microbiol *155*, 720-25.

Bathen, K.H. (1972). On the seasonal changes in the depth of the mixed layer in the north Pacific Ocean. J Geophys Res *77*, 7138-150.

Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014). Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol

Baumdicker, F., Hess, W.R., and Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol *4*, 443-456.

Berg, G.M., Shrager, J., van Dijken, G., Mills, M.M., Arrigo, K.R., and Grossman, A.R. (2011). Responses of psbA, hli and ptox genes to changes in irradiance in marine *Synechococcus* and *Prochlorococcus*. AQUATIC MICROBIAL ECOLOGY *65*, 1-14.

Berube, P.M., Biller, S.J., Kent, A.G., Berta-Thompson, J.W., Roggensack, S.E., Roache-Johnson, K.H., Ackerman, M., Moore, L.R., Meisel, J.D., et al. (2014). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. ISME J

Bhaya, D., Dufresne, A., Vaulot, D., and Grossman, A. (2002). Analysis of the hli gene family in marine and freshwater cyanobacteria. FEMS Microbiol Lett *215*, 209-219.

Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., Awad, L., Roache-Johnson, K.H., Ding, H., Giovannoni, S.J., et al. (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *1*, 140034.

Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015). *Prochlorococcus*: the structure and function of collective diversity. Nat Rev Microbiol *13*, 13-27.

de Boyer Montégut, C., Madec, G., Fischer, A.S., Lazar, A., and Iudicone, D. (2004). Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. J Geophys Res *109*, n/a-a.

Brainerd, K.E., and Gregg, M.C. (1995). Surface mixed and mixing layer depths. Deep Sea Research Part I: Oceanographic Research Papers *42*, 1521-543.

Chidgey, J.W., Linhartová, M., Komenda, J., Jackson, P.J., Dickman, M.J., Canniffe, D.P., Koník, P., Pilny, J., Hunter, C.N., and Sobotka, R. (2014). A Cyanobacterial Chlorophyll Synthase-HliD Complex Associates with the Ycf39 Protein and the YidC/Alb3 Insertase. Plant Cell *26*, 1267-279.

Chung, S.Y., and Subbiah, S. (1996). A structural explanation for the twilight zone of protein sequence homology. Structure *4*, 1123-27.

Clokie, M.R., and Mann, N.H. (2006). Marine cyanophages and light. Environ Microbiol *8*, 2074-082.

Clokie, M.R., Millard, A.D., and Mann, N.H. (2010). T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology. Virol J *7*, 291.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-23.

Coleman, M.L., and Chisholm, S.W. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. Trends Microbiol *15*, 398-407.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. Science *311*, 1768-770.

Daddy, S., Zhan, J., Jantaro, S., He, C., He, Q., and Wang, Q. (2015). A novel high light-inducible carotenoid-binding protein complex in the thylakoid membranes of *Synechocystis* PCC 6803. Sci Rep *5*, 9480.

Das, A., and Yanofsky, C. (1989). Restoration of a translational stop-start overlap reinstates translational coupling in a mutant trpB'-trpA gene pair of the *Escherichia coli* tryptophan operon. Nucleic Acids Res *17*, 9333-340.

Dauta, A., Devaux, J., Piquemal, F., and Boumnich, L. (1990). Growth rate of four freshwater algae in relation to light and temperature. *207*, 221-226-.

Denman, K.L., and Gargett, A.E. (1983). Time and space scales of vertical mixing and advection of phytoplankton in the upper ocean. Limnol Oceanogr *28*, 801-815.

Dolganov, N.A., Bhaya, D., and Grossman, A.R. (1995). Cyanobacterial protein with similarity to the chlorophyll a/b binding proteins of higher plants: evolution and regulation. Proc Natl Acad Sci U S A *92*, 636-640.

Dreher, T.W., Brown, N., Bozarth, C.S., Schwartz, A.D., Riscoe, E., Thrash, C., Bennett, S.E., Tzeng, S.C., and Maier, C.S. (2011). A freshwater cyanophage whose genome indicates close relationships to photosynthetic marine cyanomyophages. Environ Microbiol *13*, 1858-874.

Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., Paulsen, I.T., de Marsac, N.T., Wincker, P., et al. (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol *9*, R90.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792 -11797.

Eggink, L.L., and Hoober, J.K. (2000). Chlorophyll binding to peptide maquettes containing a retention motif. J Biol Chem *275*, 9087-090.

Engelken, J., Brinkmann, H., and Adamska, I. (2010). Taxonomic distribution and origins of the extended LHC (light-harvesting complex) antenna protein superfamily. BMC Evol Biol *10*, 233.

Ernst, A., Becker, S., Wollenzien, U.I., and Postius, C. (2003). Ecosystem-dependent adaptive radiations of picocyanobacteria inferred from 16S rRNA and ITS-1 sequence analysis. Microbiology *149*, 217-228.

Eyre-Walker, A. (1995). The distance between *Escherichia coli* genes is related to gene expression levels. J Bacteriol *177*, 5368-69.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proceedings of the National Academy of Sciences *110*, 9824-29.

Fragoso, G.M., Neale, P.J., Kana, T.M., and Pritchard, A.L. (2014). Kinetics of photosynthetic response to ultraviolet and photosynthetically active radiation in *Synechococcus* WH8102 (cyanobacteria). Photochem Photobiol *90*, 522-532.

Funk, C., and Vermaas, W. (1999). A cyanobacterial gene family coding for single-helix proteins resembling part of the light-harvesting proteins from higher plants. Biochemistry *38*, 9397-9404.

Giovannoni, S.J., and Vergin, K.L. (2012). Seasonality in ocean microbial communities. Science *335*, 671-76.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. Science *309*, 1242-45.

Goericke, R., and Welschmeyer, N.A. (1993). The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep Sea Research Part I: Oceanographic Research Papers *40*, 2283-294.

Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol *59*, 307 -3321.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009). Mechanisms of change in gene copy number. Nat Rev Genet *10*, 551-564.

Havaux, M., Guedeney, G., He, Q., and Grossman, A.R. (2003). Elimination of high-light-inducible polypeptides related to eukaryotic chlorophyll a/b-binding proteins results in aberrant photoacclimation in *Synechocystis* PCC6803. Biochim Biophys Acta *1557*, 21-33.

He, Q., Dolganov, N., Bjorkman, O., and Grossman, A.R. (2001). The high light-inducible polypeptides in *Synechocystis* sp. PCC6803. Expression and function in high light. J Biol Chem *276*, 306-314.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. Science *278*, 609-614.

Hess, W.R., Rocap, G., Ting, C.S., Larimer, F., Stilwagen, S., Lamerdin, J., and Chisholm, S.W. (2001). The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. Photosynth Res *70*, 53-71.

[NO STYLE for: Finn 2010].

Huang, L., McCluskey, M.P., Ni, H., and LaRossa, R.A. (2002). Global gene expression profiles of the cyanobacterium *Synechocystis* sp. strain PCC 6803 in response to irradiation with UV-B and white light. J Bacteriol *184*, 6845-858.

Huang, S., Wang, K., Jiao, N., and Chen, F. (2011). Genome sequences of siphoviruses infecting marine *Synechococcus* unveil a diverse cyanophage group and extensive phage-host genetic exchanges. Environ Microbiol

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119.

Ignacio-Espinoza, J.C., and Sullivan, M.B. (2012). Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes. Environ Microbiol

Jansson, S., Andersson, J., Kim, S.J., and Jackowski, G. (2000). An *Arabidopsis thaliana* protein homologous to cyanobacterial high-light-inducible proteins. Plant Mol Biol *42*, 345-351.

Jantaro, S., Ali, Q., Lone, S., and He, Q. (2006). Suppression of the lethality of high light to a quadruple HLI mutant by the inactivation of the regulatory protein PfsR in *Synechocystis* PCC 6803. J Biol Chem *281*, 30865-874.

Johnson, Z.I., and Chisholm, S.W. (2004). Properties of overlapping genes are conserved across microbial genomes. Genome Res *14*, 2268-272.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M., and Chisholm, S.W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science *311*, 1737-740.

Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev *33*, 376-393.

Kana, T.M., and Glibert, P.M. (1987). Effect of irradiances up to 2000 μE m−2 s−1 on marine *Synechococcus* WH7803–I. Growth, pigmentation, and cell composition. Deep Sea Research Part A. Oceanographic Research Papers *34*, 479-495.

Kana, T.M., and Glibert, P.M. (1987). Effect of irradiances up to 2000 μE m−2 s−1 on marine *Synechococcus* WH7803–II. Photosynthetic responses and mechanisms. Deep Sea Research Part A. Oceanographic Research Papers *34*, 497-516.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science *344*, 416-420.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol *30*, 772-780.

Kelly, L., Ding, H., Huang, K.H., Osburne, M.S., and Chisholm, S.W. (2013). Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. ISME J *7*, 1827-841.

Kelly, L., Huang, K.H., Ding, H., and Chisholm, S.W. (2012). ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. Nucleic Acids Res *40*, D632-640.

Kettler, G.C. (2011). Genetic diversity and its consequences for light adaptation in *Prochlorococcus*. MIT PhD Thesis

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., et al. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet *3*, e231.

Knoppová, J., Sobotka, R., Tichy, M., Yu, J., Konik, P., Halada, P., Nixon, P.J., and Komenda, J. (2014). Discovery of a Chlorophyll Binding Protein Complex Involved in the Early Steps of Photosystem II Assembly in *Synechocystis*. Plant Cell

Kondrashov, F., Rogozin, I., Wolf, Y., and Koonin, E. (2002). Selection in the evolution of gene duplications. Genome Biology *3*, research0008.1-.9.

Kopecná, J., Komenda, J., Bucinská, L., and Sobotka, R. (2012). Long-term acclimation of the cyanobacterium *Synechocystis* sp. PCC 6803 to high light is accompanied by an enhanced production of chlorophyll that is preferentially channeled to trimeric photosystem I. Plant Physiol *160*, 2239-250.

Kufryk, G., Hernandez-Prieto, M.A., Kieselbach, T., Miranda, H., Vermaas, W., and Funk, C. (2008). Association of small CAB-like proteins (SCPs) of *Synechocystis* sp. PCC 6803 with Photosystem II. Photosynthesis Research *95*, 135-145.

Labrie, S.J., Frois-Moniz, K., Osburne, M.S., Kelly, L., Roggensack, S.E., Sullivan, M.B., Gearin, G., Zeng, Q., Fitzgerald, M., et al. (2013). Genomes of marine cyanopodoviruses reveal multiple origins of diversity. Environ Microbiol *15*, 1356-376.

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res *32*, 11-16.

Latifi, A., Ruiz, M., and Zhang, C.C. (2009). Oxidative stress in cyanobacteria. FEMS Microbiol Rev 33, 258-278.

Li, W.K.W. (1994). Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. Limnol Oceanogr *39*, 169-175.

Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T., Kettler, G., Sullivan, M.B., Steen, R., et al. (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature *449*, 83-86.

Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. Nature *438*, 86-89.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A *101*, 11013-18.

Liu, H.B., Nolla, H.A., and Campbell, L. (1997). *Prochlorococcus* growth rate and contribution to primary production in the equatorial and subtropical North Pacific Ocean. AQUATIC MICROBIAL ECOLOGY *12*, 39-47.

Long, S.P., Humphries, S., and Falkowski, P.G. (1994). Photoinhibition of Photosynthesis in Nature. Annual Review of Plant Physiology and Plant Molecular Biology *45*, 633-662.

Luo, H., Friedman, R., Tang, J., and Hughes, A.L. (2011). Genome reduction by deletion of paralogs in the marine cyanobacterium *Prochlorococcus*. Mol Biol Evol *28*, 2751-760.

Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. ISME J 4, 1252-264.

Mann, N.H., Clokie, M.R., Millard, A., Cook, A., Wilson, W.H., Wheatley, P.J., Letarov, A., and Krisch, H.M. (2005). The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. J Bacteriol *187*, 3188-3200.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003). Marine ecosystems: bacterial photosynthesis genes in a virus. Nature *424*, 741.

Martiny, A.C., Tai, A.P., Veneziano, D., Primeau, F., and Chisholm, S.W. (2009). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. Environ Microbiol *11*, 823-832.

Martínez, L., Morán, A., and García, A. (2012). Effect of light on *Synechocystis* sp. and modelling of its growth rate as a response to average irradiance. *24*, 125-134-.

Mary, I., Tu, C., Grossman, A., and Vaulot, D. (2004). Effects of high light on transcripts of stress-associated genes for the cyanobacteria *Synechocystis* sp. PCC 6803 and *Prochlorococcus* MED4 and MIT9313. Microbiology *150*, 1271-281.

Millard, A.D., Zwirglmaier, K., Downey, M.J., Mann, N.H., and Scanlan, D.J. (2009). Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of *Synechococcus* host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. Environ Microbiol *11*, 2370-387.

Montané, M.H., and Kloppstech, K. (2000). The family of light-harvesting-related proteins (LHCs, ELIPs, HLIPs): was the harvesting of light their primary function? Gene *258*, 1-8.

Moore, L.R., and Chisholm, S.W. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus* : ecotypic differences among cultured isolates. Limnol Oceanogr *44*, 628-638.

Moore, L.R., Coe, A., Zinser, E.R., Saito, M.A., Sullivan, M.B., Lindell, D., Frois-Moniz, K., Waterbury, J., and Chisholm, S.W. (2007). Culturing the marine cyanobacterium *Prochlorococcus*. Limnology and Oceanography: Methods *5*, 353-362.

Moore, L.R., Goericke, R., and Chisholm, S.W. (1995). Comparative Physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absoprtive propeties. Marine Ecology Progress Series *116*, 259-275.

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature *393*, 464-67.

Morel, A., Huot, Y., Gentili, B., Werdell, P.J., Hooker, S.B., and Franz, B.A. (2007). Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. Remote Sensing of Environment *111*, 69-88.

Morris, J.J., Johnson, Z.I., Szul, M.J., Keller, M., and Zinser, E.R. (2011). Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. PLoS ONE *6*, e16805.

Morris, J.J., Kirkegaard, R., Szul, M.J., Johnson, Z.I., and Zinser, E.R. (2008). Facilitation of robust growth of *Prochlorococcus* colonies and dilute liquid cultures by "helper" heterotrophic bacteria. Appl Environ Microbiol *74*, 4530-34.

Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio *3*

Muramatsu, M., and Hihara, Y. (2012). Acclimation to high-light conditions in cyanobacteria: from gene expression to physiological responses. J Plant Res *125*, 11-39.

Mühling, M. (2012). On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. Environ Microbiol *14*, 567-579.

Nishiyama, Y., Allakhverdiev, S.I., and Murata, N. (2006). A new paradigm for the action of reactive oxygen species in the photoinhibition of photosystem II. Biochimica et Biophysica Acta (BBA) - Bioenergetics *1757*, 742-49.

Nixon, P.J., Michoux, F., Yu, J., Boehm, M., and Komenda, J. (2010). Recent advances in understanding the assembly and repair of photosystem II. Ann Bot , mcq059.

Oppenheim, D.S., and Yanofsky, C. (1980). Translational coupling during expression of the tryptophan operon of *Escherichia coli*. Genetics *95*, 785-795.

Ottesen, E.A., Young, C.R., Gifford, S.M., Eppley, J.M., Marin, R., Schuster, S.C., Scholin, C.A., and DeLong, E.F. (2014). Ocean microbes. Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. Science *345*, 207-212.

Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E.E., et al. (2003). The genome of a motile marine *Synechococcus*. Nature *424*, 1037-042.

Palenik, B., Ren, Q., Dupont, C.L., Myers, G.S., Heidelberg, J.F., Badger, J.H., Madupu, R., Nelson, W.C., Brinkac, L.M., et al. (2006). Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. Proc Natl Acad Sci U S A *103*, 13555-59.

Pallejà, A., Harrington, E.D., and Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics *9*, 335.

Papp, B., Pál, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194-97.

Partensky, F., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. Ann Rev Mar Sci *2*, 305-331.

Partensky, F., Blanchot, J., and Vaulot, D. (1999). Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. Bulletin de l'Institut Oceanographique, Monaco *19*

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev *63*, 106-127.

Partensky, F., Hoepffner, N., Li, W., Ulloa, O., and Vaulot, D. (1993). Photoacclimation of *Prochlorococcus* sp. (Prochlorophyta) Strains Isolated from the North Atlantic and the Mediterranean Sea. Plant Physiol *101*, 285-296.

Pál, C., Papp, B., and Lercher, M.J. (2006). An integrated view of protein evolution. Nat Rev Genet *7*, 337-348.

Pope, W.H., Weigele, P.R., Chang, J., Pedulla, M.L., Ford, M.E., Houtz, J.M., Jiang, W., Chiu, W., Hatfull, G.F., et al. (2007). Genome sequence, structural proteins, and capsid organization of the cyanophage Syn5: a "horned" bacteriophage of marine *Synechococcus*. J Mol Biol *368*, 966-981.

Powles, S.B. (1984). Photoinhibition of Photosynthesis Induced by Visible Light. Annu Rev Plant Physiol *35*, 15-44.

Price, M.N., Arkin, A.P., and Alm, E.J. (2006). The life-cycle of operons. PLoS Genet *2*, e96.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Promnares, K., Komenda, J., Bumba, L., Nebesarova, J., Vacha, F., and Tichy, M. (2006). Cyanobacterial Small Chlorophyll-binding Protein ScpD (HliB) Is Located on the Periphery of Photosystem II in the Vicinity of PsbH and CP47 Subunits. J. Biol. Chem. *281*, 32705-713.

Raghava, G.P., and Barton, G.J. (2006). Quantification of the variation in percentage identity for protein sequence alignments. BMC Bioinformatics *7*, 415.

Rocap, G., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol *68*, 1180-191.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature *424*, 1042-47.

Rodrigue, S., Malmstrom, R.R., Berlin, A.M., Birren, B.W., Henn, M.R., and Chisholm, S.W. (2009). Whole genome amplification and de novo assembly of single bacterial cells. PLoS One *4*, e6864.

Rodrigue, S., Materna, A.C., Timberlake, S.C., Blackburn, M.C., Malmstrom, R.R., Alm, E.J., and Chisholm, S.W. (2010). Unlocking Short Read Sequencing for Metagenomics. PLoS ONE *5*, e11840.

Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Engineering *12*, 85-94.

Sabehi, G., Shaulov, L., Silver, D.H., Yanai, I., Harel, A., and Lindell, D. (2012). A novel lineage of myoviruses infecting cyanobacteria is widespread in the oceans. Proc Natl Acad Sci U S A *109*, 2037-042.

Saito, M.A., Moffett, J.W., Chisholm, S.W., and Waterbury, J.B. (2002). Cobalt limitation and uptake in *Prochlorococcus*. Limnol Oceanogr *47*, 1629-636.

Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J. (2000). Operons in *Escherichia coli*: genomic analyses and predictions. Proc Natl Acad Sci U S A *97*, 6652-57.

Scanlan, D.J., Hess, W.R., Partensky, F., Newman, J., and Vaulot, D. (1996). High degree of genetic variation in *Prochlorococcus* (Prochlorophyta) revealed by RFLP analysis. European Journal of Phycology *31*, 1-9.

Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., Post, A.F., Hagemann, M., Paulsen, I., and Partensky, F. (2009). Ecological Genomics of Marine Picocyanobacteria. Microbiology and Molecular Biology Reviews : MMBR *73*, 249-299.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068-69.

Sher, D., Thompson, J.W., Kashtan, N., Croal, L., and Chisholm, S.W. (2011). Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. ISME J *5*, 1125-132.

Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., Calteau, A., Cai, F., Tandeau de Marsac, N., et al. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. Proc Natl Acad Sci U S A *110*, 1053-58.

Sinha, R.K., Komenda, J., Knoppová, J., Sedlářová, M., and Pospíšil, P. (2012). Small CAB-like proteins prevent formation of singlet oxygen in the damaged photosystem II complex of the cyanobacterium *Synechocystis* sp. PCC 6803. Plant Cell Environ *35*, 806-818.

Six, C., Thomas, J.C., Brahamsha, B., Lemoine, Y., and Partensky, F. (2004). Photophysiology of the marine cyanobacterium *Synechococcus* sp. WH8102, a new model organism. Aquat Microb Ecol *35*, 17-29.

Six, C., Thomas, J.C., Garczarek, L., Ostrowski, M., Dufresne, A., Blot, N., Scanlan, D.J., and Partensky, F. (2007). Diversity and evolution of phycobilisomes in marine *Synechococcus* spp.: a comparative genomics study. Genome Biol *8*, R259.

Sobotka, R. (2013). Making proteins green; biosynthesis of chlorophyll-binding proteins in cyanobacteria. Photosynth Res

Sobotka, R., McLean, S., Zuberova, M., Hunter, C.N., and Tichy, M. (2008). The C-terminal extension of ferrochelatase is critical for enzyme activity and for functioning of the tetrapyrrole pathway in *Synechocystis* strain PCC 6803. J Bacteriol *190*, 2086-095.

Sommaruga, R., Hofer, J.S., Alonso-Sáez, L., and Gasol, J.M. (2005). Differential sunlight sensitivity of picophytoplankton from surface Mediterranean Coastal Waters. Appl Environ Microbiol *71*, 2154-57.

Steglich, C., Futschik, M., Rector, T., Steen, R., and Chisholm, S.W. (2006). Genome-wide analysis of light sensing in *Prochlorococcus*. J Bacteriol *188*, 7796-7806.

Stoddard, L.I., Martiny, J.B., and Marston, M.F. (2007). Selection and characterization of cyanophage resistance in marine *Synechococcus* strains. Appl Environ Microbiol *73*, 5516-522.

Storm, P., Hernandez-Prieto, M.A., Eggink, L.L., Hoober, J.K., and Funk, C. (2008). The small CAB-like proteins of *Synechocystis* sp. PCC 6803 bind chlorophyll. Photosynthesis Research *98*, 479-488.

Storm, P., Tibiletti, T., Hall, M., and Funk, C. (2013). Refolding and enzyme kinetic studies on the ferrochelatase of the cyanobacterium *Synechocystis* sp. PCC 6803. PLoS One *8*, e55569.

Strehl, B., Holtzendorff, J., Partensky, F., and Hess, W.R. (1999). A small and compact genome in the marine cyanobacterium *Prochlorococcus marinus* CCMP 1375: lack of an intron in the gene for tRNA (Leu)(UAA) and a single copy of the rRNA operon. FEMS Microbiol Lett *181*, 261-66.

Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005). Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. PLoS Biol *3*, e144.

Sullivan, M.B., Huang, K.H., Ignacio-Espinoza, J.C., Berlin, A.M., Kelly, L., Weigele, P.R., DeFrancesco, A.S., Kern, S.E., Thompson, L.R., et al. (2010). Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol *12*, 3035-056.

Sullivan, M.B., Krastins, B., Hughes, J.L., Kelly, L., Chase, M., Sarracino, D., and Chisholm, S.W. (2009). The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. Environ Microbiol *11*, 2935-951.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol *4*, e234.

Sullivan, M.B., Waterbury, J.B., and Chisholm, S.W. (2003). Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature *424*, 1047-051.

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A *102*, 13950-55.

Thompson, A.W., Huang, K., Saito, M.A., and Chisholm, S.W. (2011). Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. ISME J *5*, 1580-594.

Thompson, L.R., Zeng, Q., Kelly, L., Huang, K.H., Singer, A.U., Stubbe, J., and Chisholm, S.W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A *108*, E757-764.

Tikkanen, M., Grieco, M., Nurmi, M., Rantala, M., Suorsa, M., and Aro, E.M. (2012). Regulation of the photosynthetic apparatus under fluctuating growth light. Philos Trans R Soc Lond B Biol Sci *367*, 3486-493.

Ting, C.S., Rocap, G., King, J., and Chisholm, S.W. (2002). Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. Trends Microbiol *10*, 134-142.

Tolonen, A.C., Aach, J., Lindell, D., Johnson, Z.I., Rector, T., Steen, R., Church, G.M., and Chisholm, S.W. (2006). Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. Mol Syst Biol *2*, 53.

Urbach, E., and Chisholm, S.W. (1998). Genetic diversity in *Prochlorococcus* populations flow cytometrically sorted from the Sargasso Sea and the Gulf Stream. Limnology and Oceanography *43*, 1615-630.

Urbach, E., Scanlan, D.J., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (1998). Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). J Mol Evol *46*, 188-201.

van de Guchte, M., Kok, J., and Venema, G. (1991). Distance-dependent translational coupling and interference in *Lactococcus lactis*. Mol Gen Genet *227*, 65-71.

Vavilin, D., Yao, D., and Vermaas, W. (2007). Small Cab-like proteins retard degradation of photosystem II-associated chlorophyll in *Synechocystis* sp. PCC 6803: kinetic analysis of pigment labeling with 15N and 13C. J Biol Chem *282*, 37660-68.

Walker, M., Pavlovic, V., and Kasif, S. (2002). A comparative genomic method for computational identification of prokaryotic translation initiation sites. Nucleic Acids Res *30*, 3181-191.

Wang, Q., Jantaro, S., Lu, B., Majeed, W., Bailey, M., and He, Q. (2008). The high light-inducible polypeptides stabilize trimeric photosystem I complex under high light conditions in *Synechocystis* PCC 6803. Plant Physiol *147*, 1239-250.

Waterbury, J.B., Watson, S.W., Valois, F.W., and Franks, D.G. (1986). Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. Canadian Bulletin of Fisheries and Aquatic Sciences *214*, 71-120.

Weigele, P.R., Pope, W.H., Pedulla, M.L., Houtz, J.M., Smith, A.L., Conway, J.F., King, J., Hatfull, G.F., Lawrence, J.G., and Hendrix, R.W. (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. Environ Microbiol *9*, 1675-695.

West, N.J., and Scanlan, D.J. (1999). Niche-Partitioning of *Prochlorococcus* Populations in a Stratified Water Column in the Eastern North Atlantic Ocean. Appl Environ Microbiol *65*, 2585-591.

Xu, H., Vavilin, D., and Vermaas, W. (2002). The presence of chlorophyll b in *Synechocystis* sp. PCC 6803 disturbs tetrapyrrole biosynthesis and enhances chlorophyll degradation. J Biol Chem *277*, 42726-732.

Xu, H., Vavilin, D., Funk, C., and Vermaas, W. (2002). Small Cab-like proteins regulating tetrapyrrole biosynthesis in the cyanobacterium *Synechocystis* sp. PCC 6803. Plant Mol Biol *49*, 149-160.

Xu, H., Vavilin, D., Funk, C., and Vermaas, W. (2004). Multiple deletions of small Cab-like proteins in the cyanobacterium *Synechocystis* sp. PCC 6803: consequences for pigment biosynthesis and accumulation. J Biol Chem *279*, 27971-79.

Yao, D., Kieselbach, T., Komenda, J., Promnares, K., Prieto, M.A.H., Tichy, M., Vermaas, W., and Funk, C. (2007). Localization of the small CAB-like proteins in photosystem II. J Biol Chem *282*, 267-276.

Yao, D.C., Brune, D.C., Vavilin, D., and Vermaas, W.F. (2012). Photosystem II component lifetimes in the cyanobacterium *Synechocystis* sp. strain PCC 6803: small Cab-like proteins stabilize biosynthesis intermediates and affect early steps in chlorophyll synthesis. J Biol Chem *287*, 682-692.

Yooseph, S., Nealson, K.H., Rusch, D.B., McCrow, J.P., Dupont, C.L., Kim, M., Johnson, J., Montgomery, R., Ferriera, S., et al. (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. Nature *468*, 60-66.

Yu, J.S., Madison-Antenucci, S., and Steege, D.A. (2001). Translation at higher than an optimal level interferes with coupling at an intercistronic junction. Molecular Microbiology *42*, 821-834.

Zeng, Q., and Chisholm, S.W. (2012). Marine viruses exploit their host's two-component regulatory system in response to resource limitation. Curr Biol *22*, 124-28.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol Evol *18*, 292-98.

Zinser, E., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D., and Chisholm, S.W. (2007). Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. Limnol Oceanogr *52*, 2205-220.

Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J., and Chisholm, S.W. (2006). *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol *72*, 723-732.

Zinser, E.R., Lindell, D., Johnson, Z.I., Futschik, M.E., Steglich, C., Coleman, M.L., Wright, M.A., Rector, T., Steen, R., et al. (2009). Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. PLoS ONE *4*, e5135.

Supplemental Information

Supplemental Table S3.1. ***Prochlorococcus*** **genomes used in this study**

Genome	Ecotype	Genome size (bp)	Number of Contigs	Percent GC	*Hli* gene count[1]	Number of proteins[2]	Citation	Accession
MIT9313	LLIV	2,410,873	1	51	10	2,551	Rocap et al., 2003	NC_005071.1
MIT9303	LLIV	2,682,675	1	50	10	2,732	Kettler et al., 2007	NC_008820.1
MIT0701	LLIV	2,592,571	53	51	8	2,666	Biller et al., 2014	JNBA00000000
MIT0702	LLIV	2,583,057	61	51	8	2,659	Biller et al., 2014	JNBB00000000
MIT0703	LLIV	2,575,057	61	51	8	2,643	Biller et al., 2014	JNBC00000000
MIT1313	LLIV	2,590,341	28	50	11	2,625	Chapter IV	unpublished[3]
MIT1306	LLIV	2,498,944	12	51	10	2,514	Chapter IV	unpublished[3]
MIT1320	LLIV	2,500,454	26	50	11	2,604	Chapter IV	unpublished[3]
MIT1323	LLIV	2,440,679	26	51	11	2,503	Chapter IV	unpublished[3]
MIT1318	LLIV	2,584,744	27	50	11	2,627	Chapter IV	unpublished[3]
MIT1312	LLIV	2,561,499	53	51	10	2,656	Chapter IV	unpublished[3]
MIT1327	LLIV	2,591,587	34	50	10	2,627	Chapter IV	unpublished[3]
MIT1342	LLIV	2,548,000	27	50	10	2,610	Chapter IV	unpublished[3]
MIT0601	LLIII	1,707,342	6	37	14	1,863	Biller et al., 2014	JNAU00000000
MIT9211	LLIII	1,688,963	1	38	13	1,852	Kettler et al., 2007	NC_009976.1
LG	LLII	1,754,063	14	36	14	1,906	Biller et al., 2014	JNAT00000000
SS2	LLII	1,752,772	19	36	14	1,910	Biller et al., 2014	JNAY00000000
SS35	LLII	1,751,015	9	36	14	1,905	Biller et al., 2014	JNAZ00000000
SS51	LLII	1,746,977	12	36	14	1,900	Biller et al., 2014	JNBD00000000
SS52	LLII	1,754,053	22	36	14	1,909	Biller et al., 2014	JNBE00000000
SS120	LLII	1,751,080	1	36	14	1,902	Dufresne et al. 2003	NC_005042.1
MIT0603	LLII	1,752,482	7	36	15	1,913	Biller et al., 2014	JNAW00000000
MIT0602	LLII	1,750,918	9	36	15	1,913	Biller et al., 2014	JNAV00000000
NATL1A	LLI	1,864,731	1	35	43	2,149	Kettler et al., 2007	NC_008819.1
NATL2A	LLI	1,842,899	1	35	43	2,108	Kettler et al., 2007	NC_007335.2
PAC1	LLI	1,841,163	20	35	25	2,162	Biller et al., 2014	JNAX00000000
MIT0801	LLI	1,929,203	1	35	40	2,190	Biller et al., 2014	CP007754
Med4	HLI	1,657,990	1	31	24	1,916	Rocap et al., 2003	NC_005072.1
EQPAC1	HLI	1,654,739	8	31	20	1,913	Biller et al., 2014	JNAG00000000

Genome	Ecotype	Genome size (bp)	Number of Contigs	Percent GC	*Hli* gene count[1]	Number of proteins[2]	Citation	Accession
MIT9515	HLI	1,704,176	1	31	24	1,898	Kettler et al., 2007	NC_008817.1
MIT9201	HLII	1,672,416	21	31	21	1,957	Biller et al., 2014	JNAL00000000
MIT9123	HLII	1,697,748	18	31	18	1,937	Biller et al., 2014	JNAK00000000
MIT9116	HLII	1,685,398	22	31	18	1,921	Biller et al., 2014	JNAJ00000000
MIT9107	HLII	1,699,937	13	31	18	1,931	Biller et al., 2014	JNAI00000000
GP2	HLII	1,624,310	11	31	18	1,832	Biller et al., 2014	JNAH00000000
MIT9401	HLII	1,666,808	17	31	17	1,916	Biller et al., 2014	JNAR00000000
MIT9322	HLII	1,657,550	11	31	17	1,902	Biller et al., 2014	JNAQ00000000
MIT9321	HLII	1,658,664	10	31	17	1,904	Biller et al., 2014	JNAP00000000
UH18301	HLII	1,654,648	18	31	19	1,975	Morris et al. 2011	PRJNA47033
MIT9312	HLII	1,709,204	1	31	26	1,919	Coleman et al., 2006	NC_007577.1
MIT9311	HLII	1,711,064	17	31	26	1,913	Biller et al., 2014	JNAN00000000
MIT9302	HLII	1,745,343	17	31	23	1,969	Biller et al., 2014	JNAM00000000
MIT9215	HLII	1,738,790	1	31	20	1,982	Kettler et al., 2007	NC_009840.1
MIT9202	HLII	1,690,387	6	31	20	1,958	Thompson et al., 2011a	ACDW01000001.1
MIT0604	HLII	1,780,061	1	31	17	2,065	Biller et al., 2014	CP007753
AS9601	HLII	1,669,886	1	31	22	1,889	Kettler et al., 2007	NC_008816.1
SB	HLII	1,669,823	4	32	19	1,892	Biller et al., 2014	JNAS00000000
MIT9314	HLII	1,690,556	16	31	20	1,938	Biller et al., 2014	JNAO00000000
MIT9301	HLII	1,641,879	1	31	17	1,873	Kettler et al., 2007	NC_009091.1

[1] The number of *hli* family genes in each genome, based on the multiple hidden Markov model approach employed in this study.

[2] Number of protein coding genes, based on reannotation using the Prokka pipeline described in this study (Seeman, 2014).

[3] Eight LLIV clade genomes included in this study have not yet been published. The isolation and sequencing of these strains is described in Chapter II of this thesis; efforts towards their publication are underway.

Supplemental Table S3.2. Marine *Synechococcus* genomes used in this study

Strain	Cluster	Genome size	Number of contigs	Percent GC	*hli* gene count	Number of proteins	Accession	Citation
WH5701	5.2	3,043,834	135	65	16	3,170	AANO01000001.1	Dufresne et al., 2008
PCC6307[2]	5.2	3,342,364	1	69	11	3,354	NC_019675.1	Shih et al., 2013

Strain	Cluster	Genome size	Number of contigs	Percent GC	*hli* gene count	Number of proteins	Accession	Citation
PCC7001[2]	5.2	2,832,697	18	69	12	2,811	ABSE01000001.1	JCVI[3]
CB0101	5.2	2,686,395	94	64	11	2,955	ADXL01000001.1	JCVI[3]
CB0205	5.2	2,427,308	78	63	9	2,677	ADXM01000001.1	JCVI[3]
RCC307	5.3	2,224,914	1	61	10	2,534	NC_009482.1	Dufresne et al., 2008
CC9311	5.1	2,606,748	1	52	17	2,865	NC_008319.1	Palenik et al., 2006
WH8102	5.1	2,434,428	1	59	9	2,703	NC_005070.1	Palenik et al., 2003
CC9605	5.1	2,510,659	1	59	10	2,882	NC_007516.1	Dufresne et al., 2008
BL107	5.1	2,283,377	6	54	11	2,498	AATZ01000001.1	Dufresne et al., 2008
CC9902	5.1	2,234,828	1	54	9	2,442	NC_007513.1	Dufresne et al., 2008
WH8016	5.1	2,694,843	16	54	11	2,979	AGIK01000001.1	JGI[3]
WH8109	5.1	2,111,515	1	60	9	2,395	CP006882.1	JCVI[3]
MITS9508	5.1	2,502,434	23	56	17	2,817	unpublished	Andrés Cubillos Ruiz[1]
MITS9509	5.1	3,087,928	33	55	20	3,507	unpublished	Andrés Cubillos Ruiz[1]
MITS9504	5.1	3,087,293	34	55	20	3,506	unpublished	Andrés Cubillos Ruiz[1]
RS9916	5.1	2,664,465	4	60	9	2,822	AAUA01000001.1	Dufresne et al., 2008
RS9917	5.1	2,579,542	9	65	8	2,692	AANP01000001.1	Dufresne et al., 2008
WH7805	5.1	2,620,367	13	58	9	2,725	AAOK01000001.1	Dufresne et al., 2008
WH7803	5.1	2,366,980	1	60	10	2,544	NC_009481.1	Dufresne et al., 2008
KORDI-100	5.1	2,789,000	1	58	11	3,000	CP006269.1	MBRD[3]
KORDI-49	5.1	2,585,813	1	61	10	2,651	CP006270.1	MBRD[3]
KORDI-52	5.1	2,572,069	1	59	9	2,760	CP006271.1	MBRD[3]
CC9616	5.1	2,644,310	17	57	12	2,856	AZXL01000001.1	JGI[3]

[1] These are currently in preparation for publication; used here courtesy of Andrés Cubillos Ruiz, Chisholm Lab.

[2] These two strains are sometimes classified as members of the genus *Cyanobium*, other times *Synechococcus* (e.g. Ernst et al., 2003, Urbach and Chisholm, 1998). PCC 6307 is freshwater, PCC 7001 marine. They are phylogenetically closely related to some marine *Synechococcus*, and so were included in this study, relevant to the larger evolutionary context of the *Synechococcus-Prochlorococcus* picocyanobacterial clade (Urbach and Chisholm, 1998, Ernst et al., 2003, Shih et al.,

2003, Scanlan et al., 2009). They are sometimes placed with *Synechococcus* cluster 5.2 (e.g. Scanlan et al., 2009, Shih et al., 2013) and sometimes assigned to their own *Cyanobium* cluster (e.g. Ernst et al., 2003).

[3] JCVI refers to J. Craig Venter Institute, Rockland, MD, JGI to the US Department of Energy Joint Genome Institute, Walnut Creek, CA, and MBRD to the Marine Biotechnology Research Division of the Korea Institute of Ocean Science and Technology, Sangnok-gu, South Korea . These sequence centers produced and made public these genomes, but have not described them in publication form.

Supplemental Table S3.3. Marine cyanophage genomes used in this study

Phage	Family	Isolation Host	Genome size	Genome gc content	*hli* count	Number of genes	Accession	Citation
KBS-M-1A	myovirus	*Synechococcus*	171,744	41	2	213	NC_020836.1	
S-CAM1	myovirus	*Synechococcus*	198,013	43	2	229	NC_020837.1	
S-CAM8-BI06	myovirus	*Synechococcus*	171,407	39	2	209	NC_021530.1	
S-CAM8-SB47	myovirus	*Synechococcus*	222,057	41	2	267	JF974299.1	
S-CBM2	myovirus	*Synechococcus*	180,892	40	2	203	HQ633061.1	
S-CRM01	myovirus	*Synechococcus*	178,563	40	1	296	NC_015569.1	
S-IOM18	myovirus	*Synechococcus*	171,797	41	2	211	NC_021536.1	
S-MbC100	myovirus	*Synechococcus*	170,438	39	2	201	NC_023584.1	
S-MbCM25	myovirus	*Synechococcus*	176,044	39	2	205	KF156339.1	
S-MbCM6	myovirus	*Synechococcus*	176,043	39	2	216	NC_019444.1	
S-MbCM7	myovirus	*Synechococcus*	189,311	40	2	214	NC_023587.1	
S-PM2	myovirus	*Synechococcus*	196,280	38	2	229	NC_006820	Mann et al., 2005
S-RIM2R1	myovirus	*Synechococcus*	175,430	42	2	205	NC_020859.1	
S-RIM2R21	myovirus	*Synechococcus*	175,430	42	2	206	HQ317290.1	
S-RIM2R9	myovirus	*Synechococcus*	175,419	42	2	204	HQ317291.1	
S-RIM8_AHR1	myovirus	*Synechococcus*	171,211	41	2	213	NC_020486.1	
S-RIM8_AHR3	myovirus	*Synechococcus*	171,211	41	2	213	JF974289.1	
S-RIM8_AHR5	myovirus	*Synechococcus*	168,327	41	2	206	HQ317385.1	
S-RSM4	myovirus	*Synechococcus*	194,454	41	2	218	NC_013085	Millard et al., 2009
S-ShM2	myovirus	*Synechococcus*	179,563	41	2	205	NC_015281	Sullivan et al., 2010
S-SM1	myovirus	*Synechococcus*	174,079	41	2	216	NC_015282	Sullivan et al., 2010

Phage	Family	Isolation Host	Genome size	Genome gc content	*hli* count	Number of genes	Accession	Citation
S-SM2	myovirus	*Synechococcus*	190,789	40	2	244	NC_015279	Sullivan et al., 2010
S-SSM2	myovirus	*Synechococcus*	179,980	41	2	200	JF974292	Kelly et al., 2013
S-SSM4	myovirus	*Synechococcus*	182,801	39	4	232	NC_020875	Kelly et al., 2013
S-SSM5	myovirus	*Synechococcus*	176,184	40	2	206	NC_015289	Sullivan et al., 2010
S-SSM7	myovirus	*Synechococcus*	232,878	39	2	297	NC_015287	Sullivan et al., 2010
S-TIM5	myovirus	*Synechococcus*	161,440	40	2	176	JQ245707	Sabehi et al., 2012
Syn1	myovirus	*Synechococcus*	191,195	41	2	206	NC_015288	Sullivan et al., 2010
Syn10	myovirus	*Synechococcus*	177,103	41	2	208	HQ634191	Kelly et al., 2013
Syn19	myovirus	*Synechococcus*	175,230	41	2	201	NC_015286	Sullivan et al., 2010
Syn2	myovirus	*Synechococcus*	175,596	41	2	204	HQ634190	Kelly et al., 2013
Syn30	myovirus	*Synechococcus*	178,807	40	2	215	NC_021072	Kelly et al., 2013
Syn33	myovirus	*Synechococcus*	174,285	40	2	209	NC_015285	Sullivan et al., 2010
Syn9	myovirus	*Synechococcus*	177,300	41	2	198	NC_008296	Weigele et al., 2007
Med4-213	myovirus	*Prochlorococcus*	180,977	38	4	229	NC_020845	Kelly et al., 2013
P-HM1	myovirus	*Prochlorococcus*	181,044	38	4	225	NC_015280	Sullivan et al., 2010
P-HM2	myovirus	*Prochlorococcus*	183,806	38	4	224	NC_015284	Sullivan et al., 2010
P-RSM1	myovirus	*Prochlorococcus*	177,211	40	3	213	NC_021071	Kelly et al., 2013
P-RSM3	myovirus	*Prochlorococcus*	178,750	37	4	223	HQ634176	Kelly et al., 2013
P-RSM4	myovirus	*Prochlorococcus*	176,428	38	4	226	NC_015283	Sullivan et al., 2010
P-RSM6	myovirus	*Prochlorococcus*	192,497	39	3	217	NC_020855	Kelly et al., 2013

Phage	Family	Isolation Host	Genome size	Genome gc content	*hli* count	Number of genes	Accession	Citation
P-SSM2	myovirus	*Prochlorococcus*	252,407	36	6	327	GU071092	Sullivan et al., 2005
P-SSM3	myovirus	*Prochlorococcus*	179,063	37	4	235	NC_021559	Kelly et al., 2013
P-SSM4	myovirus	*Prochlorococcus*	178,249	37	4	219	NC_006884	Sullivan et al., 2005
P-SSM5	myovirus	*Prochlorococcus*	252,013	36	6	332	HQ632825	Kelly et al., 2013
P-SSM7	myovirus	*Prochlorococcus*	182,180	37	5	217	NC_015290	Sullivan et al., 2010
P60	podovirus	*Synechococcus*	47,872	53	0	84	NC_003390	Chen and Lu, 2002
S-CBP2	podovirus	*Synechococcus*	92,473	51	0	126	JF974303.1	
S-CBP3	podovirus	*Synechococcus*	47,375	47	1	58	HQ633062.1	
S-CBP4	podovirus	*Synechococcus*	41,824	44	1	54	HM559717.1	
S-RIP1	podovirus	*Synechococcus*	44,892	43	1	69	NC_020867.1	
S-RIP2	podovirus	*Synechococcus*	45,728	47	1	53	NC_020838.1	
SCBP42	podovirus	*Synechococcus*	43,241	54	1	59	JF974300.1	
Syn5	podophage	*Synechococcus*	46,214	55	0	57	NC_009531	Pope et al., 2007
P-GSP1	podovirus	*Prochlorococcus*	44,945	40	1	60	NC_020878	Labrie et al., 2013
P-HP1	podovirus	*Prochlorococcus*	47,536	40	1	65	NC_016659	Labrie et al., 2013
P-RSP2	Podo	*Prochlorococcus*	42,257	34	2	53	HQ332139	Labrie et al., 2013
P-RSP5	Podo	*Prochlorococcus*	47,741	39	2	67	GU071102	Labrie et al., 2013
P-SSP10	Podo	*Prochlorococcus*	47,325	39	1	56	NC_020835	Labrie et al., 2013
P-SSP2	Podo	*Prochlorococcus*	45,890	38	1	57	GU071107	Labrie et al., 2013
P-SSP3	Podo	*Prochlorococcus*	46,198	38	1	56	HQ332137	Labrie et al., 2013
P-SSP5	Podo	*Prochlorococcus*	47,055	39	1	56	Proportal	Proportal
P-SSP6	Podo	*Prochlorococcus*	47,039	39	1	61	HQ634152	Labrie et al., 2013

Phage	Family	Isolation Host	Genome size	Genome gc content	*hli* count	Number of genes	Accession	Citation
P-SSP7	Podo	*Prochlorococcus*	44,970	39	1	54	NC_006882	Sullivan et al. 2005
P-SSP9	Podo	*Prochlorococcus*	46,997	40	1	53	HQ316584	Labrie et al., 2013
KBS-S-2A	Sipho	*Synechococcus*	40,658	49	0	56	NC_020857	
S-CBS1	Sipho	*Synechococcus*	30,332	59	0	44	NC_016164.1	
S-CBS2	Sipho	*Synechococcus*	72,332	54	1	111	NC_015463.1	
S-CBS3	Sipho	*Synechococcus*	33,004	61	0	51	NC_015465.1	
S-CBS4	Sipho	*Synechococcus*	105,580	52	0	169	NC_016766.1	
S-SKS1	Sipho	*Synechococcus*	208,007	36	2	301	NC_020851.1	Moore
P-HS1	Sipho	*Prochlorococcus*	38,834	37	0	68	NC_020857	
P-HS2	Sipho	*Prochlorococcus*	38,327	37	0	64	NC_020847	
P-HS3	Sipho	*Prochlorococcus*	37,830	37	0	62	unpublished	Katya Frois-Moniz and Simon Labrie
P-HS4	Sipho	*Prochlorococcus*	37,851	37	0	64	unpublished	Katya Frois-Moniz and Simon Labrie
P-HS5	Sipho	*Prochlorococcus*	37,850	37	0	63	unpublished	Katya Frois-Moniz and Simon Labrie
P-HS6	Sipho	*Prochlorococcus*	37,983	37	0	65	unpublished	Katya Frois-Moniz and Simon Labrie
P-HS7	Sipho	*Prochlorococcus*	36,546	36	0	65	unpublished	Katya Frois-Moniz and Simon Labrie

Phage	Family	Isolation Host	Genome size	Genome gc content	*hli* count	Number of genes	Accession	Citation
P-HS8	Sipho	*Prochlorococcus*	36,533	36	0	64	unpublished	Katya Frois-Moniz and Simon Labrie
P-SS2	Sipho	*Prochlorococcus*	107,530	52	0	128	unpublished	Katya Frois-Moniz and Simon Labrie

Supplementary Table S3.4. Single Cell genomes

Cell designation	Origin location	Sample Date	Depth	Description	Eco-type/ clade	Total bp in assembly	# of con-tigs	Largest Contig	Average contig length	N50	GC%
497_B10	BATS	Feb 8th 2009	60m	Deeply mixed water column	HLI	396,005	112	55,885	3,535	31,607	37
497_G2	BATS	Feb 8th 2009	60m	Deeply mixed water column	HLI	1,246,512	80	401,106	15,581	155,475	31
497_K6	BATS	Feb 8th 2009	60m	Deeply mixed water column	LLI	1,597,212	117	229,170	13,651	86,095	35
498_E5	BATS	Feb 8th 2009	60m	Deeply mixed water column	LLI	1,214,688	262	187,183	4,636	14,211	35
518_A2	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	HLI	1,572,845	117	480,856	13,443	168,628	31
518_A6	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	HLI	1,318,625	91	151,899	14,490	113,102	32
518_G2	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	812,316	64	343,725	12,692	202,238	36
519_B6	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	189,682	74	33,606	2,563	8,979	34
519_C6	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	757,663	84	119,505	9,019	55,126	35
520_C9	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	1,034,192	154	244,142	6,715	66,684	35
521_C9	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	HLI	736,574	144	225,607	5,115	35,308	31
521_J22	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	1,102,603	132	142,968	8,353	65,133	34
521_L15	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	806,253	151	129,930	5,339	29,063	35
521_L18	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	1,653,100	173	354,975	9,555	93,381	35
521_O2	BATS	Apr 1st 2009	60m	Stratified spring water column, in mixed layer	LLI	1,130,708	249	174,394	4,540	51,927	35

Cell designation	Origin location	Sample Date	Depth	Description	Eco-type/ clade	Total bp in assembly	# of con-tigs	Largest Contig	Average contig length	N50	GC%
527_E8	BATS	Nov 8th 2008	60m	Stratified autumn water column, in mixed layer	HLI	693,391	105	96,492	6,603	46,835	31
815_J16	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	1,439,086	156	308,726	9,224	86,845	36
816_E23	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	1,187,917	287	81,281	4,139	31,526	34
816_E5	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	1,202,360	140	168,585	8,588	39,761	34
818_A6	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	1,333,603	144	211,131	9,261	71,045	35
818_E18	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	1,092,017	135	152,164	8,089	84,943	34
818_E20	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	418,705	109	37,870	3,841	15,457	36
818_J15	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	879,653	181	104,090	4,859	27,083	35
818_J21	HOT-ALOHA	July 2-6 2009	100m	Stratified summer water col., below mixed	LLI	231,893	36	65,215	6,441	21,536	36
OMZ_B6	ETSP OMZ	Nov, 2010	55m	Secondary chlorophyll max below oxycline	HLII	1,204,253	125	441,687	9,634	232,113	31
OMZ_E11	ETSP OMZ	Nov, 2010	55m	Secondary chlorophyll max below oxycline	LLI	539,579	117	59,485	4,611	11,946	34
OMZ_F8	ETSP OMZ	Nov, 2010	55m	Secondary chlorophyll max below oxycline	LLI	802,557	236	29,880	3,400	9,525	35
OMZ_H3	ETSP OMZ	Nov, 2010	55m	Secondary chlorophyll max below oxycline	LLI	644,717	126	64,721	5,116	14,768	35
OMZ_M9	ETSP OMZ	Nov, 2010	55m	Secondary chlorophyll max below oxycline	LLI	562,190	143	25,498	3,931	11,004	35
OMZ_N22	ETSP OMZ	Nov, 2010	55m	Secondary chlorophyll max below oxycline	LLI	1,099,828	253	57,529	4,347	13,806	34

Supplementary Table S3.5 Sources for previously published *hlis* used in building hidden markov models for *hli* searches

Genus or viral family	Strain	Clade/ ecotype	NCBI accession number	Number of genes in hli-containing V3 COGs	Number of annotated hlis on ncbi genomes	Genome reference
Synechococcus	CC9311	5.1A I	CP000435	13	17	Palenik et al., 2006
Synechococcus	CC9605	5.1A II	CP000110	9	8	Dufresne et al., 2008
Synechococcus	WH8102	5.1A III	BX548020	8	9	Palenik et al., 2003
Synechococcus	BL107	5.1A IV	DS022298	11	10	Dufresne et al., 2008

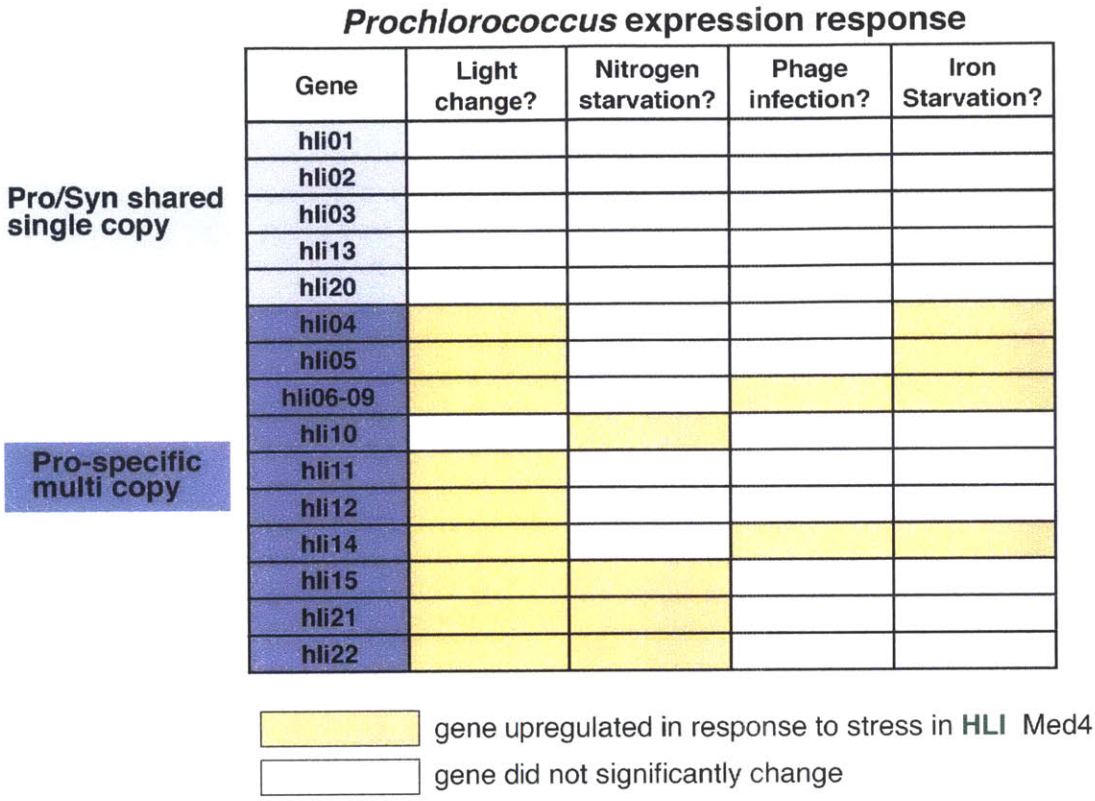
Genus or viral family	Strain	Clade/ ecotype	NCBI accession number	Number of genes in hli-containing V3 COGs	Number of annotated hlis on ncbi genomes	Genome reference
Synechococcus	CC9902	5.1A IV	CP000097	9	8	Dufresne et al., 2008
Synechococcus	RSS9917	5.1B VIII	CH724158	7	8	Dufresne et al., 2008
Synechococcus	RSS9916	5.1B IX	DS022299	8	8	Dufresne et al., 2008
Synechococcus	WH7805	5.1B VI	CH724168	9	6	Dufresne et al., 2008
Synechococcus	WH7803	5.1B V	CT971583	10	10	Dufresne et al., 2008
Synechococcus	RCC307	5	CT978603	10	9	Dufresne et al., 2008
Synechococcus	WH5701	5	AANO01	7	12	Dufresne et al., 2008
Prochlorococcus	MIT9313	LLIV/e9313	NC_005071	9	10	Rocap et al., 2003
Prochlorococcus	MIT9303	LLIV/e9313	NC_008820	10	9	Kettler et al., 2007
Prochlorococcus	SS120	LLII/eSS120	NC_005042	12	13	Dufresne et al., 2003
Prochlorococcus	MIT9211	LLIII/e9211	NC_009976	11	5	Kettler et al., 2007
Prochlorococcus	NATL1A	LLI/eNATL	NC_008819	17	7	Kettler et al., 2007
Prochlorococcus	NATL2A	LLI/eNATL	NC_007335	16	30	Kettler et al., 2007
Prochlorococcus	MIT9515	HLI/eMed4	NC_008817	17	20	Kettler et al., 2007
Prochlorococcus	Med4	HLI/eMed4	NC_005072	16	22	Rocap et al., 2003
Prochlorococcus	AS9601	HLII/e9312	NC_008816	20	18	Kettler et al., 2007
Prochlorococcus	MIT9312	HLII/e9312	NC_007577	20	22	Kettler et al., 2007
Prochlorococcus	MIT9202	HLII/e9312	NZ_DS999537	17	15	Thompson et al., 2011a
Prochlorococcus	MIT9301	HLII/e9312	NC_009091	14	11	Kettler et al., 2007
Prochlorococcus	MIT9215	HLII/e9312	NC_009840	18	10	Kettler et al., 2007
Podovirus	Syn5	MPP-A	NC_009531	0	0	Pope et al., 2007
Podovirus	P60	MPP-A	NC_003390	0	0	Chen and Lu, 2002
Podovirus	P-SSP9	MPP-A	HQ316584	0	0	Labrie et al., 2013
Podovirus	P-HP1	MPP-B1	NC_016659	1	0	Labrie et al., 2013

Genus or viral family	Strain	Clade/ ecotype	NCBI accession number	Number of genes in hli-containing V3 COGs	Number of annotated hlis on ncbi genomes	Genome reference
Podovirus	P-SSP10	MPP-B1	NC_020835	1	0	Labrie et al., 2013
Podovirus	P-SSP11/P-SSP6	MPP-B1	HQ634152	n/a	0	Labrie et al., 2013
Podovirus	P-RSP5	MPP-B2	GU071102	1	0	Labrie et al., 2013
Podovirus	P-GSP1	MPP-B2	NC_020878	1	1	Labrie et al., 2013
Podovirus	P-SSP2	MPP-B2	GU071107	1	0	Labrie et al., 2013
Podovirus	P-SSP3	MPP-B2	HQ332137	n/a	0	Labrie et al., 2013
Podovirus	P-SSP7	MPP-B2	NC_006882	1	1	Sullivan et al. 2005
Podovirus	P-RSP2	unclassified	HQ332139	0	0	Labrie et al., 2013
Podovirus	P-SSP5	unclassified	unpublished	1	n/a	on Proportal
Myovirus	Syn1	I	NC_015288	2	2	Sullivan et al., 2010
Myovirus	Syn10	III	HQ634191	2	2	Kelly et al., 2013
Myovirus	Syn19	III	NC_015286	2	2	Sullivan et al., 2010
Myovirus	Syn2	III	HQ634190	2	0	Kelly et al., 2013
Myovirus	Syn30	III	NC_021072	1	0	Kelly et al., 2013
Myovirus	Syn33	III	NC_015285	2	2	Sullivan et al., 2010
Myovirus	Syn9	III	NC_008296	2	2	Weigele et al., 2007
Myovirus	Med4-213	unknown	NC_020845	1	1	Kelly et al., 2013
Myovirus	P-HM1	unknown	NC_015280	4	4	Sullivan et al., 2010
Myovirus	P-HM2	unknown	NC_015284	4	4	Sullivan et al., 2010
Myovirus	P-RSM1	III	NC_021071	1	0	Kelly et al., 2013
Myovirus	P-RSM3	III	HQ634176	2	2	Kelly et al., 2013
Myovirus	P-RSM4	III	NC_015283	4	4	Sullivan et al., 2010
Myovirus	P-RSM6	unknown	NC_020855	2	0	Kelly et al., 2013
Myovirus	P-SSM2	II	GU071092	6	3	Sullivan et al., 2005

Genus or viral family	Strain	Clade/ ecotype	NCBI accession number	Number of genes in hli-containing V3 COGs	Number of annotated hlis on ncbi genomes	Genome reference
Myovirus	P-SSM3	II	NC_021559	2	2	Kelly et al., 2013
Myovirus	P-SSM4	III	NC_006884	4	4	Sullivan et al., 2005
Myovirus	P-SSM5	III	HQ632825	2	3	Kelly et al., 2013
Myovirus	P-SSM7	III	NC_015290	5	5	Sullivan et al., 2010
Myovirus	S-PM2	I	NC_006820	2	2	Mann et al., 2005
Myovirus	S-RSM4	unknown	NC_013085	2	2	Millard et al., 2009
Myovirus	S-ShM2	III	NC_015281	1	1	Sullivan et al., 2010
Myovirus	S-SM1	III	NC_015282	2	2	Sullivan et al., 2010
Myovirus	S-SM2	II	NC_015279	1	1	Sullivan et al., 2010
Myovirus	S-SSM2	III	JF974292	1	0	Kelly et al., 2013
Myovirus	S-SSM4	III	NC_020875	2	0	Kelly et al., 2013
Myovirus	S-SSM5	III	NC_015289	2	2	Sullivan et al., 2010
Myovirus	S-SSM7	II	NC_015287	2	2	Sullivan et al., 2010
Myovirus	S-TIM5	unknown	JQ245707	n/a	2	Sabehi et al., 2012

Unknown refers to strains not part of the Sullivan et al., 2008 Myovirus phylogenetic classification scheme based on the g20 protein.
Unclassified refers to podoviruses that didn't fit into the main groups described in Labrie et al., 2013.
For comparison and illustration of the automatic annotation challenge, and to supplement orthlogy based hmm training set with independent methods.
Goal wasn't to be comprehensive, just to get a large set of hlis with some homology evidence to use for searching (hopefully generating a comprehensive set down the line)
These COG counts do not match number published numbers of hlis per genome, and are not intended to represent a full review of past work. COGs do not contain full sets of identical proteins.
Clusters are probably a good collection of well-supported homology hlis, but illustrate the challenges of clustering these - a few tiny clusters and a few giant clusters.
In many cases publications with genomes describe hlis, but they were not included in published annotations. These are just what's present in database
g20 portal protein defined phage clusters.
*These additional genomes were searched, but had not annotated hlis (they do have hlis - added based on orthlog clusters - just not formally annotated in published genomes)
The benefit is that these rely on diverse past methods for annotation -occasionally catch different genes
For building search clusters, being comprehensive isn't necessary - just need a few representatives from each deeply branching sequence cluster, which we'll use to recruit more of their own.
Also illustrates challenges with automated annoation of these genes.

Supplemental Figure 3.1 Relationships between phage and host *hli* genes UPGMA cluster
(A) This is a different display of the clustergram in Figure 3.10 to highlight the deeply branching phage clade. Star indicates this deeply branching cluster of phage *hlis*, predominantly from *Synechococcus*-isolated phage for which it is difficult to assign relationships to host clusters. In the tree, phage clade nodes are in darker blue for myophage, lighter blue for podophage, magenta for siphophage, and host clade nodes are in black. Podophage same from a subset of the same larger cluster that myophage sample from. The labels at the leaves of the tree are colored as either *Prochlorococcus* (green), *Synechococcus* (red), or phage (blue). **(B)** Zoom of the deeply-branching all-phage clade, with taxa labels showing that most are isolated from *Synechococcus*, based on S-nomenclature for isolation host.



Supplementary Figure 3.2. It is the *Prochlorococcus*-specific, phage-shared *hlis* that respond to stress in *Prochlorococcus*: summary of microarray expression experiments in *Prochlorococcus*

Following the authors' individual significance cutoffs. The core, Synechococcus-shared genes do not respond to stress, the multicopy ones do. This figure represents the meta-analysis performed in Kettler 2011, using microarray data from Steglich et al., 2006, Tolonen et al., 2006, Lindell et al., 2007, Thompson et al, 2011a. In Med4, the array containing hli06-09 is identical at the DNA level to hli16-19, and so could not be distinguished using microarrays.

Chapter IV. Abundance, distribution and physical properties of *Prochlorococcus* of the South East Pacific: dramatic variation over gradients in nutrients and light

Jessie W. Berta-Thompson[1,2] and Sallie W. Chisholm[1,3]

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
[2]Microbiology Graduate Program, Massachusetts Institute of Technology
[3]Department of Biology, Massachusetts Institute of Technology

Abstract

The South Pacific contains some of the most productive waters in the world, in the Humboldt current upwelling region off the coast of Chile, and some of the least productive waters in the world, at the center of the South Pacific Subtropical Gyre. We had an opportunity to sample in this region across a transect spanning gradients in nutrient concentrations, temperature, water optical qualities, and community structure, including the extremely oligotrophic waters near Easter Island and the permanent oxygen minimum zone near the coast. Here we present the distribution of *Prochlorococcus* populations over this transect, measured through flow cytometry, and analyze these data in the context of features of the water column and transect, particularly light. Concentrations of *Prochlorococcus* cells were consistently high in the gyre, often above 10^5 cells ml^{-1} at their peak depth, and more variable in abundance (though always observed) near the coast and at transitional sites, a pattern often observed in *Prochlorococcus* distributions. The depth of the euphotic zone in this region changes dramatically from the relatively murky waters of the high productivity region to the extremely clear waters of the gyre, where light travels hundreds of meters down. We observed an overall trend of increasing depth of *Prochlorococcus* populations from coast to open ocean, largely tracking changes in light attenuation through the water column. In analyzing the per cell chlorophyll fluorescence, a property of both genetics and cellular acclimation to different light conditions, most samples map onto the light environment in a consistent way; chlorophyll fluorescence per cell is related to light extinction over the water column in the same way across wide variations in other conditions. However, at the two extremes of coastal and oligotrophic samples, different relationships between cellular chlorophyll fluorescence and light suggest that additional factors are at work governing this property of the cells, such as the influence of light spectral qualities and genetic variation. For one site, we obtained high resolution samples, over several days, which allow us to assess some of the challenges of typically sparse oceanographic sampling, and to observe the smooth transitions that *Prochlorococcus* populations make from small surface-adapted strains acclimated to high light with low chlorophyll content, to deep populations composed of larger cells with high chlorophyll content for optimized light gathering. This work has served as a starting place for other analyses from the same cruise, assisting in sample processing decisions and interpretation of other data exploring *Prochlorococcus* populations. The *Prochlorococcus* of the South Pacific show a population-level resilience, maintaining large populations across broad changes in conditions over space and depth, emergent from a combination of each cell's ability to acclimate to diverse conditions, a vast array of diversity within populations and likely region-specific genetic adaptations, which will be an exciting topic for future explorations.

4.1 Introduction

The South Pacific: a collection of remarkable habitats

The South Pacific covers a vast expanse of our planet containing some of the most oligotrophic waters in the world's oceans at its center, and some of the most productive in the upwelling regions off the South American coast. The gradient in productivity, from coast to gyre, in this region can be viewed from space, through ocean chlorophyll sensing satellites, which show the dramatic shift from high-chlorophyll, high-productivity regions, to very low-chlorophyll, low-productivity regions (Figure 4.1). The South Pacific Gyre is distant from equatorial, coastal and polar upwelling and continental nutrient input, making it is the largest and lowest nutrient ocean gyre in the world (Claustre et al., 2008a,b). Nutrient input essential for primary production (and all life), primarily nitrogen, phosphorus, and iron, comes primarily from small-scale vertical mixing events and dust deposition (Bonnet et al., 2008). In the western South Pacific, there are so-called high nutrient, low chlorophyll (HNLC) regions, characterized by iron-limitation resulting in low phytoplankton biomass, and a build-up of unused nitrogen and phosphorus (Claustre et al., 2008a, Moisander et al., 2011). The South Pacific gyre (central and eastern South Pacific oligotrophic open ocean region) contains low levels of nitrogen (5-10 nM ammonium) and dissolved iron (0.1 nM), but higher levels of phosphorus, (>110nM phosphate) (Bonnet et al., 2008, Moutin et al., 2008, Van Mooy et al., 2009, Fitzsimmons et al., 2014); there, nitrogen is generally the limiting nutrient for primary production (Bonnet et al., 2008, Moisander et al., 2011). This ocean is historically sparsely sampled, with no equivalent to the established time-series stations in the North Pacific and North Atlantic that have taught us much of what we know about the oligotrophic open oceans (Bonnet et al., 2008, Claustre et al., 2008b, Karl and Church, 2014, Giovannoni and Vergin, 2012). The low nutrient conditions result in low total biomass, but a rich community of microbial life on a small scale makes a living here, and it is a major *Prochlorococcus* habitat.

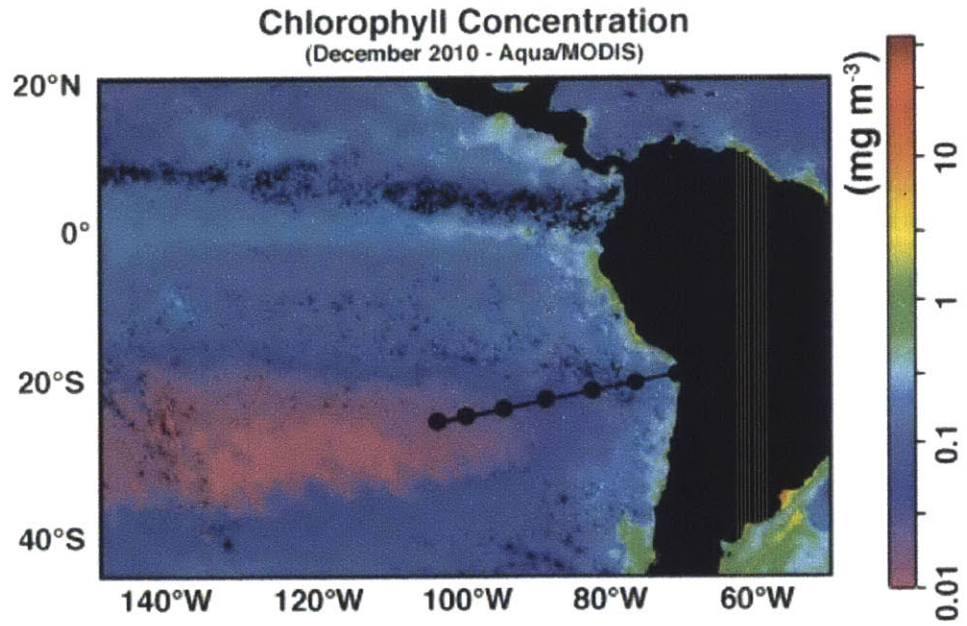


Figure 4.1. A transect across dramatic gradients in the Eastern South Pacific
The South Pacific Gyre represents the least productive, lowest chlorophyll region of the world's oceans. Satellite chlorophyll data (from the Aqua-MODIS instrument) shows the dramatic transect in productivity from coast to gyre. The Big-RAPA cruise transect described in this chapter is marked with a line, dots for sample stations.

An opportunity to study *Prochlorococcus* across South Pacific gradients in environmental conditions

We had an opportunity to sample across the Eastern South Pacific, over a 3,500 km transect from the coast of northern Chile to Easter Island, setting out with the goal of studying its *Prochlorococcus* populations on several levels (Figure 4.1). This transect crosses important ecological gradients, including in nutrient concentrations, temperature, and light, that govern microbial community structure (Figure 4.1, Figure 4.3, Bonnet et al., 2008, Duhamel et al, 2012, Fitzsimmons et al., 2014, Boiteau et al., 2013). We sampled from several trophic regions, including the eutrophic oxygen minimum zone, a transitional mesotrophic regime, and highly oligotrophic waters.

South Pacific *Prochlorococcus*

Prochlorococcus occurs at high concentrations in the South Pacific Gyre, based on the limited but growing collection of measurements available for the this region (Bouman et al., 2006, Grob et al., 2007, Zwirglmaier et al., 2008, Flombaum et al., 2013). Thanks to its many adaptations to low nutrient conditions, including its small size, small genome and nitrogen-thrifty photosynthetic antennae, *Prochlorococcus* thrives in all the subtropical oligotrophic waters of the world (Partensky et al., 1999, Partensky and Garczarek, 2010, Morel et al., 1993, Chisholm, 1992, Hess et al., 2001). It does not live in colder, richer waters nearer the poles (Partensky et al., 1999, Flombaum et al., 2013, Zinser et al., 2007). Adaptation to low nutrient conditions is a useful framework for describing *Prochlorococcus* as a whole, but the wide global distribution of *Prochlorococcus*, and its ability to live across the full range of light over the euphotic zone, are also enabled by wide-ranging diversity within the group. A few of the adaptations varying among *Prochlorococcus* studied to date include traits for coping with different nutrient conditions, toxicity, temperature, changing light, and light intensity (Coleman et al., 2006, Thompson et al., 2011, Berube et al., 2014, Moore et al.,2002, Mann et al., 2002, Moore et al., 1999, Zinser et al., 2007, Martiny et al., 2009, Malmstrom et al., 2010). The dramatic gradients over the South Pacific represent many diverse environmental conditions, and *Prochlorococcus* populations across this region likely contain wide functional variation.

The clearest waters in the world

In the extremely oligotrophic South Pacific Gyre, low productivity results in the clearest waters in the world, where light penetrates a blue ocean for hundreds of meters, resulting in a deeper distribution of phytoplankton (Morel et al., 2007a,b,c, Claustre et al., 2008b). Ordinarily, phytoplankton, other planktonic biomass and colored dissolved organic matter interfere with the passage of light through water, changing its color and attenuation rate with depth as a function of the quantity and spectral qualities of light-interacting material in the water (Morel et al., 2010). The amount of light-interacting material is itself largely a function of productivity and in some places land-origin input of colored dissolved organic matter (Morel et al., 2010, Morel et al., 2007a,c). When light enters the water column, it attenuates exponentially with depth through absorption and backscattering; red light is preferentially scattered by water, so with depth the light remaining becomes more blue. In extremely clear waters without other major influences on water color, this effect is strongest, and the water bluest at depth. UV radiation, with its potent chemistry and biologically damaging effects, is usually rapidly attenuated in seawater, but can penetrate tens of meters into these very clear waters (Osburne et al., 2010, Tedetti and Sempéré, 2006, Morel et al. 2007a, b), an unusual feature which may have consequences for the distributions and properties and adaptations of *Prochlorococcus* populations (Osburne et al., 2010).

The Oxygen Minimum Zone and *Prochlorococcus*: a remarkable habitat

The edge of the Eastern Tropical South Pacific has a permanent shallow oxygen minimum zone

(ETSP OMZ) off the coast of Chile and Peru. OMZ features occur when heavy upwelling or other nutrient input (e.g. agricultural) results in high primary productivity (photosynthesis), and subsequently high secondary productivity (respiration) (Wyrtki, 1962, Ulloa et al., 2012). As material sinks, sources of oxygen (air-water interface and photosynthesis in the euphotic zone) become spatially separated from respiration, drawing subsurface oxygen levels down to zero, even near the surface (Wyrtki, 1962, Ulloa et al., 2012). In most locations, with low or modest productivity, this basic process of surface light fueled primary production and subsequent sinking of biological material results in oxygen profiles that are somewhat reduced deep in the ocean, well below the euphotic zone, and never reach anoxia (Wyrtki, 1962). In strong OMZs, like the one in the Easter Tropical South Pacific, the anoxic region can reach so far toward the surface that it overlaps the euphotic zone, so that anoxic waters can contain organisms performing oxygenic photosynthesis, (Goericke et al., 2000, Ulloa et al., 2012). In these features, the chlorophyll profile over the euphotic zone sometimes takes on an unusual double-peaked form, one in the oxygenated well-lit zone, and a second, smaller peak, the secondary chlorophyll maximum, below the oxycline at very low light, and this peak is dominated by *Prochlorococcus* (Goericke et al., 2000, Lavin et al., 2010). Environmental DNA analyses from these populations have revealed the exciting presence of unique, as of yet uncultured *Prochlorococcus* clades in these secondary chlorophyll maximum populations, the LLV and LLVI clades, as well as other members from previously characterized *Prochlorococcus* ecotypes (Lavin et al., 2010, Astorga Eló, 2015). A distinctive anoxic microbial community lives here, very different from the usual *Prochlorococcus*-containing community and performing very different biogeochemical process than occur in oxygenated euphotic zones (Canfield et al., 2010, Stewart et al., 2011, Bryant et al., 2012, Ganesh et al., 2014). Why do we find *Prochlorococcus* in this unusual habitat? The hypotheses are numerous, and the answer likely complex, but for a start, eukaryotic grazers, and eukaryotic phytoplankton, cannot survive these waters, so it is a niche with reduced predation and competition (Goericke et al., 2000, Lavin et al., 2010). The fact that unique *Prochlorococcus* ecotypes are observed in these habitats (and nowhere else to date) suggests that these lineages could have adapted to the unique chemical conditions of the OMZ, including different forms of nitrogen sources (Goericke et al., 2000, Lavin et al., 2010, Astorga-Eló, 2015).

Studying *Prochlorococcus* with the aid of flow cytometry to answer first order questions

The expert macroecologist can identify the plant they specialize in at a glance based on its leaves, flowers or bark, easy to assess physical properties. This method does not work well for microbes, even with a microscope, so in most cases microbial ecologists rely on DNA sequence, in a variety of methodological forms, to identify and enumerate organisms of interest. As *Prochlorococcus* researchers, we are in some respects closer to the field botanist, in that we can look at seawater, with the aid of a powerful instrument known as a flow cytometer, and identify and count all the *Prochlorococcus* (Chisholm et al., 1988, Olson et al., 1990). Flow cytometry allows us to measure optical properties of individual particles, on the micron scale. The light scattering and fluorescence properties of *Prochlorococcus* are unique among all the particles in seawater; it is the smallest kind of chlorophyll-containing cell in the sea. The quantitative details of these scattering and fluorescence properties relate to the genetic diversity, the light acclimation state of the cells, and even mixing history (Moore et al., 1998, Urbach et al., 1998, Dusenbury 1999, Dusenbury et al., 2000, Crosbie et al., 2001).

In this work, we use flow cytometry to characterize how South Pacific *Prochlorococcus* populations change over these ecological gradients, from coast to gyre. Obtaining *Prochlorococcus* cell counts and characterizing individual cell fluorescence and light scattering properties, we consider basic questions about the light acclimation properties and genetic groups of samples collected across the transect. How does the abundance of *Prochlorococcus* change over these dramatic differences in ecological conditions? How are *Prochlorococcus* distributed over depth for regions with different light quality? We also examine the two extreme habitats this cruise gave us access to, the secondary chlorophyll maximum of the OMZ and the ultraoligotrophic waters of the gyre.

4.2 Materials and Methods

Cruise and sampling

All samples described here were collected as part of the CMORE BiGRAPA cruise, from November 18 to December 14, 2010. Detailed information on this cruise and access to ancillary data is available through: http://cmore.soest.hawaii.edu/cruises/big_rapa/plan.htm

For the flow cytometry analyses described here, water was collected from nine depths per station in CTD Niskin rosette casts with a depth sampling profile determined by the depths of the 10%, 1%, 0.1% light levels, the mixed layer depth and the chlorophyll maximum depth. Samples were taken at these features and in increments around them to evenly fill the depth profile, attempting to reach depths where *Prochlorococcus* populations started to thin at the base of the euphotic zone. We sampled from seven stations evenly distributed over the transect. We also collected two extra, high resolution casts from around the deep chlorophyll maximum at Station 7 (nearest Easter Island). In figures, Cast 1 (=Big RAPA Cast 57), December 8, 12:45, Cast 2 (=Big RAPA Cast 63), December 9, 18:20, Cast 3 (=Big RAPA Cast 69), December 10, 18:20.

From each of these water samples, aliquots were taken for qPCR, flow cytometry and single-cell genomics preservation methods. Metagenomic samples were taken in separate large-volume samples at the same locations, so they represent more approximate pairings. These samples collected on this cruise include viral and bacterial metagenome fractions and ecotype q-PCR, which are currently being analyzed by Libusha Kelly and Paul Berube. For flow cytometry samples, 1.0 ml of seawater was mixed (by shaking, not stirring) with 5 ul of 25% glutaraldehyde, a fixative commonly used for preserving *Prochlorococcus* cells, for a final concentration of 0.125%. Samples were then flash frozen in liquid nitrogen and stored in liquid nitrogen dewars until transport to the lab, then in -80 freezers, until time of analysis (within 1 year of sampling). Samples were preserved in cryovials to withstand flash freezing. Duplicate flow cytometry samples were taken from each water sample. Only one of these sampling replicates was run for these counts, because the analysis is destructive and our analysis was specifically focused on *Prochlorococcus* populations and different methods and instrument parameterizations would be required to view other populations, so we wanted to leave those options to other researchers.

The figures describing nutrient measurements and CTD data presented in Figure 4.2 and Figure 4.3 were produced through the cruise data repository: http://hahana.soest.hawaii.edu/cmorebigrapa/bigrapa.html

Flow cytometry data collection

Flow cytometry data was collected on a Becton-Dickinson/Cytopeia Influx Cell sorting instrument, using 10g NaCl per L MilliQ water as sheath fluid. Two lasers were used for excitations, a 488nm and 457nm aligned along the same path, an approach thought to maximize sensitivity to low chlorophyll samples. Gain to all photomultiplier tubes was set to optimize for sensitivity across the range of *Prochlorococcus* properties in these samples, and the same settings were used for all data collected. *Prochlorococcus* populations in many surface samples had very low chlorophyll autofluorescence, overlapping background readings from heterotrophic populations, a common phenomenon, even with the best instruments (e.g. Hartmann et al., 2014). The larger and brighter fluorescing *Synechococcus* were out of range on these sensitivity settings at some depths and stations, so we do not enumerate *Synechococcus* here. Two-micron Fluorescebrite YG fluorescent beads (Polysciences, Warrington, PA) were used as a standard, to maintain instrument alignment during data collection and to normalize cellular fluorescence values during analysis. These standards are larger and more fluorescent than *Prochlorococcus* cells, so they do not interfere with cellular signals, but are small enough to be visible in the same sensitivity settings appropriate for

Prochlorococcus. Where possible, samples were run long enough to collect data more than 10,000 *Prochlorococcus* cells, to reduce the Poisson counting noise to below 1%. For a few samples, however, there were not 10,000 cells in our 1ml sample, so we ran as much volume as possible. All samples were run unstained. Data collection was triggered off forward scatter signal. Two technical replicates were performed for each sample, running the same seawater twice through the flow cytometry data collection process. Samples were all run undiluted, with 5 or 10 ul of concentrated 2um fluorescent beads added to as much of the 1ml samples as could be recovered from the cryovials (950-990ul). This data was analyzed in FlowJo, extracting counts, chl/cell and fsc/cell information. Analysis was performed in the FloJo software package, gating *Prochlorococcus* counts on their chorophyll (488ex, 695/40 bandpass em) and forward small angle scatter properties, keeping an eye on the phycoerythrin channel (488 ex, 580/30 em) to make sure *Prochlorococcus* and *Synechococcus* populations were not mistaken.

Analysis of PAR data

The photosynthetically active radiation (PAR) data used for analysis of light environment gathered on this cruise presented a challenge in analysis. There were two sensors, one on the CTD and one on the ship. The Surface PAR sensor always reads higher than the CTD PAR sensor, and there is no measurement with both instruments at the same place, the CTD only collected data under water, Surface PAR only on the ship. Normalizing CTD data to near-surface data sometimes resulted in values slightly greater than 1. This could be explained by a small mismatch in calibration. We attempted several normalization variants, but in the end settled on simply presenting the data as is. The same sensors were used for all data, and all measurements occurred within the span of one month, so they should be comparable across samples. The disparity between CTD 1m measurements (closest to surface measurements we have) and Surface PAR are on the order of 15%, smaller than the order of magnitude variations of light discussed.

Data sourcing for figures from major databases

For introductory figures describing region, data was downloaded from the Aqua/MODIS satellite data repository, and map rendered with sampling locations using MATLAB; data accessible here: http://oceancolor.gsfc.nasa.gov/cms/. For global KdPAR map, image was downloaded from the Global Ocean Color database (http://www.globcolour.info/data_access_full_prod_set.html), representing an amalgamation of data from many satellites, for the month of December, 2010 around the time of our sampling.

Single cell sorting and ITS sequencing

Single cells were sorted under clean conditions as described in (Rodrigue et al 2009), using Epicenter phi29 reagent kits and random hexamer primers with a thiophosphate 3' modification from IDT. We sorted and amplified the DNA of single cells from the South Pacific using the method described in Rodrigue et al., 2009 and Zhang et al., 2006. We sorted of one 2ml sample (several replicates remain archived) of cells preserved in 10% glycerol collected at Station 1, 55m - the secondary chlorophyll max where the euphotic zone overlaps the oxygen minimum zone. To identify single cells of interest, we PCR amplified and sequenced the ITS rRNA marker gene from our single cell amplified DNA libraries, as described in Rodrigue et al., 2009. Because initial rounds resulted in low yields of positive sequences, we performed additional reactions with 1:10, 1:100, 1:1000 dilutions of MDA products, which improved overall yield somewhat, and we used cyanobacterial primers from both Rodrigue et al., 2009 and Iteman et al., 2000, which bind slightly different conserved regions outside the ITS. PCR reactions were performed using Phusion polymerase (NEB). Sequencing was performed at the Massachusetts General Hospital DNA core facility, Cambridge, MA.

4.3 Results and Discussion

4.3.1 The transect

We sampled the Eastern South Pacific with the Center for Microbial Oceanography Research and Education's BiG RAPA cruise (Biogeochemical Gradients: Role in Arranging Planktonic Assemblages) in the early austral summer of 2010, from Nov. 18 to Dec 14. We traveled from east to west, and slightly south, spanning 3,500km from the coast of Chile near Iquique to Easter Island, sampling at seven approximately evenly spaced stations spanning the dramatic biological, chemical and physical gradients described above, from Station 1 in the east to Station 7 in the west (Figure 4.2).

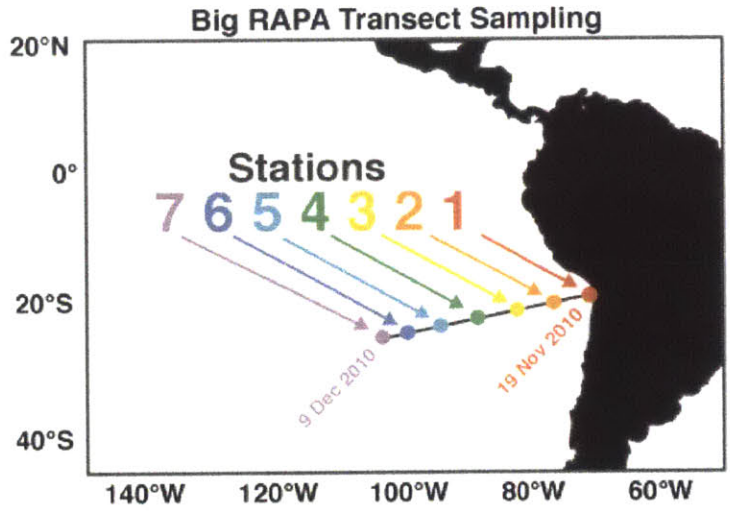


Figure 4.2. Transect sampling and Stations
Sampling order, locations and reference numbers for the stations from which our samples came, across the eastern South Pacific. Station 7 is near Easter Island, Station 1 is near Iquique, Chile. Refer to Figure 1 for productivity context - the map and positional dots are identical in that figure. Refer to Figure 4.3 for physical and chemical properties of these stations. The station colors here are used for marking data from the corresponding station throughout this chapter.

Based on measurements taken during the BiG RAPA transect, in parallel with our *Prochlorococcus*-sampling, temperature increased towards the gyre, while chlorophyll decreased and moved deeper into the water column (Figure 4.3A and B). From the combined nitrite+nitrate, phosphorus and chlorophyll measurements (Figure 4.3C, D, and E), it is clear that Station 7 is the most oligotrophic station, Stations 4, 5 and 6 are also oligotrophic, Station 3 is perhaps transitional, Station 2 is mesotrophic and Station 1 is eutrophic (Figure 4.3b, Figure 4.3D, Figure 4.3E, BiG RAPA data repository, 2011). The intense OMZ at Station 1 can be observed in the very low oxygen concentrations that occur just below the surface, and by comparison with nitrogen and phosphorus data in the same location, the high nutrient concentrations driving this phenomenon (Figure 4.3C, Figure 4.3D, Figure 4.3B). These features are largely consistent with measurements from the best existing comparable dataset previously taken for this region during the BIOSOPE expedition of 2004, which sampled at the same time of year, but along a transect from Easter Island to Concepción, Chile, 2,000 km South of Iquique, sampling a more subtropical slice of this ocean (Claustre et al., 2008b).

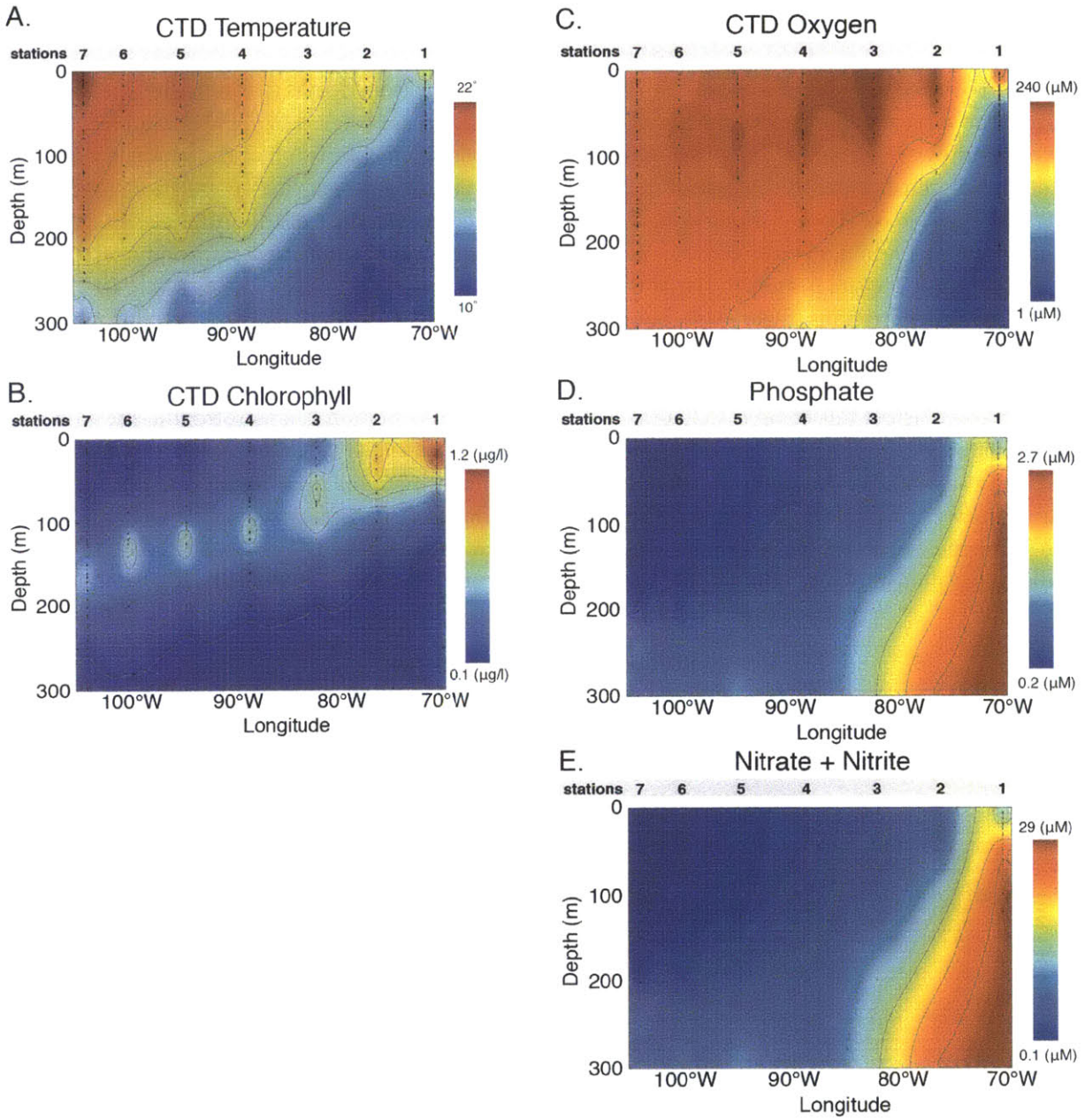


Figure 4.3. Basic characterization of the BiG RAPA transect
CTD refers to the "Conductivity, temperature, depth" instrument which surveys the water column in conjunction with sample collection, resulting in measurements of temperature, chlorophyll fluorescence (calibrated to concentration), and dissolved oxygen. These measurements were taken on different sampling casts than our samples for biological analysis in the following sections, but they were taken from the same locations and usually on the same day. The depth scales for all these plots (0-300m) approximately represent the relevant range for the *Prochlorococcus* habitat, plus a bit deeper. The phosphate and nitrate+nitrite concentrations are from analytical measurements. All of this data and the plot rendering was accessed through the BiG RAPA data repository described in Materials and Methods.

4.3.2 *Prochlorococcus* abundances over geography and depth over a South East Pacific transect

We collected samples of seawater for *Prochlorococcus* analysis from depths spanning the euphotic zone, from the surface to the base of the phytoplankton population as observed by CTD chlorophyll traces. Sampling depths were determined not by a fixed depth scale but by a combination of the irradiance levels, and chlorophyll signals to reach the base of the euphotic zone and of phytoplankton populations. We enumerated *Prochlorococcus* populations using flow cytometry, describing their distributions across space and depth (Figure 4.4). Overall, *Prochlorococcus* concentrations were high across the transect, except at stations 1 and 3. At Station 1, the site nearest the coast with high-nutrient waters and an OMZ, there were overall low *Prochlorococcus* concentrations (Figure 4.4), but we did observe *Prochlorococcus* in a secondary chlorophyll maximum below the oxycline (see Figure 4.14), the unusual OMZ-*Prochlorococcus* feature reported previously (Goericke et al., 2000, Lavin et al., 2010). The highest concentration of *Prochlorococcus* in a single water sampled observed across this transect (330,000 cells ml^{-1}) occurred at Station 2, a mesotrophic site with a relatively shallow euphotic zone (Figure 4.1). *Prochlorococcus* concentrations at Station 3 were very low. Moving into the gyre, *Prochlorococcus* populations were consistently high, reaching deeper into the water column as the water became more oligotrophic to the west. The deepest samples we took were from 250m at Station 7, the most oligotrophic station, where a distinct *Prochlorococcus* population was observed (1,000 cell/ml), deeper than at typical oligotrophic sites (Johnson et al., 2006, Malmstrom et al., 2010).

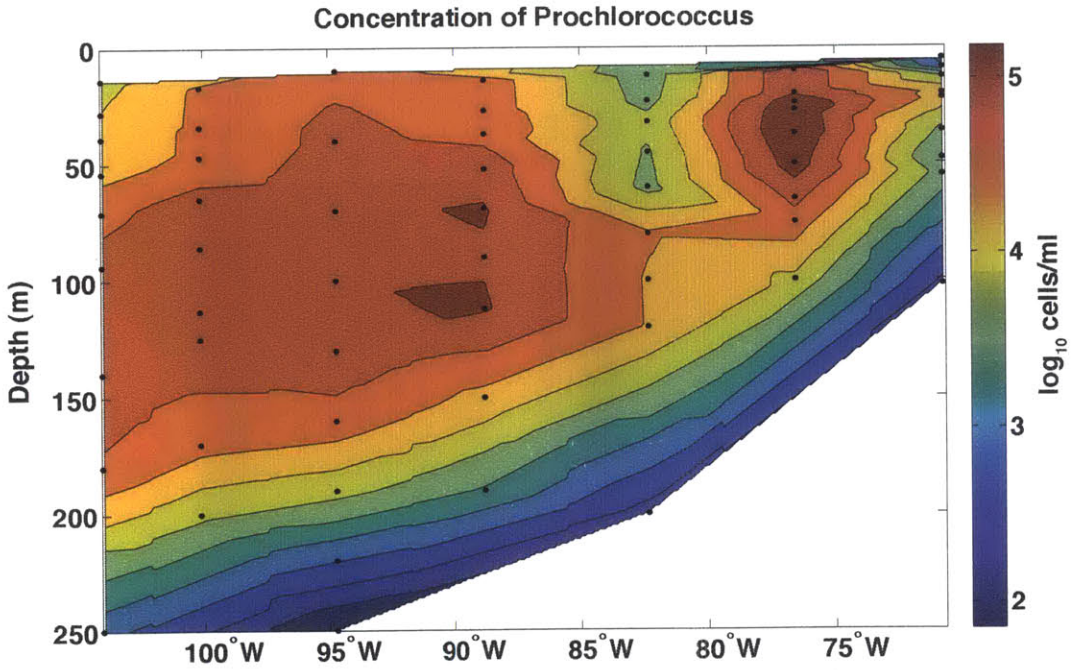


Figure 4.4. ***Prochlorococcus*** **abundances across a South Pacific Transect**
Concentration of *Prochlorococcus* cells, measured by flow cytometry, over depth and over the BiG RAPA transect. There was a small north-south component to this sampling as well (see Table 4.1 for coordinates). Interpolated data, 9 depths per station, from Station 1 at right, in the oxygen minimum zone, to Station 7 at left, in the South Pacific Gyre. Black dots represent samples.

Stations 1-3 are snapshots taken hundreds of miles apart from a complex oceanographic region, the Humboldt current, full of complex eddies and upwelling events changing nutrients and temperature conditions on the scale of tens of kilometers; each of our points is taken at random out of this patchy distribution, which could explain why we see high *Prochlorococcus* concentrations at Station 2 and low concentrations at Station 3 (Shaffer et al., 1995, Claustre et al., 2008b). The very low concentrations of *Prochlorococcus* at Station 1 are consistent with global observations that *Prochlorococcus* are not found at high abundances immediately off coasts, a phenomenon that is not fully understood (Biller et al., 2015, Flombaum et al., 2013). Stations 4-7 come from a more stable, spatially homogenous environment, so each sample is more likely representative of their surrounding waters and together they are more likely to describe true regional gradients (Claustre et al., 2008b). We displayed the data as interpolated in Figure 4.4 to emphasize the general coast to gyre trends, but this is an extremely coarse spatial sampling scheme, and many other features in *Prochlorococcus* distribution may be occurring between samples over this transect. This transect took a month to traverse - all the samples were not collected at the same time - but we believe these patterns largely represent changes over space, at least for open ocean samples, because conditions are stable on a month time scale in the open ocean gyre (see Supplement Figure 4.1), though for the same underlying reasons of dynamism in the Humboldt, this is not necessarily the case for Stations 1-3.

Our concentration measurements were in many ways similar to *Prochlorococcus* counts from the 2004 BIOSOPE expedition (a South Pacific transect, further south) - we both observe deep populations in the gyre and low coastal concentrations, but we generally report higher surface *Prochlorococcus* populations (Grob et al., 2007). This could be a feature of the region, but surface *Prochlorococcus* populations are notoriously difficult to count (Hartmann et al., 2014), so measurements using identical methods would be required to determine the nature of the longitudinal gradients in *Prochlorococcus* abundance in this region.

The environment of South Pacific *Prochlorococcus*: a geographic gradient in water clarity

Prochlorococcus abundance is dictated by many interacting factors in a water parcel including physical and chemical properties, history and biological community, together determining the cell division rates and mortality. The most dramatic trend in the *Prochlorococcus* distributions observed along the transect (Figure 4.4) is the increasing depth of *Prochlorococcus* populations towards more oligotrophic sites; this led us to better understand the relationship between depth and light in these water columns. Light is a very important factor, probably the most important, driving the evolution and environmental distributions of *Prochlorococcus* (Rocap et al., 2003, Biller et al., 2015, Zinser et al., 2007). The relationship between light and depth changes from the rich coastal waters (Station 1-2) to the hyper-oligotrophic, clear waters of Station 7 (Figure 4.5, Morel et al., 2007). The light environment of our samples spans a wide range of extinction coefficients (Table 4.1). Ultimately, water clarity and color vary because particulate matter (biologic in origin, e.g. plankton; or abiotic, e.g, dust) and dissolved material (colored DOM; e.g. biologic breakdown products) influence the way light travels through the water column (Morel et al., 2007, Stramski et al., 2008, Falkowski and Raven, 2007). In productive regions, like the chilean coast, water is murky - light only penetrates a little. In the clear waters of the deep blue sea, light travels far. Since the materials that create turbidity (particles, colored dissolved organic matter and phytoplankton) usually change light quality as well as light quantity, we have a very different light-environment at the extremes of this transect.

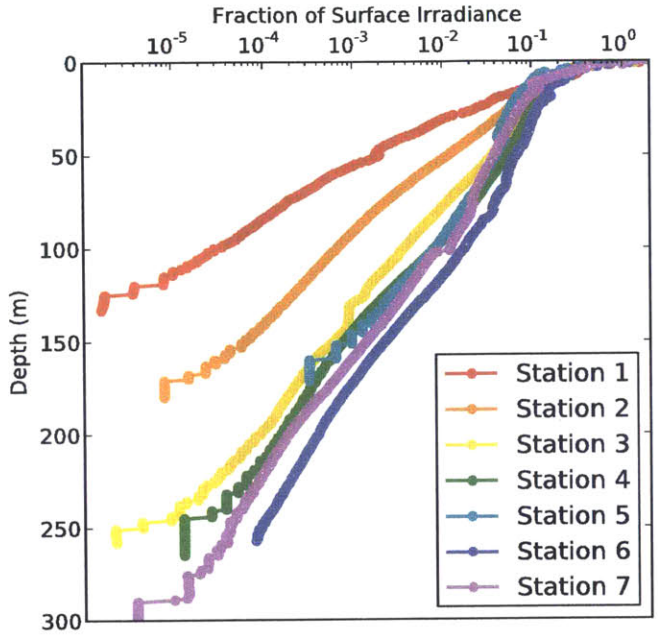


Figure 4.5. Fraction of surface irradiance as a function of depth along the transect.
The coastal station has much higher extinction coefficient than the deeper waters. Behavior at the surface is complex, showing the furthest deviation from simple exponential extinction, and simplifies below 10m, showing steeper slopes for clearer waters. The CTD PAR/ Surface PAR ratio at the surface is not unity (see Methods). Station 5 was measured during low absolute light (cloudy), so light the extinction could not be measured as deep as for other samples.

To help place our *Prochlorococcus* measurements the oceanographic context of each station, we summarized (i) the photosynthetically active radiation extinction coefficients (K_d(PAR)), which describe the attenuation of light with depth, (ii) mixed layer depths describing the depth to which surface waters are vertically homogenized by turbulent mixing, and (iii) the depth of the chlorophyll maximum, based on total chlorophyll fluorescence from all phytoplankton for each site, (Table 4.1). As we move from coast to gyre, the mixed layer depth increases, K_d(PAR) decreases (slower attenuation of light with depth) and the depth of the chlorophyll maximum increases (Table 4.1).

Table 4.1. Basic Properties of water column at each sampling station across the South East Pacific transect

Station	Coordinates	Mixed Layer Depth[1]	k_d PAR light extinction coefficient[2]	Chlorophyll Fluorescence Maximum[3]
1	20° 05' S, 70° 48' W	3m	0.091 m^{-1}	24m
2	21° 11' S, 76° 34' W	20m	0.057 m^{-1}	34m
3	22° 16' S, 82° 21' W	23m	0.041 m^{-1}	61m
4	23° 28' S, 88° 46' W	39m	0.04 m^{-1}	114m
5	24° 34' S, 94° 43' W	28m	0.036 m^{-1}	126m
6	25° 33' S, 100° 01' W	47m	0.034 m^{-1}	118m
7	26° 15' S, 103° 58' W	50m	0.036 m^{-1}	177m

[1]Mixed layer depth is the last depth before temperature decline begins (rough estimate). [2]KdPAR calculated by fitting an exponential curve to relationship between depth and the CTD PAR data normalized to a surface PAR sensor aboard the ship. [3]Chlorophyll max is depth at maximum chlorophyll fluorescence, except where single point outliers did not match trend; at some sites the shape of chlorophyll peak was somewhat irregular. These measurements come from the casts samples were taken from. *Station 5 measured at low absolute light, limited depth.

The K_d(PAR) values for the water columns we sampled are typical of the region (Table 4.1, Figure 4.6). Values of K_d(PAR) 0.06 m^{-1} and below are generally considered oligotrophic (Saulquin et al., 2013, Morel et al., 2007b); Stations 3-7 are below 0.06 m^{-1}, Station 2 is close to 0.06 m^{-1}, Station 1 well above, consistent with descriptions trophic status above based on nutrient and chlorophyll data. Stations 5, 6 and 7, have some of the lowest K_d(PAR) values, or clearest waters, in the world (Figure 4.6). This pattern in light attenuation, combined with surface nutrient depletion, may explain the unusually deep *Prochlorococcus* populations we observed at Stations 5, 6 and 7 - there is still light at 250m (1/10,000 of surface light, Figure 4.5). Although most of the populations are concentrated higher in the water column, at 250m nutrient concentrations are beginning to rise (Figure 4.3D and 4.3E), perhaps providing a benefit to cells capable of gathering sufficient energy from the scarce photons available in that habitat.

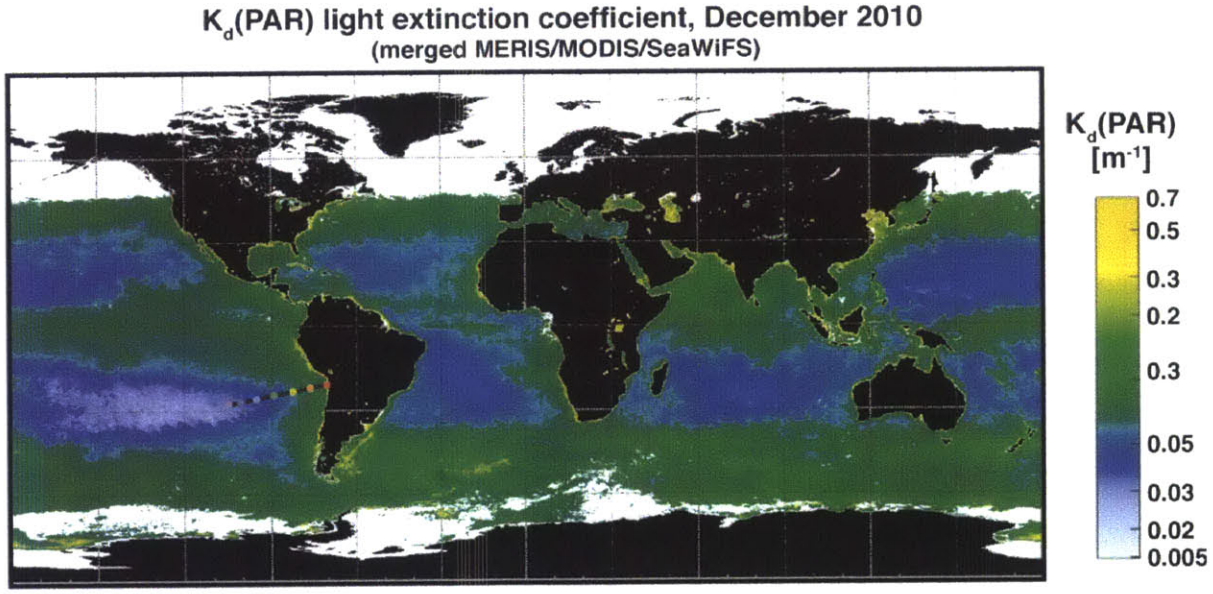


Figure 4.6. Satellite Kd(PAR) values across the global oceans
Global ocean estimates of Kd(PAR), the diffuse attenuation coefficient for photosynthetically active radiation, averaged over the month of December 2010 (roughly during the BiG RAPA Cruise). Map obtained from the Global Ocean Color Database, which integrates data over several satellites (see Materials and Methods). The South Pacific has the lowest Kd(PAR) values, clearest waters, in the world. BiG RAPA transect is marked. Note that the BiG RAPA transect does not extend all the way to the most oligotrophic waters. The gyre continues for the next few thousand kilometers west, getting even more oligotrophic at the center.

Distribution of *Prochlorococcus* populations over depth as a function of light intensity

The relationship between depth and light varies across our transect as extinction coefficient varies (Table 4.1, Figure 4.5). The distribution of *Prochlorococcus* with depth, the shape of the curve, maximum abundance and deepest occurrence, is complex. Stations 4, 5, 6 and 7 have similar profiles, but Stations 1, 2 and 3 are each unique (Figure 4.7A). Examination of the relationship between cell abundance and light intensity (Figure 4.7B) reveals that these populations thrive in waters spanning a full 4 orders of magnitude of light, enabled by diversity in light harvesting strategies across these populations (Moore et al., 1998, Paul Berube, BiG RAPA ecotype qPCR data, personal communication). This highlights that the traditional operational definition of the euphotic zone as above the 1% light level (Falkowski and Raven,

2007) does not describe the full range of where phototrophs live, especially for *Prochlorococcus*-dominated oligotrophic environments. Light penetrates much farther than the 1% light level, and some phototrophs, including *Prochlorococcus*, can live below that 1% light level (e.g. Zinser et al., 2007, Figure 4.7B). The peaks in *Prochlorococcus* abundance over depth range from the 5% (Station 1, 2) to the 0.05% light level (Station 7). So, the peak and range of *Prochlorococcus* abundance over the water column is not only a function of depth or light, but is a complex feature of the environment and community, an emergent property of growth rate, mortality and history at every point. These rates depend on light, nutrients, temperature, mixing, chemical toxicity, predation by phage and grazing, and possibly other unmeasured factors.

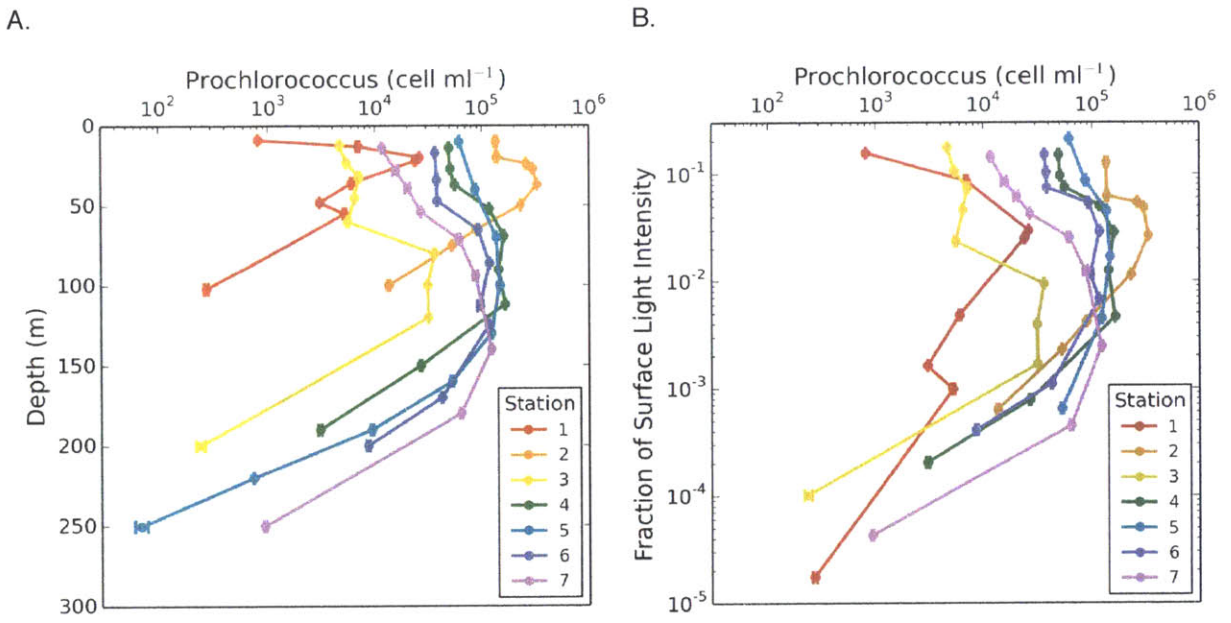


Figure 4.7. Comparing distribution of Prochlorococcus abundances over depth and light
Prochlorococcus concentrations station by station (1 = OMZ, coast, 7 = farthest into the gyre), in relation to depth (A) and light (B). The profiles at Stations 1 and 2 are very shallow; at farther stations cells reach deeper into the water column, with *Prochlorococcus* maxima between 100 and 150m. Depth refers to CTD dbar of pressure, and fraction of surface light intensity is readings from a CTD PAR sensor at depth normalized to surface par sensor aboard the vessel.

4.3.3 *Prochlorococcus* individual cell characteristics

In addition to counting *Prochlorococcus*, flow cytometry lets us look at the simple optical properties of cells; the mean chlorophyll fluorescence and light scatter properties of a population gives us information about the chlorophyll content and size of the cells respectively (Figure 4.8, Figure 4.9). The chlorophyll fluorescence per cell of the *Prochlorococcus* populations increases with depth (or decreasing light in the water column), through a combination of the cells' photoacclimation strategy – expanding the photosynthetic antennae with more chlorophyll to gather more light – and, at the population level, the ecotypic shift from HL to LL *Prochlorococcus*, which display more chlorophyll fluorescence per cell generally (Binder et al., 1996, Moore et al., 1998, Chisholm et al., 1988). These principles can be easily observed in the depth profiles: cells in

the deeper samples have more forward light scatter and chlorophyll fluorescence per cell (Figure 4.8, Figure 4.9). For the most part, this is a simple function of light - all the complexity of the population numbers data presented above falls away. The per cell chlorophyll fluorescence shows a beautiful progression across the transect as we move away from the coast, the depth at which a given value of chlorophyll fluorescence occurs moves deeper (as the water becomes clearer).

However, not all stations follow the same relationship between chlorophyll fluorescence per cell and light (Figure 4.9). The different environments and different properties or responses of cells in these environments result in more chlorophyll fluorescence at a given light level for onshore stations compared to clear waters. Stations 2, 3, 4, 5 and 6 largely follow one pattern (Figure 4.9 b), and Stations 1 and 6 are outliers. One possible hypothesis this invites is that light color preferences might make PAR an inappropriate measure for the bioavailable light the cell actually receives. The differences could also reflect very different genetic populations with different pigments and acclimation strategies, or populations with different acclimation histories content.

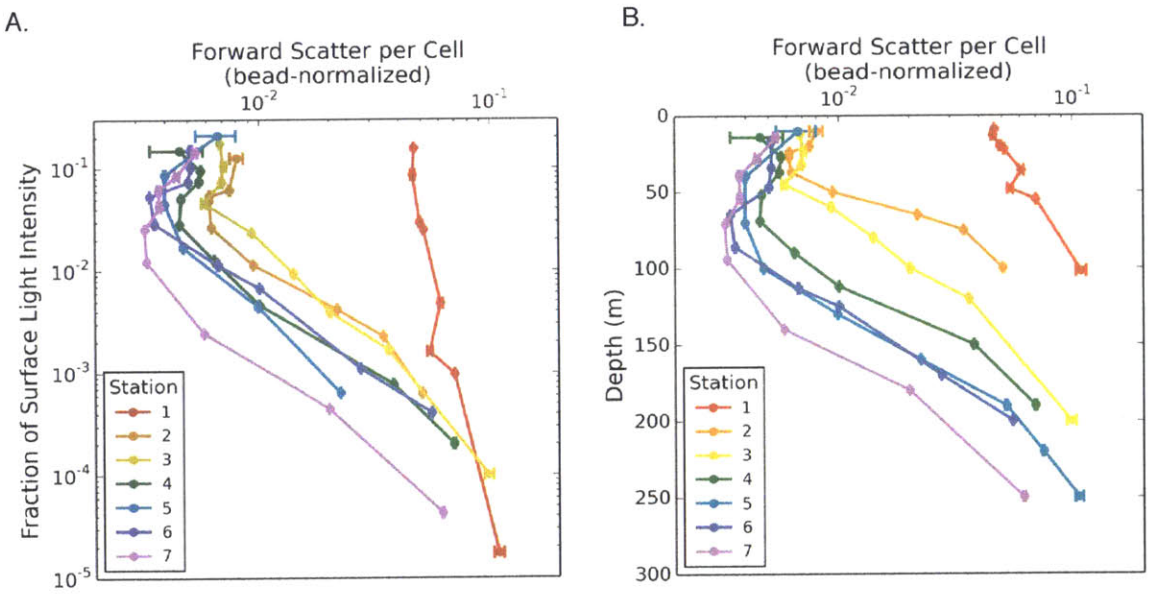


Figure 4.8. ***Prochlorococcus*** **light scattering as a function of depth and light**
Forward angle scatter is related to cell size and other properties of cell contents and pigmentation. In *Prochlorococcus*, low light adapted ecotypes have much higher forward scatter than high light adapted, and this parameter also changes with acclimation, increasing with acclimation to lower light like chlorophyll, but to a lesser extent.

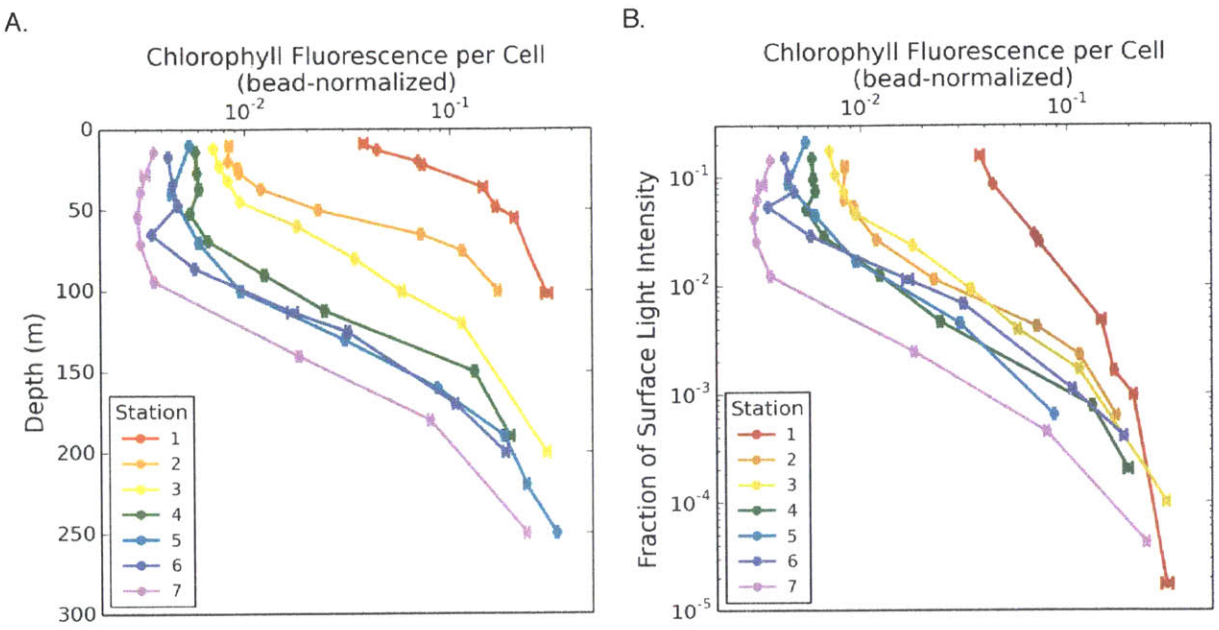


Figure 4.9. Mean chlorophyll fluorescence per cell as a function of depth, and fraction of surface light intensity. *Prochlorococcus* chlorophyll fluorescence per cell increase linearly with depths (A), at dramatically different depths as we move from turbid coast (Sta1) to clear gyre waters (Sta7). Two replicate flow cytometry runs for each station are plotted against the ratio of CTD measurements of PAR and Surface PAR (uncorrected; approximately fractional irradiance). Chlorophyll fluorescence per cell is not merely a function of irradiance (B).

Contrasting communities and changing *Prochlorococcus* populations

The *Prochlorococcus* populations and whole microbial communities at the oligotrophic (e.g. Station 7) and mesotrophic sites (e.g. Station 2) are different in many ways. A few of these differences can be appreciated by looking at the whole seawater in flow cytometry space, beyond the simple count data above (Figure 4.10). The most oligotrophic, clear water sites have *Prochlorococcus* populations with barely detectable chlorophyll fluorescence in the surface mixed layer - making a living off photosynthesis with very little chlorophyll and lots of light (Figure 4.10, Station 7). At oligotrophic stations, *Prochlorococcus* is abundant and there are no other large populations of phototrophs in its size class, only small populations of *Synechococcus* and picoeukaryotes (Figure 4.10, Station 7). By contrast, at Station 2, the *Prochlorococcus* populations are surrounded by other chlorophyll-containing populations, again, in the size class surveyed with our instrument, *Synechococcus* and picoeukaryotes, but this time at high concentrations (Figure 4.10, Station 2). The very different chlorophyll fluorescence of *Prochlorococcus* cells in the mixed layer at these two sites (bright cells at Station 2, barely detectable cells at Station 7) could be a matter of genetic differences or it could perhaps be explained by differences in light spectrum, because *Prochlorococcus* is less able to use green light - so for equivalent PAR (defined as all light 400-700nm), water with more yellow, green or orange light (typical of mesotrophic sites) would appear dimmer to the *Prochlorococcus* photosystem than the pure blue water (of the most oligotrophic sites to which the organism is adapted), and thus would require more chlorophyll for optimized growth). Detailed genetic work could explore whether light quality differences may be responsible for the different distributions of cells here.

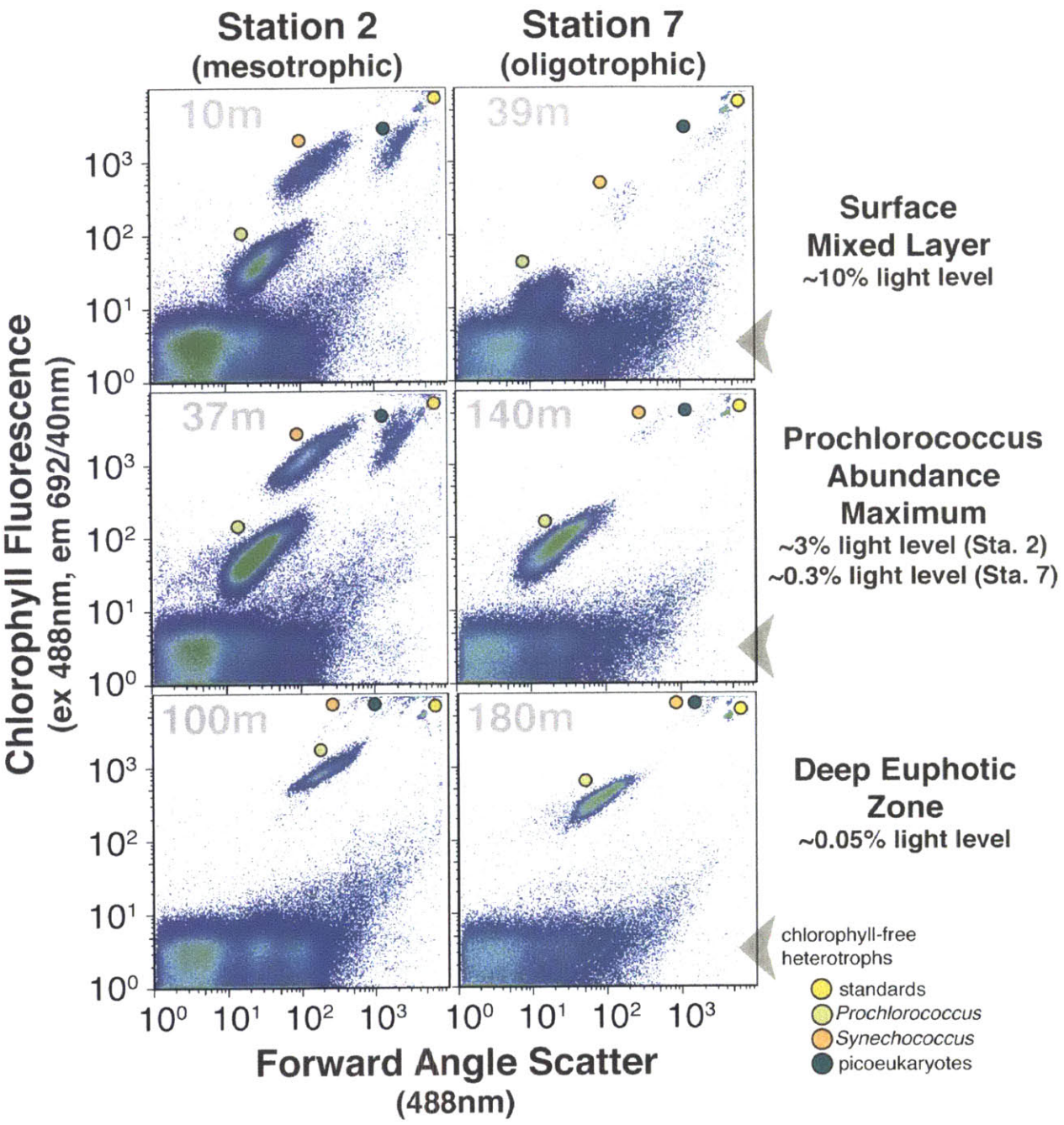


Figure 4.10. Seawater by flow cytometry: contrasting communities and picophytoplankton at two stations
Seawater viewed on a flow cytometer: forward scatter (also called small angle scatter) is roughly proportional to size (although other material features of particles can influence it as well, and shapes complicate it). These dot plots represent all data collected for each sample (0.25-0.8 ml of seawater), stacked and colored by relative abundance at each position.

4.3.4 High resolution sampling over depth in the middle of a chlorophyll maximum

For the oligotrophic environment of Station 7, with *Prochlorococcus* populations stretched over a spatially elongated euphotic zone, we sampled additional high resolution casts. Our primary goal was to assess the fine transitions in *Prochlorococcus* cell populations and properties at this site, but this repeated and higher resolution sampling also allowed us to evaluate the spatial and temporal nature of our sampling approaches. We took samples from three separate casts from Station 7, the one shown in data above (at low resolution over the full euphotic zone), and two additional ones, sampling at higher resolution around the deep chlorophyll maximum. We can watch acclimation happen, as cells gain increasingly more chlorophyll fluorescence per cell with depth (Figure 4.11, Figure 4.13), and we can probably also see the transition between HL and LL ecotypes in these profiles (Figure 4.11, Figure 4.13). Based on past genetic characterization of flow cytometric populations of *Prochlorococcus*, bimodal or complex scatter plot *Prochlorococcus* populations represent the transition from HL ecotype-dominated *Prochlorococcus* populations to LL ecotype-dominated populations, although both ecotypes are present in varying amounts throughout most water columns. With this high resolution sampling, we can see this transition happen between 120 and 175m where populations become more complex in their shapes (Figure 4.11), and where a sharp transition occurs in chlorophyll fluorescence per cell (Figure 4.13) consistent with ecotype transition; in this case we have molecular measurements of these ecotypes to confirm this transition depth (PM Berube, personal communication).

These repeated, high-resolution samples also allowed us to address two major issues in the nature of oceanographical sampling, consequences of the movement of water: space and time. First, we can assess the stability of sampling at the same geographic site over three days. All parameters measured were relatively close across the three days, with a few exceptions (Figure 4.12, Figure 4.13). This was relatively comforting for the larger interpretation of our data. The spatial component comes with depth: how many depth samples should we be taking to build a good description of a *Prochlorococcus* profile? The high resolution cast would be a valuable approach for capturing ecotype transitions without the aid of molecular tools, which could be applied for any science questions that would benefit from ecotype-specific sampling at sea or from flow cytometry populations (e.g. targeted isolations, targeted metatrascriptomics, single cell genomics). This transition is viewed best in the chlorophyll fluorescence per cell plot of Figure 4.12, where the high resolution cast (Cast 2) shows the sharp transition in chlorophyll marking the ecotype transition, which is lost in the lower resolution sampling (Cast 1). This high resolution spatial sampling also reveals spatial inhomogeneities - there are wobbles, on the scale of 5 meters, in the *Prochlorococcus* concentrations over depth, which we would never have observed through typical sampling. The open ocean habitat may be generally mixed and homogenous by some measures and on some scales, but there is still a great deal of complexity on fine scales that we do not fully understand.

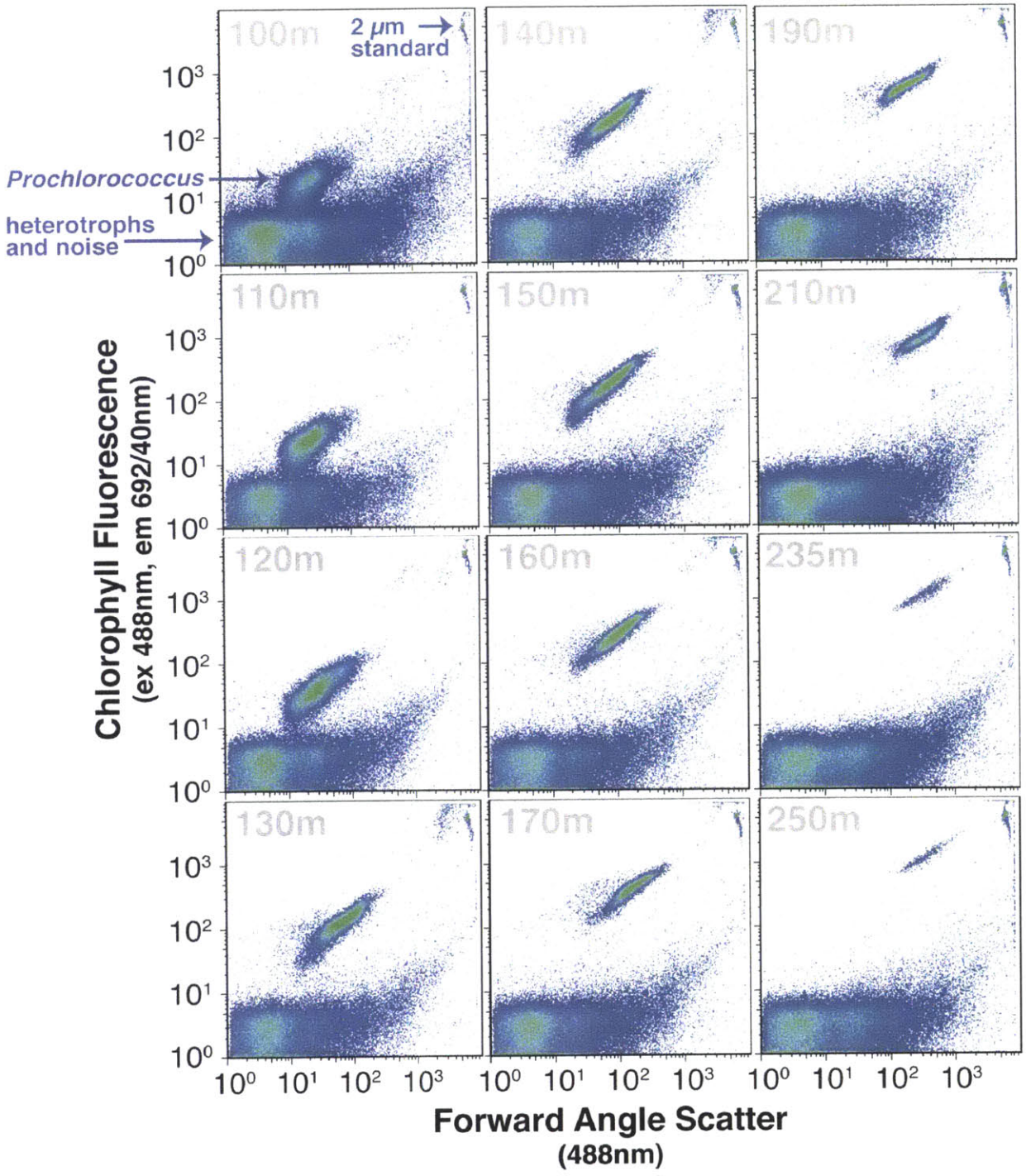


Figure 4.11. Transitions with depth in *Prochlorococcus* populations, viewed at high resolution

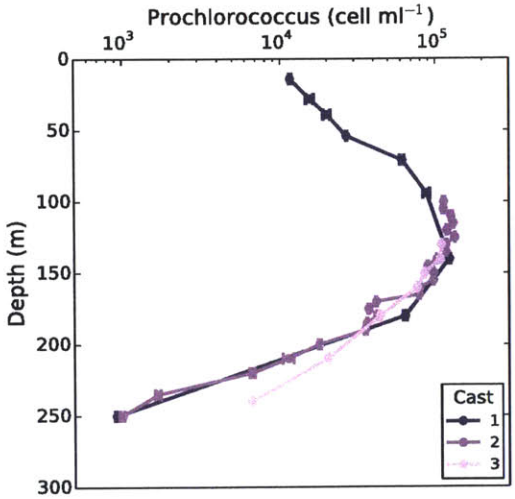


Figure 4.12. ***Prochlorococcus*** **abundance measurements from profiles from three days at the same site**
Station 7
Cast 1 - December 8, 12:45, data shown above, collected the same was as for all stations.
Cast 2 - December 9, 18:20, high resolution sampling around chlorophyll peak.
Cast 3 - December 10, 18:20, repeated sampling of chlorophyll peak.

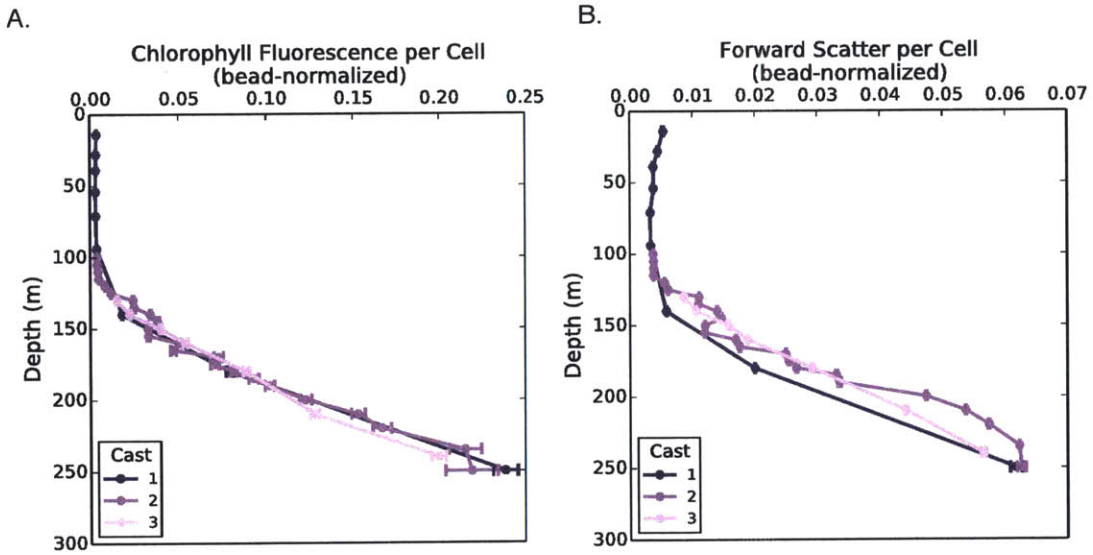


Figure 4.13. Profiles Station 7, over 3 days: chlorophyll fluorescence and forward scatter per cell
Plotted on a linear scale to show linear relationship between these parameters and depth for populations below 100m (see supplemental Figure S4.2 for log version showing surface features). Values represent means for whole *Prochlorococcus* population in each sample, error bars represent range of duplicate technical replicates.

4.3.5 *Prochlorococcus* in a secondary chlorophyll maximum in the oxygen minimum zone

The oxygen minimum zone *Prochlorococcus*

At Station 1 we observed *Prochlorococcus* in a secondary chlorophyll maximum below the oxycline in the OMZ (Figure 4.14). This is a unique *Prochlorococcus* habitat, observed in other OMZs and in this region previously (Goericke et al., 2000, Lavin et al., 2010). These OMZ secondary chlorophyll features are defined based on bulk CTD chlorophyll fluorescence, which has a complex relationship to biomass, cell number, species, pigment per cell, and photosynthetic rates (Huot et al, 2007). As chlorophyll fluorescence per cell increases with depth due to acclimation, fewer cells contribute more to the bulk chlorophyll signal at the base of the euphotic zone (Figure 4.14A).

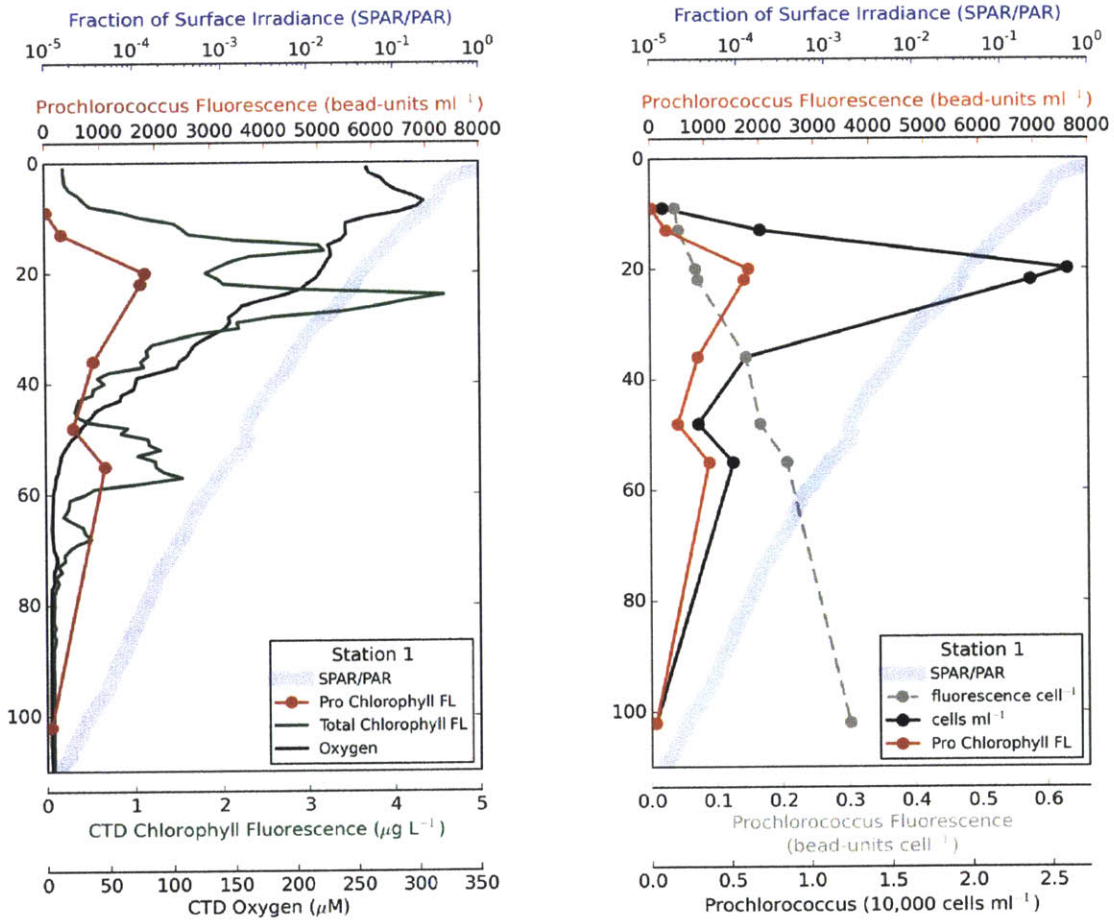


Figure 4.14. ***Prochlorococcus*** **contributes to the secondary chlorophyll maximum in the OMZ at Station 1.** [A] CTD measurements show a peak of chlorophyll fluorescence at 50-60m, below the decline in O_2 concentration. To show the *Prochlorococcus* populations in the context of this limited oceanographic measurement, we use the product of *Prochlorococcus* cell concentration and the *Prochlorococcus* chlorophyll fluorescence per cell to create a *Prochlorococcus*-fluorescence unit which is conceptually comparable to bulk CTD total chlorophyll fluorescence. Although the units are unrelated, this allows us to see the relative fluorescence values of the *Prochlorococcus* secondary peaks at the same depth as the OMZ secondary chlorophyll max. Flow cytometry measurements indicate a peak at the same depth in *Prochlorococcus*' fluorescence contribution (red). [B] The *Prochlorococcus*-specific chlorophyll depth profile is calculated as the product of the cell density (black) and the per-cell fluorescence (gray dashed).

Single cell sorting of *Prochlorococcus* cells in the OMZ

We wanted to know whether our samples collected from the oxygen-minimum zone off the coast of Chile contain representatives of uncultured *Prochlorococcus* clades previously reported to inhabit this region (Lavin et al., 2010). Ultimately we would like cultures from these clades, but another strategy to learn more about these in the future will be to obtain partial genome sequences from these clades, through single cell sequencing, to explore how they differ from cultured *Prochlorococcus* cousins, and look for adaptations to life in the oxygen minimum zone. In an attempt to do this, we flow cytometrically sorted several hundred individual *Prochlorococcus* cells (Figure 4.15), copied their DNA using multiple displacement amplification and PCR amplified and sequenced their 16S-23S rRNA Internal Transcribed Spacer region (Figure 4.16). Some promising initial results from this sequencing (tree below) showed that we indeed captured one of the uncultured OMZ-associated clades, the LLV (Figure 4.16). Unfortunately, whole genome sequencing on these particular cells failed, so this sorting effort will have to be repeated to address the question of what LLV genomes contain, although at least we know now that these samples contain the cells of interest. Many of the cells in this sample come from the LLI clade, which is of particular interest to the work in Chapter III of this thesis. Sequencing for most of these cells succeeded, and they are beginning to contribute to our understanding of evolution in this clade.

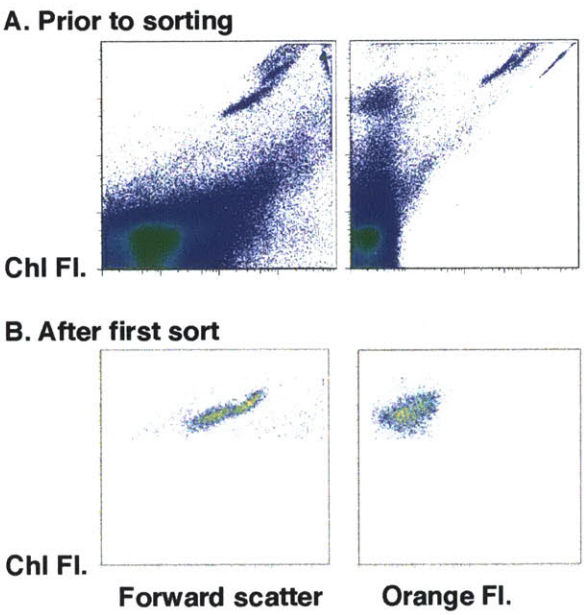


Figure 4.15. Single-cell sorting from the OMZ secondary chlorophyll maximum

A. Flow cytometry from standard glutaraldehyde preserved samples shows a distinct *Prochlorococcus* population in the secondary chlorophyll max, as well as two *Synechococcus* populations and likely eukaryotes. B. Midway through sorting our single cells, separated from the *Synechococcus* based on the absence of phycoerythrin, this snapshot shows how in the sorted, glycerol preserved sample the *Prochlorococcus* population appears to be composed of two overlapping populations. There were not enough cells in this sample to afford sorting them separately, but it is possible, given the results of genetic analysis, that these represent HL and LL ecotypes. Forward Scatter refers to forward angle scatter, a rough proxy for size. Orange Fl. refers to 580/30 fluorescence (488nm excitation), which shows phycoerythrin contents of *Synechococcus* cells, but nor *Prochlorococcus*. Chl FL. y-axis for all plots show chlorophyll fluorescence, 680/40 emission, 488nm excitation.

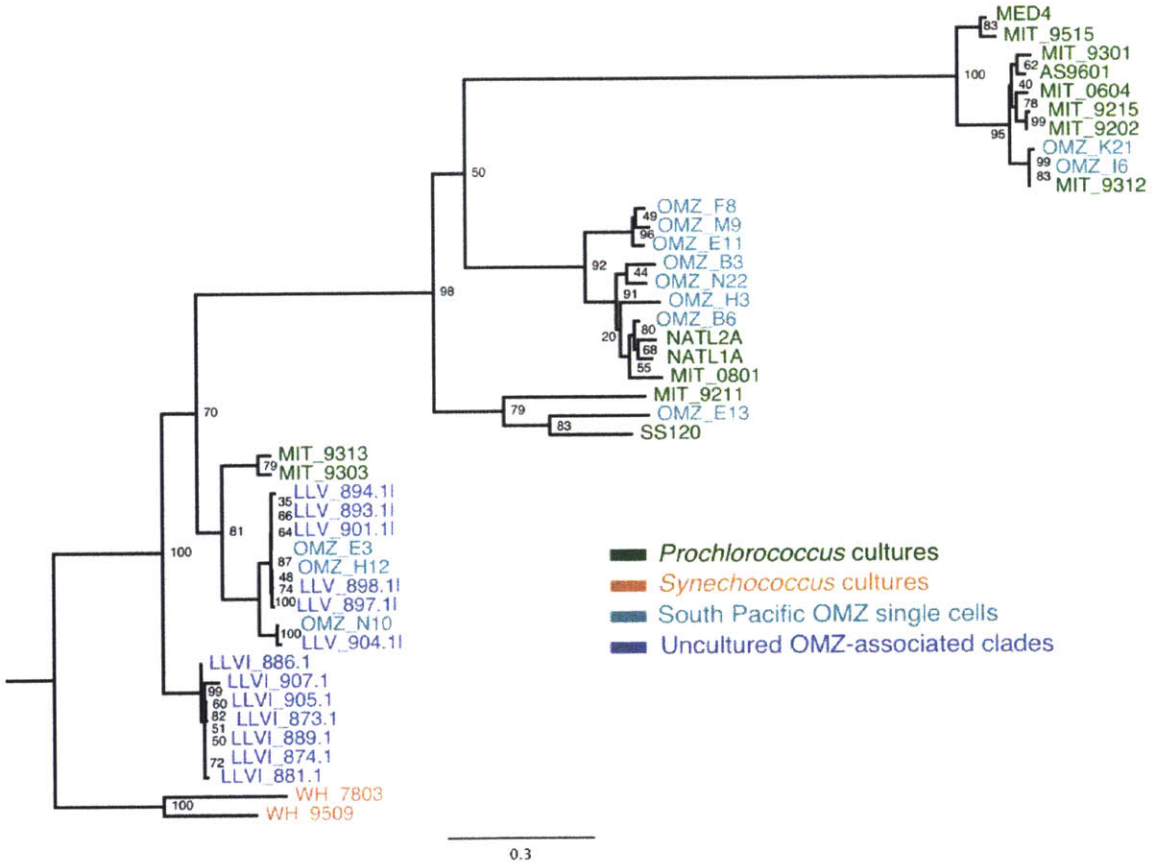


Figure 4.16. Phylogeny of ITS sequences from BiG RAPA Chile OMZ single-cell amplified DNA
Some of the secondary chlorophyll maximum single-cell amplified ITS sequences cluster with one of the two uncultured clades found in OMZ samples from the same region (Lavin et al., 2010). Other cells cluster with the LLI, LLII/III and HLII clades.

4.4 Conclusions and Future Directions

We surveyed *Prochlorococcus* populations across an Eastern South Pacific transect in November and December 2010. We add these measurements of *Prochlorococcus* concentrations at these times and places to the ever-growing collection of this kind of data, which is necessarily limited by the scale of the oceans, but ongoing observation is an important part of our regional and global understanding of *Prochlorococcus*. Our concentration measurements are broadly similar to those taken several years earlier, for a more southerly transect in the same region during the BIOSOPE cruise, a major microbial oceanography expedition traversing the entire South Pacific from Chile to the Marquesas to the west, which we have relied heavily on for all background references in this chapter. We found high concentrations of *Prochlorococcus* extending to depths of 150m in the open ocean waters, and closer to the coast we found shallower euphotic zones and wide variations in concentrations of *Prochlorococcus*. While *Prochlorococcus* populations in stratified open ocean gyres are consistently found with high abundance, consistently over time and space, *Prochlorococcus* populations are more variable closer to land and at the edges of gyres. What could explain the patchiness we observed of *Prochlorococcus* populations for the 1,000 miles off the coast of Chile? We know that temperature and mixing play a role in structuring *Prochlorococcus* ecotype distributions. One possibility is that when conditions are locally and briefly warmer, more stratified, or lower nutrient like the open ocean, *Prochlorococcus* populations on the edge of their habitat range can establish. Then, when conditions are cold and rich, unlike the open oceans, other phytoplankton fill the waters. The question of what happens to a population at the edge of its habitat range is an important one for understanding basic biology of the organisms and the causal factors controlling global populations. For *Prochlorococcus*, it may be profitable to pursue targeted, high-resolution sampling of *Prochlorococcus* not in their favorite habitats, but at the edges, where we know less.

The work described in this chapter supports and guides ongoing work by other researchers using samples from Big RAPA, providing baseline estimates of *Prochlorococcus* abundance and distribution in the water column. qPCR methods have been applied to samples from the transect, mapping out the distribution of *Prochlorococcus* ecotypes over the Eastern South Pacific (Paul Berube, in preparation for publication). Extensive metagenomic analyses efforts are also underway for phage and bacterial size fractions from this transect, which have so far confirmed the ecotype trends seen in qPCR. They also reveal that our samples from the chlorophyll maximum at the oligotrophic Station 7 seem to represent an active phage infection, with a population of closely related cyanomyophage dominating the phage fraction, and LLI *Prochlorococcus* abundant in the host fraction, a remarkable opportunity to observe population structure on an infective burst in the wild (Libusha Kelly, in preparation for publication). It will be interesting to look for adaptations in light related genome content to the ultraoligotrophic habitat of the South Pacific gyre, to see whether cells have different tools to cope with the high light, the particularly blue nature of the light or the increased UV load.

Acknowledgements

Many thanks to Paul Berube who performed much of the planning and sampling for this cruise, and is leading the effort to process other samples for this cruise and tell this story more fully. Thanks to Allison Coe, Jake Waldbauer and Anne Thompson for detailed advice on getting the most out of *Prochlorococcus* field sample flow cytometry. Thanks to the crew and science organizing team for the BiG RAPA cruise. Thanks to Ken Doggett and Ger Van Den Engh for helpful discussions about flow cytometry of *Prochlorococcus*. Many thanks to Zachory Berta-Thompson for assistance with plot rendering and statistical advice in data analysis. This work was funded by grants to Sallie W Chisholm from the Center for Microbial Oceanography Research and Education (a National Science Foundations Science and Technology Center) and the Evolutionary Biology section of the National Science Foundation, and by the Gordon and Betty Moore Foundation Marine Microbiology initiative.

References

Astorga-Eló, M., Ramírez-Flandes, S., DeLong, E.F., and Ulloa, O. (2015). Genomic potential for nitrogen assimilation in uncultivated members of *Prochlorococcus* from an anoxic marine zone. ISME J

Berube, P.M., Biller, S.J., Kent, A.G., Berta-Thompson, J.W., Roggensack, S.E., Roache-Johnson, K.H., Ackerman, M., Moore, L.R., Meisel, J.D., et al. (2014). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. ISME J

Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015). *Prochlorococcus*: the structure and function of collective diversity. Nat Rev Microbiol *13*, 13-27.

Binder, B.J.J., Chisholm, S.W.W., Olson, R.J.J., Frankel, S.L.L., and Worden, A.Z.Z. (1996). Dynamics of picophytoplankton, ultraphytoplankton and bacteria in the central equatorial Pacific. Deep Sea Research Part II: Topical Studies in Oceanography *43*, 907-931.

Boiteau, R.M., Fitzsimmons, J.N., Repeta, D.J., and Boyle, E.A. (2013). Detection of iron ligands in seawater and marine cyanobacteria cultures by high-performance liquid chromatography-inductively coupled plasma-mass spectrometry. Anal Chem *85*, 4357-362.

Bonnet, S., Guieu, C., Bruyant, F., Prá\vsil, O., Van Wambeke, F., Raimbault, P., Moutin, T., Grob, C., Gorbunov, M.Y., et al. (2008). Nutrient limitation of primary productivity in the Southeast Pacific (BIOSOPE cruise). Biogeosciences *5*, 215-225.

Bouman, H.A., Ulloa, O., Scanlan, D.J., Zwirglmaier, K., Li, W.K., Platt, T., Stuart, V., Barlow, R., Leth, O., et al. (2006). Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. Science *312*, 918-921.

Bryant, J.A., Stewart, F.J., Eppley, J.M., and DeLong, E.F. (2012). Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. Ecology *93*, 1659-673.

Canfield, D.E., Stewart, F.J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E.F., Revsbech, N.P., and Ulloa, O. (2010). A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. Science *330*, 1375-78.

Chisholm, S.W. (1992). Phytoplankton size. In Primary Productivity and Biogeochemical Cycles in the Sea, P. Falkowski, and A.D. Woodhead, eds. (New York: Plenum Press).

Claustre, H., Huot, Y., Obernosterer, I., Gentili, B., Tailliez, D., and Lewis, M. (2008). Gross community production and metabolic balance in the South Pacific Gyre, using a non intrusive bio-optical method. Biogeosciences *5*, 463-474.

Claustre, H., Sciandra, A., and Vaulot, D. (2008). Introduction to the special section bio-optical and biogeochemical conditions in the South East Pacific in late 2004: the BIOSOPE program. Biogeosciences *5*, 679-691.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. Science *311*, 1768-770.

Crosbie, N.D., and Furnas, M.J. (2001). Abundance, distribution and flow-cytometric characterization of picophytoprokaryote populations in central (17°S) and southern (20°S) shelf waters of the Great Barrier Reef. Journal of Plankton Research *23*, 809-828.

Duhamel, Björkman, and Karl (2012). Light dependence of phosphorus uptake by microorganisms in the subtropical North and South Pacific Ocean. Aquat Microb Ecol *67*, 225-238.

Dusenberry, J.A.A. (1999). Frequency distributions of phytoplankton single-cell fluorescence and vertical mixing in the surface ocean. Limnol Oceanogr *44*, 431-35.

Dusenberry, J.A.A., Olson, R.J.J., and Chisholm, S.W.W. (2000). Field observations of oceanic mixed layer dynamics and picophytoplankton photoacclimation. JOURNAL OF MARINE SYSTEMS *24*, 221-232.

Falkowski, P.G., and Raven, J.A. (2007). Aquatic Photosynthesis (Princeton University Press).

Fitzsimmons, J.N., Boyle, E.A., and Jenkins, W.J. (2014). Distal transport of dissolved hydrothermal iron in the deep South Pacific Ocean. Proc Natl Acad Sci U S A *111*, 16654-661.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proceedings of the National Academy of Sciences *110*, 9824-29.

Ganesh, S., Parris, D.J., Delong, E.F., and Stewart, F.J. (2014). Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. ISME J *8*, 187-211.

Giovannoni, S.J., and Vergin, K.L. (2012). Seasonality in ocean microbial communities. Science *335*, 671-76.

Goericke, R., Olson, R.J., and Shalapyonok, A. (2000). A novel niche for *Prochlorococcus* sp. in low-light suboxic environments in the Arabian Sea and the Eastern Tropical North Pacific. Deep Sea Research Part I: Oceanographic Research Papers *47*, 1183-1205.

Grob, C., Ulloa, O., Claustre, H., Huot, Y., Alarcón, G., and Marie, D. (2007). Contribution of picoplankton to the total particulate organic carbon concentration in the eastern South Pacific. Biogeosciences *4*, 837-852.

Hartmann, M., Gomez-Pereira, P., Grob, C., Ostrowski, M., Scanlan, D.J., and Zubkov, M.V. (2014). Efficient CO2 fixation by surface *Prochlorococcus* in the Atlantic Ocean. ISME J *8*, 2280-89.

Hess, W.R., Rocap, G., Ting, C.S., Larimer, F., Stilwagen, S., Lamerdin, J., and Chisholm, S.W. (2001). The photosynthetic apparatus of *Prochlorococcus*: Insights through comparative genomics. Photosynth Res *70*, 53-71.

Huot, Y., Babin, M., Bruyant, F., Grob, C., Twardowski, M.S., and Claustre, H. (2007). Relationship between photosynthetic parameters and different proxies of phytoplankton biomass in the subtropical ocean. Biogeosciences *4*, 853-868.

Iteman, I., Rippka, R., Tandeau De Marsac, N., and Herdman, M. (2000). Comparison of conserved structural and regulatory domains within divergent 16S rRNA-23S rRNA spacer sequences of cyanobacteria. Microbiology *146 (Pt 6)*, 1275-286.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M., and Chisholm, S.W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science *311*, 1737-740.

Karl, D.M., and Church, M.J. (2014). Microbial oceanography and the Hawaii Ocean Time-series programme. Nat Rev Microbiol *12*, 699-713.

Lavin, P., González, B., Santibáñez, J.F., Scanlan, D.J., and Ulloa, O. (2010). Novel lineages of *Prochlorococcus* thrive within the oxygen minimum zone of the eastern tropical South Pacific. Environmental Microbiology Reports *2*, 728-738.

Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P.J., Joewono, I., Rusch, D., Chisholm, S.W., and van der Donk, W.A. (2010). Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. Proc Natl Acad Sci U S A *107*, 10430-35.

Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. ISME J *4*, 1252-264.

Mann, E.L., Ahlgren, N., Moffett, J.W., and Chisholm, S.W. (2002). Copper toxicity and cyanobacteria ecology in the Sargasso Sea. Limnol Oceanogr *47*, 976-988.

Martiny, A.C., Tai, A.P., Veneziano, D., Primeau, F., and Chisholm, S.W. (2009). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. Environ Microbiol *11*, 823-832.

Moisander, P.H., Zhang, R., Boyle, E.A., Hewson, I., Montoya, J.P., and Zehr, J.P. (2011). Analogous nutrient limitations in unicellular diazotrophs and *Prochlorococcus* in the South Pacific Ocean. ISME J

Moore, L.R., and Chisholm, S.W. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus* : ecotypic differences among cultured isolates. Limnol Oceanogr *44*, 628-638.

Moore, L.R., Post, A.F., Rocap, G., and Chisholm, S.W. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. Limnology and Oceanography *47*, 989-996 .

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature *393*, 464-67.

Morel, A., Ahn, Y.-H., Partensky, F., Vaulot, D., and Claustre, H. (1993). *Prochlorococcus* and *Synechococcus*: A comparative study of their optical properties in relation to their size and pigmentation. Journal of Marine Research *51*, 617-649.

Morel, A., Claustre, H., and Gentili, B. (2010). The most oligotrophic subtropical zones of the global ocean: similarities and differences in terms of chlorophyll and yellow substance. Biogeosciences *7*, 3139-151.

Morel, A., Claustre, H., Antoine, D., and Gentili, B. (2007). Natural variability of bio-optical properties in Case 1 waters: attenuation and reflectance within the visible and near-UV spectral domains, as observed in South Pacific and Mediterranean waters. Biogeosciences *4*, 913-925.

Morel, A., Huot, Y., Gentili, B., Werdell, P.J., Hooker, S.B., and Franz, B.A. (2007). Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach. Remote Sensing of Environment *111*, 69-88.

Morel, A.M., Gentili, B.G., Claustre, H.C., Babin, M.B., Bricaud, A.B., Ras, J.R., and Tièche, F.T. (2007). Optical Properties of the "clearest" natural waters. Limnology and Oceanography *52*, 217-229.

Moutin, T., Karl, D.M., Duhamel, S., Rimmelin, P., Raimbault, P., Van Mooy, B.A.S., and Claustre, H. (2008). Phosphate availability and the ultimate control of new nitrogen input by nitrogen fixation in the tropical Pacific Ocean. Biogeosciences *5*, 95-109.

Olson, R.J., Chisholm, S.W., Zettler, E.R., Altabet, M.A., and Dusenberry, J.A. (1990). Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. Deep Sea Research Part A. Oceanographic Research Papers *37*, 1033-051.

Osburne, M.S., Holmbeck, B.M., Frias-Lopez, J., Steen, R., Huang, K., Kelly, L., Coe, A., Waraska, K., Gagne, A., and Chisholm, S.W. (2010). UV hyper-resistance in *Prochlorococcus* MED4 results from a single base pair deletion just upstream of an operon encoding nudix hydrolase and photolyase. Environ Microbiol *12*, 1978-988.

Partensky, F., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. Ann Rev Mar Sci *2*, 305-331.

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev *63*, 106-127.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature *424*, 1042-47.

Rodrigue, S., Malmstrom, R.R., Berlin, A.M., Birren, B.W., Henn, M.R., and Chisholm, S.W. (2009). Whole genome amplification and de novo assembly of single bacterial cells. PLoS One *4*, e6864.

Saulquin, B., Hamdi, A., Gohin, F., Populus, J., Mangin, A., and d'Andon, O.F. (2013). Estimation of the diffuse attenuation coefficient KdPAR using MERIS and application to seabed habitat mapping. Remote Sensing of Environment *128*, 224 - 233.

Shaffer, G., Salinas, S., Pizarro, O., Vega, A., and Hormazabal, S. (1995). Currents in the deep ocean off Chile (30°S). Deep Sea Research Part I: Oceanographic Research Papers *42*, 425 - 436.

Sommaruga, R., Hofer, J.S., Alonso-Sáez, L., and Gasol, J.M. (2005). Differential sunlight sensitivity of picophytoplankton from surface Mediterranean Coastal Waters. Appl Environ Microbiol *71*, 2154-57.

Stewart, F.J., Ulloa, O., and Delong, E.F. (2011). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. Environ Microbiol

Stramski, D., Reynolds, R.A., Babin, M., Kaczmarek, S., Lewis, M.R., Röttgers, R., Sciandra, A., Stramska, M., Twardowski, M.S., et al. (2008). Relationships between the surface concentration of particulate organic carbon and optical properties in the eastern South Pacific and eastern Atlantic Oceans. Biogeosciences *5*, 171-201.

Tedetti, M., and Sempéré, R. (2006). Penetration of Ultraviolet Radiation in the Marine Environment. A Review. Photochem Photobiol *82*, 389-397.

Thompson, A.W., Huang, K., Saito, M.A., and Chisholm, S.W. (2011). Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. ISME J *5*, 1580-594.

Ulloa, O., Canfield, D.E., Delong, E.F., Letelier, R.M., and Stewart, F.J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proc Natl Acad Sci U S A *109*, 15996-16003.

Urbach, E., and Chisholm, S.W. (1998). Genetic diversity in *Prochlorococcus* populations flow cytometrically sorted from the Sargasso Sea and the Gulf Stream. Limnology and Oceanography *43*, 1615-630.

Van Mooy, B.A., Fredricks, H.F., Pedler, B.E., Dyhrman, S.T., Karl, D.M., Koblizek, M., Lomas, M.W., Mincer, T.J., Moore, L.R., et al. (2009). Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature *458*, 69-72.

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

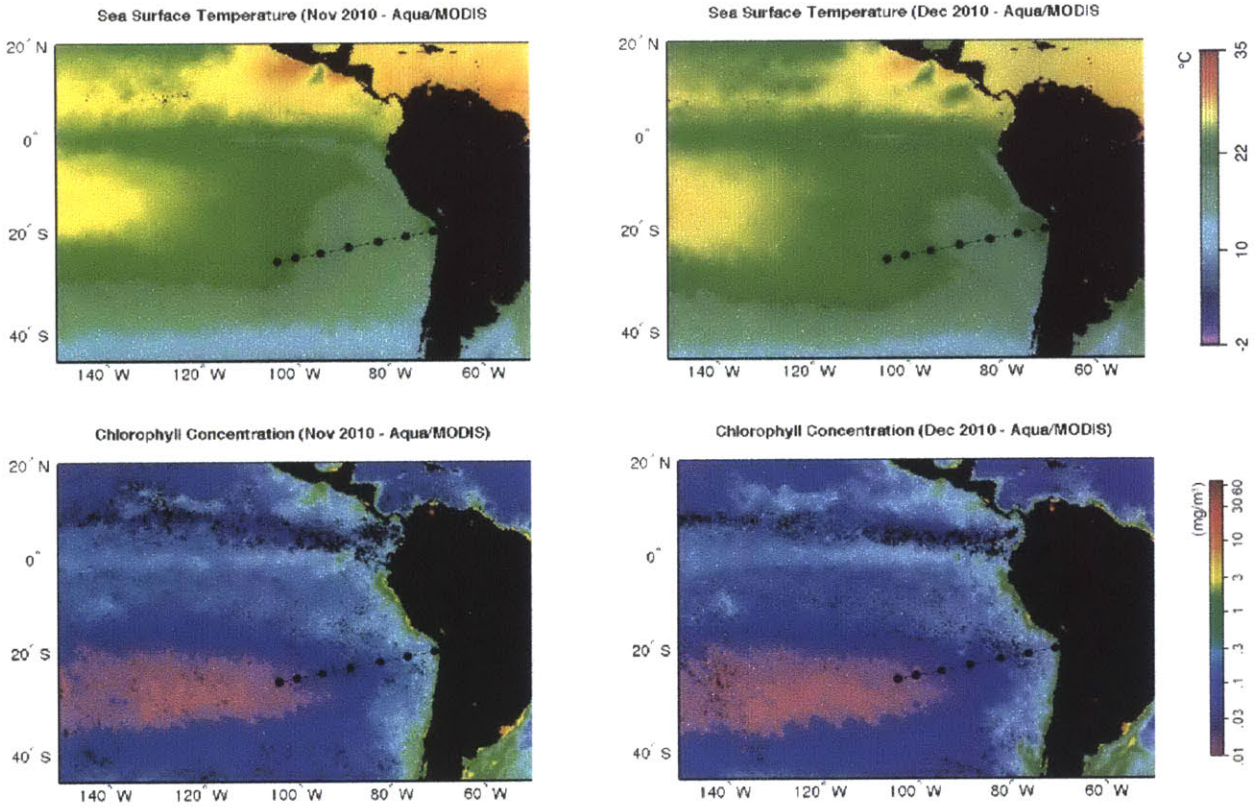
Zhang, K., Martiny, A.C., Reppas, N.B., Barry, K.W., Malek, J., Chisholm, S.W., and Church, G.M. (2006). Sequencing genomes from single cells by polymerase cloning. Nat Biotechnol *24*, 680-86.

Zinser, E., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D., and Chisholm, S.W. (2007). Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. Limnol Oceanogr *52*, 2205-220.

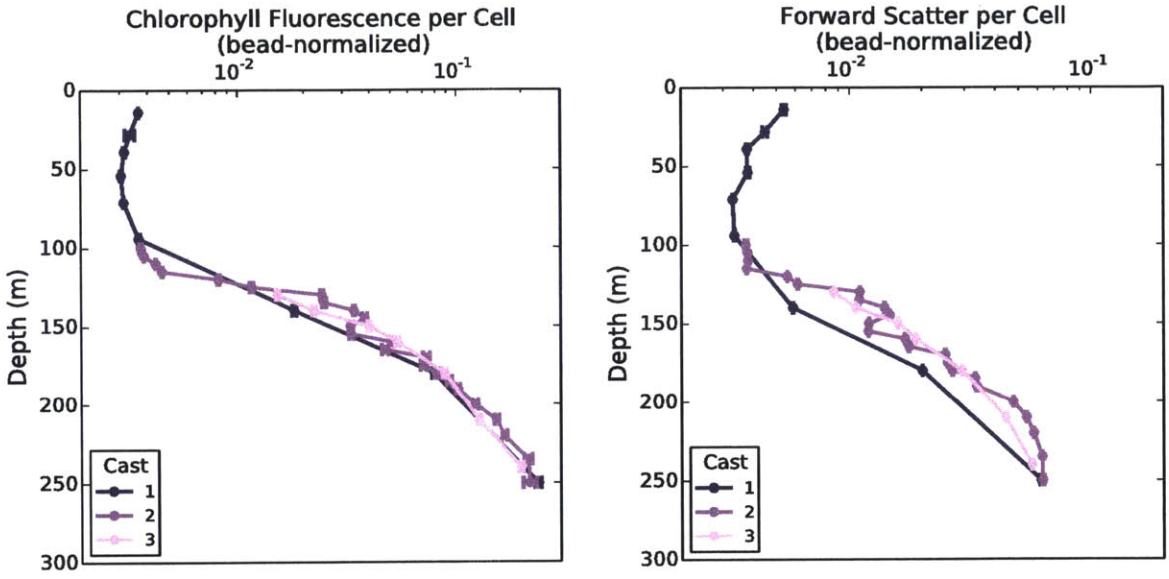
Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F., Massana, R., Ulloa, O., and Scanlan, D.J. (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. Environ Microbiol *10*, 147-161.

Supplemental

Supplemental Figure S4.1. A transect across dramatic gradients in the South Pacific
Since our transect spanned November-December, and satellite data is conveniently accessible smoothed on a monthly basis, we looked at the integrated satellite data for both months, which show that this region is stable and which gives some supports to the idea that our transect through time can be interpreted largely as a transect through space. The sea surface temperature changes over this transect are not dramatic relative to the global scale, as these are all subtropical waters, there is some slight warming towards the gyre, and variability near the coast. The full temperature depth profile measured on the transect (Figure 4.3A) gives a more complete picture for what *Prochlorococcus* across the water column experience.



Supplemental Figure S4.2. High resolution profiles of station 7: 3 days, chlorophyll and forward scatter per cell Same data as in Figure 4.13, cell properties changing over depth, viewed on a log scale to show the rapid transition between surface and deep populations in more detail. Variation in parameters near the surface may be an artifact from the reduction in accuracy for surface populations with very low chlorophyll content. Error bars represent range of technical replicates.



Chapter V. Conclusions and Future Directions

Productivity and Potential of the *Prochlorococcus* System

The power of the *Prochlorococcus* system for understanding larger questions in microbial ecology and evolution comes from our ability to study it with relative ease and accessibility in the field and lab, combining approaches from field oceanography, like rate measurements, perturbation experiments, flow cytometry observation and whole-community sequencing, with culture-based physiology, molecular biology and genomics (Coleman and Chisholm, 2007). *Prochlorococcus* diverged from its common ancestor with *Synechococcus* around 150 million years ago, based on 16S molecular clock estimates (Dufresne et al., 2005, Ochman et al., Ochman et al., 1987), or around 500 million years ago, based on fossil calibrated cyanobacterial genome phylogenies (Blank et al., 2010). With our genomic datasets, we can study the evolution of *Prochlorococcus* on multiple time scales over this long history, from fine-scale changes between recently diverged genomes, to the ancient divisions differentiating ecotypes (Kashtan et al., 2014, Rocap et al., 2003, Biller et al., 2015). Over the course of 30 years of multifaceted research, *Prochlorococcus* has taught us a great deal of basic biology, about how phytoplankton adapt to their many niches, how organisms in marine ecosystems interact, how carbon flows through the oceans, how microbial populations are structured and how microbial genomes evolve (Chisholm et al., 1988, Partensky and Garczarek, 2010, Zubkov et al., 2003, Becker et al., 2014, Coleman et al., 2006, Martiny et al., 2009, Morris et al., 2011, Biller et al., 2015). In the spirit of this history, this thesis represents an education by *Prochlorococcus*. We have presented here a body of work spanning a number of projects, approaches and questions, towards understanding the ecology and evolution of this important organism, contributing new cultures, new genomes, new support for old ideas and new ideas.

Targeted isolation of low-light adapted *Prochlorococcus* (Chapter II)

Even after about 25 years of cultivation, *Prochlorococcus* is relatively difficult to isolate and purify, although progress has been made towards understanding these challenges in recent years (Moore et al., 2007, Morris et al., 2011, Berube et al., 2014, Biller et al., 2014). There are relatively few strains in culture representing the low-light adapted *Prochlorococcus*, a few for each LL ecotype, but these genomes are highly diverse and of interest to numerous areas of *Prochlorococcus* research (Kettler et al., 2007, Scanlan et al., 2009, Partensky and Garczarek, 2010, Biller et al., 2014, Biller et al., 2015). We performed targeted isolations of LL *Prochlorococcus*, taking advantage of existing cruise plans to obtain enrichment samples from the well-studied Station ALOHA (Karl and Church, 2014). Through an enrichment program targeted for low-light adapted *Prochlorococcus*, combining traditional enrichment techniques with more recently developed dilution-to-extinction purification methods, we successfully isolated many new *Prochlorococcus* strains from the wild (e.g. Figure 5.1), sequenced their genomes, and began to explore their unique traits. For future field work, it is useful to keep in mind that the initial work setting up enrichments requires very little sea water and time during the cruise; most of the work happens in the lab, in small increments spread out over months. It would be easy to integrate a small isolation program into any cruise plan, and we should do this especially when given access to less sampled regions of the world's oceans without cultured representatives.

Figure 5.1 Growing *Prochlorococcus* LLIV *Prochlorococcus* strain MIT1313, one of the new cultures presented in this study, as viewed under phase contrast microscopy, 100X objective, showing typical rounded cells (cocci) and a few dividing cells.

We isolated many new strains from the LLIV clade, the deepest branching clade with the largest genomes and most strain-to-strain genomic content variation among *Prochlorococcus*, and one strain from the HLII clade, the most abundant in the oceans. Previously there were five published LLIV cultures and genomes, from two ocean basins, the North Atlantic and South Atlantic (Biller et al., 2015, Rocap et al., 2003, Moore et al., 1998, Kettler et al., 2007); now we have 12 more LLIV strains and genomes, from a different ocean (the North Pacific), spanning a range of diversity within the clade. Previously there was only one fully purified axenic (free from other bacteria) LLIV strain, MIT9313ax (Moore et al., 2005); now we have eight more axenic strains, which will enable a broader range of physiology experiments, for example simple inference in nutrient usage work. The first five LLIV genomes all encoded the ability to make secondary metabolites, the prochlorosin lantipeptides (Li et al., 2010). From the new genomes, we now know this trait has a patchy distribution across the clade, which may have interesting implications for its evolutionary history and ecological function. The isolates obtained in these efforts are all sympatric, coexisting prior to isolation in a single place (in fact, all the successful isolations came from a single water sample from 150m), which may be interesting for studying genetic variation within the framework of a basic shared ecology. We isolated HL and LL strains from the same water (as in Moore et al, 1998) and multiple LLIV strains, including both divergent strains from different subclades and fine-scale variants (like the backbone subpopulations of Kashtan et al., 2014). This gives us the power to answer questions about genomic and functional variation on several evolutionary time scales within this clade.

Laboratory cultures play a large role in the study of *Prochlorococcus*, enabling whole-genome sequencing, easy access to biomass for the study of DNA, RNA and protein, and strain-by-strain comparative physiology, but our culture collection is still limited relative to the vast diversity of *Prochlorococcus* we know exists in the wild (reviewed in Biller et al., 2015). Using all that we know about *Prochlorococcus*, a rapidly expanding body of information, we can continue to improve and refine our isolation efforts. One critical factor in the isolation process described here, likely determining the success of a few enrichments out of many attempts, was providing early enrichments attention on the time scale of the organisms' slow growth (weekly to monthly) by observing low-density growth with flow cytometry followed quickly by transfers to fresh media to maintain healthily growing *Prochlorococcus* populations. Over the course of isolation efforts in this work we took advantage of known properties of LL *Prochlorococcus* habitat, cell size and light preferences. Elsewhere, for example, chemical conditions of the media have been used to select for *Prochlorococcus* with the ability to use specific forms of nitrogen (Berube et al., 2014). Moving forward, we could continue to expand our culture collection in a targeted fashion, by sampling from locations with

known high abundances of uncultured clades, for example from iron limited regions (HLIII, IV, V; Rusch et al., 2010, Huang et al., 2011, Malmstrom et al., 2013) or oxygen minimum zones (LLV and LLVI; Lavin et al., 2010) or simply the base of the euphotic zone (LLVII, Martiny et al., 2009, Biller et al., 2015), and by using selection for additional known and hypothesized properties of *Prochlorococcus* lineages, for example light shock tolerance (Malmstrom et al., 2010) and phosphonate usage (Martinez et al., 2010, Martinez et al., 2011). The particular approach applied here for choosing sampling depth, looking at past ecotype abundance data, would only be useful at a few well-characterized sites. However, it is possible to apply the rich body of knowledge about ecotype biological differences and distributions relative to light and temperature to perform targeted sampling and enrichment strategies (Moore et al., 1999, Johnson et al., 2006, Zinser et al., 2007). This work contributes to the expansion of the *Prochlorococcus* culture collection and will significantly enrich future study of the LLIV clade.

Light shock and the high-light-inducible genes of *Prochlorococcus* (Chapter III)

The high-light-inducible gene family appears repeatedly in different areas of *Prochlorococcus* research, in viruses, in the study of *Prochlorococcus* diversity at multiple scales, and in almost every transcriptome perturbation study. In this thesis, we explore the relationship between the physiological response of diverse *Prochlorococcus* cultures to light shock, the number of *hli* genes in their genomes and the complex evolutionary history of the *hli* gene family. Marine *Synechococcus*, HL-adapted *Prochlorococcus* and the LLI clade of *Prochlorococcus* easily recover from severe transient light shock. The LLIV clade does not, and the LLII/III clade is intermediate in its response, to the extent that we can measure it in non-axenic cultures. Change over time on several scales, including seasonality, was one of the niche dimensions that Hutchinson proposed as a solution to the paradox of the plankton (Hutchinson, 1961). By showing for an expanded sample set that light-shock tolerance is an ecotype-linked trait, we have contributed support to the hypothesis that fluctuation in light such as that during seasonal mixing events plays a role in the differentiation of ecotypes (Malmstrom et al., 2010).

The number of *hli* family genes per genome varies by ecotype, a pattern observed in early genomes that holds up well in our examination to the recently expanded available genome data. *hlis* represent an exception to the general trend of loss of paralogs through genome streamlining over the course of *Prochlorococcus* evolution (Luo et al., 2011). For this gene family, copy number variation and the evolution of new paralogs are evolutionary tools used in the refinement of the *Prochlorococcus* flexible genome. Marine *Synechococcus* have 8-20 *hli* genes per genome, LLIV *Prochlorococcus* have 8-11 *hlis* per genome, LLII/III genomes have 12-14 *hlis*, LLI have 25-43, and HL have 17-26. The variation between and within ecotypes relates to light physiology for both growth and light shock tolerance, as we understand it for these ecotypes, consistent with the idea that these genes may play a role in light shock tolerance. By organizing the *hli* gene family of *Prochlorococcus* into deeply branching sequence clusters, we found that some of the deeply branching *hli* sequence variants specific to *Prochlorococcus* were already present in the common ancestor of all *Prochlorococcus* and most were present in the common ancestor of the LLII/III, LLI, and HL ecotypes. Although LLIV *Prochlorococcus* and *Synechococcus* have similar numbers of *hli* family genes, our clustering analysis shows that only a small number of genes (five) are shared between them, and each has their own distinctive pool of *hli* genes, which include both core genes conserved across each genus and flexible genes that vary in the presence or copy number within each genus. Expansion in this gene family to the very high numbers observed in LLI and HL *Prochlorococcus* occurred through duplication of a few of these existing sequence variants. We found that these multicopy *hli* genes are arranged in tandem arrays on the genome, likely operons, with each composed of several divergent *hli* sequence variants. Closely related genomes can differ by units of whole arrays, and the contents of arrays can change over time.

Why are there so many different *hli* genes *Prochlorococcus*? Placing this evolutionary explosion of *hli* proteins in the context of *Prochlorococcus* evolution, part of the answer may lie in the major changes to the *Prochlorococcus* photosystem and loss of other mechanisms of protection from light. If these *hli* proteins are functioning in safe delivery of chlorophyll to apoproteins, a function which has recently been shown for some distantly related *hli* genes (Chidgey et al., 2014, Knoppová et al., 2014), then when *Prochlorococcus* switched from the phycobilisome to a unique chlorophyll-based light antennae using the prochlorophyte binding proteins and divinyl chlorophylls (Ting et al., 2002, Zhang et al., 2007), it may have required a new suite of chlorophyll traffickers to match. Many of the insights in *hli* biochemistry have come from studying their interactions with other proteins; they are small proteins that act in complexes with each other and with other proteins, associated with the photosystems and chlorophyll synthesis machinery. To test the functional roles of *hlis* in *Prochlorococcus*, this would be a good place to start, looking for which other proteins the different *hli* sequence variants bind and whether tandem arrays produce physically associated multimeric protein complexes.

Now that we have built a large dataset of improved *hli* annotations, there is also more these genes can tell us through future evolutionary analyses. Data of several types that we now have in hand, including high similar and highly divergent genome pairs (Biller et al., 2014), a population genomics single cell dataset of many closely related strains (Kashtan et al., 2014) and additional environmental sequence data, has the power to start asking more nuanced molecular evolution questions. For subsets of the *hli* family that can be reliably aligned and at the DNA or protein level, and for which phylogeny can be reliably inferred, gene-tree species-tree reconciliation methods may help us resolve individual duplication and horizontal transfer events (Maddison, 1997, Koonin, 2005). Given the evidence of transfers and rearrangements we have observed so far, it would also be useful to look for evidence of recombination, to quantify the effect of that process on the history of hlis. Although there are no crystal structures available for *hli* proteins, there are structures available from their distant plant homologs, the light-harvesting complex or chlorophyll A/B binding proteins, which may be conserved enough in folds and chlorophyll-binding sites to allow structural modeling of *hli* genes, to inform hypotheses about the relationship between sequence variation and functional variation (Engelken et al., 2010). To look for patterns of selection and atypical evolutionary processes, for whole genes and residues within each *hli* gene cluster, it could be helpful to apply, to appropriate subsets of our *hli* data, codon substitution models (Yang et al., 2002), and Fst and site frequency spectrum analyses (Nielsen, 2005, Kashtan et al., 2014). These tools will allow us to measure relative rates of evolution between gene clusters (and compared to *Prochlorococcus* core genes), to understand which are conserved and which are changing rapidly, and identify parts of the proteins undergoing positive, neutral and negative selection, which may inform further our hypotheses about function to guide future work. Through our study of *hli* genes we have shed some light on the evolution of a gene family that plays a critical role in niche adaptation in *Prochlorococcus*, in terms of the distribution of different members of these gene family and different numbers of each genotype, and the relationship between genotype and changing light conditions over time.

Distribution and physical properties of South Pacific *Prochlorococcus* (Chapter IV)

Taking advantage of a rare sampling opportunity to study *Prochlorococcus* populations in the South Pacific, in Chapter IV we enumerated *Prochlorococcus* populations across a long oceanographic gradient spanning diverse marine ecosystems, and explored some the ecological patterns they form across this interesting region of the ocean. We used flow cytometry to identify *Prochlorococcus* in preserved seawater samples and to characterize these populations with respect to cellular light scattering and fluorescence properties, which change over depth as a product of acclimation to different light intensity and genetic variation within populations.

We detected consistently high *Prochlorococcus* abundances in the South Pacific gyre. This region is highly oligotrophic, which results in unusually clear water. Under these conditions we observe *Prochlorococcus* populations extending deep into the water column, beyond 200m, and the peak abundance of these populations often occurs at greater than 100m. For the first few hundred miles offshore, *Prochlorococcus* abundance was patchy, detectable in most samples but with variable abundance. The range of *Prochlorococcus* populations over depth grows shallower toward the coast, consistent with the changing water clarity in a gradient from coast to open ocean, although many other factors contribute to the full complexity of the *Prochlorococcus* depth profiles.

Cell properties change in a consistent way with light, except for at the extremes of the transect where the relationship between cell properties and light does not match the rest of the samples, indicating possible roles for genetic variation and light quality differences. Metagenomic sequencing currently under way will enable future work studying the genetic differences in *Prochlorococcus* populations across this transect, and light spectral data collected during the same cruise may enable more explicit treatment of the relationships between light color and adaptations in the genomes of *Prochlorococcus* of these different environments. For example, we could look for evidence of differences between among in their ability to produce different accessory pigments or in UV protection and damage response genes.

Simplicity and complexity in the *Prochlorococcus* system

Prochlorococcus is the smallest and most abundant free-living phototroph on the planet, with unique photosynthetic machinery and pigments (Partensky and Garczarek, 2010, Ting et al., 2002). Its vast diversity enables the success of different *Prochlorococcus* across a wide range of habitats over depth and geography, spanning various conditions in light, chemistry, mixing regimes, community structure and temperature (Biller et al., 2015). In some sense, *Prochlorococcus* is a minimal phototroph, with its small genomes and small cells, and at the same time, it is also highly complex, innovative and exploratory, with vast diversity between lineages and a large pan-genome, of which we have only scratched the surface (Kettler et al., 2007, Scanlan et al., 2009, Partensky and Garczarek, 2010, Baumdicker et al., 2012, Biller et al., 2015). The larger genomes of some other marine microbes allow individuals to exploit changing resources and withstand diverse stressors; in *Prochlorococcus*, individuals have a relatively limited repertoire, with loss of function over the course of genome streamlining limiting the ability for any one lineage to survive many conditions, a viable strategy in the relatively stable conditions of the open ocean (Giovannoni et al., 2005, Scanlan et al., 2009, Lauro et al., 2009, Morris et al, 2012). All *Prochlorococcus* are together adapted to a few of the constants in their environment, relatively low nutrients and blue light (Partensky et al., 1999b, Biller et al., 2015). Different subsets of *Prochlorococcus* are adapted to variations on this general environment, for example which nutrients are most limiting or light intensity (Coleman et al., 2010, Moore et al., 1999). *Prochlorococcus* populations, containing multiple ecotypes and many finer scale variants within ecotypes, can withstand a tremendous range of conditions, so we see high abundances of *Prochlorococcus* persisting across the oceans, while individuals are limited in niche (Johnson et al., 2006, Kashtan et al., 2014). Each genome is small, but the global meta-populations are not restricted by the abilities of any individual; different *Prochlorococcus* explore many different specializations.

Ecotype clades differentially adapted to light, mixing and temperature, the major divisions within the *Prochlorococcus* radiation, have given us a powerful framework for understanding the distribution of genetic diversity along light gradients and the ancient evolutionary trajectory of the clade. Looking broadly at the properties and evolutionary history of ecotypes, including some of the findings of this thesis, we can start to imagine the evolutionary arc of *Prochlorococcus*, starting with the evolution of the ancestor of all *Prochlorococcus* from a more *Synechococcus*-like ancestor. We know the ancestral *Prochlorococcus* underwent a dramatic transition in photosynthetic light gathering strategy, thylakoid structure, cell size, and genome

remodelling, which over eons of time gave rise to the vast diversity and abundance we see today. Based on the branching order of the ecotypes existing today, the first *Prochlorococcus* were more likely adapted to low light, giving up some of the flexibility of the cyanobacterial phycobilisome in return for access to a new habitat at the base of the oligotrophic euphotic zone. This habitat has low nutrient concentrations compared to coastal or upwelling locations, giving an advantage to streamlined cells, genomes and photosystems in early *Prochlorococcus*, but still higher nutrient concentrations than oligotrophic surface waters, representing an intermediate place in the process of *Prochlorococcus* adaptation to oligotrophic conditions. In many features, the HL and LL ecotypes are two distinct groups, indicating major changes in the HL common ancestor, giving them advantages in surface waters. Looking closely at ecotype distributions, light shock phenotypes and genomic adaptations to light, however, we can also see a pattern of gradual change within the LL ecotypes. From the first hypothetical deep-adapted ancestor, progressively more recently derived ecotypes show a gradual process of adaptation to higher light and lower nutrients (e.g. smaller genomes), working their way up the water column, along the way carving out niches spanning the full range of light conditions over the water column. Most recently (in ecotype scale) this trajectory gave rise to the HL strains that handle both high light and low nutrients, but have lost the ability to grow at the vanishingly low light levels that support deep branching *Prochlorococcus* at the base of the euphotic zone.

Our knowledge of deep branching *Prochlorococcus* primarily comes from the LLIV clade, which has been well-studied in culture and in the field. However we know of several even more deeply branching groups, but the LLV and LLVI oxygen minimum zone-associated clades, restricted in observations to date to low-light, low-oxygen habitats which are chemically separated from the rest of the euphotic zone; more study of these might inform our understanding of early *Prochlorococcus* and the common ancestor of all *Prochlorococcus*. On top of this picture of ancient ecotype history, in the time since the divergence events giving rise to each ecotype, *Prochlorococcus* in every lineage has undergone continual selection and change, visible in genomic islands. Much of the work exploring within ecotype differences has focused on nutrient acquisition strategies, which appear to be driven more by recent local selection than ancient ecotype history. Over the course of this work, one question that appeared and reappeared, and may be of interest in future work, was whether there is adaptation to light within ecotypes. Although light physiology to date largely maps onto ecotype differences, looking the field data, it seems possible that there could also be local, recent selection of light-related traits, for example, within the HLII clades that span several orders of magnitude in light at a single site, or for members of the same ecotype that live in habitats with different light spectral qualities. This could be addressed through a combination of carefully chosen sets of single cells and metagenomes, pairing members of the same ecotypes from different light habitats, and physiology work exploring similarities and differences in light physiology for strains in the same ecotype, perhaps guided by specific hypotheses from genomic variation (like *hli* and other light-related gene content variation).

A water sample from any sunlit part from the warmer half of the world's open oceans, away from coasts, dark depths and the poles, contains *Prochlorococcus* (Partensky et al., 1999a, Bouman et al., 2006, Zwirglmaier et al., 2008, Flombaum, et al., 2013). It is an organism of global significance and an important part of marine ecosystems, and we have only begun to explore its complexity. We look forward to the discovery of many more traits in the *Prochlorococcus* pangenome, in *Prochloroccus* physiology and in *Prochlorococcus* field biology which will reveal more niche axes of this beautiful organism.

References

Baumdicker, F., Hess, W.R., and Pfaffelhuber, P. (2012). The infinitely many genes model for the distributed genome of bacteria. Genome Biol Evol *4*, 443-456.

Becker, J.W., Berube, P.M., Follett, C.L., Waterbury, J.B., Chisholm, S.W., Delong, E.F., and Repeta, D.J. (2014). Closely related phytoplankton species produce similar suites of dissolved organic matter. Front Microbiol *5*, 111.

Berube, P.M., Biller, S.J., Kent, A.G., Berta-Thompson, J.W., Roggensack, S.E., Roache-Johnson, K.H., Ackerman, M., Moore, L.R., Meisel, J.D., et al. (2014). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. ISME J

Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., Awad, L., Roache-Johnson, K.H., Ding, H., Giovannoni, S.J., et al. (2014). Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *1*, 140034.

Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015). *Prochlorococcus*: the structure and function of collective diversity. Nat Rev Microbiol *13*, 13-27.

Blank, C.E., and Sánchez-Baracaldo, P. (2010). Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. Geobiology *8*, 1-23.

Bouman, H.A., Ulloa, O., Scanlan, D.J., Zwirglmaier, K., Li, W.K., Platt, T., Stuart, V., Barlow, R., Leth, O., et al. (2006). Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. Science *312*, 918-921.

Chidgey, J.W., Linhartová, M., Komenda, J., Jackson, P.J., Dickman, M.J., Canniffe, D.P., Koník, P., Pilny, J., Hunter, C.N., and Sobotka, R. (2014). A Cyanobacterial Chlorophyll Synthase-HliD Complex Associates with the Ycf39 Protein and the YidC/Alb3 Insertase. Plant Cell *26*, 1267-279.

Chisholm, S.W., Olson, R.J., Zettler, E.R., Waterbury, J.B., Goericke, R., and Welschmeyer, N. (1988). A novel free-living prochlorophyte occurs at high cell concentrations in the oceanic euphotic zone. Nature *334*, 340-43.

Coleman, M.L., and Chisholm, S.W. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. Trends Microbiol *15*, 398-407.

Coleman, M.L., and Chisholm, S.W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. Science *311*, 1768-770.

Dufresne, A., Garczarek, L., and Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol *6*, R14.

Engelken, J., Brinkmann, H., and Adamska, I. (2010). Taxonomic distribution and origins of the extended LHC (light-harvesting complex) antenna protein superfamily. BMC Evol Biol *10*, 233.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proceedings of the National Academy of Sciences *110*, 9824-29.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. Science *309*, 1242-45.

Huang, S., Wilhelm, S.W., Harvey, H.R., Taylor, K., Jiao, N., and Chen, F. (2011). Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. ISME J

Hutchinson, G.E. (1961). The Paradox of the Plankton. The American Naturalist *95*, 137-145.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M., and Chisholm, S.W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science *311*, 1737-740.

Karl, D.M., and Church, M.J. (2014). Microbial oceanography and the Hawaii Ocean Time-series programme. Nat Rev Microbiol *12*, 699-713.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science *344*, 416-420.

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., et al. (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. PLoS Genet *3*, e231.

Knoppová, J., Sobotka, R., Tichy, M., Yu, J., Konik, P., Halada, P., Nixon, P.J., and Komenda, J. (2014). Discovery of a Chlorophyll Binding Protein Complex Involved in the Early Steps of Photosystem II Assembly in Synechocystis. Plant Cell

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet *39*, 309-338.

Lauro, F.M., McDougald, D., Thomas, T., Williams, T.J., Egan, S., Rice, S., DeMaere, M.Z., Ting, L., Ertan, H., et al. (2009). The genomic basis of trophic strategy in marine bacteria. Proc Natl Acad Sci U S A *106*, 15527-533.

Lavin, P., González, B., Santibáñez, J.F., Scanlan, D.J., and Ulloa, O. (2010). Novel lineages of *Prochlorococcus* thrive within the oxygen minimum zone of the eastern tropical South Pacific. Environmental Microbiology Reports *2*, 728-738.

Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P.J., Joewono, I., Rusch, D., Chisholm, S.W., and van der Donk, W.A. (2010). Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. Proc Natl Acad Sci U S A *107*, 10430-35.

Luo, H., Friedman, R., Tang, J., and Hughes, A.L. (2011). Genome reduction by deletion of paralogs in the marine cyanobacterium *Prochlorococcus*. Mol Biol Evol *28*, 2751-760.

Maddison, W.P. (1997). Gene Trees in Species Trees. Syst Biol *46*, 523-536.

Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. ISME J *4*, 1252-264.

Malmstrom, R.R., Rodrigue, S., Huang, K.H., Kelly, L., Kern, S.E., Thompson, A., Roggensack, S., Berube, P.M., Henn, M.R., and Chisholm, S.W. (2013). Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. ISME J *7*, 184-198.

Martinez, A., Osburne, M.S., Sharma, A.K., Delong, E.F., and Chisholm, S.W. (2011). Phosphite utilization by the marine picocyanobacterium *Prochlorococcus* MIT9301. Environ Microbiol

Martinez, A., Tyson, G.W., and Delong, E.F. (2010). Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. Environ Microbiol *12*, 222-238.

Martiny, A.C., Tai, A.P., Veneziano, D., Primeau, F., and Chisholm, S.W. (2009). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. Environ Microbiol *11*, 823-832.

Moore, L.R., and Chisholm, S.W. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus* : ecotypic differences among cultured isolates. Limnol Oceanogr *44*, 628-638.

Moore, L.R., Coe, A., Zinser, E.R., Saito, M.A., Sullivan, M.B., Lindell, D., Frois-Moniz, K., Waterbury, J., and Chisholm, S.W. (2007). Culturing the marine cyanobacterium *Prochlorococcus*. Limnology and Oceanography: Methods *5*, 353-362.

Moore, L.R., Ostrowski, M., Scanlan, D.J., Feren, K., and Sweetsir, T. (2005). Ecotypic variation in phosphorus-acquisition mechanisms within marine picocyanobacteria. AQUATIC MICROBIAL ECOLOGY *39*, 257-269.

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature *393*, 464-67.

Morris, J.J., Johnson, Z.I., Szul, M.J., Keller, M., and Zinser, E.R. (2011). Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. PLoS ONE *6*, e16805.

Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. MBio *3*

Nielsen, R. (2005). Molecular signatures of natural selection. Annu Rev Genet *39*, 197-218.

Ochman, H., and Wilson, A.C. (1987). Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol *26*, 74-86.

Partensky, F., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. Ann Rev Mar Sci *2*, 305-331.

Partensky, F., Blanchot, J., and Vaulot, D. (1999). Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. Bulletin de l'Institut Oceanographique, Monaco *19*

Partensky, F., Hess, W.R., and Vaulot, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. Microbiol Mol Biol Rev *63*, 106-127.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature *424*, 1042-47.

Rusch, D.B., Martiny, A.C., Dupont, C.L., Halpern, A.L., and Venter, J.C. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. Proceedings of the National Academy of Sciences *107*, 16184 -116189.

Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., Post, A.F., Hagemann, M., Paulsen, I., and Partensky, F. (2009). Ecological Genomics of Marine Picocyanobacteria. Microbiology and Molecular Biology Reviews : MMBR *73*, 249-299.

Ting, C.S., Rocap, G., King, J., and Chisholm, S.W. (2002). Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies. Trends Microbiol *10*, 134-142.

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol *19*, 908-917.

Zhang, Y., Chen, M., Zhou, B.B., Jermiin, L.S., and Larkum, A.W. (2007). Evolution of the inner light-harvesting antenna protein family of cyanobacteria, algae, and plants. J Mol Evol *64*, 321-331.

Zinser, E., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D., and Chisholm, S.W. (2007). Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. Limnol Oceanogr *52*, 2205-220.

Zubkov, M.V., Fuchs, B.M., Tarran, G.A., Burkill, P.H., and Amann, R. (2003). High rate of uptake of organic nitrogen compounds by *Prochlorococcus* cyanobacteria as a key to their dominance in oligotrophic oceanic waters. Appl Environ Microbiol *69*, 1299-1304.

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F., Massana, R., Ulloa, O., and Scanlan, D.J. (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. Environ Microbiol *10*, 147-161.

Appendices

Co-authored papers:

A. Sher, D., **Thompson, J. W.**, Kashtan, N., Croal, L., and Chisholm, S. W. (2011) Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria, *ISME J 5*, 1125-1132.

B. Kashtan, N., Roggensack, S. E., Rodrigue, S., **Thompson, J. W.**, Biller, S. J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R. R., Stocker, R., Follows, M. J., Stepanauskas, R., and Chisholm, S. W. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*, *Science 344*, 416-420.

C. Biller, S. J., Berube, P. M., **Berta-Thompson, J. W.**, Kelly, L., Roggensack, S. E., Awad, L., Roache-Johnson, K. H., Ding, H., Giovannoni, S. J., Rocap, G., Moore, L. R., and Chisholm, S. W. (2014) Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*, *Scientific Data 1*, 140034.

D. Berube, P. M., Biller, S. J., Kent, A. G., **Berta-Thompson, J. W.**, Roggensack, S. E., Roache-Johnson, K. H., Ackerman, M., Moore, L. R., Meisel, J. D., Sher, D., Thompson, L. R., Campbell, L., Martiny, A. C., and Chisholm, S. W. (2014) Physiology and evolution of nitrate acquisition in *Prochlorococcus*, *International Society of Microbial Ecology Journal.*

Additional Appendices:

E. *Prochlorococcus* fluorescent and light microscopy methods, applications and images

F. *Synechococcus* of the MIT culture collection

The ISME Journal (2011), 1–8

www.nature.com/ismej

ORIGINAL ARTICLE

Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria

Daniel Sher[1], Jessie W Thompson, Nadav Kashtan, Laura Croal and Sallie W Chisholm
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Interactions between microorganisms shape microbial ecosystems. Systematic studies of mixed microbes in co-culture have revealed widespread potential for growth inhibition among marine heterotrophic bacteria, but similar synoptic studies have not been done with autotroph/heterotroph pairs, nor have precise descriptions of the temporal evolution of interactions been attempted in a high-throughput system. Here, we describe patterns in the outcome of pair-wise co-cultures between two ecologically distinct, yet closely related, strains of the marine cyanobacterium *Prochlorococcus* and hundreds of heterotrophic marine bacteria. Co-culture with the collection of heterotrophic strains influenced the growth of *Prochlorococcus* strain MIT9313 much more than that of strain MED4, reflected both in the number of different types of interactions and in the magnitude of the effect of co-culture on various culture parameters. Enhancing interactions, where the presence of heterotrophic bacteria caused *Prochlorococcus* to grow faster and reach a higher final culture chlorophyll fluorescence, were much more common than antagonistic ones, and for a selected number of cases were shown to be mediated by diffusible compounds. In contrast, for one case at least, temporary inhibition of *Prochlorococcus* MIT9313 appeared to require close cellular proximity. Bacterial strains whose 16S gene sequences differed by 1–2% tended to have similar effects on MIT9313, suggesting that the patterns of inhibition and enhancement in co-culture observed here are due to phylogenetically cohesive traits of these heterotrophs.
The ISME Journal advance online publication, 17 February 2011; doi:10.1038/ISMEJ.2011.1
Subject Category: microbe–microbe and microbe–host interactions
Keywords: heterotrophic bacteria; interactions; phylogeny; *Prochlorococcus*

Introduction

Interactions, such as symbiosis, competition and allelopathy are a central feature of microbial communities (Bassler and Losick, 2006; Azam and Malfatti, 2007; Hibbing *et al.*, 2009). Even in dilute oceanic environments, microbial interactions abound: antagonistic interactions can promote biodiversity (Czaran *et al.*, 2002; Bidle and Falkowski, 2004; Pernthaler, 2005), and synergistic interactions can provide sources of sustenance in complex communities (Azam *et al.*, 1983; Boetius *et al.*, 2000; Croft *et al.*, 2005; Azam and Malfatti, 2007; Amin *et al.*, 2009; Tripp *et al.*, 2010). Although marine microbial interactions often occur on scales of nanometers or microns (Blackburn *et al.*, 1998; Stocker *et al.*, 2008; Malfatti and Azam, 2009; Seymour *et al.*, 2010), they ultimately affect entire ecosystems and global biogeochemical cycles (Azam and Malfatti, 2007).

Heterotrophic bacteria have been shown to both enhance and inhibit the growth of marine and freshwater algae (Grossart *et al.*, 2006; Grossart and Simon, 2007; Mayali *et al.*, 2008) and cyanobacteria (Bratbak and Thingstad, 1985; Manage *et al.*, 2000; Morris *et al.*, 2008) in liquid culture and on solid media. Through these and similar studies we have come to recognize specific mechanisms of interaction, which can occur in the marine environment, such as facilitation of iron uptake (Amin *et al.*, 2009; D'Onofrio *et al.*, 2010), transfer of essential vitamins (Croft *et al.*, 2005), inter- and intra-specific communication (Bassler and Losick, 2006; Vardi *et al.*, 2006) and allelopathy (Mayali *et al.*, 2008; Hibbing *et al.*, 2009). Hypothesizing that bacterium–bacterium antagonistic interactions shape microbial community structure at the microscale, Long and Azam (2001) analyzed interactions among 86 pairs of co-isolated marine bacteria on solid media, revealing the widespread distribution of the potential for growth inhibition among these bacterial strains (Long and Azam, 2001; Grossart *et al.*, 2004; Rypien *et al.*, 2009). More recently, several strains of

Correspondence: SW Chisholm, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 15 Vassar Street, Cambridge, MA 02139, USA.
E-mail: chisholm@mit.edu
[1]Current address: Department of Marine Biology, Leon H Charney School of Marine Sciences, University of Haifa, Mount Carmel, 31905, Israel.
Received 5 August 2010; revised 2 December 2010; accepted 6 December 2010

heterotrophic bacteria have been shown to enhance the growth of a number of ecotypes of *Prochlorococcus*—the dominant phototroph in temperate and tropical oceans (Coleman and Chisholm, 2007; Partensky and Garczarek, 2010)—at low cell concentrations on solid and liquid media (Morris *et al.*, 2008). It was shown that the mechanism of enhancement in this case was the reduction of oxidative stress, explaining in part long-standing anecdotal observations that culturing *Prochlorococcus* is usually more robust when indigenous bacterial contaminants are present.

While *Prochlorococcus* have been extensively studied *vis-à-vis* the role of environmental factors, such as light, temperature and nutrient availability in shaping their ecology (Moore *et al.*, 1998, 2002; Bouman *et al.*, 2006; Johnson *et al.*, 2006; Coleman and Chisholm, 2007), and 'top down' processes, such as predation and viral lysis have also been studied to some degree (Lindell *et al.*, 2005, 2007; Sullivan *et al.*, 2005; Frias-Lopez *et al.*, 2009), systematic studies of their interaction with heterotrophic bacteria are limited to that of Morris and Zinser (Morris *et al.*, 2008) described above, who focused on the growth-enhancing role of bacteria in low-density cultures of *Prochlorococcus.* Inspired by this work, and by systemic analyses of Long and Azam (2001), we undertook a broad-based and quantitative analysis of co-cultures of two axenic *Prochlorococcus* ecotypes (Saito *et al.*, 2002; Moore *et al.*, 2005) with hundreds of diverse heterotrophic bacteria, examining the response of the *Prochlorococcus* cells to the presence of bacteria over the entire growth curve of the cultures.

We chose two strains of *Prochlorococcus*, one adapted to low light (MIT9313) and one adapted to high light (MED4), for these studies because they are ecologically and phylogenetically distinct. Additionally, MIT9313 is known to produce a diverse array of secondary metabolites of unknown function, whereas the genes encoding this system are absent in MED4 (Li *et al.*, 2010). We paired each strain with each of 344 strains of heterotrophic bacteria isolated from an oligotrophic marine environment. We asked: (1) how does the presence of added heterotrophic bacteria influence the growth of each *Prochlorococcus* strain over the course of its growth curve? (2) Do the two ecotypes respond differently to the presence of the same heterotroph? (3) Do different strains of heterotrophs have different effects, and are they related to the phylogeny of the heterotrophs? (4) Are the observed interactions mediated by soluble compounds or do they require close cellular proximity or contact?

Although the experimental system does not mimic the natural environment in many ways (Supplementary Information), it reveals some fundamental differences between the responses of two *Prochlorococcus* ecotypes to co-culture with hundreds of bacteria—differences that may hold clues to factors governing their realized niches in the ocean. It further highlights a strong correlation of the outcome of co-culture with the phylogeny of the heterotrophic bacteria, yielding hypotheses for further study on the mechanisms of these interactions and their potential role in marine microbial communities.

Materials and methods

We isolated heterotrophic bacteria from the Hawaii Ocean Time Series (HOT) station ALOHA (22°45′ N, 158° W), one of the most comprehensively studied sites in the ocean, with a microbial community dominated by *Prochlorococcus* and characterized in some detail (DeLong *et al.*, 2006). The heterotrophs were re-streaked for purity three times, and the final library was preserved at −80 °C in 25% glycerol. *Prochlorococcus* strains MIT9313 and MED4 were isolated from the Gulf Stream and the Mediterranean Sea, respectively (Rocap *et al.*, 2003), and were maintained in the lab at 20 °C and 27 μE constant cold white illumination. Co-culture was initiated by adding 2 μl of an overnight culture of each heterotroph from the library to 200 μl of *Prochlorococcus* culture (10^6 cells ml^{-1}) in 96-well plates. The culture media was Pro99 (Moore *et al.*, 2007) with the addition of 0.01% w/v pyruvate, acetate, lactate and glycerol as well as a vitamin mix (Morris *et al.*, 2008). The co-culture plates were maintained for 42 days at 20 °C and 27 μE constant cold white illumination, and the bulk chlorophyll fluorescence (FL) (ex440 em680) measured almost daily using a Bio-Tek Synergy HT plate reader. The resulting curves were filtered to retain consistent curves, defined as those in which the Euclidian distance between normalized curves fell within the range defined by 95% of the between-plate replicates of axenic curves. The growth parameters were extracted from the growth curves using macros written in Excel VBA, which are available from the investigators on request. Hierarchical Clustering was performed in Matlab. For detailed materials and methods see Supplementary Information.

Results and Discussion

Differences between Prochlorococcus *MIT9313 and MED4 in outcome of co-culture*

To determine what kinds of interactions occur when *Prochlorococcus* is grown in co-culture with many different strains of bacteria, we constructed a 'library' of 344 heterotrophic bacterial isolates from seawater collected in the open ocean, at the HOT station ALOHA (22°45′ N, 158° W) (Supplementary Figure 1). The heterotrophic strains were isolated on solid media (see Supplementary Information) and consist of at least 65 unique ribotypes (based on partial 16S ribosomal DNA sequences) clustering into 23, 13, 8 and 6 distinct OTUs at 1%, 3%, 5%

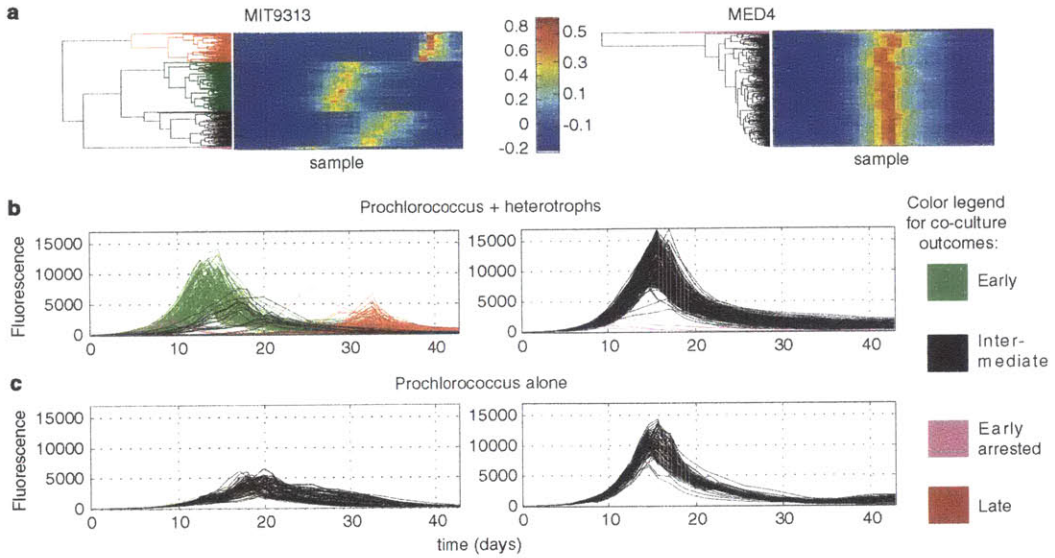


Figure 1 Features of *Prochlorococcus* MED4 and MIT9313 growth patterns in response to co-culture with 250 different strains of heterotrophic bacteria. (**a**) Heat maps of the normalized FL of all 338 growth curves (250 co-cultures and 88 controls) as clustered using hierarchical clustering (HC). A clearly different pattern can be seen between four major clusters in MIT9313 but only two in MED4. (**b**, **c**) FL curves of the 250 co-cultures (**b**) and 88 axenic *Prochlorococcus* cultures (**c**). The curves are colored as shown in the legend based on the clustering results in panel (**a**). Four different types of curves, which differ in their growth timing and maximal FL, can be observed for MIT9313, whereas only two clusters are observed for MED4. Note the similarity in the shape of the 'early arrested' outcome between MIT9313 and MED4.

and 7% ribosomal DNA sequence divergence, respectively (Supplementary Figure 1). The strains belong to the gamma-proteobacteria (primarily Alteromonas, Marinobacter and Alcanivorax) and alpha-proteobacteria (Rhodobacter) classes. Each of the 344 heterotrophic strains was inoculated into co-culture with axenic *Prochlorococcus* strains MED4 and MIT9313 in 96-well plates (under our conditions the outcome of co-culture does not depend on the initial number of heterotrophs inoculated—see Supplementary Information, Supplementary Figure 2). We measured the bulk *in vivo* chlorophyll FL of the cultures, which is widely used (Grossart, 1999; Mayali *et al.*, 2008; Malmstrom *et al.*, 2010) to follow the dynamics of phytoplankton cultures in a non-invasive manner. Although FL is only proportional to cell number when the cultures are in balanced growth (log phase, see Supplementary Information), the shape of the FL curve can reveal differences between the bulk behavior of the cultures throughout the culture period.

From the hundreds of co-cultures analyzed, only a few general types of co-culture outcomes emerged, as defined by the shape of the FL curves (Figure 1). Fifty-seven percent of the MIT9313 co-cultures fell into the group described as 'early' (green, Figure 1b) as these cultures entered exponential growth earlier, and reached higher maximal FL than the heterotroph-free MIT9313 cultures (Figure 1c). A small fraction of the co-cultures (3%) displayed the same initial timing as the 'early' group, but FL stopped increasing at an early stage and then declined rapidly ('early arrested', purple, Figure 1b). Thirty-four percent of the cultures stopped increasing in FL after 2–3 days, declined to undetectable levels, and then increased again much later (the 'late' group, red, Figure 1b). Finally, only 6 % of the co-cultures with MIT9313 behaved similarly to the heterotroph-free cultures ('intermediate', black, Figure 1b).

The synoptic response of MED4 to co-culture with the same library of bacterial strains was dramatically different from that of MIT9313. Ninety-eight percent of the heterotroph culture collection revealed no clear effect on the growth of MED4—as evidenced by their 'intermediate' growth patterns, which are very similar to the heterotroph-free cultures. The growth of *Prochlorococcus* MED4 in the remaining 2% of the co-cultures was arrested early, displaying strong inhibition by the presence of these heterotrophs (Figure 1b). The heterotrophic bacterial strains that inhibited MED4 were the same strains that defined the 'early arrested' group in the MIT9313 cultures.

Quantifying the parameter space of the MED4 and MIT9313 co-culture outcomes

To provide a quantitative estimate of the effect of the microbial interactions can have on *Prochlorococcus*

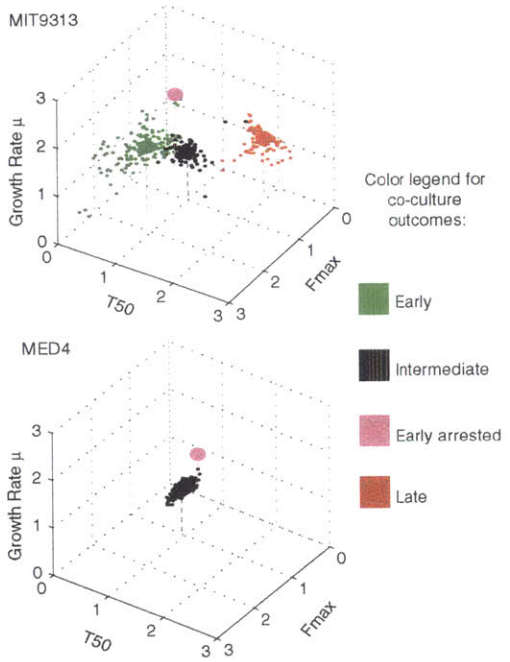


Figure 2 The quantitative three-dimensional parameter spaces defining the effect co-culture on *Prochlorococcus* MIT9313 and MED4. A three-dimensional parameter space is shown, with the axes being the maximum growth rate (μ), the time it took the cultures to reach half of the maximal FL (T_{50}), and the maximum FL (F_{max}). The parameter spaces shown includes both the co-cultures and the control axenic cultures, and are normalized to axenic wells on the same plates (that is, values larger than one represent an increase in the relevant parameter compared with axenic culture, smaller than one represent a decrease). The data points are colored based on the clustering shown in Figure 1. Large circles represent the median coordinates of each co-culture outcome.

culture dynamics, we extracted from the FL curves shown in Figure 1 biologically relevant descriptive parameters (similar to those used by Warringer *et al.* (2008)): the maximum growth rate (μ), the time it took the cultures to reach half of the maximal FL (T_{50}), and the maximum FL (F_{max}). As was clear in the qualitative analysis, the parameter space is not homogenously covered (Figure 2; Supplementary Figures 3 and 4). Rather, parts of the parameter space are densely populated, whereas others regions are empty or sparse, representing parameter combinations that are not observed in our co-culture curves (for example, co-cultures in which the log phase growth rate was significantly reduced compared with heterotroph-free cultures).

While the growth rate in log phase was influenced by the presence of bacteria in most of the MIT9313 co-cultures, the median of this parameter actually increased in most of the types of co-culture outcomes compared with the heterotroph-free cultures (Supplementary Figure 3) even when the overall effect was clearly one of much later onset of growth. Therefore, in agreement with other studies (Warringer *et al.*, 2008), our results suggest that a combination of different growth parameters is necessary in order to fully describe the complex effect of microbial interactions.

As described above, the most striking is the difference between the large parameter space inhabited by MIT9313 co-cultures and the much more limited space inhabited by MED4 co-cultures (Figure 2). The suite of heterotrophic bacteria that strongly influences the growth of MIT9313, decreasing some parameters up to 10-fold or increasing them up to 4-fold has minimal, if any, impact on MED4.

Heterotroph phylogeny and co-culture outcome

We next asked whether closely related bacteria, as defined by their partial 16S ribosomal DNA sequence (ribotype), affect the growth of *Prochlorococcus* cultures similarly. As shown in Figure 3, the heterotroph ribotypes, which induced 'early', 'early arrested' and 'late growth' phenotypes were significantly different for MIT9313 (UniFrac test with Bonferroni correction, $P \leqslant 0.06$; Lozupone and Knight, 2005), as were the groups that induced 'intermediate' and 'early inhibited' for MED4 ($P \leqslant 0.01$). For example, all but two of the heterotrophic strains, which induced a 'late' outcome of MIT9313 belong to two well-defined clades of Alteromonads (Figure 3, Supplementary Figure 1). Similarly, the same strains induced the 'early arrested' outcome in both MED4 and MIT9313, and all of these strains belong to a well-defined clade of Rhodobacters, similar to *Marinovum algicola* and *Ruegeria* sp. In most of these cases, the differentiation between strains, which inhibit *Prochlorococcus* in co-culture and strains, which do not is relatively deep-rooted, within the resolution afforded by our cultured collection of heterotrophs. For example, two Alteromonad clades differing by 1–2% in their partial 16S sequence both inhibit MIT9313, whereas a third clade, which differs by 4–5% from these two clades enhances MIT9313. Similarly, the clade of Rhodobacters inducing 'early arrested' phenotype differs from the most closely related strains in our collection that do not induce this phenotype by about 4% in their 16S. This level of divergence corresponds to one commonly used to delineate species or genus level differentiation (Schloss and Handelsman, 2005).

Co-culture outcome and proximity of cells

Although many interactions between microorganisms are mediated by diffusible soluble compounds, some have also been observed to occur when cells live in close proximity or even necessitate direct

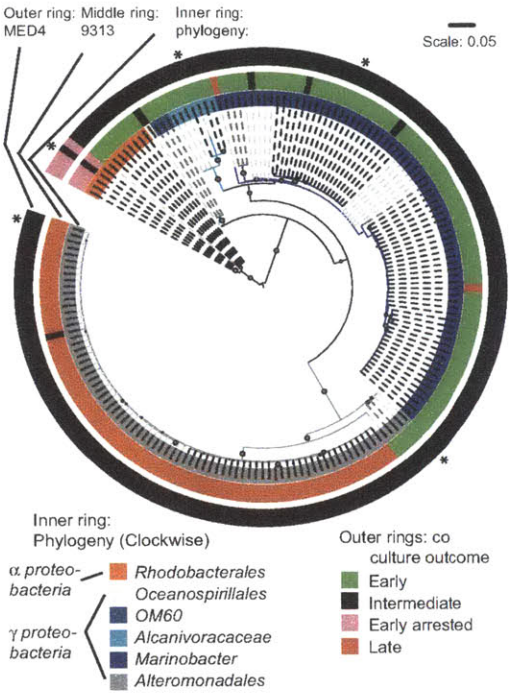


Figure 3 The relationship between patterns observed in co-cultures and the ribotype of the heterotroph. A maximum likelihood (ML) tree (partial 16S rDNA sequences) is shown, with the co-culture outcome (as defined in Figure 1) shown for MIT9313 (middle ring) and MED4 (outer ring). Spheres on the tree branches denote >80% approximate Likelihood Ratio Test (aLRT) confidence (Anisimova and Gascuel, 2006). Different shading of the branches of the tree denotes operational taxonomic units (OTUs) at 0.01 resolution (see also Supplementary Figure 1 and Supplementary Table 1). Asterisks denote the phylogenetic position of strains used for the experiments shown in Figure 4.

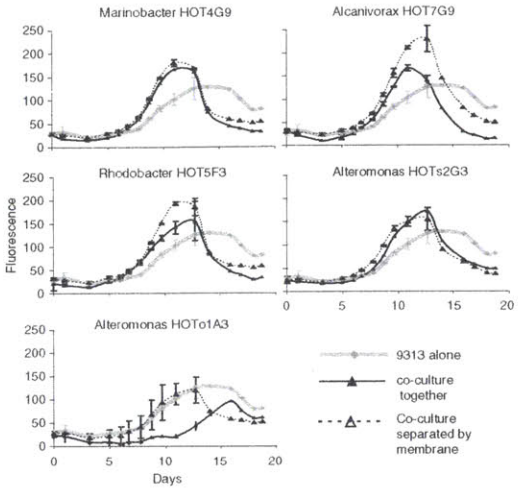


Figure 4 Comparison of outcomes of co-culture experiments between *Prochlorococcus* MIT9313 and five strains of bacteria grown together in mixed culture or separated by a 0.4 μm permeable membrane. Results shown are averages and s.d. ($n=4$ for axenic MIT9313, $n=2$ for co-cultures). The phylogenetic positions of the strains used in this experiment are shown in Figure 3 as asterisks.

cell–cell contact (Mayali and Azam, 2004; Croft *et al.*, 2005). To test whether close cell–cell proximity is necessary for the different co-culture outcomes observed with MIT9313, we selected five heterotroph strains representing different phylogenetic clades and co-culture outcomes, and co-cultured them with MIT9313 either separated by a membrane permeable to small molecules or mixed together as in the experiment presented above. As shown in Figure 4, when the FL of the co-cultures increased earlier than that of the axenic cultures this happened regardless of whether or not the heterotrophic bacteria were separated from MIT9313 by a membrane. Thus, the 'early' outcome of *Prochlorococcus* cultures is likely mediated in these cases by soluble, diffusible compounds, although we cannot preclude the possibility that the small number of heterotrophic bacteria than can cross the membrane during these 19-day long experiments (see Supplementary Information) may also directly impact the growth of MIT9313. In contrast, the late co-culture outcome occurred only when MIT9313 and Alteromonas strain HOTo1A3 were grown in close proximity and not when they were separated by a membrane.

Potential mechanisms underlying different co-culture outcomes

MIT9313 and MED4 represent two taxonomic extremes within the *Prochlorococcus* lineage, differing by ~3% in their 16S rRNA sequence. MED4 is a small cell with a highly streamlined genome, and is a member of the high-light adapted clade of *Prochlorococcus*. MIT9313, in contrast, is a slightly larger cell with a larger genome, and is better adapted for growth at the low light levels found deeper in the water column (Moore *et al.*, 1998, 2002; Rocap *et al.*, 2003; Bouman *et al.*, 2006; Johnson *et al.*, 2006; Coleman and Chisholm, 2007). Both strains are growing in these experiments below their respective temperature and light optima (although closer to those of MIT9313, (Rocap *et al.*, 2003; Zinser *et al.*, 2007)), but have been pre-acclimated to the experimental conditions for >7 months (~120 generations) and thus the difference in co-culture outcome is likely not caused by a general stress response in one strain because of culture conditions.

The 'early' culture outcome is the one most commonly observed with MIT9313, is widely distributed among the different phylogenetic groups,

and in all cases tested is caused by soluble, diffusible molecules. This is consistent with a 'helper' effect where the growth of *Prochlorococcus* increases as a result of basic attributes common to many lineages of heterotrophic cells, as suggested by Morris *et al.* (2008). Such attributes may include scavenging of reactive oxygen species (Morris *et al.*, 2008), increasing carbon dioxide concentration (Moore *et al.*, 2007) or cycling waste products. MED4 as a high-light adapted strain, may be better adapted to deal with oxidative stress (often generated during photosynthesis) than MIT9313, thus the latter strain may benefit more from interacting with heterotrophs. Notably, however, MED4 can readily form colonies on solid media only with the help of heterotrophs, and thus this strain is not immune to the effect of co-occurring bacteria (Morris *et al.*, 2008).

In contrast, inhibition of MIT9313 (early arrested or late outcomes) was observed mainly in co-cultures with two well-defined groups of bacteria belonging to the Alteromonads and Rhodobacters, with the latter group being the only one to clearly affect the growth of MED4 under our conditions. Related bacteria have previously been shown to inhibit other microbes through the production of secreted allelochemicals (for example, Mayali and Azam, 2004; Gram *et al.*, 2009). An intriguing observation is that inhibition of MIT9313 by an Alteromonas strain required proximity between the heterotrophic bacteria and MIT9313—that is, the effect could not be mimicked when the cells were kept apart by a semi-permeable membrane. Recently, close physical association (cell–cell contact) has been observed in natural seawater samples between *Synechococcus* cells, which are closely related to *Prochlorococcus*, and heterotrophic bacteria of unknown taxonomy (Malfatti and Azam, 2009; Malfatti *et al.*, 2010). These observations suggest the potential for close or contact-mediated interactions even in tiny picoplankton cells.

Conclusions

Although some features of our experimental system limit extrapolation of our results to the experience of wild *Prochlorococcus*—for example, the co-cultured strains were not co-isolated and the cell densities were higher than found in the wild (see also Supplementary Information)—our study has revealed some properties of these microbial interaction that likely have ecological relevance. First, the two *Prochlorococcus* ecotypes display fundamentally different responses to the presence of bacteria, both in terms of general patterns, and in terms of specific responses to specific bacterial strains. These differences could influence the connectivity of these two strains within the microbial network in the wild. If so, MIT9313 may be more susceptible to changes in the microbial community than MED4. Similar trends have been suggested for other marine bacterioplankton based on network analysis of patterns of co-occurrence in the oceans (Fuhrman and Steele, 2008).

Second, both the antagonistic and enhancing interactions in our system revealed a clear phylogenetic signature, with closely related bacteria causing similar responses in the co-cultured *Prochlorococcus*. Furthermore, only a handful of different interaction types, as measured through their effect on *Prochlorococcus* growth curves, were observed. The heterotroph culture collection we used represents only a fraction of diversity found in the oceans, and does not include many of the most common lineages. Future work with a wider diversity of bacteria may either reveal additional types of interactions or highlight unknown constrains on the types of interactions, which can affect cells in the aquatic environment.

Considering the high levels of microheterogeneity in both marine microbial populations (Thompson *et al.*, 2005; Hunt *et al.*, 2008) and their environment (Blackburn *et al.*, 1998; Azam and Malfatti, 2007; Stocker *et al.*, 2008; Seymour *et al.*, 2010), the task of understanding how complex microbial populations interact in the oceans is a daunting one. Although it is encouraging, as we seek general patterns, that the co-culture outcomes we observe are not random with respect to the phylogeny of the heterortophs, the opposite has been observed in cultures of interacting heterotrophic bacteria (Long and Azam, 2001). Clearly expanded and in depth study of the network of possible interactions between microbial groups is essential, if we ever wish to incorporate microbial interactions into our understanding of marine microbial communities.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Daniele Veneziano for help with statistical analyses and two anonymous referees for many constructive remarks. This study was supported by grants from the Gordon and Betty Moore Foundation, the NSF, and the US DOE-GTL (to SWC). DS was supported by postdoctoral fellowships from the Fullbright Foundation and the United States—Israel Binational Agricultural Research and Development Fund (Vaadia-BARD Postdoctoral Fellowship Award No. FI-399-2007). NK was supported by a postdoctoral fellowship from the Rothschild Yad Hanadiv Foundation and LC was supported by a postdoctoral research fellowship in biology from the National Science Foundation.

References

Amin SA, Green DH, Hart MC, Kupper FC, Sunda WG, Carrano CJ. (2009). Photolysis of iron-siderophore chelates promotes bacterial-algal mutualism. *Proc Natl Acad Sci USA* **106**: 17071–17076.

Anisimova M, Gascuel O. (2006). Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. *Syst Biol* **55**: 539–552.

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

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

Bassler BL, Losick R. (2006). Bacterially speaking. *Cell* **125**: 237–246.

Bidle KD, Falkowski PG. (2004). Cell death in planktonic, photosynthetic microorganisms. *Nat Rev Microbiol* **2**: 643–655.

Blackburn N, Fenchel T, Mitchell J. (1998). Microscale nutrient patches in planktonic habitats shown by chemotactic bacteria. *Science* **282**: 2254–2256.

Boetius A, Ravenschlag K, Schubert CJ, Rickert D, Widdel F, Gieseke A *et al.* (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**: 623–626.

Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WK, Platt T *et al.* (2006). Oceanographic basis of the global surface distribution of Prochlorococcus ecotypes. *Science* **312**: 918–921.

Bratbak G, Thingstad TF. (1985). Phytoplankton-bacteria interactions - an apparent paradox - analysis of a model system with both competition and commensalism. *Mar Ecol Prog Ser* **25**: 23–30.

Coleman ML, Chisholm SW. (2007). Code and context: Prochlorococcus as a model for cross-scale biology. *Trends Microbiol* **15**: 398–407.

Croft MT, Lawrence AD, Raux-Deery E, Warren MJ, Smith AG. (2005). Algae acquire vitamin B-12 through a symbiotic relationship with bacteria. *Nature* **438**: 90–93.

Czaran TL, Hoekstra RF, Pagie L. (2002). Chemical warfare between microbes promotes biodiversity. *Proc Natl Acad Sci USA* **99**: 786–790.

D'Onofrio A, Crawford JM, Stewart EJ, Witt K, Gavrish E, Epstein S *et al.* (2010). Siderophores from neighboring organisms promote the growth of uncultured bacteria. *Chem Biol* **17**: 254–264.

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

Frias-Lopez J, Thompson A, Waldbauer J, Chisholm SW. (2009). Use of stable isotope-labelled cells to identify active grazers of picocyanobacteria in ocean surface waters. *Environ Microbiol* **11**: 512–525.

Fuhrman JA, Steele JA. (2008). Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquat Microb Ecol* **53**: 69–81.

Gram L, Melchiorsen J, Bruhn J. (2010). Antibacterial activity of marine culturable bacteria collected from a global sampling of ocean surface waters and surface swabs of marine organisms. *Mar Biotechnol* **12**: 439–451.

Grossart HP. (1999). Interactions between marine bacteria and axenic diatoms (Cylindrotheca fusiformis, Nitzschia laevis, and Thalassiosira weissflogii) incubated under various conditions in the lab. *Aquat Microb Ecol* **19**: 1–11.

Grossart HP, Czub G, Simon M. (2006). Algae-bacteria interactions and their effects on aggregation and organic matter flux in the sea. *Environ Microbiol* **8**: 1074–1084.

Grossart HP, Schlingloff A, Bernhard M, Simon M, Brinkhoff T. (2004). Antagonistic activity of bacteria isolated from organic aggregates of the German Wadden Sea. *Fems Microbiol Ecol* **47**: 387–396.

Grossart HP, Simon M. (2007). Interactions of planktonic algae and bacteria: effects on algal growth and organic matter dynamics. *Aquat Microb Ecol* **47**: 163–176.

Hibbing ME, Fuqua C, Parsek MR, Peterson SB. (2009). Bacterial competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol* **8**: 15–25.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. (2008). Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**: 1081–1085.

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

Li B, Sher D, Kelly L, Shi Y, Huang K, Knerr PJ *et al.* (2010). Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc Natl Acad Sci USA* **107**: 10430–10435.

Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T *et al.* (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**: 83–86.

Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**: 86–89.

Long RA, Azam F. (2001). Antagonistic interactions among marine pelagic bacteria. *Appl Environ Microbiol* **67**: 4975–4983.

Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.

Malfatti F, Azam F. (2009). Atomic force microscopy reveals microscale networks and possible symbioses among pelagic marine bacteria. *Aquat Microb Ecol* **58**: 1–14.

Malfatti F, Samo TJ, Azam F. (2010). High-resolution imaging of pelagic bacteria by atomic force microscopy and implications for carbon cycling. *ISME J* **4**: 427–439.

Malmstrom RR, Coe A, Kettler GC, Martiny AC, Frias-Lopez J, Zinser ER *et al.* (2010). Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. *ISME J* **4**: 1252–1264.

Manage PM, Kawabata Z, Nakano S. (2000). Algicidal effect of the bacterium Alcaligenes denitrificans on Microcystis spp. *Aquat Microb Ecol* **22**: 111–117.

Mayali X, Azam F. (2004). Algicidal bacteria in the sea and their impact on algal blooms. *J Eukaryot Microbiol* **51**: 139–144.

Mayali X, Franks PJ, Azam F. (2008). Cultivation and ecosystem role of a marine roseobacter clade-affiliated cluster bacterium. *Appl Environ Microbiol* **74**: 2595–2603.

Moore LR, Coe A, Zinser ER, Saito MA, Sullivan MB, Lindell D *et al.* (2007). Culturing the marine cyanobacterium Prochlorococcus. *Limnol Oceanogr Methods* **5**: 353–362.

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

Moore LR, Post AF, Rocap G, Chisholm SW. (2002). Utilization of different nitrogen sources by the marine cyanobacteria Prochlorococcus and Synechococcus. *Limnol Oceanogr* **47**: 989–996.

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

Morris JJ, Kirkegaard R, Szul MJ, Johnson ZI, Zinser ER. (2008). Facilitation of robust growth of Prochlorococcus colonies and dilute liquid cultures by 'helper' heterotrophic bacteria. *Appl Environ Microbiol* **74**: 4530–4534.

Partensky F, Garczarek L. (2010). Prochlorococcus: advantages and limits of minimalism. *Ann Rev Mar Sci* **2**: 305–331.

Pernthaler J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* **3**: 537–546.

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

Rypien KL, Ward JR, Azam F. (2009). Antagonistic interactions among coral-associated bacteria. *Environ Microbiol* **12**: 28–39.

Saito MA, Moffett JW, Chisholm SW, Waterbury JB. (2002). Cobalt limitation and uptake in Prochlorococcus. *Limnol Oceanogr* **47**: 1629–1636.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.

Seymour JR, Simo R, Ahmed T, Stocker R. (2010). Chemoattraction to dimethylsulfoniopropionate throughout the marine microbial food web. *Science* **329**: 342–345.

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

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3**: e144.

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J *et al.* (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**: 1311–1313.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**: 90–94.

Vardi A, Formiggini F, Casotti R, De Martino A, Ribalet F, Miralto A *et al.* (2006). A stress surveillance system based on calcium and nitric oxide in marine diatoms. *PLoS Biol* **4**: e60.

Warringer J, Anevski D, Liu B, Blomberg A. (2008). Chemogenetic fingerprinting by analysis of cellular growth dynamics. *BMC Chem Biol* **8**: 3.

Zinser ER, Johnson ZI, Coe A, Karaca E, Veneziano D, Chisholm SW. (2007). Influence of light and temperature on Prochlorococcus ecotype distributions in the Atlantic Ocean. *Limnol Oceanogr* **52**: 2205–2220.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*

Nadav Kashtan *et al.*

Science **344**, 416 (2014);
DOI: 10.1126/science.1248575

Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*

Nadav Kashtan,[1]* Sara E. Roggensack,[1] Sébastien Rodrigue,[1,2] Jessie W. Thompson,[1] Steven J. Biller,[1] Allison Coe,[1] Huiming Ding,[1,3] Pekka Marttinen,[4] Rex R. Malmstrom,[5] Roman Stocker,[1] Michael J. Follows,[6] Ramunas Stepanauskas,[7] Sallie W. Chisholm[1,3]*

Extensive genomic diversity within coexisting members of a microbial species has been revealed through selected cultured isolates and metagenomic assemblies. Yet, the cell-by-cell genomic composition of wild uncultured populations of co-occurring cells is largely unknown. In this work, we applied large-scale single-cell genomics to study populations of the globally abundant marine cyanobacterium *Prochlorococcus*. We show that they are composed of hundreds of subpopulations with distinct "genomic backbones," each backbone consisting of a different set of core gene alleles linked to a small distinctive set of flexible genes. These subpopulations are estimated to have diverged at least a few million years ago, suggesting ancient, stable niche partitioning. Such a large set of coexisting subpopulations may be a general feature of free-living bacterial species with huge populations in highly mixed habitats.

The cyanobacterium *Prochlorococcus* is the smallest and most abundant photosynthetic cell in the oligotrophic oceans, contributing substantially to global photosynthesis (*1*). A single species by traditional measures, *Prochlorococcus* can be divided into several major clades, or ecotypes, defined by the intergenic transcribed spacer (ITS) region of their ribosomal RNA (rRNA) genes. These ecotypes are physiologically distinct (*2*–*4*); display distinctive seasonal, depth, and geographic patterns (*3*); and, like other microorganisms (*5*–*10*), embody tremendous genotypic and phenotypic diversity (*4*). To begin to understand the scope and limits of ecologically meaningful diversity within the canonical *Prochlorococcus* ecotypes, we examined cell-by-cell genomic diversity within a small sample of seawater and explored how it shifts in a dynamic environment.

We applied single-cell genome sequencing (*11*–*14*) to wild *Prochlorococcus* cells from samples collected at the Bermuda-Atlantic Time-series Study (BATS) site at three separate times of year (November 2008, February 2009, and April 2009) (Fig. 1A) (*15*). Because light, temperature, nutrients, and co-occurring communities change with winter deep mixing (*15*, *16*) (Fig. 1A), cells experience substantial environmental changes over tens of generations, enough to cause shifts in abundance of ITS-defined ecotypes (*2*, *15*, *17*).

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology (MIT), 77 Massachusetts Avenue, Cambridge, MA 02139, USA. [2]Département de Biologie, Université de Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke, Québec J1K 2R1, Canada. [3]Department of Biology, MIT, 77 Massachusetts Ave, Cambridge, MA 02139, USA. [4]Helsinki Institute for Information Technology, Department of Information and Computer Science, Aalto University, Post Office Box 15400, FI-00076 Aalto, Finland. [5]Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [6]Department of Earth, Atmospheric and Planetary Sciences, MIT, 77 Massachusetts Avenue, Cambridge, MA 02139, USA. [7]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA.

*Corresponding author. E-mail: chisholm@mit.edu (S.W.C.); nadav.kashtan@gmail.com (N.K.)

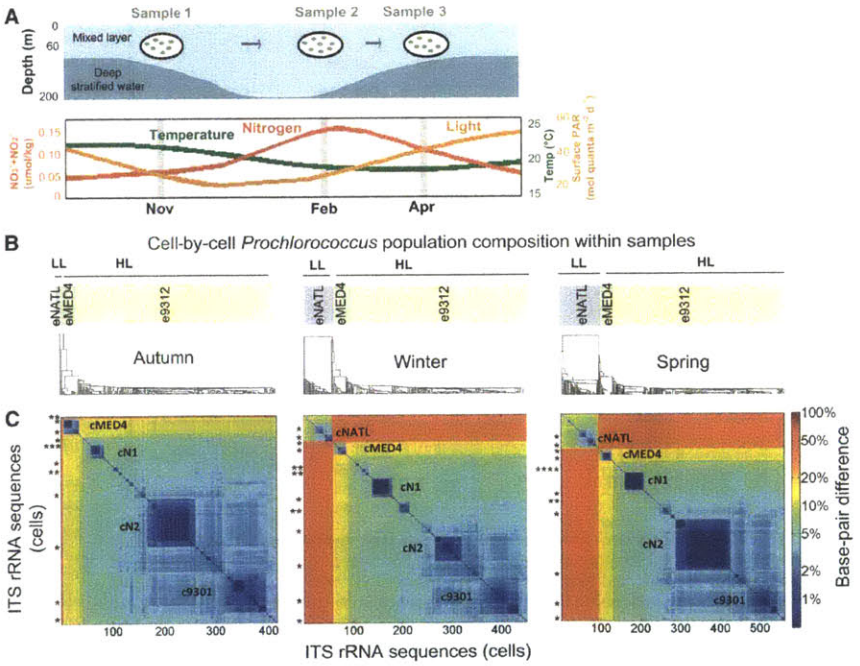


Fig. 1. Cell-by-cell *Prochlorococcus* population composition in samples from three separate times of year at the BATS site. Cells were collected within the mixed layer at 60 m depth in November 2008, February 2009, and April 2009 [see (*15*)]. **(A)** Schematic of seasonal dynamics at BATS and sampling design. (Top) A typical mixed-layer depth profile and context of our three samples. (Bottom) Typical average dynamics of light [smoothed mean surface photosynthetically active radiation (PAR) from 2004 to 2009], temperature, and nitrogen (within the mixed layer, averaged from 1999 to 2009) experienced by cells (*15*). Winter deep mixing brings cold nutrient-rich water to the surface. **(B)** Phylogenetic trees from pairwise genetic distances of ITS-rRNA sequences of individual cells from each sample [based on multiple alignment (*15*)]. The relevant sub-tree range of the known ecotypes (*2*) is marked above each tree if cells belonging to that ecotype were found, as is the division into low-light–adapted (LL) and high-light–adapted (HL) groups (*2*). **(C)** Heat maps describing the pairwise distance matrix between ITS-rRNA sequences of individual cells from each sample. Rows and columns are arranged according to the order of leaves of the trees shown in (B). The color map represents genetic distances as a percentage of base substitutions per site (log scale), such that the blue blocks identify very closely related ITS ribotypes. ITS sequences from cultured isolates with completely sequenced genomes are denoted by asterisks centered on the relevant line. Names of the largest clusters are marked in bold (e.g., cN2).

Flow sorting and DNA amplification (*11*–*14*) of more than 1000 co-occurring *Prochlorococcus* cells allowed us to explore the cell-by-cell genomic composition of these wild populations. We were able to identify coherent subpopulations at the whole-genome level and their relationship to those defined by the ITS region, explore finely resolved diversity patterns within and between subpopulations, and examine shifting abundances with seasonal changes in the habitat.

We first examined the population composition by sequencing the ITS regions of hundreds of *Prochlorococcus* cells in each sample, revealing the presence of finely resolved clusters within the broadly defined ecotypes (Fig. 1B). The populations were composed of tens to hundreds of nearly identical ITS clusters (>98% similar) within the coarse-grained ecotypes (Fig. 1, B and C). The relative abundance of cells belonging to the different clusters changed with season (Fig. 1, A to C) (*15*), suggesting shifts in their relative fitness in response to environmental changes.

To study the fine-scale genomic variation and compare it with the ITS-defined clusters, we sequenced the partial genomes (representing, on average, 70% of the total genome) of 90 individual cells (30 per sample) from the largest nearly identical ITS cluster, cN2 (Figs. 1C and 2), as well as 6 cells from two other clusters, cN1 and c9301. For each time of year, cells were randomly selected for genome sequencing from within the major ITS ribotypes (>99% similar) within cluster cN2 (C1 to C5) (30 cells), as well as from c9301-C8 and cN1-C9 (one cell each), as detailed in (*15*). We used a modified mediator genome reference assembly approach (*15*, *18*) to analyze between-cell variation in the partial genomes recovered. The topologies of the ITS and genomic trees were highly congruent (Fig. 2), indicating that ITS sequences can serve as a proxy for genome sequences in *Prochlorococcus* at a much finer level of resolution than previously demonstrated (*4*, *19*). The genomic data further revealed that the largest cluster cN2 is divided into five major clades [C1 to C5 (Fig. 2)] and a few additional minor clades represented by only one cell each. The delineation of clades C1 to C5 was highly robust and also observed in trees constructed from genomic position subsets (figs. S1 and S2).

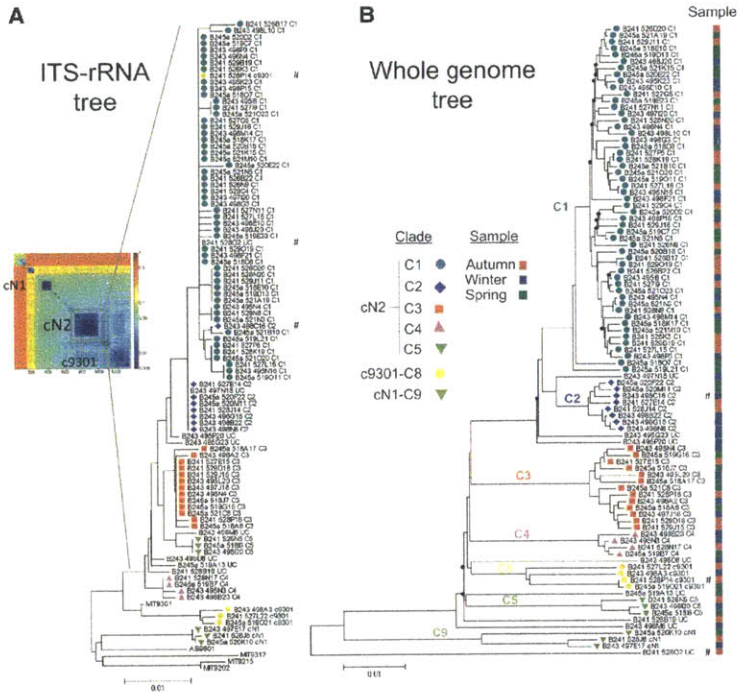


Fig. 2. ITS-rRNA sequence and whole-genome neighbor-joining phylogenetic trees at a fine resolution of diversity. **(A)** Phylogenetic tree based on ITS-rRNA sequences of 96 single cells (90 cN2 ribotypes, three cN1 ribotypes, and three c9301 ribotypes), as well as additional five high-light–adapted cultured strains. **(B)** Phylogenetic tree of the 96 single cells based on whole-genome sequences. The colored symbols to the left of the leaf labels in (A) and (B) represent the different clades depicted from the deep branches observed in the whole-genome tree. The sample origin of each cell is marked with red, blue, and green squares (representing autumn, winter, and spring, respectively) on the right. Distance units are base substitutions per site (see scale bar) (*15*). Bootstrap values <80 are marked as black dots on the internal nodes in (B) (fig. S1). Cells marked with # fall into an ITS clade that differs from the genome-defined clade. Neighbor-joining trees in (A) and (B) were constructed using p-distance.

To explore the evolutionary forces that shaped the cN2 C1 to C5 clades, we examined differences in nucleotide sequences within and between clades. For example, the C1 and C3 subpopulations (Fig. 2B) differ in 52,885 dimorphic single-nucleotide polymorphisms (SNPs), which represent 3.2% of their genomes (Fig. 3A, blue). The dimorphic SNPs between C1 and C3 are scattered across the genomes, occurring in 1519 out of 1974 genes (most of them core genes); 8% of these SNPs are found in intergenic regions (9% of the genome is noncoding). Of the intragenic SNPs, 37% are nonsynonymous, thus affecting the amino acid sequences of the proteins they encode. In contrast to the scattered nature of the sequence variation between the C1 and C3 clades, the polymorphism within them is confined to a few regions of the genome (Fig. 3A, black), indicating that most regions along the genome are conserved within clades and are different between them (*15*), which is true for all pairwise comparisons within C1 to C5 (figs. S3 and S4).

This emerging pattern was further supported by a standard measure of genetic differentiation between populations, the fixation index (F_{ST}) (*20*), applied at gene-by-gene resolution to the five cN2 clades, C1 to C5 (Fig. 3, B and C). Seventy-five percent of the core genes had high F_{ST} values (>0.8), (Fig. 3, B and C) (*15*), meaning different clades contained significantly different alleles. Some of the differentiated core genes have functions involved in the interaction between the cell and environmental stimuli [e.g., transporters, genes that affect oxidative stress responses, and cell surface biosynthesis and modification (Data S1)]; that is, they are not all simply "housekeeping genes" that control central metabolism. For example, alleles of phosphoglucosamine mutase, which is involved in the biosynthesis of outer membrane lipopolysaccharides (*21*), differ by an average of 10% of their amino acid sequences (Fig. 3C), with substitutions in the hydrophilic center of the enzyme (*21*), possibly affecting its specificity and kinetics.

We next asked whether different clade subpopulations carry distinct sets of flexible genes. Using de novo assemblies to capture regions unmapped by the reference assemblies (*15*), we found that each subpopulation carries a small set of distinct genes, typically in the form of cassettes within genomic islands (Table 1). Cassettes containing genes in the glycosyltransferase family account for much of the gene content variation between these clade subpopulations (Table 1 and table S1). The gene content in these cassettes suggests involvement in outer membrane modifications, possibly affecting phage attachment (*22*), recognition by grazers (*23*), cell-to-cell communication, or interactions with other bacteria (*24*).

We conclude that these clade subpopulations have distinct "genomic backbones" (and are

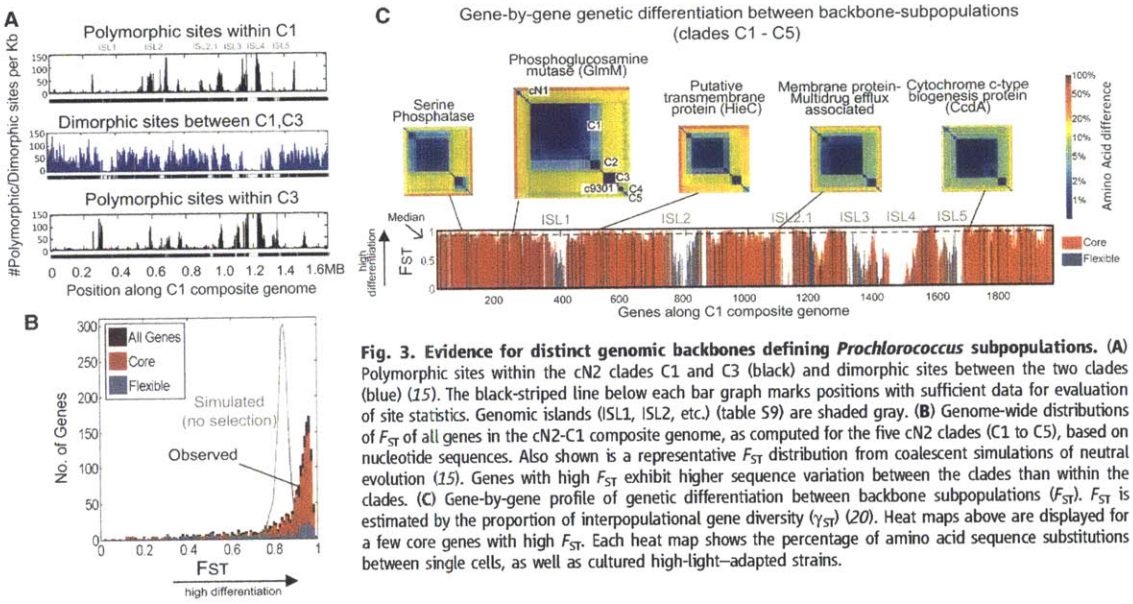


Fig. 3. Evidence for distinct genomic backbones defining *Prochlorococcus* subpopulations. (A) Polymorphic sites within the cN2 clades C1 and C3 (black) and dimorphic sites between the two clades (blue) (*15*). The black-striped line below each bar graph marks positions with sufficient data for evaluation of site statistics. Genomic islands (ISL1, ISL2, etc.) (table S9) are shaded gray. **(B)** Genome-wide distributions of F_{ST} of all genes in the cN2-C1 composite genome, as computed for the five cN2 clades (C1 to C5), based on nucleotide sequences. Also shown is a representative F_{ST} distribution from coalescent simulations of neutral evolution (*15*). Genes with high F_{ST} exhibit higher sequence variation between the clades than within the clades. **(C)** Gene-by-gene profile of genetic differentiation between backbone subpopulations (F_{ST}). F_{ST} is estimated by the proportion of interpopulational gene diversity (γ_{ST}) (*20*). Heat maps above are displayed for a few core genes with high F_{ST}. Each heat map shows the percentage of amino acid sequence substitutions between single cells, as well as cultured high-light–adapted strains.

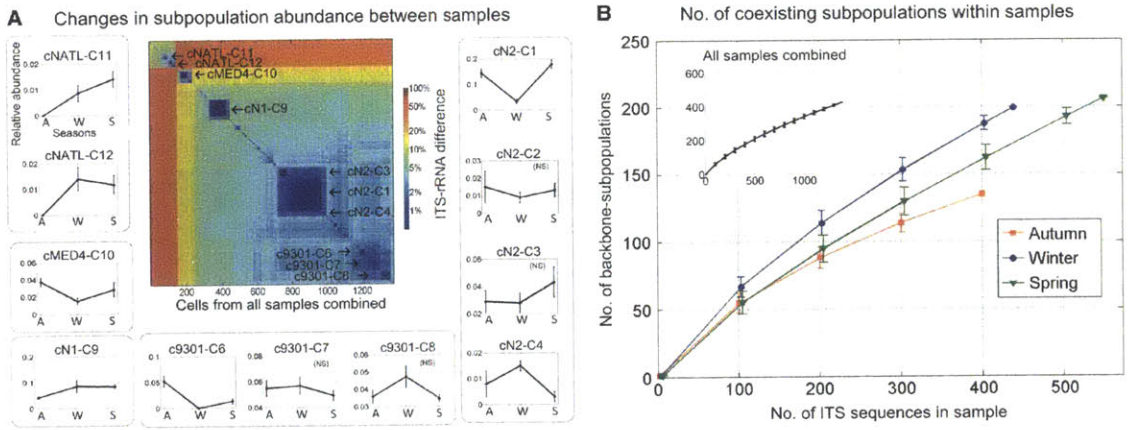


Fig. 4. Abundance profiles of backbone subpopulations and the estimated number of coexisting subpopulations within samples. (A) Relative abundance profiles of the 11 largest backbone subpopulations in our samples within the ITS clusters cNATL, cMED4, cN1, cN2, and c9301. A, autumn; W, winter; S, spring. Backbone names are marked near the relevant cluster on the ITS heat map. Backbone subpopulations were predicted by 99% ITS similarity for the full set of 1381 ITS sequences. Error bars represent SEM. NS, no significant changes between seasons (false discovery rate, $\alpha = 0.05$) (*15*). **(B)** Rarefaction curves estimating the number of coexisting backbone subpopulations within samples (*15*). Backbones predicted as in (A). Error bars represent 95% confidence intervals. (Inset) Rarefaction curve of all samples combined.

henceforth referred to as "backbone subpopulations") consisting of highly conserved (within subpopulation) alleles of the majority of core genes and a small distinct set of flexible genes that is linked with a particular backbone. This covariation between the core alleles and flexible gene content, and its fine scale resolution, represents a new dimension of microdiversity within wild *Prochlorococcus* populations. It is noteworthy that similar patterns have been identified in cultured isolates and metagenomic assemblies within coexisting members of a few other microbial species with very different ecologies (*5*–*10*, *25*), suggesting that differentiated genomic backbones may be a feature of diverse types of microbial populations.

At a finer resolution of diversity, we observed that cells within the five cN2 backbone subpopulations differ by 19,000 nucleotide positions on average, in comparison to 77,000 positions between backbone subpopulations (equivalent to 1.2 and 4.7% of the genome, respectively) (Fig. 2B). The most similar pairs of individual cell genomes in our samples differ in a few hundred base pairs [close to the detection limit when one considers single-cell processing and sequencing error (*15*)]; some of these pairs likely have identical

Table 1. Flexible gene cassettes associated with different cN2 backbone subpopulations highlighting gene content that may contribute to ecological differentiation. GT, glycosyltransferase; ABC-T, adenosine triphosphate–binding cassette (ABC) transporter; HLIP, high-light–inducible protein; CO, Cytochrome oxidase c subunit VIb; HlpA, histone-like protein; CpsL, polysaccharide biosynthesis protein. In the "Selected gene annotations" column, numbers before gene annotation refer to number of that type of gene. A complete list of the genes in each cassette is described in table S1 (*15*).

Clade	Cassette ID	Position	No. of genes in cassette	Selected gene annotations	Cassette function
cN2-C1	CST_I	Island 2.1	4	HLIP, CO	Redox stress response
	CST_II	Island 4	7	3GT, ABC-T	Outer membrane modification
cN2-C2	CST_II	Island 4	7	3GT, ABC-T	Outer membrane modification
cN2-C3	CST_III	Island 1	2	2GT	Outer membrane modification
cN2-C4	CST_I	Island 2.1	4	HLIP, CO	Redox stress response
	CST_IV	Island 4	14	3GT, HlpA, CpsL	Outer membrane modification
cN2-C5	CST_V	Island 4	5	2GT	Outer membrane modification

gene content (*15*). Except for these few pairs, each cell carries at least one gene cassette not found in any other. In some cases, a few closely related cells (a subclade) within backbones share a distinct gene cassette. Among these genes are, again, glycosyltransferase genes, as well as transporters and genes involved in nucleotide binding and processing. In a few cases, cells from different backbone subpopulations carry similar flexible gene cassettes [e.g., high-light–related genes (Table 1) and phosphonate related genes], demonstrating the combinatorial nature of backbones and flexible genes.

If backbone subpopulations have differential fitness, we would expect their relative abundance to change with changing environmental conditions (Fig. 1). Accordingly, the majority of the largest subpopulations exhibited significant seasonal abundance variation (Fig. 4A), higher than expected by chance (*15*), consistent with the hypothesis that this reflects selection, but more data are needed to draw that conclusion. Backbone subpopulations maintain their genomic composition between seasons (tested for C1) (*15*), which we would expect, as the establishment of new mutations and the acquisition and loss of genes are not likely to be in play on these time scales (*15*).

The congruency of genomic and ITS phylogenies in *Prochlorococcus* at both coarse (*4*, *19*) and fine resolution (Fig. 2) suggests that ITS-ribotype clusters coincide, in most cases, with distinct genomic backbones (*15*). This allowed us to estimate the number of coexisting backbone subpopulations in our samples through rarefaction analysis, revealing at least hundreds of coexisting subpopulations with distinct backbones (Fig. 4B) in each sample. These backbone subpopulations are estimated to have diverged at least a few million years ago (*15*), suggesting ancient, stable niche partitioning. That they have different alleles of genes associated with environmental interactions, carry a distinct set of flexible genes, and differ in relative abundance profiles as the environment changes suggests strongly that they are ecologically distinct.

Enormous population sizes and immense physical mixing probably played a role in the evolution of diverse genomic backbones in *Prochlorococcus*. A simple fluid mechanics model bridging the micrometer and kilometer scales for a typical ocean suggests that just-divided cells will be centimeters apart within minutes, tens of meters apart within an hour, and a few kilometers apart within a week (*15*). Thus, *Prochlorococcus* populations are expected to be well mixed over large water parcels (~10 km^2 area by 3 m depth) on ecologically relevant time scales (~1 week) (*15*). This mixing and a stable collective *Prochlorococcus* population density of 10^7 to 10^8 cells $liter^{-1}$ (*17*) make the size of each backbone subpopulation in such parcels enormous (>10^{13} cells) (*15*). The effective population size is arguably close to this census population size (*15*), implying that *Prochlorococcus* evolution is governed by selection, not genetic drift [based on population genetics theory (*26*)]. Consistent with this argument, the difference in the observed F_{ST} distribution from that estimated for no selection (Fig. 3B) provides further evidence that the differentiation of genomic backbones in *Prochlorococcus* is a product of selection (*15*).

The correlation between phylogeny and flexible gene content (Table 1, tables S1 and S13, and fig. S5) leads us to propose that the emergence of a genomic backbone is initiated by the acquisition of a beneficial flexible gene cassette, followed by slow fine-adjustment of the core gene alleles to the new niche dimension afforded by the acquired cassette. Given the huge effective population size, even extremely weak fitness differentials among alleles (*27*) can facilitate fine-adjustment of core genes (*15*) over the millions of years of evolution after divergence.

The diverse set of hundreds of subpopulations with distinct genomic backbones probably plays an important role in the dynamic stability of the *Prochlorococcus* "collective" in the global oceans (fig. S6). Small fitness differentials, niche differentiation, and selective phage and grazer predation, in the context of temporal and spatial environmental variation, help to explain their coexistence (*28*, *29*). On seasonal time scales, the *Prochlorococcus* collective maintains a relatively stable population size through temporal and local adjustments in the relative abundance of backbone subpopulations (Figs. 1C and 4A and fig. S6D). On longer time scales (decades to millions of years), the collective may respond to shifting selective pressures through the exchange of gene cassettes between and within backbone subpopulations, and through the evolution of the backbones themselves. The coherence of the collective population holds as long as subpopulations do not diverge to the point where they are no longer able to exchange flexible genes and backbone extinction and emergence rates are relatively balanced. If *Prochlorococcus* backbone subpopulations were designated as distinct species (*30*), it would imply that the global collective is an assortment of thousands of species. It is likely that such a large set of coexisting subpopulations with distinct genomic backbones is a characteristic feature of free-living bacterial species with very large population sizes living in highly mixed habitats.

References and Notes

1. F. Partensky, W. R. Hess, D. Vaulot, *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999).
2. L. R. Moore, G. Rocap, S. W. Chisholm, *Nature* **393**, 464–467 (1998).
3. Z. I. Johnson *et al.*, *Science* **311**, 1737–1740 (2006).
4. G. C. Kettler *et al.*, *PLOS Genet.* **3**, e231 (2007).
5. J. Grote *et al.*, *MBio.* **3**, e00252-12 (2012).
6. D. E. Hunt *et al.*, *Science* **320**, 1081–1085 (2008).
7. S. L. Simmons *et al.*, *PLOS Biol.* **6**, e177 (2008).
8. H. Cadillo-Quiroz *et al.*, *PLOS Biol.* **10**, e1001265 (2012).
9. A. Gonzaga *et al.*, *Genome Biol. Evol.* **4**, 1360–1374 (2012).
10. R. T. Papke *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14092–14097 (2007).
11. S. Rodrigue *et al.*, *PLOS ONE* **4**, e6864 (2009).
12. T. Kalisky, P. Blainey, S. R. Quake, *Annu. Rev. Genet.* **45**, 431–445 (2011).
13. R. Stepanauskas, *Curr. Opin. Microbiol.* **15**, 613–620 (2012).
14. R. S. Lasken, *Nat. Rev. Microbiol.* **10**, 631–640 (2012).
15. Materials and methods are available as supplementary materials on *Science* Online.
16. A. F. Michaels *et al.*, *Deep Sea Res. Part I* **41**, 1013–1038 (1994).
17. R. R. Malmstrom *et al.*, *ISME J.* **4**, 1252–1264 (2010).
18. O. Wurtzel, M. Dori-Bachash, S. Pietrokovski, E. Jurkevitch, R. Sorek, *PLOS ONE* **5**, e15628 (2010).
19. M. Mühling, *Environ. Microbiol.* **14**, 567–579 (2012).
20. M. Nei, in *Human Genetics, Part A: The Unfolding Genome*, B. Bonné-Tamir, T. Cohen, R. M. Goodman, Eds. (Alan R. Liss, New York, 1982), p. 167.
21. R. Mehra-Chaudhary, J. Mick, L. J. Beamer, *J. Bacteriol.* **193**, 4081–4087 (2011).
22. S. Avrani, O. Wurtzel, I. Sharon, R. Sorek, D. Lindell, *Nature* **474**, 604–608 (2011).
23. J. Pernthaler, *Nat. Rev. Microbiol.* **3**, 537–546 (2005).
24. F. Malfatti, F. Azam, *Aquat. Microb. Ecol.* **58**, 1–14 (2009).
25. U. Dobrindt, B. Hochhut, U. Hentschel, J. Hacker, *Nat. Rev. Microbiol.* **2**, 414–424 (2004).
26. J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).

27. R. D. Barrett, D. Schluter, *Trends Ecol. Evol.* **23**, 38–44 (2008).
28. A. D. Barton, S. Dutkiewicz, G. Flierl, J. Bragg, M. J. Follows, *Science* **327**, 1509–1511 (2010).
29. F. Rodriguez-Valera *et al.*, *Nat. Rev. Microbiol.* **7**, 828–836 (2009).
30. C. C. Thompson *et al.*, *Microb. Ecol.* **66**, 752–762 (2013).

Acknowledgments: We thank S. Itzkovitz, P. H. R. Calil, D. Sher, R. Milo, P. M. Berube, A. P. Yelton, R. Braakman, and particularly M. F. Polz for comments on the manuscript. We thank the Bermuda Atlantic Time-series Study for sample collection, the Bigelow Laboratory Single Cell Genomics Center for single-cell sorting and whole-genome amplification, and the BioMicroCenter facility at MIT for their contributions to the generation of genomic data. N.K. acknowledges the Rothschild Foundation (Yad Hanadiv) and the National Oceanic and Atmospheric Administration "Climate and Global Change" Postdoctoral Research Fellowships. This work was supported by grants to S.W.C. from the NSF Evolutionary Biology Section and Biological Oceanography Section, the NSF Center for Microbial Oceanography Research and Education (C-MORE), the U.S. Department of Energy (DOE)–GTL, and the Gordon and Betty Moore Foundation Marine Microbiology Initiative; to R. Stepanauskas from the NSF Biological Oceanography Section; and to R.R.M from the DOE (contract number DE-AC02-05CH11231). Genomic data have been deposited in National Center for Biotechnology Information GenBank under accession numbers KJ477896 to KJ479276 and JFKN00000000 to JFOE00000000. Additional data files have been deposited to Dryad (doi:10.5061/dryad.9r0p6).

Supplementary Materials
www.sciencemag.org/content/344/6182/416/suppl/DC1
Materials and Methods
Figs. S1 to S21
Tables S1 to S13
References (*31–91*)
Data S1

18 November 2013; accepted 20 March 2014
10.1126/science.1248575

Structure-Guided Transformation of Channelrhodopsin into a Light-Activated Chloride Channel

Andre Berndt,[1]* Soo Yeun Lee,[1]* Charu Ramakrishnan,[1] Karl Deisseroth[1,2,3]†

Using light to silence electrical activity in targeted cells is a major goal of optogenetics. Available optogenetic proteins that directly move ions to achieve silencing are inefficient, pumping only a single ion per photon across the cell membrane rather than allowing many ions per photon to flow through a channel pore. Building on high-resolution crystal-structure analysis, pore vestibule modeling, and structure-guided protein engineering, we designed and characterized a class of channelrhodopsins (originally cation-conducting) converted into chloride-conducting anion channels. These tools enable fast optical inhibition of action potentials and can be engineered to display step-function kinetics for stable inhibition, outlasting light pulses and for orders-of-magnitude-greater light sensitivity of inhibited cells. The resulting family of proteins defines an approach to more physiological, efficient, and sensitive optogenetic inhibition.

The microbial opsins (*1–3*) used for optical control of genetically targeted cellular activity (*4–7*) include light-activated proton and Cl^- pumps and the cation channels called channelrhodopsins (ChRs). ChRs are derived from algae (*3*, *8–10*) and, when expressed in neurons, can elicit precise action potential (AP) firing (*11–15*). ChRs conduct K^+, Na^+, protons, and Ca^{2+} (*3*, *10*, *16*, *17*); because of this nonselective cation-conductance, ChRs display reversal potentials (V_{rev}) near 0 mV under physiological conditions and therefore depolarize neurons, leading to AP generation (*18*).

Direct light-triggered inhibition of neuronal activity is possible with inward-pumping Cl^--transporting opsins and outward-pumping proton-transporting opsins (*10*); hyperpolarization to −150 mV or beyond can be achieved (*18–20*). However, pumps are inefficient in neural systems because only one ion is moved per photon and no input resistance decrease is elicited (failing to recruit the most potent mechanism of spiking inhibition). Moreover, because the pumps use energy to transport ions against electrochemical gradients, the creation of abnormal gradients is more likely (*18*). Last, pumps cannot take advantage of certain molecular engineering opportunities to achieve light sensitivity and long-term photocurrent stability enhanced by many orders of magnitude (but which depend on formation of a transmembrane pore) (*21–23*). Therefore, the creation of inhibitory channels has long been a central goal of optogenetics.

Given typical ion balance in neural systems, identification or creation of light-activated K^+ or Cl^- channels could give rise to inhibitory optogenetic tools. ChRs can be engineered to alter kinetics, spectrum, and selectivity among cations (*10*, *24*, *25*). However, V_{rev} has not been shifted sufficiently for nondepolarizing spike inhibition in neurons. We have designed a family of ChRs for Cl^- permeability and capability to inhibit APs without depolarizing neurons to or beyond the AP-generation threshold.

Building on the high-resolution crystal structure of the ChR chimera C1C2 (*24*), we noted that the ion-selectivity pore of ChR is less ordered as compared with the well-defined symmetry of tetrameric K^+-selective channels such as KcsA and NaK2K (*26–31*). Therefore, we speculated that the specific cation selectivity of ChR is rather a result of negative electrostatic potential surrounding the pore and vestibule; for instance, the C1C2 structure shows seven glutamates framing the conduction pathway (*24*). We hypothesized that systematic replacement of such residues within or close to the pore according to structure-guided electrostatic modeling could reverse this polarity and create an inhibitory ChR, if it were possible to maintain proper protein folding, membrane expression, optical activation, and pore gating.

We initiated a broad structure-guided screen by introducing single site-directed mutations into C1C2 (Fig. 1A). We expressed all variants in cultured rat hippocampal neurons and tested photocurrents using whole-cell patch-clamp so as to ensure proper function in neurons (external/internal $[Cl^-]$, 147 mM/4 mM). We quantified stationary photocurrent amplitudes across a range of holding potentials (Fig. 1B), with particular attention to V_{rev}, in order to identify permeability variants (Fig. 1C). C1C2 exhibits V_{rev} of −7 mV under these conditions, which is typical for nonspecific cation channels (*16*, *26*, *32*, *33*). Certain mutations with powerful effects on V_{rev} displayed concomitant adverse effects on photocurrent sizes (such as E136R and E140K) (Fig. 1B), and were not studied further (*34*). More promising mutations, such as N297Q and H173R, exhibited both potent currents and altered V_{rev} (Fig. 1C) and were combined in a series of increasingly integrated mutations. The fivefold mutation T98S/E129S/E140S/E162S/T285N and fourfold mutation V156K/H173R/V281K/N297Q both displayed prominently-shifted V_{rev} (in the range of −40 mV) while maintaining functionality (Fig. 1, D and E).

We next combined these constructs to generate a ninefold mutated variant with contiguous shifts in expected electrostatic potential distribution (Fig. 2A and fig. S1) (*24*). We expressed the ninefold variant in human embryonic kidney (HEK) 293 cells to test both V_{rev} and permeability under controlled ion composition and optimized voltage clamp settings (Fig. 2B). We mapped photocurrents over a broad range of membrane potentials (Fig. 2C) (from −75 mV to +55 mV) (*35*). Under these conditions (external/internal $[Cl^-]$, 147mM/4mM), the combined ninefold mutation exhibited V_{rev} of −61 mV, which is far more negatively shifted than was the C1C2 backbone or either parental 4× or 5× construct (Fig. 2D). Despite this major change in functionality, both peak and stationary photocurrents remained fast and robust (predicting suitability for optogenetics, especially because this channel could also recruit a reduced-membrane resistance mechanism for spiking inhibition), and the original blue light–activation spectrum of C1C2

[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA. [2]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA. [3]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA.
*These authors contributed equally to this work.
†Corresponding author. E-mail: deissero@stanford.edu

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Environmental microbiology

» Genomics

Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*

Steven J. Biller[1], Paul M. Berube[1], Jessie W. Berta-Thompson[1,2], Libusha Kelly[1,†], Sara E. Roggensack[1], Lana Awad[1], Kathryn H. Roache-Johnson[3], Huiming Ding[1,4], Stephen J. Giovannoni[5], Gabrielle Rocap[6], Lisa R. Moore[3] & Sallie W. Chisholm[1,4]

Received: 10 June 2014
Accepted: 19 August 2014
Published: 30 September 2014

The marine cyanobacterium *Prochlorococcus* is the numerically dominant photosynthetic organism in the oligotrophic oceans, and a model system in marine microbial ecology. Here we report 27 new whole genome sequences (2 complete and closed; 25 of draft quality) of cultured isolates, representing five major phylogenetic clades of *Prochlorococcus*. The sequenced strains were isolated from diverse regions of the oceans, facilitating studies of the drivers of microbial diversity—both in the lab and in the field. To improve the utility of these genomes for comparative genomics, we also define pre-computed clusters of orthologous groups of proteins (COGs), indicating how genes are distributed among these and other publicly available *Prochlorococcus* genomes. These data represent a significant expansion of *Prochlorococcus* reference genomes that are useful for numerous applications in microbial ecology, evolution and oceanography.

Design Type(s)	observation design • individual genetic characteristics comparison design • strain comparison design
Measurement Type(s)	genome sequencing
Technology Type(s)	next generation sequencing
Factor Type(s)	
Sample Characteristic(s)	*Prochlorococcus* • ocean biome

[1]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Microbiology Graduate Program, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Department of Biological Sciences, University of Southern Maine, Portland, Maine, USA. [4]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Microbiology, Oregon State University, Corvallis, Oregon, USA. [6]School of Oceanography, Center for Environmental Genomics, University of Washington, Seattle, Washington, USA. †Present address: Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, New York, USA.
Correspondence and requests for materials should be addressed to S.J.B. (email: sbiller@mit.edu) or to S.W.C. (email: chisholm@mit.edu).

Background & Summary

As the smallest (< 1 μm diameter) and most abundant (3×10^{27} cells) photosynthetic organism on the planet[1], *Prochlorococcus* has a unique status in the microbial world. This unicellular marine cyanobacterium is found throughout the euphotic zone of the open ocean between ~45 °N and 40 °S, where it carries out a notable fraction of global photosynthesis[1,2]. The group, which would be considered a single microbial 'species' by the traditional measure of >97% 16S rRNA similarity, is composed of multiple phylogenetically distinct clades (Figure 1) (as defined by either rRNA internal transcribed spacer (ITS)[3] or whole-genome sequences[4]) which are physiologically distinct. Adaptations for optimal growth at different light intensities differentiate deeply branching groups of *Prochlorococcus* into high light (HL) and low light (LL) adapted clades[3,5–8].

Prochlorococcus have the smallest genomes of any known free-living photosynthetic cell, ranging from ~1.6 to 2.7 Mbp[4]. While they all share a core set of genes present in all strains, there exists remarkable

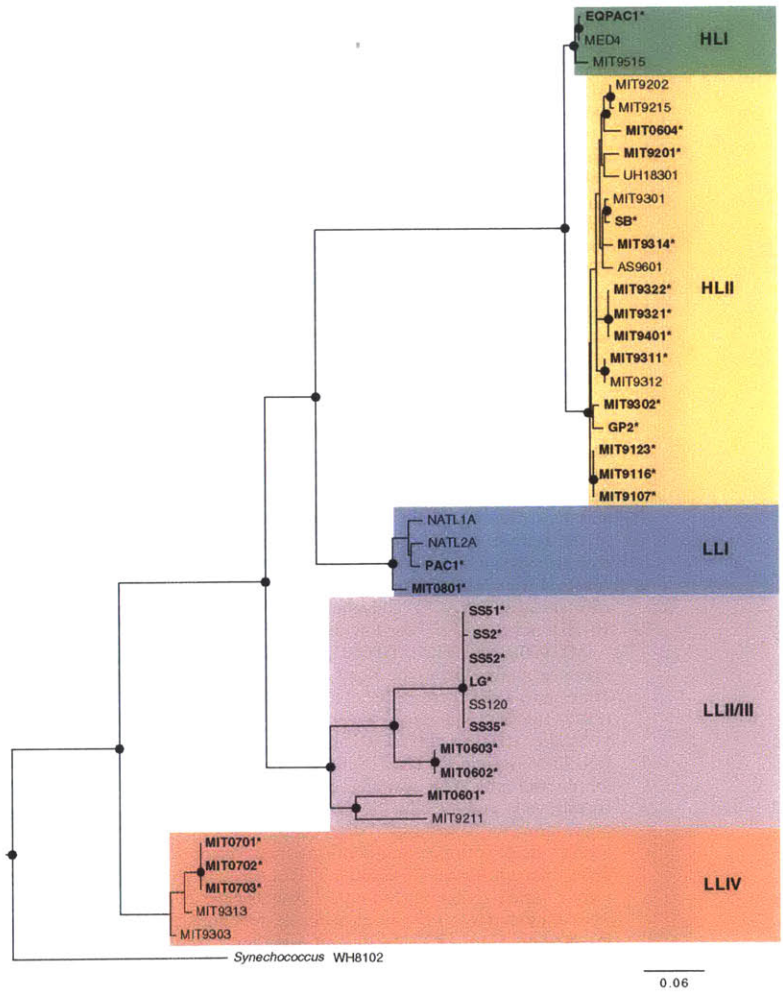


Figure 1. *Prochlorococcus* strains sequenced in this work. ITS-based phylogeny of the strains included in this data set (names in bold, with *) in relation to previously sequenced *Prochlorococcus*. Phylogenetic clade affiliation[4,6] is indicated at right; closed circles indicate nodes with bootstrap support >75%. HL—High light adapted; LL—Low light adapted, as determined by physiological studies of some of the isolates[3,5,7].

diversity in gene content among isolates. The group has an 'open' pan-genome, i.e. each newly sequenced genome typically contains many new genes never before seen in *Prochlorococcus*[4]. Given the abundance of *Prochlorococcus*, studies of their genomic and metagenomic features have provided numerous insights into features of ocean ecosystems[9–17]. In addition, the group has proven to be a valuable system for studying microbial evolution[18,19], genome streamlining[20,21], and the relationship between genotypic, phenotypic and ecological variation in marine populations[3,7,22]. Since *Prochlorococcus* is abundant in surface waters, these reference genomes have also been extremely valuable for interpreting marine metagenomic and metatranscriptomic datasets[14,23–28].

To advance our understanding of *Prochlorococcus* genetic diversity, we sequenced the genomes of 27 *Prochlorococcus* strains from a variety of ocean environments. The strains sequenced included both previously reported strains as well as eight new isolates (Table 1). The newly isolated strains come from ocean regions that previously only had few or no cultured representatives and substantially expand the number of cultured *Prochlorococcus* available for five major clades. These results demonstrate the applicability of high-throughput dilution-to-extinction cultivation approaches[29] to *Prochlorococcus*.

The genome sequences reported here represent a notable increase in the number of genome sequences available from the major phylogenetic clades with existing cultured representatives. While many genomes differed greatly in gene content, other sets are very closely related and differ primarily by single nucleotide polymorphisms (e.g., LG, SS2, SS35, SS51, SS52, SS120; and MIT0701, MIT0702, and MIT0703). Thus, this dataset encompasses a broad range of pairwise genomic diversity among *Prochlorococcus* strains.

Most genomes were sequenced to draft status; two were closed (Table 2). We used two annotation methods to identify the potential functions of genes in the genomes. Genes were first called and annotated by the RAST pipeline[30]. To expand on these predictions—especially for the myriad genes of unknown function—we also derived annotations from an independent pipeline, Argot2[31]. To facilitate the utility of these genomes for comparative genomics and evolutionary studies, we define a set of pre-computed orthologous gene clusters for *Prochlorococcus*. All cluster data are supplied in this data set (Data Citation 1 and Data Citation 2).

These genomes should be useful to researchers interested in many aspects of marine microbial ecology and evolution. Since the genomes are from cultured isolates, hypotheses generated from these data can be tested in laboratory experiments. The genomes will also greatly facilitate the interpretation of transcriptomic and proteomic studies, as well as meta-'omic' data from field studies where *Prochlorococcus* is a dominant phototroph.

Methods

Culturing and strain isolations

Many of the strains sequenced have been previously described[3,5,6,32–36] (Table 1); 8 are reported here for the first time. All cultures were unialgal; this was initially determined crudely by flow cytometry profiles, and then more specifically by confirming the presence of only one cyanobacterial 16S rRNA ITS sequence in the culture. All cultures except SB and MIT0604 contained heterotrophic bacteria. Cultures were maintained in acid-washed glassware in Pro99 media[37] prepared with 0.2 μm filtered, autoclaved seawater collected from Vineyard Sound, MA or the Sargasso Sea under either a 14:10 light:dark cycle at 24 °C or constant light flux at 21 °C. Light levels were 30–40 μmol Q m^{-2} s^{-1} for high-light adapted strains, and 10–20 μmol Q m^{-2} s^{-1} for low-light adapted strains.

MIT0601, MIT0602, MIT0603, and MIT0604 were derived from enrichment cultures initiated with seawater obtained from the North Pacific Ocean at Station ALOHA (22.75°N, 158°W) on Hawai'i Ocean Time-series (HOT) cruise 181. The seawater was amended with nitrogen, phosphorous and trace metals (PRO2 nutrient additions[37], except all nitrogen sources were replaced by 0.217 mM sodium nitrate).

Strains MIT0701, MIT0702, and MIT0703 were isolated from the South Atlantic (CoFeMUG cruise KN192-05, station 13, 13.45 °S, 0.04 °W) at 150 m using a high throughput culturing method[29] adapted for phototrophs. The seawater used for isolations was first filtered through a 1 μm filter with no amendments and kept in the dark at 18–20 °C for 21 days. The total red fluorescing phytoplankton population (1×10^5 cells ml^{-1} determined with a Guava EasyCyte flow cytometer) was diluted in PRO3V media[37] made with the same South Atlantic water that had been filtered through a 0.1 μm Supor 142 mm filter, then autoclaved to sterilize. This media contained 100 μM NH_4Cl, 10 μM NaH_2PO_4, PRO2 trace metals[37] and f/2 vitamins (0.1 μg l^{-1} cyanocobalamin, 20 g l^{-1} thiamin and 1 μg l^{-1} biotin[38,39]). Ten cells were dispensed into 1 ml volumes in a 48-well polystyrene multiwell culture plate and incubated at 20 °C in ~20 μmol Q m^{-2} s^{-1} (14:10 light:dark) for 2 months.

MIT0801 was isolated in a similar manner, but from seawater obtained from 40 m depth at the Bermuda Atlantic Time-series station (BATS; 31.67 °N, 64.16 °W) that had been sitting in the dark for 5 days. The same PRO3V media recipe was made with 0.1 μm filtered and autoclaved BATS seawater, and 2.5 cells (on average) were dispensed in 5 ml volume in Teflon plates (prepared as described[29]). Cells were detected within 1 month of enrichment.

DNA sequencing and assembly

Genomes were sequenced from genomic DNA collected from 20 ml laboratory cultures. Cells were collected by centrifugation (10,000 *g*, 10 min), the pellet transferred into a 2 ml tube and

Strain	Alternate Name	Ecotype/Clade [4,57]	Isolation location	Isolation (Lat/Lon)	Isolation depth (m)	Isolation date	Strain reference
EQPAC1	RCC278	eMED4/HLI	Equatorial Pacific	0°N 180°W	30		Roscoff Culture Collection
GP2		eMIT9312/HLII	Western Pacific	8°N 136°E	150	Sep-1992	32
MIT0604		eMIT9312/HLII	Station ALOHA/ North Pacific	22.75°N 158°W	175	May-2006	This work
MIT9107		eMIT9312/HLII	Tropical Pacific	15°S 135°W	25	8-Aug-1991	33
MIT9116		eMIT9312/HLII	Tropical Pacific	15°S 135°W	25	8-Aug-1991	6
MIT9123		eMIT9312/HLII	Tropical Pacific	15°S 135°W	25	8-Aug-1991	6
MIT9201		eMIT9312/HLII	Tropical Pacific	12°S 145.42°W	Surface	26-Sep-1992	5
MIT9302		eMIT9312/HLII	Sargasso Sea	34.76°N 66.19°W	100	15-Jul-1993	3
MIT9311		eMIT9312/HLII	Gulf stream	37.51°N 64.24°W	135	17-Jul-1993	6
MIT9314		eMIT9312/HLII	Gulf stream	37.51°N 64.24°W	180	17-Jul-1993	6
MIT9321		eMIT9312/HLII	Equatorial Pacific	1°N 92°W	50	12-Nov-1993	6
MIT9322		eMIT9312/HLII	Equatorial Pacific	0.27°N 93°W	Surface	16-Nov-1993	6
MIT9401		eMIT9312/HLII	Sargasso Sea	35.5°N 70.4°W	Surface	May-1994	6
SB		eMIT9312/HLII	Western Pacific	35°N 138.3°E	40	1-Oct-1992	32
MIT0801	HTCC 1603	eNATL/LLI	BATS/Sargasso Sea	31.67°N 64.17°W	40	25-Mar-2008	This work
PAC1		eNATL/LLI	Station ALOHA/ North Pacific	22.75°N 158°W	100	1992	34,35
LG		eSS120/LLII,III	Sargasso Sea	28.98°N 64.35°W	120	30-May-1988	36
MIT0601		eMIT9211/LLII,III	Station ALOHA/ North Pacific	22.75°N 158°W	125	17-Nov-2006	This work
MIT0602		eSS120/LLII,III	Station ALOHA/ North Pacific	22.75°N 158°W	125	17-Nov-2006	This work
MIT0603		eSS120/LLII,III	Station ALOHA/ North Pacific	22.75°N 158°W	125	17-Nov-2006	This work
SS2		eSS120/LLII,III	Sargasso Sea	28.98°N 64.35°W	120	30-May-1988	6
SS35		eSS120/LLII,III	Sargasso Sea	28.98°N 64.35°W	120	30-May-1988	6
SS51		eSS120/LLII,III	Sargasso Sea	28.98°N 64.35°W	120	30-May-1988	6
SS52		eSS120/LLII,III	Sargasso Sea	28.98°N 64.35°W	120	30-May-1988	6
MIT0701	HTCC 1600	eMIT9313/LLIV	South Atlantic	13.45°S 0.04°W	150	1-Dec-2007	This work
MIT0702	HTCC 1601	eMIT9313/LLIV	South Atlantic	13.45°S 0.04°W	150	1-Dec-2007	This work
MIT0703	HTCC 1602	eMIT9313/LLIV	South Atlantic	13.45°S 0.04°W	150	1-Dec-2007	This work

Table 1. Origin of the *Prochlorococcus* strains sequenced in this study.

frozen at −80 °C. Genomic DNA was isolated using the QIAamp DNA mini kit (Qiagen). 2 µg of DNA was then used to construct an Illumina sequencing library as previously described[40], except that the bead: sample ratios in the double solid phase reversible immobilization (dSPRI) size-selection step were 0.7 followed by 0.15, resulting in fragments with an average size of ~340 bp (range: 200–600 bp). PAC1 and

Strain	Clade[4]	Assembly size (bp)	%GC	No. contigs	N50 (bp)	No. coding sequences	NCBI accession*
EQPAC1	HLI	1,654,739	30.8	8	328,627	1,954	JNAG00000000
GP2	HLII	1,624,310	31.2	11	416,038	1,884	JNAH00000000
MIT0604	HLII	1,780,061	31.2	1	1,780,061	2,085	CP007753
MIT9107	HLII	1,699,937	31.0	13	170,362	1,991	JNAI00000000
MIT9116	HLII	1,685,398	31.0	22	117,620	1,972	JNAJ00000000
MIT9123	HLII	1,697,748	31.0	18	137,374	2,005	JNAK00000000
MIT9201	HLII	1,672,416	31.3	21	145,955	1,989	JNAL00000000
MIT9302	HLII	1,745,343	31.1	17	242,124	2,015	JNAM00000000
MIT9311	HLII	1,711,064	31.2	17	189,094	1,983	JNAN00000000
MIT9314	HLII	1,690,556	31.2	16	221,824	1,990	JNAO00000000
MIT9321	HLII	1,658,664	31.2	10	259,210	1,956	JNAP00000000
MIT9322	HLII	1,657,550	31.2	11	367,597	1,959	JNAQ00000000
MIT9401	HLII	1,666,808	31.2	17	110,519	1,972	JNAR00000000
SB	HLII	1,669,823	31.5	4	1,237,529	1,933	JNAS00000000
MIT0801	LLI	1,929,203	34.9	1	1,929,203	2,287	CP007754
PAC1	LLI	1,841,163	35.1	20	182,484	2,264	JNAX00000000
LG	LLII,III	1,754,063	36.4	14	326,623	1,973	JNAT00000000
MIT0601	LLII,III	1,707,342	37.0	6	547,047	1,934	JNAU00000000
MIT0602	LLII,III	1,750,918	36.3	9	511,704	1,998	JNAV00000000
MIT0603	LLII,III	1,752,482	36.3	7	434,668	2,015	JNAW00000000
SS2	LLII,III	1,752,772	36.4	19	187,268	1,989	JNAY00000000
SS35	LLII,III	1,751,015	36.4	9	446,270	1,977	JNAZ00000000
SS51	LLII,III	1,746,977	36.4	12	232,789	1,974	JNBD00000000
SS52	LLII,III	1,754,053	36.4	22	124,224	1,987	JNBE00000000
MIT0701	LLIV	2,592,571	50.6	53	84,463	3,079	JNBA00000000
MIT0702	LLIV	2,583,057	50.6	61	76,101	3,066	JNBB00000000
MIT0703	LLIV	2,575,057	50.6	61	81,186	3,054	JNBC00000000

Table 2. Genome characteristics and assembly statistics. *For the Whole Genome Shotgun projects deposited at DDBJ/EMBL/GenBank: the version described in this paper is version JN**01000000.

EQPAC1 libraries were constructed using dSPRI bead:sample ratios of 0.9 followed by 0.21, yielding an average size of ~220 bp. DNA libraries were sequenced on an Illumina GAIIx, producing 200+200 nt paired reads, at the MIT BioMicro Center. An average of 1.6 million paired-end reads were obtained for each genome.

Low quality regions of sequencing data were removed from the raw Illumina data using quality_trim (V3.2, from the CLC Assembly Cell package; CLC bio) with default settings (at least 50% of the read must be of a minimum quality of 20). Paired-end reads were overlapped using the SHE-RA algorithm[41], keeping any resulting overlapping sequences with an overlap score >0.5. For all genomes except PAC1 and EQPAC1, the overlapped reads, as well as the trimmed paired-end reads that did not overlap, were assembled using the Newbler assembler (V2.6; 454/Roche) with the following parameters: '-e 200 -rip.' Contigs < 1 Kbp were discarded at this stage.

Reads for PAC1 and EQPAC1 were assembled using clc_novo_assemble (V3.2, from the CLC Assembly Cell package; CLC bio) with a minimum contig length of 500 bp and automatic wordsize determination enabled. These initial contigs were searched against a custom database of marine microbial genomes[9] using BLAST[42] to identify contigs with a closest match to *Prochlorococcus*. Sequencing reads belonging to the putative *Prochlorococcus* contigs were then identified by mapping the raw sequences to these contigs using clc_ref_asssemble_long (CLC bio). The *Prochlorococcus*-like reads were then re-assembled using clc_novo_assemble using the same parameters as above to produce the final assembly, now largely free of heterotrophic sequences.

MIT0604 and MIT0801 were completed to finished quality with no gaps by directed PCR reactions to sequence contig junctions, combined with Pacific Biosciences long sequencing reads. Contigs were ordered into putative scaffolds based on their similarity to closely related closed *Prochlorococcus* genomes, as determined by Mauve[43]. PCR primers specific to the ends of putatively adjacent contigs were designed and used to amplify the junctions between contigs. Purified PCR products were sequenced by Sanger chemistry at the MGH DNA core facility, and the resulting sequences used to join contigs in Consed[44]. This resulted in a highly improved but still incomplete assembly. To span difficult repeat regions in MIT0801, we obtained long Pacific Biosciences sequences. We obtained DNA from 25 ml cultures using the Epicentre Masterpure kit (Epicentre) and sequenced this at the Yale Center for Genome Analysis. We combined this set of long but low quality reads with the high quality Illumina short reads obtained previously using the PacBioToCA software[45], to produce assemblies with a reduced number of contigs. These contigs were aligned to the PCR-improved contigs described above, and the final gaps were closed with a small number of additional directed PCR reactions (as described above) using the Geneious sequence analysis package (V6.1, Biomatters), until the genomes were closed.

As most of the *Prochlorococcus* cultures sequenced were known to contain heterotrophs, we identified the most '*Prochlorococcus*-like' contigs from non-axenic cultures by searching each resulting contig against a custom database of sequenced marine microbial genomes[9] using BLAST[42]. Contigs with a best match to a non-*Prochlorococcus* genome were removed from the assembly. Subsequent examination of these contig sets indicated that a number of shorter sequences (generally < 10 kbp) with significant heterotroph-like stretches had passed through the initial filtering steps. To remove these questionable contigs from the assemblies, we manually examined each < 10 kbp contig using the RAST annotation server (see below), and only kept those contigs with clear homology to previously sequenced and closed *Prochlorococcus* or *Synechococcus* genomes. Although these filtering steps may have removed a small amount of true *Prochlorococcus* sequence from the final assembly, we considered missing a few genes preferable to misrepresenting heterotroph sequences as *Prochlorococcus*.

Examination of the non-cyanobacterial 16S rRNA genes found within these data indicate that the most abundant heterotrophs in the cultures were members of the *Alteromonadales*, *Flavobacteriales*, *Rhodospirillales*, *Halomonadaceae*, and *Sphingobacteriales*. We have included a separate data file containing all of the assembled contigs—including those from co-cultured heterotrophs—for anyone who is interested (Data File 4).

Genome annotation

The assembled contigs for each genome were annotated using the RAST server[30] against FIGfam release 49. Additional functional annotation for all genes called by RAST were generated by the Argot2 server[31], using default settings.

To confirm the rRNA-based phylogeny of these strains, rRNA ITS sequences were aligned in ARB[46] and maximum likelihood phylogenies calculated in PhyML version 20120412[47], using the HKY85 model of nucleotide substitution, a fixed proportion of invariable sites, and non-parametric bootstrap analysis with 100 replicates.

Clusters of orthologous groups of proteins (COGs) were computed, as described elsewhere[48], on a data set comprised of previously sequenced *Prochlorococcus* and *Synechococcus* strains[4,10,16,17,49–53], the new *Prochlorococcus* genomes described here, 11 *Prochlorococcus* single-cell genomes[12] and two consensus metagenomic assemblies[14] (Data Citation 1). To facilitate comparisons among genomes, we re-annotated 16 previously sequenced *Prochlorococcus* genomes (Table 3) with the RAST pipeline as described above; this ensured that a uniform methodology for gene calling and functional annotation was used. Single cell genomes[12] were not re-annotated due to difficulties encountered using this pipeline on such fragmented contigs; instead, we utilized the ORFs previously defined in GenBank. Detailed information regarding these updated annotations is provided (Data Citation 1 and Data Citation 2).

Orthologous gene clusters were defined based on reciprocal best blastp scores (with an e-value cutoff of 1e−5); the sequence alignment length had to be at least 75% of the shorter protein, with at least a 35% identity. Additional orthologous genes that did not pass this criterion were added to clusters based on HMM profiles constructed from automated MUSCLE[54] alignments of orthologous sequences within each cluster using HMMER[55]. The clusters described here are noted as 'V4' CyCOGs in the associated Data Records and on the ProPortal website[48] (Data Citation 1).

Data Records

The complete dataset is available from the *Prochlorococcus* Portal website (Data Citation 1) and Dryad (Data Citation 2). The 27 *Prochlorococcus* genome sequences have also been deposited at DDBJ/EMBL/GenBank (Data Citations 3–29) under the accession numbers indicated in Table 2.

Name	Genome source	Clade	Assembly size (bp)	%GC	No. coding sequences*	NCBI accession	Sequence reference
MED4	Cultured isolate	HLI	1,657,990	30.8	1,959	BX548174	10
MIT9515	Cultured isolate	HLI	1,704,176	30.8	1,951	CP000552	4
AS9601	Cultured isolate	HLII	1,669,886	31.3	1,944	CP000551	4
MIT9202	Cultured isolate	HLII	1,691,453	31.1	2,000	DS999537	49
MIT9215	Cultured isolate	HLII	1,738,790	31.1	2,035	CP000825	4
MIT9301	Cultured isolate	HLII	1,641,879	31.3	1,925	CP000576	4
MIT9312	Cultured isolate	HLII	1,709,204	31.2	1,982	CP000111	16
UH18301	Cultured isolate	HLII	1,654,648	31.2	1,947	PRJNA47033	50
W6	Single cell amplified genome	HLII	385,307	31.3	646	ALPK00000000	12
HNLC2	Metagenomic assembly	HLIII	1,484,494	30.3	1,701	GL947595	14
W3	Single cell amplified genome	HLIII	339,045	30.7	529	ALPC00000000	12
W5	Single cell amplified genome	HLIII	99,467	29.8	212	ALPL00000000	12
W7	Single cell amplified genome	HLIII	905,221	30.7	989	ALPE00000000	12
W8	Single cell amplified genome	HLIII	841,756	31.4	917	ALPF00000000	12
W9	Single cell amplified genome	HLIII	420,150	30.7	638	ALPG00000000	12
HNLC1	Metagenomic assembly	HLIV	1,569,623	29.8	1,830	GL947594	14
W10	Single cell amplified genome	HLIV	561,998	30.8	892	ALPH00000000	12
W11	Single cell amplified genome	HLIV	766,829	30.6	929	ALPI00000000	12
W12	Single cell amplified genome	HLIV	423,437	29.6	602	ALPJ00000000	12
W2	Single cell amplified genome	HLIV	1,266,767	30.5	1,374	ALPB00000000	12
W4	Single cell amplified genome	HLIV	765,485	29.9	819	ALPD00000000	12
NATL1A	Cultured isolate	LLI	1,864,731	35.0	2,242	CP000553	4
NATL2A	Cultured isolate	LLI	1,842,899	35.1	2,194	CP000095	4
MIT9211	Cultured isolate	LLII,III	1,688,963	38.0	1,943	CP000878	4
SS120	Cultured isolate	LLII,III	1,751,080	36.4	1,973	AE017126	17
MIT9303	Cultured isolate	LLIV	2,682,675	50.0	3,253	CP000554	4
MIT9313	Cultured isolate	LLIV	2,410,873	50.7	2,993	BX548175	10

Table 3. Previously sequenced *Prochlorococcus* genomes included in the cyanobacterial clusters of orthologous groups of proteins (CyCOG) definitions. *For the cultured isolate and metagenomic assembly genomes, this value represents the number of coding sequences as predicted in this study using the RAST pipeline; these values may differ from those previously published for this reason. Re-annotation data is included in this dataset (Data Citation 1 and Data Citation 2).

Datasets deposited at Dryad and ProPortal

Sequence, gene annotations, and COG definitions for *Prochlorococcus* genomes.

File 1—Tab-delimited file containing cluster assignments and annotation metadata for genes in the newly sequenced *Prochlorococcus* genomes described in this work, as well as previously published genomes. Columns are as follows:

Genome. The *Prochlorococcus* strain where the gene is found.

Gene ID. Unique ID for each *Prochlorococcus* gene, of the format 'P < strain>_####'. Note that, due to the re-annotation of previously published genomes, these names (and the underlying gene boundaries) may not necessarily correspond to those in Genbank.

NCBI ID. For the new genome sequences presented here, the systematic NCBI locus_tag identifier for that gene. For previously published genomes, this column contains the corresponding Genbank locus ID (noted as an 'Alternative locus ID' for strains MED4, SS120 and MIT9313 in Genbank) from Kettler *et al.* (2007)[4].

V1 CyCOG. Where applicable, the cyanobacterial cluster of orthologous groups of proteins (CyCOG) definition from Kettler *et al.* (2007)[4].

V3 CyCOG. Where applicable, the CyCOG definition from Kelly *et al.* (2013)[56].

V4 CyCOG. Number indicating the CyCOG to which this gene belongs, as defined in this work.

RAST annotation. Predicted functional annotation description, as supplied by the RAST annotation pipeline. Note that this text may differ slightly from the annotation in Genbank, due to changes imposed by NCBI annotation formatting guidelines.

GO annotation. Gene Ontology categorization for the gene, when available.

Argot2 annotation. Functional annotation prediction from the Argot2 pipeline, when available.

File 2 – Full RAST gene/protein sequence and annotation results. ZIP format file archive of individual tab-delimited files. Files are supplied for the new genome sequences presented here, as well as re-annotations of previously published genomes included in the CyCOG definitions. Columns are as follows:

contig_id. The name of the sequence contig on which the gene is found.

gene_id. The unique Gene ID code for that feature.

feature_id. Unique RAST-generated identifier for that feature.

type. peg: protein encoding gene; rna: RNA molecule.

location. Ordered location code for the position on the genome merging contig_id, start, and stop position.

start. Start location on contig, bp.

stop. Stop location on contig, bp.

strand. Orientation of gene on contig (+: on forward strand; –: on reverse).

function. The predicted function of the feature, if known.

aliases. Alternative names for the predicted function.

figfam. FigFAM membership for that feature.

evidence_codes. Code indicating the reason for the annotation. See http://www.nmpdr.org/FIG/wiki/view.cgi/FIG/EvidenceCode for more details.

nucleotide_sequence. The nucleotide sequence of the predicted gene.

aa_sequence. The protein (amino acid) sequence of the predicted gene.

File 3 – Set of nucleotide FASTA-formatted files containing the new *Prochlorococcus* genome assemblies described in this work.

File 4 – Set of nucleotide FASTA files containing all assembled contigs (>500 bp) from each culture (i.e., both *Prochlorococcus* and heterotrophs) sequenced in this work. Each file contains the set of contigs assembled from the raw sequencing data, before any filtering to separate *Prochlorococcus* from heterotroph contigs. These files are provided for reference, but due to the known heterotroph sequences in these files, they should be used with caution.

File 5 – Set of nucleotide FASTA files containing the predicted nucleotide sequence for all open reading frames (ORFs) in each genome. This file includes ORFs from both the new genomes presented here as well as the re-annotation of previously released *Prochlorococcus* genomes.

File 6 – Set of protein FASTA files containing the predicted amino acid translation for all ORFs in each genome. This file includes ORFs from both the new genomes presented here as well as the re-annotation of previously released *Prochlorococcus* genomes.

Technical Validation

Phylogenetic analysis of the ITS sequences obtained from these cultured isolate genomes (Figure 1) group these strains into the expected clades[57] as previously determined from directed sequencing of the ITS sequences[6]. We were only able to obtain a single cyanobacterial ITS sequence from the assembled genome contigs, again consistent with these strains being unialgal. *Prochlorococcus* genome size and %GC content are typically quite similar for strains found within the same ITS-defined clade[4], and both the draft and closed genomes are consistent with previously sequenced strains for these measures as well (Table 2).

The quality of the genome assemblies was assessed in multiple ways. Re-mapping of the original Illumina sequencing reads to the final assembled contigs showed that the reads were distributed evenly along the length of the assembly, ruling out some categories of major assembly errors (such as duplicated regions). Whole-genome alignments of contigs against closely related closed reference *Prochlorococcus* genomes indicated that the overall gene order of these contigs was broadly consistent with known sequences, indicating that the sequences do not contain obvious chimeras or other artifacts. We also estimated the completeness of the draft genomes by examining the core gene content of the strains, based on the set of genes shared by all closed *Prochlorococcus* genomes. We found that all of the draft genome assemblies contained >98% of the genes universally present in the 13 previously published closed *Prochlorococcus* genomes, indicating that these contigs represent most (or perhaps all) of the genome sequence.

The final closed sequences of the MIT0604 and MIT0801 genomes were verified in two additional ways. First, we compared the experimentally observed PCR product sizes from directed contig joining reactions to the distances predicted from the final assembled sequence to confirm the assembly. Second, we mapped the original (quality trimmed) Illumina sequencing reads against the final assembly. These alignments indicated that the final closed assembly was fully consistent with the original short-read sequence data. In addition, we confirmed that the per-base SNP frequency was not above the expected error frequency.

References

1. Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl Acad. Sci.* **110,** 9824–9829 (2013).
2. Partensky, F., Hess, W. R. & Vaulot, D. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63,** 106–127 (1999).
3. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393,** 464–467 (1998).
4. Kettler, G. C. *et al.* Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genetics* **3,** e231 (2007).
5. Moore, L. & Chisholm, S. Photophysiology of the marine cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. *Limnol. and Oceanogr.* **44,** 628–638 (1999).
6. Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl. Environ. Microbiol.* **68,** 1180–1191 (2002).
7. Zinser, E. R. *et al.* Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol. Oceanogr.* **52,** 2205–2220 (2007).
8. Johnson, Z. I. *et al.* Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311,** 1737–1740 (2006).
9. Coleman, M. L. & Chisholm, S. W. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc. Natl. Acad. Sci.* **107,** 18634–18639 (2010).
10. Rocap, G. *et al.* Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424,** 1042–1047 (2003).
11. Martiny, A. C., Huang, Y. & Li, W. Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ. Microbiol.* **11,** 1340–1347 (2009).
12. Malmstrom, R. R. *et al.* Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. *ISME J.* **7,** 184–198 (2013).
13. Martinez, A., Tyson, G. W. & Delong, E. F. Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ. Microbiol.* **12,** 222–238 (2010).
14. Rusch, D. B., Martiny, A. C., Dupont, C. L., Halpern, A. L. & Venter, J. C. Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc. Natl Acad. Sci.* **107,** 16184–16189 (2010).
15. Martiny, A. C., Coleman, M. L. & Chisholm, S. W. Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc. Nat. Acad. Sci.* **103,** 12552–12557 (2006).
16. Coleman, M. L. *et al.* Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311,** 1768–1770 (2006).
17. Dufresne, A. *et al.* Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl Acad. Sci.* **100,** 10020–10025 (2003).
18. Zhaxybayeva, O., Doolittle, W. F., Papke, R. T. & Gogarten, J. P. Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. *Genome Biol. Evol.* **1,** 325–339 (2009).
19. Baumdicker, F., Hess, W. R. & Pfaffelhuber, P. The infinitely many genes model for the distributed genome of bacteria. *Genome Biol. Evol.* **4,** 443–456 (2012).
20. Dufresne, A., Garczarek, L. & Partensky, F. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6,** R14 (2005).
21. Sun, Z. & Blanchard, J. L. Strong genome-wide selection early in the evolution of *Prochlorococcus* resulted in a reduced genome through the loss of a large number of small effect genes. *PLoS ONE* **9,** e88837 (2014).
22. Kashtan, N. *et al.* Single-cell genomics reveals hundreds of coexisting subpopulations in wild. *Prochlorococcus*. *Science* **344,** 416–420 (2014).
23. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304,** 66–74 (2004).
24. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5,** e77–e431 (2007).
25. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci.* **105,** 3805–3810 (2008).
26. Ghai, R. R. *et al.* Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* **4,** 1154–1166 (2010).

27. Shi, Y., Tyson, G. W., Eppley, J. M. & Delong, E. F. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J.* **5,** 999–1013 (2011).
28. Poretsky, R. S. *et al.* Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ. Microbiol.* **11,** 1358–1375 (2009).
29. Stingl, U., Tripp, H. J. & Giovannoni, S. J. Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. *ISME J.* **1,** 361–371 (2007).
30. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75 (2008).
31. Falda, M. *et al.* Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. *BMC Bioinformatics* **13,** S14 (2012).
32. Shimada, A., Nishijima, M. & Maruyama, T. Seasonal appearance of *Prochlorococcus* in Suruga Bay, Japan in 1992–1993. *J. Oceanogr.* **51,** 289–300 (1995).
33. Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Rapid diversification of marine Picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). *J. Mol. Evol.* **46,** 188–201 (1998).
34. Parpais, J., Marie, D., Partensky, F., Morin, P. & Vaulot, D. Effect of phosphorus starvation on the cell cycle of the photosynthetic prokaryote *Prochlorococcus* spp. *Mar. Ecol. Prog. Ser.* **132,** 265–274 (1996).
35. Penno, S., Campbell, L. & Hess, W. R. Presence of phycoerythrin in two strains of *Prochlorococcus* (Cyanobacteria) isolated from the subtropical North Pacific Ocean. *J. Phycol.* **36,** 723–729 (2000).
36. Chisholm, S. W. *et al. Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll *a* and *b*. *Arch. Microbiol.* **157,** 297–300 (1992).
37. Moore, L. *et al.* Culturing the marine cyanobacterium *Prochlorococcus*. *Limnol. Oceanogr. Methods* **5,** 353–362 (2007).
38. Guillard, R. R. & Ryther, J. H. Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (cleve) Gran. *Can. J. Microbiol.* **8,** 229–239 (1962).
39. Guillard, R. R. in *Culture of Marine Invertebrate Animals* 26–60 (Plenum Press, 1975).
40. Rodrigue, S. *et al.* Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE* **4,** e6864 (2009).
41. Rodrigue, S. *et al.* Unlocking short read sequencing for metagenomics. *PLoS ONE* **5,** e11840 (2010).
42. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
43. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5,** e11147 (2010).
44. Gordon, D. & Green, P. Consed: a graphical editor for next-generation sequencing. *Bioinformatics* **29,** 2936–2937 (2013).
45. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30,** 693–700 (2012).
46. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32,** 1363–1371 (2004).
47. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systemat. Biol.* **59,** 307–321 (2010).
48. Kelly, L., Huang, K. H., Ding, H. & Chisholm, S. W. ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res.* **40,** D632–D640 (2012).
49. Thompson, A. W., Huang, K., Saito, M. A. & Chisholm, S. W. Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. *ISME J.* **5,** 1580–1594 (2011).
50. Morris, J. J., Johnson, Z. I., Szul, M. J., Keller, M. & Zinser, E. R. Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. *PLoS ONE* **6,** e16805 (2011).
51. Dufresne, A. *et al.* Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol.* **9,** R90 (2008).
52. Palenik, B. *et al.* The genome of a motile marine *Synechococcus*. *Nature* **424,** 1037–1042 (2003).
53. Palenik, B. *et al.* Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proc. Natl Acad. Sci.* **103,** 13555–13559 (2006).
54. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
55. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).
56. Kelly, L., Ding, H., Huang, K. H., Osburne, M. S. & Chisholm, S. W. Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. *ISME J.* **7,** 1827–1841 (2013).
57. Ahlgren, N. A., Rocap, G. & Chisholm, S. W. Measurement of *Prochlorococcus* ecotypes using real-time polymerase chain reaction reveals different abundances of genotypes with similar light physiologies. *Environ. Microbiol.* **8,** 441–454 (2006).

Data Citations

1. Biller, S. J. *et al. Prochlorococcus Portal* http://proportal.mit.edu/ (2014).
2. Biller, S. J. *et al. Dryad* http://dx.doi.org/10.5061/dryad.k282c (2014).
3. Biller, S. J. *et al. Genbank* JNAG00000000 (2014).
4. Biller, S. J. *et al. Genbank* JNAH00000000 (2014).
5. Biller, S. J. *et al. Genbank* CP007753 (2014).
6. Biller, S. J. *et al. Genbank* JNAI00000000 (2014).
7. Biller, S. J. *et al. Genbank* JNAJ00000000 (2014).
8. Biller, S. J. *et al. Genbank* JNAK00000000 (2014).
9. Biller, S. J. *et al. Genbank* JNAL00000000 (2014).
10. Biller, S. J. *et al. Genbank* JNAM00000000 (2014).
11. Biller, S. J. *et al. Genbank* JNAN00000000 (2014).
12. Biller, S. J. *et al. Genbank* JNAO00000000 (2014).
13. Biller, S. J. *et al. Genbank* JNAP00000000 (2014).
14. Biller, S. J. *et al. Genbank* JNAQ00000000 (2014).
15. Biller, S. J. *et al. Genbank* JNAR00000000 (2014).
16. Biller, S. J. *et al. Genbank* JNAS00000000 (2014).
17. Biller, S. J. *et al. Genbank* CP007754 (2014).
18. Biller, S. J. *et al. Genbank* JNAX00000000 (2014).
19. Biller, S. J. *et al. Genbank* JNAT00000000 (2014).
20. Biller, S. J. *et al. Genbank* JNAU00000000 (2014).
21. Biller, S. J. *et al. Genbank* JNAV00000000 (2014).
22. Biller, S. J. *et al. Genbank* JNAW00000000 (2014).
23. Biller, S. J. *et al. Genbank* JNAY00000000 (2014).
24. Biller, S. J. *et al. Genbank* JNAZ00000000 (2014).
25. Biller, S. J. *et al. Genbank* JNBD00000000 (2014).

26. Biller, S. J. *et al. Genbank* JNBE00000000 (2014).
27. Biller, S. J. *et al. Genbank* JNBA00000000 (2014).
28. Biller, S. J. *et al. Genbank* JNBB00000000 (2014).
29. Biller, S. J. *et al. Genbank* JNBC00000000 (2014).

Acknowledgements

The authors are grateful to Allison Coe for careful maintenance of the MIT *Prochlorococcus* culture collection. We thank Luke Thompson, as well as the HOT and BATS teams, for assistance with field sampling. This work was supported in part by the Gordon and Betty Moore Foundation through Grant GBMF #495.01 and the National Science Foundation through grants OCE-1153588, OCE-0425602 and DBI-0424599, the NSF Center for Microbial Oceanography: Research and Education (C-MORE) to S.W.C. L.R.M. was supported by a NSF-ROA Supplement to NSF grant OCE-0806455 (to S.J.G.); L.R.M. and K.H.R.-J. were also supported by NSF OCE-0851288. G.R. was supported by NSF grant OCE-0723866.

Author Contributions

S.J.B. sequenced, assembled and analysed genomes, and prepared the manuscript. P.M.B. sequenced and assembled the PAC1 and EQPAC1 genomes. J.W.B.-T. closed the MIT0801 genome. L.K. contributed to gene cluster generation, validation and annotation. S.E.R. contributed to closing gaps in genomes. L.A. contributed to closing gaps in genomes. K.H.R.-J. isolated strains. H.D. implemented the gene clustering pipeline. S.J.G. isolated strains. G.R. isolated and characterized strains. L.R.M. isolated and characterized strains. S.W.C. supervised the work and helped prepare the manuscript.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Biller, S. J. *et al.* Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Sci. Data* 1:140034 doi: 10.1038/sdata.2014.34 (2014).

The ISME Journal (2014), 1–13

www.nature.com/ismej

ORIGINAL ARTICLE

Physiology and evolution of nitrate acquisition in *Prochlorococcus*

Paul M Berube[1], Steven J Biller[1], Alyssa G Kent[2], Jessie W Berta-Thompson[1,3], Sara E Roggensack[1], Kathryn H Roache-Johnson[4,5], Marcia Ackerman[5], Lisa R Moore[5], Joshua D Meisel[6], Daniel Sher[7], Luke R Thompson[8], Lisa Campbell[9], Adam C Martiny[2,10] and Sallie W Chisholm[1,6]

[1]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA;* [2]*Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, USA;* [3]*Microbiology Graduate Program, Massachusetts Institute of Technology, Cambridge, MA, USA;* [4]*Department of Molecular and Biomedical Sciences, University of Maine, Orono, ME, USA;* [5]*Department of Biological Sciences, University of Southern Maine, Portland, ME, USA;* [6]*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA;* [7]*Department of Marine Biology, University of Haifa, Haifa, Israel;* [8]*Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO, USA;* [9]*Department of Oceanography, Texas A&M University, College Station, TX, USA and* [10]*Department of Earth System Science, University of California, Irvine, Irvine, CA, USA*

Prochlorococcus **is the numerically dominant phototroph in the oligotrophic subtropical ocean and carries out a significant fraction of marine primary productivity. Although field studies have provided evidence for nitrate uptake by** ***Prochlorococcus*****, little is known about this trait because axenic cultures capable of growth on nitrate have not been available. Additionally, all previously sequenced genomes lacked the genes necessary for nitrate assimilation. Here we introduce three** ***Prochlorococcus*** **strains capable of growth on nitrate and analyze their physiology and genome architecture. We show that the growth of high-light (HL) adapted strains on nitrate is ~17% slower than their growth on ammonium. By analyzing 41** ***Prochlorococcus*** **genomes, we find that genes for nitrate assimilation have been gained multiple times during the evolution of this group, and can be found in at least three lineages. In low-light adapted strains, nitrate assimilation genes are located in the same genomic context as in marine** ***Synechococcus*****. These genes are located elsewhere in HL adapted strains and may often exist as a stable genetic acquisition as suggested by the striking degree of similarity in the order, phylogeny and location of these genes in one HL adapted strain and a consensus assembly of environmental** ***Prochlorococcus*** **metagenome sequences. In another HL adapted strain, nitrate utilization genes may have been independently acquired as indicated by adjacent phage mobility elements; these genes are also duplicated with each copy detected in separate genomic islands. These results provide direct evidence for nitrate utilization by** ***Prochlorococcus*** **and illuminate the complex evolutionary history of this trait.**
The ISME Journal advance online publication, 28 October 2014; doi:10.1038/ismej.2014.211

Introduction

The unicellular cyanobacterium *Prochlorococcus* is the smallest known free-living oxygenic phototroph (Chisholm *et al.*, 1992; Partensky *et al.*, 1999; Coleman and Chisholm, 2007; Partensky and Garczarek, 2010). It is numerically dominant in the tropical and subtropical regions of the world's oceans and responsible for 5–10% of marine primary productivity (Campbell *et al.*, 1994; Partensky *et al.*, 1999; Buitenhuis *et al.*, 2012; Flombaum *et al.*, 2013). *Prochlorococcus* has undergone a process of genome reduction following divergence from its closest relatives, the marine *Synechococcus* (Rocap *et al.*, 2002; Kettler *et al.*, 2007). These streamlined genomes are often considered an adaptation to the oligotrophic environments they occupy (Dufresne *et al.*, 2003; Rocap *et al.*, 2003). Even though individual genomes are small, the collective of all *Prochlorococcus* cells possesses a vast reservoir of genetic and physiological diversity (Kettler *et al.*, 2007). *Prochlorococcus* is composed of a polyphyletic group of low-light (LL) adapted clades (LLI–LLVI and NC1), and a more recently diverged monophyletic group of high-light

Correspondence: PM Berube or SW Chisholm, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building 48-424, Cambridge, MA 02139, USA.
E-mail: pmberube@gmail.com or chisholm@mit.edu
Received 10 June 2014; revised 8 September 2014; accepted 23 September 2014

(HL) adapted clades (HLI–HLVI) (Moore *et al.*, 1998; Moore and Chisholm, 1999; Rocap *et al.*, 2002; Martiny *et al.*, 2009c; Lavin *et al.*, 2010; Shi *et al.*, 2011; Huang *et al.*, 2012; Malmstrom *et al.*, 2013). Some of these clades are known to be differentially distributed along gradients of light intensity, temperature and nutrient concentrations (Bouman *et al.*, 2006; Johnson *et al.*, 2006; Zinser *et al.*, 2006; Zwirglmaier *et al.*, 2007, 2008; Malmstrom *et al.*, 2010, 2013).

Nitrogen availability often limits primary productivity in marine systems (Tyrrell, 1999), and organisms have evolved diverse mechanisms for uptake of various chemical forms of nitrogen. Nitrate is one of the more abundant sources of inorganic nitrogen available to phytoplankton (Gruber, 2008), and the majority of cyanobacteria possess pathways for the uptake and assimilation of nitrate (Herrero *et al.*, 2001; García-Fernández *et al.*, 2004; Ohashi *et al.*, 2011). Early reports on the vertical distributions of *Prochlorococcus* noted a subsurface maximum in abundance at the base of the euphotic zone, suggesting that *Prochlorococcus* was sensitive to nitrogen depletion and might be assimilating nitrate supplied from deep waters (Olson *et al.*, 1990; Vaulot and Partensky, 1992). Therefore, it was surprising that nearly all isolates of *Prochlorococcus* could not use nitrate and lacked the genes required for this function (Moore *et al.*, 2002; Coleman and Chisholm, 2007; Kettler *et al.*, 2007), even though most isolates of *Synechococcus* are capable of using nitrate (Fuller *et al.*, 2003; Ahlgren and Rocap, 2006). Only a single *Prochlorococcus* culture, PAC1 isolated in 1992, was reported to utilize nitrate (Williams *et al.*, 1999), but because of the presence of other bacteria in that culture, direct nitrate uptake by *Prochlorococcus* could not be conclusively demonstrated.

Several pieces of evidence indicated that nitrate assimilation was a more common trait within *Prochlorococcus* populations than previously thought. Field experiments demonstrated the uptake of isotopically labeled nitrate by *Prochlorococcus* cells in the Sargasso Sea (Casey *et al.*, 2007), and nitrate assimilation genes were found to be associated with uncultivated *Prochlorococcus* genomes from many regions of the subtropical oceans (Martiny *et al.*, 2009b). A scaffold assembled from metagenomic data from the Global Ocean Sampling (GOS) expedition indicated that all the genes required for nitrate assimilation were colocalized in a specific region of the genomes of HL adapted *Prochlorococcus*. The metagenomic data primarily identified nitrate utilization genes in the HLII clade of *Prochlorococcus* as sequences from this clade comprised the majority of *Prochlorococcus*-like sequences in the GOS data set (Rusch *et al.*, 2007).

These past observations raised two important questions about nitrate assimilation in *Prochlorococcus*. (1) Can axenic strains grow on nitrate as the sole nitrogen source? (2) What is the evolutionary history of nitrate assimilation genes in this group? To address these questions, we isolated and sequenced *Prochlorococcus* strains capable of nitrate assimilation and examined their growth on different nitrogen sources. We then used comparative genomics to better understand how this trait had evolved in *Prochlorococcus*.

Materials and methods

Strains and enrichments

Five strains of *Prochlorococcus* (SB, MIT0604, PAC1, MIT9301 and MED4), one strain of *Synechococcus* (WH8102) and two *Prochlorococcus* enrichment cultures (P0902-H212 and P0903-H212) were used in this study. MIT9301, MED4 and WH8102 have previously been rendered axenic (free of heterotrophic contaminants). All axenic cultures were routinely assessed for purity by confirming a lack of turbidity after inoculation into a panel of purity test broths: ProAC (Morris *et al.*, 2008), MPTB (Saito *et al.*, 2002) and ProMM (Pro99 medium (Moore *et al.*, 2007) supplemented with 1 × Va vitamin mix (Waterbury and Willey, 1988) and 0.05% w/v each of pyruvate, acetate, lactate and glycerol). ProMM is a modified version of the PLAG medium (Morris *et al.*, 2008), but uses 100% sea water as the base.

PAC1 was enriched from sea water collected from the deep chlorophyll maximum in the North Pacific Ocean at Station ALOHA (22.75°N, 158°W) on Hawai'i Ocean Time-series (HOT) cruise 36. Sea water was passed through a 0.6 μm Nucleopore filter twice, and the filtrate was serially diluted into K/10 medium (Chisholm *et al.*, 1992), but with the following modifications for final nutrient concentrations: 5 μM urea, 5 μM ammonium and 1 μM β-glycerophosphate replacing inorganic phosphate, 0.01 μM Na_2MoO_4 and 0.05 μM $NiCl_2$. MIT0604 was derived from an enrichment culture initiated with Pro2 nutrient additions (Moore *et al.*, 2007) to sea water obtained at Station ALOHA on HOT cruise 181, but with all nitrogen sources replaced by 0.217 mM sodium nitrate. The P0902-H212 and P0903-H212 enrichments were initiated with Pro2 nutrient additions (Moore *et al.*, 2007) to sea water obtained from Station ALOHA on HOT cruise 212, but with all nitrogen sources replaced by 0.05 mM sodium nitrate.

Purification of Prochlorococcus strains

SB and MIT0604 were rendered axenic in this study using a modified dilution to extinction method. *Prochlorococcus* from exponential phase cultures were enumerated using an Influx Cell Sorter (BD Biosciences, San Jose CA, USA) or a FACSCalibur flow cytometer (BD Biosciences) as previously described (Olson *et al.*, 1985; Cavender-Bares *et al.*, 1999). Cultures consisting of >80% *Prochlorococcus* cells were serially diluted into multiple multiwell

plates at final concentrations of 1–10 cells per well in at least 200 µl of ProMM medium. Axenic *Prochlorococcus* do not grow from such low cell densities in Pro99 medium without 'helper' heterotrophic bacteria (Morris *et al.*, 2008, 2011); however, they do grow when diluted into ProMM. The main ingredient in ProMM that promotes the growth of cells from low densities is pyruvate, and we suspect that in this context pyruvate serves as a potent hydrogen peroxide scavenger (Giandomenico *et al.*, 1997). Wells contaminated with heterotrophic bacteria were identified by the appearance of turbidity. The multiwell plates were monitored by eye and by fluorometry using a Synergy HT Microplate Reader (BioTek, Winooski, VT, USA), and nonturbid wells were monitored by flow cytometry using a FACSCalibur flow cytometer. Wells that appeared green or had *Prochlorococcus* cells as determined by flow cytometry were immediately transferred to Pro99 medium directly, or into fresh ProMM medium until consistent growth was observed, at which point the cultures were introduced back into Pro99 medium. Cultures were examined for heterotrophic bacteria contaminants by flow cytometry and by inoculation into the panel of purity test broths as described above.

PCR screen for the nitrate reductase gene

Based on an alignment of GOS reads coding for the *Prochlorococcus narB* sequence (Martiny *et al.*, 2009b), degenerate primers 30narB175f (5′-TGYGTD AAAGGMGCAACAGTNTG-3′) and 30narB574r (5′-GACAYTCWGCBGTATTWGTHCC-3′) were designed to specifically amplify the *narB* gene from HLII clade *Prochlorococcus*, and degenerate primers 40narB1447f (5′-TATTGYCCAGCWTTYMGDCCDTG-3′) and 40nar B1766r (5′-AKAGGWTGYTTWGTRTARAAYTG-3′) were designed to specifically amplify the *narB* gene from LLI clade *Prochlorococcus*. PCR used annealing temperatures of 52.5 °C for the HLII *narB* sequence and 56 °C for the LLI *narB* sequence. Reactions contained 1 × PCR buffer, 2.5 mM $MgCl_2$, 0.2 mM each of dATP, dTTP, dCTP and dGTP, 0.2 µM of each primer, 1 unit of Platinum Taq DNA polymerase (Life Technologies, Grand Island, NY, USA) and 1 ng of genomic DNA prepared from *Prochlorococcus* cultures in the MIT Cyanobacteria Culture Collection (Chisholm Laboratory, MIT, Cambridge, MA, USA). DNA from *Synechococcus* WH8102, which contains a *narB* gene, was used as a negative control. Reactions were cycled 30 times at 94 °C for 15 s, the primer-specific annealing temperature for 15 s and 72 °C for 60 s. PCR products with the expected size were sequenced at the Dana-Farber/Harvard Cancer Center DNA Resource Core (Boston, MA, USA) to confirm amplification of the *narB* gene.

Growth in the presence of alternative nitrogen sources

Axenic *Prochlorococcus* strains SB, MIT0604, MIT9301 and MED4, and axenic *Synechococcus* strain WH8102 were acclimated to Pro99 medium (Moore *et al.*, 2007) prepared with sea water from the South Pacific Subtropical Gyre and grown at 24 °C and 50 µmol photons $m^{-2}s^{-1}$ continuous illumination for at least 10 generations or until growth rates were similar between successive transfers. Bulk culture fluorescence was measured as a proxy for biomass using a 10-AU fluorometer (Turner Designs, Sunnyvale, CA, USA). Triplicate cultures of each strain were initiated in Pro99 that contained 0.8 mM ammonium chloride. Once cultures had reached mid-exponential phase, they were transferred into Pro99 medium containing 0.8 mM ammonium chloride, 0.8 mM sodium nitrate, 0.8 mM sodium cyanate or no nitrogen additions as a control to monitor utilization of carry-over ammonium. Cultures were successively transferred at mid-exponential phase until growth in the cultures lacking nitrogen additions had arrested because of nitrogen limitation. Specific growth rates were estimated from the log-linear portion of the growth curve for the final transfer. Two tailed homoscedastic *t*-tests were conducted in Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) in order to evaluate the likelihood of significantly different growth rates in each strain for each pair of nitrogen sources and for strains grown on the same nitrogen source.

Genome data

A total of 41 *Prochlorococcus* and 15 *Synechococcus* genomes (Biller *et al.*, 2014) that include the genomes of the nitrate assimilating strains SB, MIT0604 and PAC1 were used in this study. Sequence data were also obtained for the P0902-H212 and P0903-H212 enrichment cultures as described in the Supplementary Methods. These enrichment assemblies had total sequence lengths approximately twice the size of previously sequenced *Prochlorococcus* genomes, suggesting the presence of at least two unique strains dominating each enrichment. Binning contigs based on average sequencing coverage yielded a subset of highly covered contigs in each assembly with a total sequence length similar to that of previously sequenced *Prochlorococcus* genomes. In the highly covered subsets for each assembly, the complete set of nitrate assimilation genes were only found on a single contig. For the purpose of this study, only these contigs were relevant and entered into our analysis.

All sequence data were annotated using the RAST server (Aziz *et al.*, 2008) with FIGfam release 49 in order to facilitate comparison between genomes by ensuring a uniform methodology for gene calling and functional annotation. Clusters of orthologous groups of proteins (COGs) were identified as previously described (Kelly *et al.*, 2012). These clusters are included in the 'V4' CyCOGs on the ProPortal website (http://proportal.mit.edu) (Kelly *et al.*, 2012; Biller *et al.*, 2014).

Genome phylogeny
We translated 537 single-copy core genes to amino acid sequences, aligned each gene individually in protein space using ClustalW (Larkin *et al.*, 2007), and then back-translated the sequences using TranslatorX (Abascal *et al.*, 2010). Using the principle previously described (Kettler *et al.*, 2007), we randomly concatenated 100 of these aligned genes and built maximum likelihood and neighbor-joining phylogenies using PHYLIP v3.69 (Felsenstein, 2005). We repeated the random concatenation and tree generation 100 times.

Estimation of gene gain and loss
Using a maximum parsimony approach (Mirkin *et al.*, 2003), the patterns of gene gain and loss were mapped onto the topology of the maximum likelihood nucleotide tree using WH5701 as an outgroup. Utilizing 13 590 non-core single-copy COGs, we reconstructed ancestral character states of gene absence and presence on our guide tree and minimized the cost of gains and losses given a gene gain equal to twice a gene loss. We used the program DendroPy to implement the tree traversal portion of the algorithm (Sukumaran and Holder, 2010).

Phylogenies of genes involved in the transport and reduction of nitrate and nitrite
COGs corresponding to the *nirA*, *narB*, *focA* and *napA* genes were aligned in protein space using ClustalW. Phylogenetic trees were estimated with PHYLIP v3.69 (Felsenstein, 2005) using the programs SEQBOOT, PROTDIST with the Jones–Taylor–Thornton matrix and a constant rate of variability among sites and NEIGHBOR on the aligned amino acid sequences with *Synechococcus* WH5701 used as an outgroup for *nirA* and *narB* and *Synechococcus* CB0101 used as an outgroup for *focA* and *napA*. We included GOS consensus sequences: GOS *nirA*, GOS *narB* and GOS *napA* (Martiny *et al.*, 2009b).

Results and discussion

Isolates of Prochlorococcus are capable of nitrate assimilation
To identify possible cultures capable of nitrate assimilation, we screened existing *Prochlorococcus* cultures for the assimilatory nitrate reductase gene, *narB*, using PCR. We found that the LL adapted PAC1 strain (Penno *et al.*, 2000) and the HL adapted SB strain (Shimada *et al.*, 1995) each contained the gene. In search of additional strains capable of utilizing nitrate, we performed selective enrichments from sea water obtained from the subtropical North Pacific Ocean using nitrate as the sole added nitrogen source. This yielded one HL adapted strain (*Prochlorococcus* MIT0604) and two mixed *Prochlorococcus* cultures (P0902-H212 and P0903-H212) with the *narB* gene (Table 1).

We then rendered SB and MIT0604 axenic and examined their growth in the presence of nitrate or ammonium. As hypothesized, both SB and MIT0604 can grow on nitrate as the sole source of nitrogen, but with a significant reduction in growth rate (18% and 17%, respectively), compared with growth on ammonium (Figure 1 and Supplementary Figure S1). Although the slower growth on nitrate could be explained by the greater amount of reducing power required to assimilate more oxidized N sources (García-Fernández *et al.*, 2004), we assume that these cultures were growing at saturating light intensities based on previous measurements of light saturating irradiances for the growth of *Prochlorococcus* (Moore and Chisholm, 1999); thus energy supply and reducing power were likely not limiting. Furthermore, recent work has shown that the growth rates and chemical composition of some marine cyanobacteria are not directly related to the oxidation state of the cells' N source (Collier *et al.*, 2012). Under light-limiting conditions, for example, the growth rate and chemical composition of *Synechococcus* grown on ammonium was the same as that on nitrate; however, under light-saturating conditions, cells grown on nitrate had a higher carbon-to-nitrogen ratio (Collier *et al.*, 2012). This perhaps suggests a bottleneck in the uptake and conversion of nitrate compared with ammonium when energy is sufficient (Collier *et al.*, 2012), and may explain the

Table 1 *Prochlorococcus* strains and enrichments capable of growth in the presence of nitrate as the sole nitrogen source

Name	Clade	Axenic	Isolation depth (m)	Isolation coordinates	Region	Isolation date	Assembly size (bp)	Contigs	% GC	Genbank accession	Reference
Unialgal cultures (complete genome sequences)											
SB	HL II	Yes	40	35°N, 138.3°E	Suruga Bay, Japan	October 1992	1 668 514	3	31.5	JNAS00000000	Shimada *et al.* (1995); Biller *et al.* (2014)
MIT0604	HL II	Yes	175	22.75°N, 158°W	North Pacific	May 2006	1 780 061	1	31.2	CP007753	This study
PAC1	LL I	No	100	22.75°N, 158°W	North Pacific	April 1992	1 825 493	15	35.1	JNAX00000000	Penno *et al.* (2000); Biller *et al.* (2014)
Mixed enrichments (partial genome assemblies)											
P0902-H212	LL I	No	175	22.75°N, 158°W	North Pacific	July 2009	501 825	1	35.4	KJ947870	This study
P0903-H212	LL I	No	200	22.75°N, 158°W	North Pacific	July 2009	291 739	1	35.2	KJ947871	This study

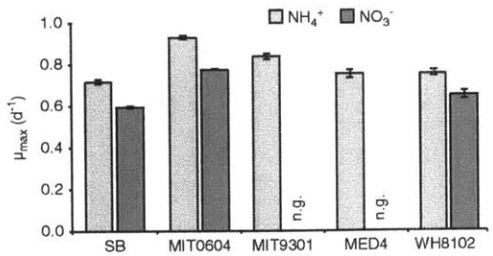


Figure 1 Maximum specific growth rates (μ max) of *Prochlorococcus* strains SB, MIT0604, MIT9301, MED4 and *Synechococcus* WH8102 in the presence of ammonium or nitrate. Values represent the mean and s.d. of three biological replicates. Growth rate differences for each strain grown on ammonium compared with nitrate as well as growth rate differences between strains on the same nitrogen source were significant ($P<0.05$) in a two-tailed homoscedastic *t*-test; n.g., no growth.

slower growth of *Prochlorococcus* on nitrate compared with ammonium.

In the early days of research on *Prochlorococcus*, the absence of cultures known to utilize nitrate resulted in a distorted view of the role of *Prochlorococcus* in marine ecosystems; ecosystem models and ecophysiological interpretations were guided by the assumption that most, if not all, *Prochlorococcus* were incapable of nitrate assimilation (García-Fernández *et al.*, 2004; Fuller *et al.*, 2005; Follows *et al.*, 2007). Why have nitrate-utilizing *Prochlorococcus* appeared so infrequently in culture collections in the past? Is it because we were selecting against them in isolations using media containing ammonium but not nitrate (Moore *et al.*, 2007)? We think not because SB and MIT0604—both *narB*-containing strains—grow at equal or better rates on ammonium compared with other HL adapted *Prochlorococcus* strains (Figure 1 and Supplementary Figure S1). An alternative explanation is that most of the early cultures of *Prochlorococcus* were isolated from environments that are relatively nitrogen replete—that is, thought to be more limited by phosphorus or iron availability (for example, the Sargasso Sea, Mediterranean Sea and the Equatorial Pacific) (Vaulot *et al.*, 1996; Mann and Chisholm, 2000; Wu *et al.*, 2000; Marty *et al.*, 2002; Kettler *et al.*, 2007; Rusch *et al.*, 2010). We now know that *Prochlorococcus* cells capable of nitrate assimilation are more likely to be found in ocean regions with lower average nitrate concentrations, such as the Caribbean Sea and Indian Ocean (Martiny *et al.*, 2009b). Indeed, PAC1 and SB (both *narB*-containing strains that were isolated on medium containing ammonium but lacking nitrate) were isolated from N-poor regions (Shimada *et al.*, 1995; Penno *et al.*, 2000; Wu *et al.*, 2000; Iwata *et al.*, 2005). Thus, we believe that the probability of obtaining a *narB*-containing strain using medium containing ammonium is in large part a function of the particular water sample used to start enrichment cultures.

Nitrate assimilation is found in diverse lineages of Prochlorococcus

What can the features of the nitrate assimilation genes in *Prochlorococcus* tell us about how they have been gained or lost during the evolution of this group? The genomes of PAC1, SB and MIT0604, along with contigs containing nitrate assimilation genes from the P0902-H212 and P0903-H212 enrichment cultures, were informative in this regard. These *Prochlorococcus* belong to both the LL adapted LLI clade (PAC1, P0902-H212 and P0903-H212) and the HL adapted HLII clade (SB and MIT0604) (Figure 2 and Supplementary Figures S2 and S3), demonstrating that nitrate utilization is found in multiple and diverse lineages of *Prochlorococcus* and suggesting a complex evolutionary history. The presence of nitrite and nitrate metabolism in *Prochlorococcus* follows that of *Synechococcus* in that some strains are able to reduce nitrite and some are able to reduce both nitrite and nitrate. Because these traits are not monophyletic, a model of gene gain and loss events provides evidence for three gains and two losses for the *narB* nitrate reductase gene and two gains and three losses for the *nirA* nitrite reductase gene (Figure 2). With the limited number of genomes available, it appears that there is evidence for multiple gains and losses of nitrogen assimilation traits through the evolution of *Prochlorococcus* and *Synechococcus*, with *narB* found in at least three distinct *Prochlorococcus* lineages.

The genomic context of the nitrate assimilation gene cluster suggests a complex evolutionary history

To look for features that might help us interpret the gains and losses of nitrate and nitrite assimilation genes in *Prochlorococcus*, we examined the local genomic context of these genes. Although the full complement of nitrate assimilation genes was predicted to be localized in a single region of the highly syntenic HLII clade genomes from metagenomic assemblies (Martiny *et al.*, 2009b), it was unclear whether this context would be found in any individual cell. Furthermore, given that these genes were found in a different region in *Prochlorococcus* compared with marine *Synechococcus*, we were curious as to whether we might find evidence for rearrangements or lateral gene transfer.

The nitrate assimilation genes in PAC1 and the P0902-H212 and P0903-H212 contigs are syntenic and also found in the same genomic region as the nitrite assimilation genes in NATL1A and the nitrate assimilation genes in *Synechococcus* WH8102 (Figure 3). This region is bounded by a pyrimidine biosynthesis gene (*pyrG*) and a polyphosphate kinase gene (*ppk*) between which many nitrogen assimilation genes are located in marine *Synechococcus*. Although gene gains and losses have been observed in this region (Scanlan *et al.*, 2009), our data indicate that the genomic location of the nitrate

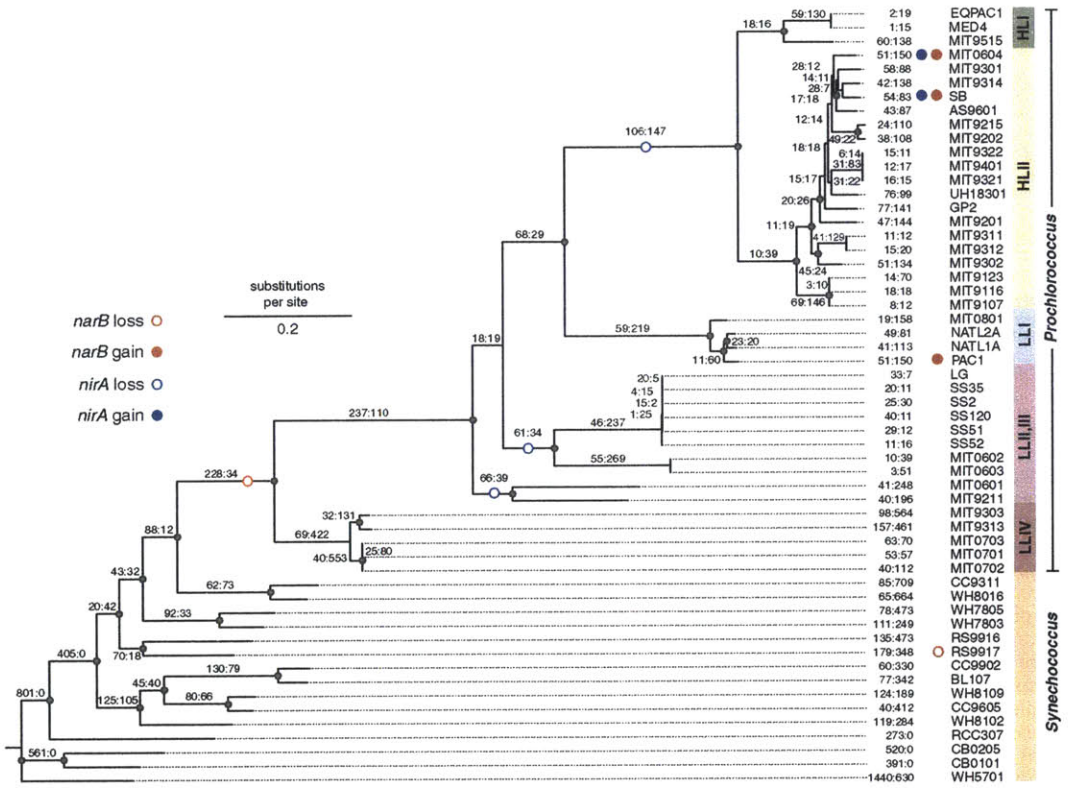


Figure 2 Maximum likelihood phylogeny of *Prochlorococcus* and *Synechococcus* based on the similarity of 100 randomly concatenated single-copy core genes. Nodes are marked by closed circles to indicate that the associated taxa clustered together in at least 75% of 100 replicate trees. Genes lost and gained in the evolution of *Prochlorococcus* and *Synechococcus* are indicated at each node by values representing losses followed by gains. Predicted losses (open circles) or gains (closed circles) of *nirA* (blue) or *narB* (orange) are labeled on their respective branches.

and nitrite assimilation genes is reasonably well fixed in LLI *Prochlorococcus* and closely related *Synechococcus*. Although our model of gene gain and loss events suggests the loss of nitrate assimilation genes early in the evolution of *Prochlorococcus* (Figure 2), the local genomic features of these genes are consistent with the interpretation that some lineages may have retained these genes following the divergence of *Prochlorococcus* from *Synechococcus*.

Analysis of metagenomic data from GOS (Martiny *et al.*, 2009b) suggested that the nitrate utilization genes in HLII *Prochlorococcus* should be located in a different genomic region compared with LLI genomes, indicating an alternative evolutionary origin. Based on a scaffold of mate-paired metagenomic reads, it was inferred that this cluster should be located ~500 kb downstream of the *pyrG-ppk* region containing the nitrate assimilation genes in WH8102 and the nitrite assimilation genes in NATL1A (Martiny *et al.*, 2009b). We found a high degree of similarity between the nitrate assimilation gene cluster in SB and the scaffold derived from GOS metagenome sequences obtained from multiple individual cells from multiple sampling stations. This similarity manifested itself not only in the gene order and chromosomal location, but also the phylogeny of the nitrate assimilation genes (Figures 3–5), placing the nitrate assimilation gene cluster in a genomic region that is syntenic with other HLII genomes and adjacent to a known genomic island (ISL3) in this clade (Figure 4). Furthermore, a partial genome from a *Prochlorococcus* single cell belonging to the HLII clade (B241-528J8; Genbank JFLE01000089.1) (Kashtan *et al.*, 2014) also possesses a nitrate assimilation gene cluster in the same location and in the same order. The striking similarity between the nitrate assimilation gene clusters of these individual *Prochlorococcus* and the GOS consensus indicates that the order and location of nitrate assimilation genes are stable within HLII genomes.

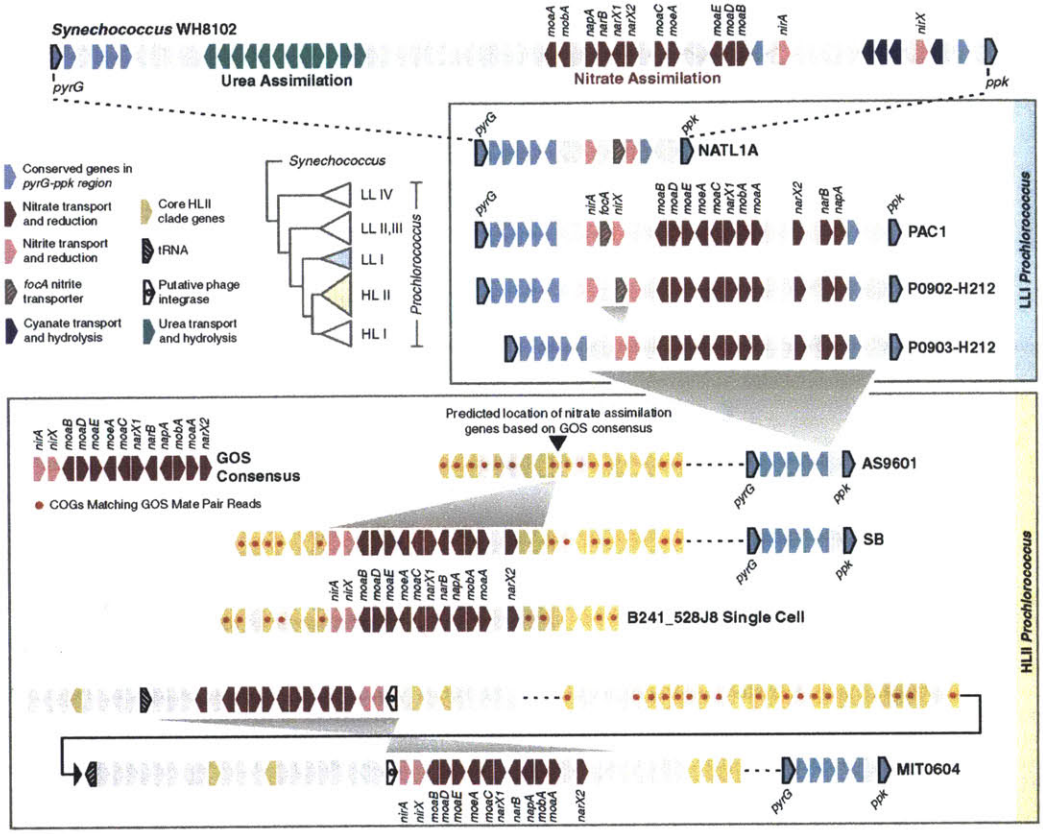


Figure 3 Architecture of the nitrite and nitrate assimilation genes in LL adapted (LLI clade) and HL adapted (HLII clade) *Prochlorococcus* relative to *Synechococcus* WH8102. Similar to *Synechococcus*, the nitrite and nitrate assimilation genes in the LLI clade of *Prochlorococcus* are found within the region between the *pyrG* (pyrimidine biosynthesis) and *ppk* (polyphosphate kinase) genes. Most LLI clade *Prochlorococcus*, with the exception of the P0903-H212 contig, possess a *focA* nitrite transporter in this region (possibly acquired from proteobacteria (Rocap *et al.*, 2003)). Metagenome data (Martiny *et al.*, 2009b), a partial genome from a single cell (B241-528J8) (Kashtan *et al.*, 2014) and a culture genome (*Prochlorococcus* SB) indicate that the nitrate assimilation genes within HLII clade *Prochlorococcus* are commonly found in a syntenic region adjacent to genomic island ISL3 (see Figure 4). *Prochlorococcus* MIT0604 is an exception in that it possesses duplicate nitrate assimilation gene clusters located within genomic islands ISL3 and ISL4 (see Figure 4), with phage integrase genes immediately adjacent to each copy of the *nirA* (nitrite reductase) gene.

The nitrate assimilation genes in strain MIT0604 had a different local genome structure compared with strain SB and the partial single-cell genome, B241-528J8. MIT0604 has duplicate clusters of these genes that are inversely oriented and located upstream and downstream of the GOS-predicted location (Figure 3 and 4). A Southern blot confirmed that MIT0604 does indeed contain two copies of *narB* whereas SB contains only one (Supplementary Figure S4), and they are located within genomic islands ISL3 and ISL4 of HLII clade *Prochlorococcus* (Figure 4). Genomic islands are common features of *Prochlorococcus* genomes, particularly within the HL adapted clades (Coleman *et al.*, 2006; Kettler *et al.*, 2007). They harbor much of the variability in gene content between members of the same clade and are hot spots for lateral gene transfer. Phage integrase genes are located proximal to both nitrate assimilation gene clusters in MIT0604, and a transfer RNA gene is adjacent to one of these clusters (Figure 3). The transfer RNA genes are known to serve as sites for insertion of phage DNA in bacteria (Williams, 2002), and thus the location of these phage integrase and transfer RNA genes suggests transduction as a possible mechanism by which MIT0604 has acquired the nitrate assimilation gene cluster. Notably, duplication of such a large region of the chromosome has not been observed previously in *Prochlorococcus*, and, thus far, MIT0604 is the only *Prochlorococcus* or *Synechococcus* strain possessing two complete copies of the genes required for nitrate assimilation.

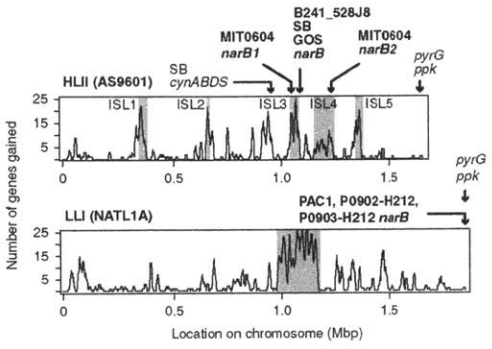


Figure 4 Locations of nitrate and cyanate assimilation genes in strains of *Prochlorococcus* capable of nitrate assimilation relative to the known genomic islands (shaded regions) observed in the HLII and LLI clades of *Prochlorococcus*; plots modified from Kettler *et al.* (2007). *Prochlorococcus* genomes are highly syntenic and genomic islands have been identified in HL adapted genomes (for example, AS9601) by conserved breaks in gene synteny among strains (Coleman *et al.*, 2006; Kettler *et al.*, 2007). Genomic islands have also been identified (for example, the large region within LLI clade genomes such as NATL1A) by predicted gene gain events along the chromosome (Kettler *et al.*, 2007).

The phylogenies of nitrate assimilation genes are similar to the phylogeny of genomes

Given the evidence for both a stable arrangement of the nitrate assimilation genes in some *Prochlorococcus* and possible gene transfer leading to acquisition of the nitrate assimilation trait in MIT0604, we were curious to know whether the phylogenies of these genes were congruent with whole genome phylogenies (Figure 2 and Supplementary Figure S2), as well as the phylogeny of GyrB (Supplementary Figure S3) that has been identified as a useful phylogenetic marker for *Prochlorococcus* (Mühling, 2012). Thus, we reconstructed the amino acid phylogenies of the NirA and NarB reductases, the FocA nitrite transporter and the NapA nitrite/nitrate transporter (Figure 5). The NirA phylogeny is largely consistent with our observations based on the GOS metagenome data (Martiny *et al.*, 2009b), such that the NirA proteins from genomes in the LLIV clade are more closely associated with marine *Synechococcus* than with other *Prochlorococcus* sequences. In all phylogenetic trees, the PAC1, P0902-H212 and P0903-H212 sequences are in a separate clade distinct from that of the SB and MIT0604 sequences, reinforcing the HL versus LL differentiation (Figure 5). The NirA and NarB sequences from SB are consistently more closely affiliated with the GOS consensus sequence (Martiny *et al.*, 2009b) than with the MIT0604 sequences. NapA sequences from SB and MIT0604 are also both closely related to the GOS NapA consensus sequence (Figure 5). Similar to the GyrB phylogeny (Supplementary Figure S3), the P0903-H212 sequences fall outside the clade containing the other LLI sequences. With the exception of the LLIV NirA sequences, the phylogenies of these nitrite and nitrate assimilation proteins (Figure 5) are congruent with whole genome and GyrB phylogenies (Figure 2 and Supplementary Figures S2 and S3) at a resolution defining the major *Prochlorococcus* clades.

Nitrate assimilating Prochlorococcus possess a diverse set of nitrogen acquisition pathways

Gene content in *Prochlorococcus* has been shown, for several traits, to reflect the selective pressures in the specific environments from which they (or their genes) were captured (Martiny *et al.*, 2006; Rusch *et al.*, 2007; Coleman and Chisholm, 2010; Feingersch *et al.*, 2012; Malmstrom *et al.*, 2013). Thus, we wondered whether other nitrogen assimilation traits might co-occur with nitrate assimilation in *Prochlorococcus*, and examined the potential for PAC1, SB and MIT0604 to access alternative sources of nitrogen based on their gene content (Supplementary Table S1 and Supplementary Figure S5).

Like other members of the LLI clade, PAC1 possesses genes for the assimilation of ammonium and urea, but lacks cyanate transporter genes. In addition to the *napA* nitrite/nitrate transporter, the *focA* nitrite transporter is found in both PAC1 and the contig from P0902-H212. However, the *focA* gene is absent from HL adapted strains SB and MIT0604, and most surface water metagenomic samples (Martiny *et al.*, 2009b). Some *Synechococcus* strains (for example, WH8102) (Supplementary Figure S5) also lack *focA*; thus, this gene is clearly subject to gain and loss. Although *focA* is also similar to formate transporters, evidence implicates its role in nitrite uptake in *Prochlorococcus*; for example, the gene is located near other nitrite assimilation genes (Figure 3), it is upregulated under nitrogen stress (Tolonen *et al.*, 2006) and it is absent from *Prochlorococcus* that cannot grow on nitrite (Moore *et al.*, 2002; Coleman and Chisholm, 2007; Kettler *et al.*, 2007) (Supplementary Figure S5). As PAC1 possesses both a nitrite transporter (*focA*) and the dual-function nitrate/nitrite transporter (*napA*), it is possible that *focA* provides some advantage to LL adapted cells that are often maximally abundant near the nitrite maxima in the oceans (Scanlan and West, 2002; Lomas and Lipschultz, 2006). LL adapted cells that possess the dual-function nitrite/nitrate transporter may benefit from having an additional transporter for nitrite. Given that HL adapted *Prochlorococcus* strains capable of nitrate utilization lack the *focA* gene, these cells may be less reliant on nitrite as a nitrogen source.

SB and MIT0604 possess urea assimilation genes and can utilize urea as a sole nitrogen source (Supplementary Figure S6). Furthermore, SB possesses cyanate transporter genes that are rare in both *Prochlorococcus* and *Synechococcus* strains (Kamennaya *et al.*, 2008), and it can indeed grow

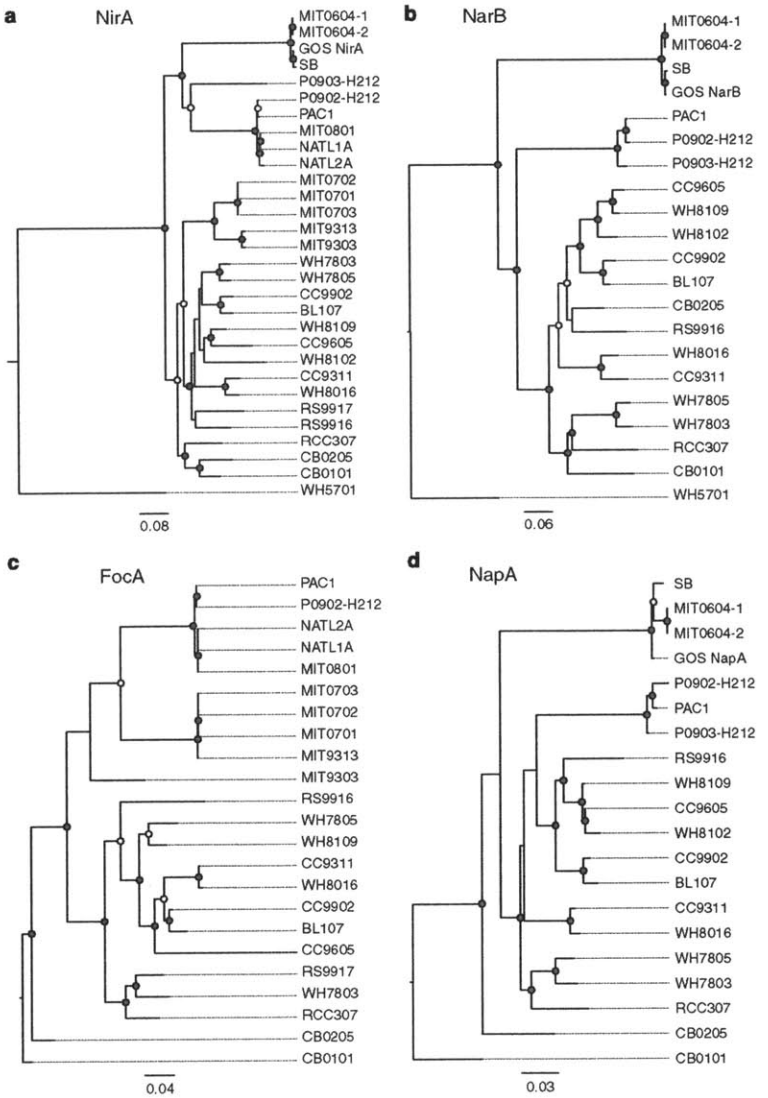


Figure 5 Neighbor-joining phylogeny of four proteins involved in the transport and reduction of nitrate and nitrite in marine cyanobacteria: (**a**) NirA; nitrite reductase, (**b**) NarB; nitrate reductase, (**c**) FocA; nitrite transporter and (**d**) NapA; nitrite/nitrate transporter. The percentage of 100 replicate trees in which the associated taxa clustered together is indicated at nodes by closed circles (>75%) or open circles (>50%). Scale bars represent substitutions per site.

utilizing cyanate (Supplementary Figure S1) as the sole source of nitrogen. Although very little is known about cyanate concentrations in marine systems, *cynA* genes (encoding the periplasmic component of the cyanate ABC-type transporter system) were relatively abundant in the seasonally stratified and nitrogen depleted waters of the northern Red Sea (Kamennaya *et al.*, 2008). The *cynA* gene of SB clusters with clones obtained from the Red Sea (Supplementary Figure S7), supporting their origin in HLII clade genomes as hypothesized by Kamennaya *et al.* (2008).

SB contains the most extensive suite of nitrogen acquisition pathways of any cultured *Prochlorococcus* strain examined to date. Why might this be? A useful analogy can be drawn from our understanding of selection pressures that have shaped *Prochlorococcus* genomes with respect to

adaptations involved in phosphorus assimilation. Individual cells and populations from phosphorus-limited environments possess accessory phosphorus acquisition genes, such as alkaline phosphatase (*phoA*) and phosphonate utilization (*phnYZ*) genes, at a higher frequency than *Prochlorococcus* from phosphorus-replete environments (Martiny *et al.*, 2006, 2009a; Coleman and Chisholm, 2010; Feingersch *et al.*, 2012). Thus, we hypothesize that the nitrogen assimilation traits present in *Prochlorococcus* SB were likely shaped by frequent nitrogen limitation in its original habitat (Iwata *et al.*, 2005); that is, cells capable of accessing a wide pool of nitrogen compounds may be at a selective advantage in nitrogen-limited environments.

Conclusions

Given the large standing stock of *Prochlorococcus* in the subtropical oceans and the extent to which nitrogen limits primary production in these regions (Tyrrell, 1999; Moore *et al.*, 2013), the absence of nitrate assimilation capabilities in cultured strains of *Prochlorococcus* has long puzzled biological oceanographers. This motivated field studies (Casey *et al.*, 2007; Martiny *et al.*, 2009b) and the use of models to help us understand the selection pressures driving the loss of nitrate assimilation genes in *Prochlorococcus* relative to *Synechococcus* (Bragg *et al.*, 2010). In this study we show unequivocally that some strains of *Prochlorococcus* are indeed capable of growth using nitrate as the sole nitrogen source. Future studies of these strains will help elucidate the physiological tradeoffs of carrying these genes and help refine the nitrogen inventory in biogeochemical models of the global ocean (Follows *et al.*, 2007). Correlations between environmental nitrate concentrations and ribotype phylogeny (Martiny *et al.*, 2009c) and the striking similarity between *Prochlorococcus* SB and the GOS consensus sequence both suggest that the trait for nitrate assimilation could be tied to distinct ribotype lineages. Still, evolution has many ways of introducing genomic complexity: the MIT0604 genome suggests that these genes are also subject to horizontal gene transfer, allowing further diversification of this trait in other lineages. This is reminiscent of the phylogenetic characteristics of phosphorus acquisition traits that are nearly independent of ribotype phylogeny (Martiny *et al.*, 2009c)—with extensive diversity in the 'leaves of the tree'. As we learn more about these layers of diversity, it will inform parameterizations of the relationship between light, temperature and nutrient acquisition traits for ocean simulation modeling.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank the captain and crew of the *R/V Kilo Moana* and members of the Hawai'i Ocean Time-series program (HOT181 and HOT212) for technical support with field operations. We also thank Robert D Harper and Hassan Shaleh (University of Southern Maine, Portland, ME, USA) for culturing assistance as well as Libusha Kelly (Albert Einstein College of Medicine, Bronx, NY, USA) for advice on bioinformatics analyses. This work was funded in part by the Gordon and Betty Moore Foundation through Grant GBMF495 to SWC and by the National Science Foundation (Grants OCE-1153588 and DBI-0424599 to SWC, OCE-0928544 to ACM, OCE-0851288 to LRM and OCE-9417071 to LC). AGK was supported by the NSF Graduate Research Fellowship Program (DGE-1321846). This article is a contribution from the NSF Center for Microbial Oceanography: Research and Education (C-MORE).

References

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

Ahlgren NA, Rocap G. (2006). Culture isolation and culture-independent clone libraries reveal new marine *Synechococcus* ecotypes with distinctive light and N physiologies. *Appl Environ Microbiol* **72**: 7193–7204.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

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

Bouman HA, Ulloa O, Scanlan DJ, Zwirglmaier K, Li WK, Platt T *et al.* (2006). Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. *Science* **312**: 918–921.

Bragg JG, Dutkiewicz S, Jahn O, Follows MJ, Chisholm SW. (2010). Modeling selective pressures on phytoplankton in the global ocean. *PLoS One* **5**: e9569.

Buitenhuis ET, Li WKW, Vaulot D, Lomas MW, Landry MR, Partensky F *et al.* (2012). Picophytoplankton biomass distribution in the global ocean. *Earth Syst Sci Data* **4**: 37–46.

Campbell L, Nolla HA, Vaulot D. (1994). The importance of *Prochlorococcus* to community structure in the central North Pacific Ocean. *Limnol Oceanogr* **39**: 954–961.

Casey JR, Lomas MW, Mandecki J, Walker DE. (2007). *Prochlorococcus* contributes to new production in the Sargasso Sea deep chlorophyll maximum. *Geophys Res Lett* **34**: L10604.

Cavender-Bares KK, Mann EL, Chisholm SW, Ondrusek ME, Bidigare RR. (1999). Differential response of equatorial Pacific phytoplankton to iron fertilization. *Limnol Oceanogr* **44**: 237–246.

Chisholm SW, Frankel SL, Goericke R, Olson RJ, Palenik B, Waterbury JB *et al.* (1992). *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll *a* and *b*. *Arch Microbiol* **157**: 297–300.

Coleman ML, Chisholm SW. (2007). Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* **15**: 398–407.

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

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF *et al.* (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768–1770.

Collier JL, Lovindeer R, Xi Y, Radway JC, Armstrong RA. (2012). Differences in growth and physiology of marine *Synechococcus* (Cyanobacteria) on nitrate versus ammonium are not determined solely by nitrogen source redox state. *J Phycol* **48**: 106–116.

Dufresne A, Salanoubat M, Partensky F, Artiguenave F, Axmann IM, Barbe V *et al.* (2003). Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci USA* **100**: 10020–10025.

Feingersch R, Philosof A, Mejuch T, Glaser F, Alalouf O, Shoham Y *et al.* (2012). Potential for phosphite and phosphonate utilization by *Prochlorococcus*. *ISME J* **6**: 827–834.

Felsenstein J. (2005). *PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the Author.* Department of Genome Sciences, University of Washington: Seattle.

Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N *et al.* (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci USA* **110**: 9824–9829.

Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. (2007). Emergent biogeography of microbial communities in a model ocean. *Science* **315**: 1843–1846.

Fuller NJ, Marie D, Partensky F, Vaulot D, Post AF, Scanlan DJ. (2003). Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol* **69**: 2430–2443.

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

García-Fernández JM, de Marsac NT, Diez J. (2004). Streamlined regulation and gene loss as adaptive mechanisms in *Prochlorococcus* for optimized nitrogen utilization in oligotrophic environments. *Microbiol Mol Biol Rev* **68**: 630–638.

Giandomenico AR, Cerniglia GE, Biaglow JE, Stevens CW, Koch CJ. (1997). The importance of sodium pyruvate in assessing damage produced by hydrogen peroxide. *Free Radical Bio Med* **23**: 426–434.

Gruber N. (2008). The marine nitrogen cycle: overview and challenges. In: Capone DG, Bronk DA, Mulholland MR, Carpenter EJ (eds) *Nitrogen in the Marine Environment*. Academic Press: Burlington, MA, pp 1–50.

Herrero A, Muro-Pastor AM, Flores E. (2001). Nitrogen control in Cyanobacteria. *Biochim Biophys Acta* **183**: 411–425.

Huang S, Wilhelm SW, Harvey HR, Taylor K, Jiao N, Chen F. (2012). Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. *ISME J* **6**: 285–297.

Iwata T, Shinomura Y, Natori Y, Igarashi Y, Sohrin R, Suzuki Y. (2005). Relationship between salinity and nutrients in the subsurface layer in the Suruga Bay. *J Oceanogr* **61**: 721–732.

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

Kamennaya NA, Chernihovsky M, Post AF. (2008). The cyanate utilization capacity of marine unicellular Cyanobacteria. *Limnol Oceanogr* **53**: 2485–2494.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A *et al.* (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**: 416–420.

Kelly L, Huang KH, Ding H, Chisholm SW. (2012). ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res* **40**: D632–D640.

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

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lavin P, González B, Santibáñez JF, Scanlan DJ, Ulloa O. (2010). Novel lineages of *Prochlorococcus* thrive within the oxygen minimum zone of the eastern tropical South Pacific. *Environ Microbiol Rep* **2**: 728–738.

Lomas MW, Lipschultz F. (2006). Forming the primary nitrite maximum: nitrifiers or phytoplankton? *Limnol Oceanogr* **51**: 2453–2467.

Malmstrom RR, Coe A, Kettler GC, Martiny AC, Frias-Lopez J, Zinser ER *et al.* (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J* **4**: 1252–1264.

Malmstrom RR, Rodrigue S, Huang KH, Kelly L, Kern SE, Thompson A *et al.* (2013). Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. *ISME J* **7**: 184–198.

Mann EL, Chisholm SW. (2000). Iron limits the cell division rate of *Prochlorococcus* in the eastern equatorial Pacific. *Limnol Oceanogr* **45**: 1067–1076.

Martiny AC, Coleman ML, Chisholm SW. (2006). Phosphate acquisition genes in *Prochlorococcus* ecotypes: evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* **103**: 12552–12557.

Martiny AC, Huang Y, Li W. (2009a). Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* **11**: 1340–1347.

Martiny AC, Kathuria S, Berube PM. (2009b). Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci USA* **106**: 10787–10792.

Martiny AC, Tai AP, Veneziano D, Primeau F, Chisholm SW. (2009c). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environ Microbiol* **11**: 823–832.

Marty J-C, Chiavérini J, Pizay M-D, Avril B. (2002). Seasonal and interannual dynamics of nutrients and phytoplankton pigments in the western Mediterranean Sea at the DYFAMED time-series station (1991-1999). *Deep Sea Res Part II* **49**: 1965–1985.

Mirkin BG, Fenner TI, Galperin MY, Koonin EV. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* **3**: 2.

Moore LR, Chisholm SW. (1999). Photophysiology of the marine cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. *Limnol Oceanogr* **44**: 628–638.

Moore LR, Coe A, Zinser ER, Saito MA, Sullivan MB, Lindell D *et al.* (2007). Culturing the marine cyanobacterium *Prochlorococcus*. *Limnol Oceanogr Meth* **5**: 353–362.

Moore LR, Post AF, Rocap G, Chisholm SW. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol Oceanogr* **47**: 989–996.

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

Moore CM, Mills MM, Arrigo KR, Berman-Frank I, Bopp L, Boyd PW *et al.* (2013). Processes and patterns of oceanic nutrient limitation. *Nat Geosci* **6**: 701–710.

Morris JJ, Johnson ZI, Szul MJ, Keller M, Zinser ER. (2011). Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. *PLoS One* **6**: e16805.

Morris JJ, Kirkegaard R, Szul MJ, Johnson ZI, Zinser ER. (2008). Facilitation of robust growth of *Prochlorococcus* colonies and dilute liquid cultures by "helper" heterotrophic bacteria. *Appl Environ Microbiol* **74**: 4530–4534.

Mühling M. (2012). On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. *Environ Microbiol* **14**: 567–579.

Ohashi Y, Shi W, Takatani N, Aichi M, Maeda SI, Watanabe S *et al.* (2011). Regulation of nitrate assimilation in cyanobacteria. *J Exp Bot* **62**: 1411–1424.

Olson RJ, Chisholm SW, Zettler ER, Altabet MA, Dusenberry JA. (1990). Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. *Deep Sea Res Part I* **37**: 1033–1051.

Olson RJ, Vaulot D, Chisholm SW. (1985). Marine phytoplankton distributions measured using shipboard flow cytometry. *Deep Sea Res Part I* **32**: 1273–1280.

Partensky F, Blanchot J, Vaulot D. (1999). Differential distribution and ecology of *Prochlorococcus* and *Synechococcus* in oceanic waters: a review. In: Charpy L, Larkum AWD (eds). *Marine Cyanobacteria* Bulletin de l Institut oceanographique de Monaco, No. spécial 19, Musée océanographique: Monaco, pp 457–475.

Partensky F, Garczarek L. (2010). *Prochlorococcus*: advantages and limits of minimalism. *Annu Rev Mar Sci* **2**: 305–331.

Penno S, Campbell L, Hess WR. (2000). Presence of phycoerythrin in two strains of *Prochlorococcus* (Cyanobacteria) isolated from the subtropical north Pacific Ocean. *J Phycol* **36**: 723–729.

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

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

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.

Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC. (2010). Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc Natl Acad Sci USA* **107**: 16184–16189.

Saito MA, Moffett JW, Chisholm SW, Waterbury JB. (2002). Cobalt limitation and uptake in *Prochlorococcus*. *Limnol Oceanogr* **47**: 1629–1636.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.

Scanlan DJ, West NJ. (2002). Molecular ecology of the marine cyanobacterial genera *Prochlorococcus* and *Synechococcus*. *FEMS Microbiol Ecol* **40**: 1–12.

Shimada A, Nishijima M, Maruyama T. (1995). Seasonal appearance of *Prochlorococcus* in Suruga Bay, Japan. *J Oceanogr* **51**: 289–300.

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

Sukumaran J, Holder MT. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**: 1569–1571.

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

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

Vaulot D, Lebot N, Marie D, Fukai E. (1996). Effect of phosphorus on the *Synechococcus* cell cycle in surface Mediterranean waters during summer. *Appl Environ Microbiol* **62**: 2527–2533.

Vaulot D, Partensky F. (1992). Cell cycle distributions of prochlorophytes in the north western Mediterranean Sea. *Deep Sea Res Part I* **39**: 727–742.

Waterbury JB, Willey JM. (1988). Isolation and growth of marine planktonic cyanobacteria. *Methods Enzymol* **167**: 100–105.

Williams EZ, Campbell L, DiTullio G. (1999). The nitrogen specific uptake of three strains of *Prochlorococcus*. Presented at the American Society of Limnology and Oceanography Aquatic Sciences Meeting; 4 February 1999, Santa Fe, NM.

Williams KP. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res* **30**: 866–875.

Wu J, Sunda W, Boyle EA, Karl DM. (2000). Phosphate depletion in the western North Atlantic Ocean. *Science* **289**: 759–762.

Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, Scanlan DJ *et al.* (2006). *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl Environ Microbiol* **72**: 723–732.

Zwirglmaier K, Heywood JL, Chamberlain K, Woodward EM, Zubkov MV, Scanlan DJ. (2007). Basin-scale distribution patterns of picocyanobacterial lineages in the Atlantic Ocean. *Environ Microbiol* **9**: 1278–1290.

Zwirglmaier K, Jardillier L, Ostrowski M, Mazard S, Garczarek L, Vaulot D *et al.* (2008). Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol* **10**: 147–161.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)

Appendix E.
Prochlorococcus light and fluorescence microscopy: methods and images

Taking pictures of *Prochlorococcus*

In the 1970s, when John Waterbury looked at seawater through a fluorescence microscope and saw signatures of chlorophyll and other pigments in marine bacteria, he discovered the marine picocyanobacteria, and the world met the marine *Synechococcus*. The other abundant picocyanobacterium in the oceans, *Prochlorococcus*, had to wait 15 years more for its discovery, when Rob Olsen and Penny Chisholm took a flow cytometer to sea. Their specially adapted flow cytometer was capable of detecting small particles based laser-excited chlorophyll fluorescence emission, that turned out to be *Prochlorococcus*, who can be operationally defined as the smallest chlorophyll-containing particles in the sea. Why didn't early microscopy efforts see *Prochlorococcus?* The genus resisted imaging primarily because its chlorophyll fluorescence bleaches very quickly - a minute under a bright fluorescence microscope's lamp and the cells go dark. High quality images of fluorescing phytoplankton benefit from moderate photographic exposure times, creating beautiful *Synechococcus* and diatom pictures, but bleaching the *Prochlorococcus*. In flow cytometry a more powerful light source provides the excitation, over a shorter length of time, and detection occurs not through an imaging exposure, but a rapid detection and quantification of emitted photons by a photomultiplier tube (highly sensitive detectors that converts photon to electronic signal) - the combination of sensitivity and rapid timing enables easy detection and quantification of *Prochlorococcus* and its chlorophyll fluorescence, unaffected by sensitivity to photobleaching. They are also small, 0.6-0.8um diameter coccoid cells (compare to 1um *Synechococcus*), the size of the wavelengths of light they absorb, and a size which does not resolve as well under traditional light or widefield fluorescence microscopy, although they are well above the ~250nm theoretical resolution limit of the technique. Here we present a protocol, based on advice from Steve Biller's experience imaging *Bacillus subtilis* using agar pads, modified to fit *Prochlorococcus*. Simon Labrie also helped, by teaching me how to properly focus and align the microscope, because all parts of the microscope needs to be functioning at their best to visualize these very small cells.

Here we present a method for efficiently imaging concentrated live *Prochlorococcus* cells for routine observations of axenicity, morphology, physical co-culture interactions, and also for outreach and communication of what a *Prochlorococcus* cell looks like. The goal is to have cells that hold still (not floating in liquid) so you can get multiple kinds of images of the same field of cells and exposure times long enough to see chlorophyll fluorescence, but have not been heavily manipulated. This basic method would probably also work with fixed cells and filter-concentrated cells, but I haven't explored these specifically, just live cultures.

A protocol for the use of agar pads for light and fluorescence visualization of live *Prochlorococcus*

1) Melt a mixture of 0.8% agar in Pro99 in the microwave. 0.8% is soft but solid; the Pro99 is so that the cells stay as happy as possible during the few hours it might take to prepare and image a set of slides.
2) Pipette ~200 ul of the warm agar onto a clean microscope slide. It should make a rounded pool in the center of the slide (amount depends on consistency and size of slide)
3) Then, there are two possible methods for making a flat agar pad; if the agar is hot it spreads thin quickly and creates a fairly flat surface, so you can just wait for the agar pool to harden on its own. Alternatively, to bias things towards a nice flat surface (important for creating a single focal plane of layer cells), you can quickly place another microscope slide on top of the wet agar (a sandwich). Once the agar sets, carefully slide one plate off, leaving a flat pad. It's important that the slides are parallel, not slanted.
4) For the images presented below I used pelleted concentrated *Prochlorococcus*: 1 ml of culture (moderately to very green log phase), spun ~10 minutes at 13,000 rpm on a benchtop centrifuge at room temperature usually makes a pellet. Spin more if needed. Pipette off most of clear supernatant (leave 20-100 ul), and resuspend cells in residual amount of liquid.
5) Pipette a few ul of pelleted cells onto the agar pad (enough to see a spot), allow it to sit a few minutes (I think some moisture sinks into the agar and the cells start to settle), then squish it with a coverslip.
6) Now they're ready to image - to see cells you'll need 100X oil immersion lenses on microscope. The center of the pellet will have many layers of cells, hard to focus, but around the edges, spread around the agar there will be a monolayer that images well.
7) To get good images of *Prochlorococcus* autofluorescence, it is necessary to focus and choose a field using light microscopy (exposing those cells to as little light as possible), then switch to fluorescence microscopy (at much higher light - cells are not visible with the naked eye), and immediately take the photographic exposure (for a long time, 30-60s). By the time the exposure is done the cells will be nearly bleached, and you move onto another field. See figures below.

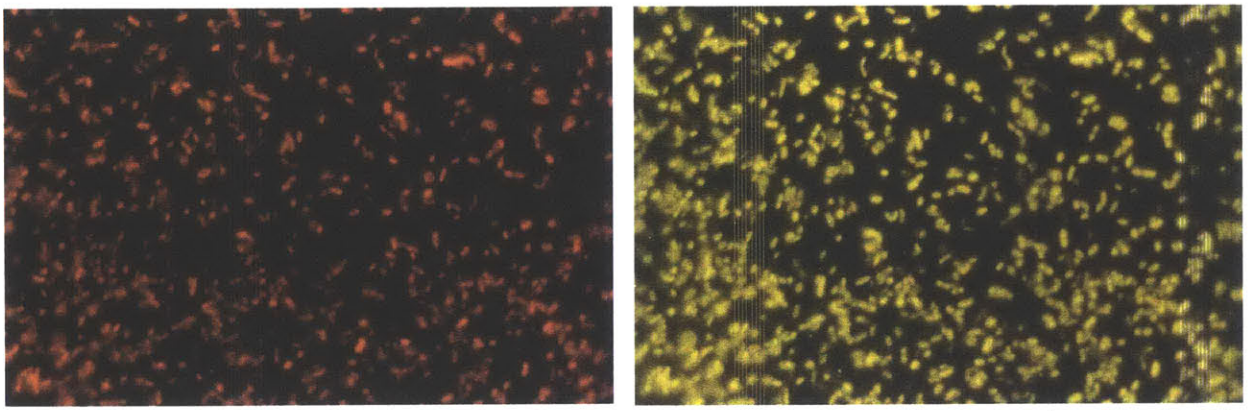


Figure E.1. *Synechococcus* 6501
The image at left is chlorophyll fluorescence and at right phycoerythrin fluorescence of the same field of cells, using our 100X oil immersion lens. The chlorophyll is a little underexposed and the phycoerythrin is a little overexposed – it took a few tries with exposure times to get a reasonable intensity. I usually start with *Synechococcus* to set up the microscope because they are easier to see. It's more difficult to image *Prochlorococcus* fluorescence because they bleach faster.

Figure E.2. NATL2A, nonaxenic
Apart from the smudge and the annoying yellow-beige background tone (an issue with auto-white-balance on the camera - clear white background to the eye), this image came out well enough, so we sent it to Jennifer Frazer for her Exploratorium exhibit. You can see almost the true color of the cells and the dramatic size, shape and color differences between the heterotrophs and *Prochlorococcus*. This image is slightly out of focus – the cells pop a bit, dark edges and pale middles –not ideal, but it actually makes it easier to see cells. I made the scale bar by imaging a hemocytomer with markings of defined sizes using the same magnification and camera settings, then measuring those markings in pixels to convert

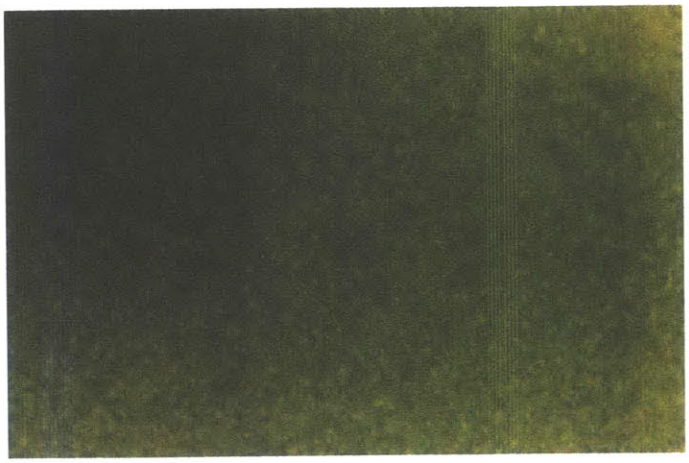
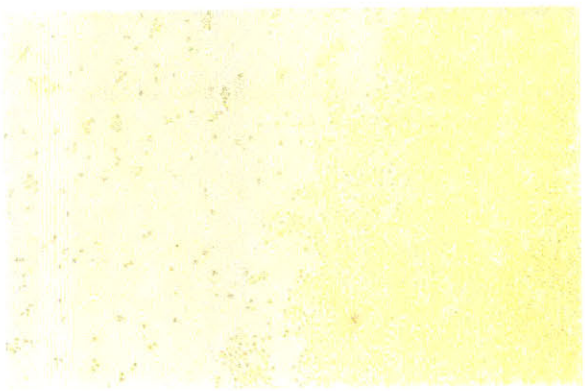


Figure E.3. NATL2A pellet
On the same slide as the above image of a cell monolayer, here is the center of a pellet – bright green and impossible to focus on a single cell layer in bright field. Gives a nice sense of color and abundance, but not very useful for seeing cells.



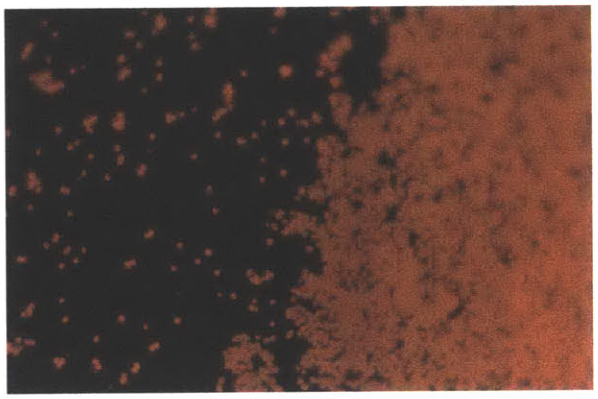


Figure E.4. NATL2A paired bright field/chlorophyll fluorescence
In the bright field image at left *Prochlorococcus* and heterotrophs contaminating this culture show clearly. Comparing between the two images you can watch the heterotrophs disappear in the chlorophyll fluorescence image.

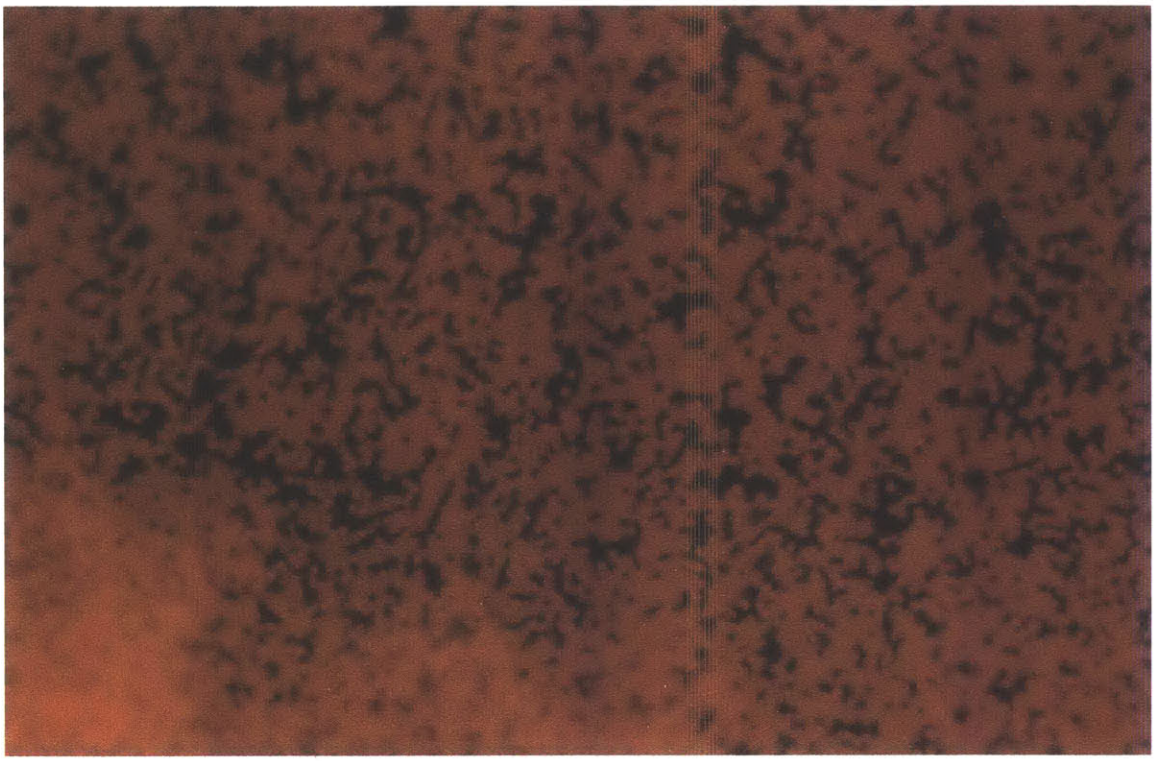


Figure E.5. NATL2A edge of a pellet
Just another favorite, still from the same prepared slide of non axenic NATL2A, showing a single layer of tightly packed cells transitioning into a pellet mountain of chlorophyll in the lower right.

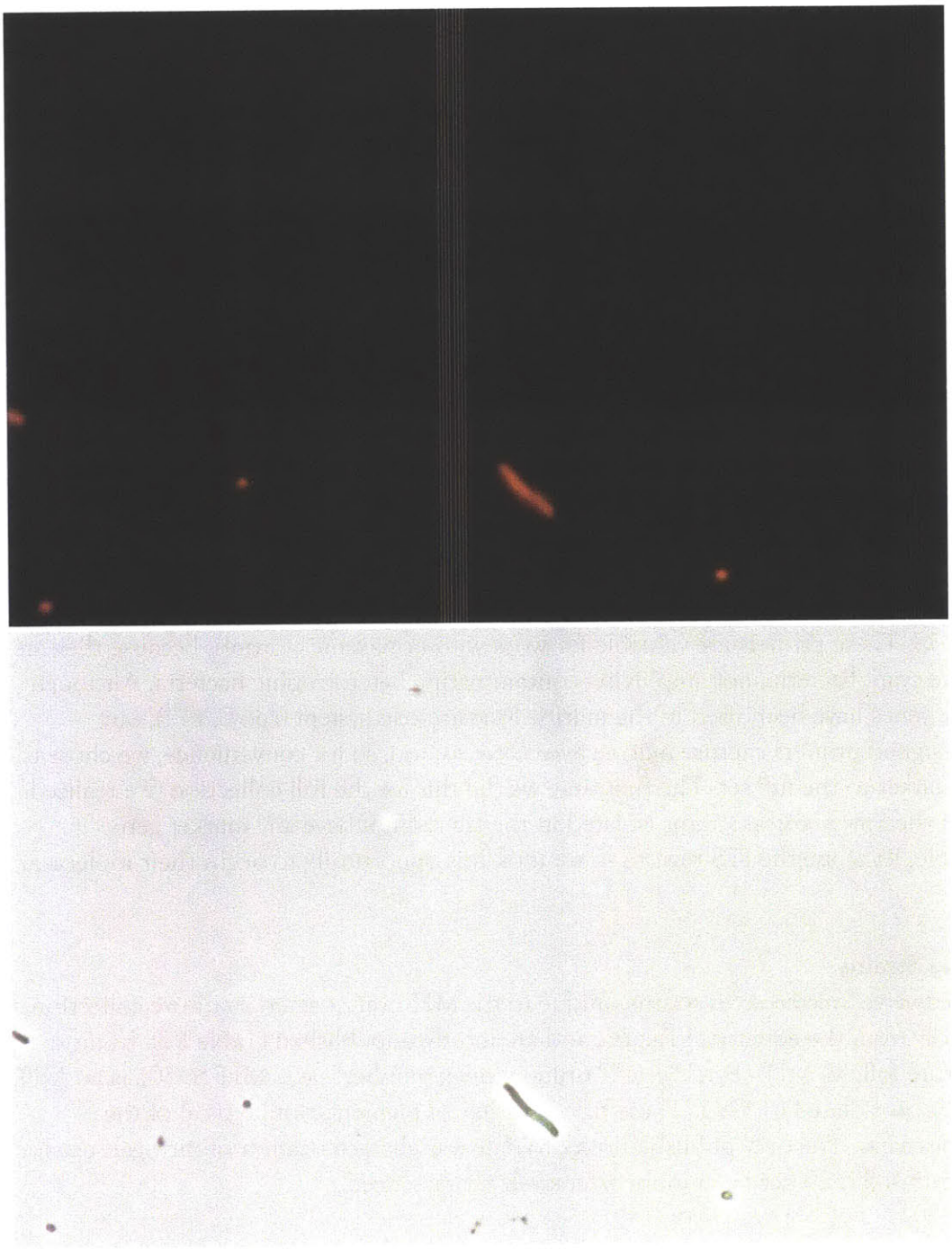


Figure E.6. The observation long *Prochlorococcus* cells
As part of our work in Chapter II, characterizing new LLIV clade isolates, we (with Duygu Kasdogan) noticed in some axenic dilution purified strains, the presence of long cells with chlorophyll fluorescence (in this case, strain 1B3; see Chapter II). We observed the phenomenon repeatedly over the course of several months, in several ecotypes. Understanding this unprecedented morphological variation will be an exciting future avenue of research.

Appendix F.
Synechococcus of the MIT culture collection

Marker gene sequencing and taxonomic classification for MIT marine *Synechococcus* strains

MIT culture collection and marker gene barcoding

The Chisholm lab maintains a collection of *Prochlorococcus* and marine *Synechococcus* strains, isolated over the years in our lab and by others, including many strains not present in national culture collections. These strains are kept both as cryopreserved stocks, which safely preserve cultures but take a very long time in this system (months) to revive to active growth, and as liquid batch cultures, which enable easy access and rapid scale up of cultures. For these batch cultures, we recently started routinely checking marker gene sequences, to ensure strain identity was stable over time, since with liquid culture there is a small risk of cross contamination. It is generally good practice to check strain identity before performing experiments, and it seemed prudent to apply this simple barcode scanning to the culture collection occasionally.

The marker gene sequence of choice in the *Prochlorococcus* system is the ITS-rRNA (internal transcribed spacer between the 16S and 23S bacterial rRNAs), because the ITS is highly variable, a useful barcode and phylogenetic marker, but can be amplified by PCR off conserved priming sites in the 16S and 23S. These primers are valuable for work with nonaxenic cultures, because they are specific to marine cyanobacteria, not amplifying contaminating heterotrophic bacteria. Although historically other genes have been used in the marine *Synechococcus* system (rpoC, pE?), our *Prochlorococcus*-designed primers capture marine *Synechococcus*, too, so for convenience, we chose to apply the same marker to the full set. The first time we did this for the full collection, we realized that a number of the *Synechococcus* strains isolated in the lab did not have any marker gene sequences available, let alone the ITS region, so we took this opportunity to begin their molecular characterization.

MIT *Synechococcus* Strains

There are twelve *Synechococcus* strains unique to the MIT cyanobacterial culture collection, which are primarily from the equatorial Pacific, and are mostly unpublished (Table F1). In most cases, nomenclature follows MIT - Syn" "year" "unique index number" - e.g. MIT S9501 is an MIT strain of *Synechococcus* isolated in 1995. These have a variety of pigmentation, typical of the diversity of *Synechococcus*. The only published work to date was characterization of nitrogen use for a subset of the strains. S9220 has been more extensively characterized.

Table F1. MIT *Synechococcus* Strains

Strain	Isolation information	Publication
MIT S9501	Equatorial Pacific, E. Mann	unpublished
MIT S9503	Equatorial Pacific, E. Mann	unpublished
MIT S9504	Equatorial Pacific, 20m, E. Mann	Moore et al., 2002
MIT S9506	Equatorial Pacific, E. Mann	unpublished
MIT S9507	Equatorial Pacific, E. Mann	unpublished
MIT S9508	Equatorial Pacific, surface water E. Mann	Moore et al., 2002
MIT 9509	Equatorial Pacific, E. Mann	unpublished
MIT 9510	Equatorial Pacific, E. Mann	unpublished
MIT S9214	South Pacific, (11°60'S, 145°25'W), surface water, B. Binder	Moore et al., 2002
MIT S9220	Equatorial Pacific (0°,40°W), surface water, B. Binder	Moore et al., 2002
Cu2B8	E. Mann	unpublished
MIT S9451	Sargasso Sea, 65m, L. Aref	Moore et al., 2002
MIT S9452	Sargasso Sea, 65m, L. Aref	Moore et al., 2002

Results

Who are these strains and what can we learn from them?

For a first pass at placing these sequences in their phylogenetic context, we built a phylogeny relating these strain to 17 marine *Synechococcus* with sequenced genomes (Figure F1). At the broadest scale of diversity, one of our strains belongs to deeply branching subcluster 5.3, which is less frequently observed across the oceans, but can be abundant at certain places at certain times (Ahlgren and Rocap, 2012). The rest belong to different subgroups of the main clade of marine *Synechococcus* in the oceans, 5.1. In some cases these strains have other close relatives in culture; in other cases they represent uncultured fine scale sub clades (Table F2). The majority of the strains in this set are closely related to each other in the CRD clade (Saito et al., including strains isolated from different years from the South Pacific, that are not close to any strains with sequenced genomes. These strains have diverse pigmentation.

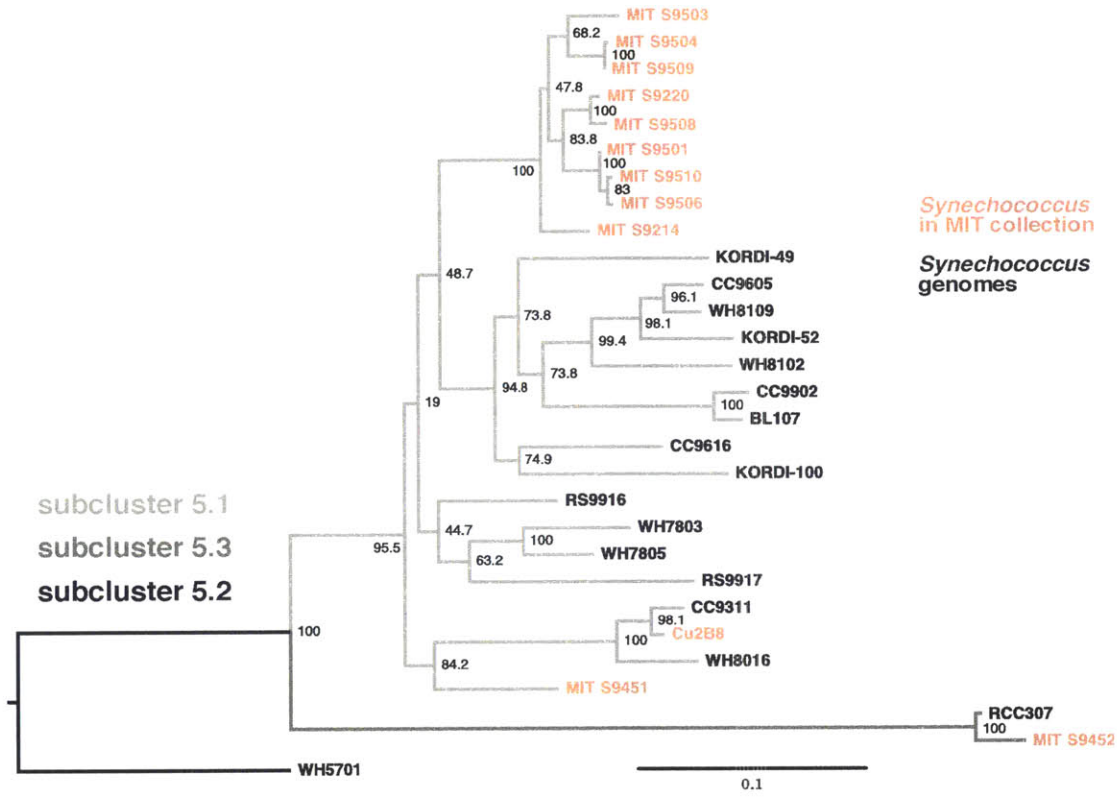


Figure F1. MIT *Synechococcus* ITS phylogeny with *Synechococcus* with sequenced genomes
ITS sequences for MIT collection and 17 marine *Synechococcus* with sequenced genomes. Aligned with mafft using gap-friendly einsi parameters, tree uses neighbor joining algorithm, tamura-nei distances and 1000 bootstrap replicates (implemented in Geneious) Bootstrap values at nodes.

Table F2. Best BLAST hits, closest relatives, cultured and uncultured for MIT Syn strains

Strain	Best blast hit in the 'non redundant' database	Best cultured blast hit
MIT S9501	uncultured CRD29, 99%, Sargasso Sea (Saito paper)	RCC1016, 98%
MIT S9503	uncultured UTK255, 99%, Equatorial pacific, Huang novel lin	RCC66, 96%
MIT S9504	RCC66, 99%	same
MIT S9506	uncultured CRD29, 99% Sargasso Sea (Saito paper)	RCC1018, 96%
MIT S9507	uncultured CRD29, 99%, Sargasso Sea (Saito paper)	RCC1016, 97%
MIT S9508	uncultured CRD12, 99% Sargasso Sea (Saito paper)	S9920, 98%
MIT 9509	RCC66, 100%	same
MIT 9510	uncultured CRD29, 99% Sargasso Sea (Saito paper)	RCC1016, 97%
MIT S9214	uncultured CRD25, 97% Sargasso Sea (Saito paper)	RCC1018 96%
MIT 9220	itself - previous published	same
Cu2B8	UW76 99%	same
MIT S9451	UW149 100%	same
MIT S9452	uncultured oc5m73, 99% Sargasso sea, march 2002, Ahlgren 2006 CUlture isolation...	KORDI-30 99%

Methods

Culture conditions

These cultures are maintained in Pro99 media, which is standard for *Prochlorococccus* composed of filtered, sterilized seawater amended with ammonia, phosphate and tracemetals. Although *Synechococcus* are generally cultured in other media (e.g.), we found that Pro99 supports the growth of all of these *Synechococcus* strains, so for convenience they are maintained alongside the *Prochlorococcus* cultures in the MIT cyanobacterial culture collection, under identical conditions.

ITS PCR and sequencing

ITS-PCR was performed as described in Rodrigue et al., 2009, using template prepared as described in Chapter II.

References

Ahlgren, N.A., and Rocap, G. (2012). Diversity and Distribution of Marine *Synechococcus*: Multiple Gene Phylogenies for Consensus Classification and Development of qPCR Assays for Sensitive Measurement of Clades in the Ocean. Front Microbiol 3, 213.

Moore, L.R., Post, A.F., Rocap, G., and Chisholm, S.W. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. Limnology and Oceanography 47, 989-996 .

Saito, M.A., Rocap, G., and Moffett, J.W. Production of cobalt binding ligands in a *Synechococcus* feature at the Costa Rica upwelling dome. Limnology and Oceanography 50, 279-290.